UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Biologie Structurale et Nanobiologie**

Arrêté ministériel : 7 août 2006

Présentée par

Ambroise DESFOSSES

Thèse dirigée par **Rob Ruigrok** et codirigée par **Irina Gutsche**

préparée au sein du « **Unit for Virus Host Cell Interactions (UVHCI) UMI 3265 UJF-EMBL-CNRS** »
dans **l'École Doctorale Chimie et Sciences du Vivant**

Analyse d'images de microscopie électronique de biopolymères hélicoïdaux flexibles

Thèse soutenue publiquement le **31 Octobre 2012**,
devant le jury composé de :

M Patrick SCHULTZ
Directeur de Recherche, IGBMC Strasbourg, Rapporteur

M Denis CHRETIEN
Directeur de Recherche, Université de Rennes, Rapporteur

M Carsten SACHSE
Chef de groupe, EMBL Heidelberg, Membre

Mme Irina GUTSCHE
Charge de recherche, UVHCI Grenoble, Membre

M Rob RUIGROK
Professeur, Université de Grenoble, Membre

M Hans GEISELMANN
Professeur, Université de Grenoble, Président

Acknowledgments

Je voudrais tout d'abord remercier mes directeurs de Thèse, Irina Gutsche et Rob Ruigrok, pour m'avoir accordé leur confiance et accepté comme thésard au sein de l'UVHCI.

Irina, j'ai démarré en Master 2 avec toi sans rien connaitre en microscope électronique, en préparation de grilles, en traitement d'image, en scripts, et tu as passé beaucoup de temps à m'apprendre les bases et plus. Je t'en suis très reconnaissant. Ces apprentissages me servent toujours ! Merci aussi d'avoir partagé une vision de la science ou on évite le « copier-coller » et ou la curiosité et le plaisir de faire soi-même prime. Pour la façon dont tu as mené la publication de nos résultats, ton écriture, ta persévérance, je t'exprime toute ma gratitude. Enfin, pour m'avoir accompagné plus au jour le jour, merci d'avoir su trouver un équilibre pour en même temps me guider et me laisser beaucoup de liberté dans mes choix. Même si on a parfois douté de cet équilibre, je crois que le résultat montre que ça a marché…alors que tu partais avec un étudiant pas facile, spécialiste de la résistance passive et de la procrastination (« oui, oui, ça va, j'avance, tu auras le chapitre dans une semaine », avait-il dit six mois plus tôt..) ! Merci donc aussi pour ta patience.

Rob, tu es pour beaucoup dans le fait que j'ai continué dans la science, alors qu'un fameux Vendredi après-midi j'allais tout arrêter : tu m'as montré qu'il était légitime de se poser des questions sur la voie que l'on veut suivre, même après des années dans le monde de la recherche, et que l'on peut concilier la vie de chercheur avec beaucoup d'autres centres d'intérêts. Et avoir beaucoup d'humour. Je sais que j'ai été source de beaucoup de « tracas », jusqu'au bout ….et donc merci d'avoir malgré tout continue à me soutenir. Merci à Guy Schoehn, en tant que responsable du groupe de microscopie électronique.

Leandro, il faudrait plus qu'une thèse pour te dire merci ! Merci de m'avoir ouvert ton amitié, elle m'est, et m'a été, extrêmement précieuse. Concernant le travail, merci d'avoir pris tant de temps pour partager tes connaissances et ta passion, ainsi que ta façon d'aborder les problèmes et de les résoudre. Chaque semaine qui passe, ce que tu m'as transmis m'est utile, et ma reconnaissance est énorme. Je n'aurais pu rêver meilleur maitre, et ami, pour ces quatre ans à Grenoble !

Cédric, c'était tout d'abord pour continuer à partager de bons moments avec toi, de science ou autres, que je t'ai « suivi » à l'UVHCI. Tu as été très bon pour rappeler qu'avant tout il fallait s'amuser pour être un bon chercheur. Sans toi, les années de Master et de thèse n'auraient pas été aussi fun !

Merci à Francine Gérard et Julien Perard pour la transmission de protocoles et les conseils en biochimie. Merci à Gregory Effantin d'avoir passé du temps à utiliser mes scripts et de me donner un feedback. Merci à Marc Jamin pour m'avoir donné une bonne raison, via l'analyse d'output de servers de prédictions de structure, d'apprendre beaucoup en grep, sed, tr, awk et autres commandes barbares.

Nico, Euripides, Ivan, avec qui nous avons aussi partagé la paillasse, merci pour l'excellente humeur et les franches rigolades. De nombreux moments resteront dans les annales, c'est certain. Merci à mes collègues de bureau pour l'ambiance plus qu'agréable et les passionnantes discussions sur CHMP2B.

Merci à tous les collègues de l'UVHCI, de l'EMBL, de l'IBS, de l'ILL et de l'ESRF, avec qui j'ai passé d'excellentes années. Vous êtes trop nombreux -tant mieux- pour tous vous nommer ici mais vous vous reconnaitrez, j'en suis sûr !

Je remercie Patrick Schultz et Denis Chretien qui ont accepté d'être les rapporteurs de ce mémoire, le président et les membres du jury qui me font l'honneur de juger ce travail.

Merci à mes parents, à mes frères et à ma belle-famille, pour vos encouragements et votre soutien précieux. Merci mes filles, Leïa, Alix, pour la force que votre amour me donne. Sandra, tu mériterais très largement de recevoir cette thèse, pour tout ce que tu as fait comme chemin avec moi, pour toute l'énergie que tu y as mis, je te suis reconnaissant plus que tu ne peux l'imaginer… merci, merci, et encore merci !

Table of content

Abbreviations

1D : one-dimension(al)

2D : two dimension(al)

Å : Angstrom

ACC : average cross-correlation

CC : cross-correlation

DNA : Deoxyribonucleic acid

EM : Electron Microscopy

EMDB : Electron Microscopy Data Bank

L : polymerase (see context)

MeV : Measles Virus

MeVD : Measles Virus digested nucleocapsids

MeVND : Measles Virus non-digested nucleocapsids

N : nucleoprotein (see context)

P : phosphoprotein (see context)

RNA : Ribonucleic acid

mRNA : messenger RNA

RSV : Respiratory Syncytial Virus

TMV : Tobacco Mosaic Virus

VSV : Vesicular Stomatitis Virus

Introduction

The work presented here took place in the Unit for Virus Host Cell Interaction (UVHCI) in Grenoble between 2007 and 2010. The main subject of the laboratory is interdisciplinary research covering virus structure, assembly and maturation, virus-host cell interactions, host and virus gene-expression mechanisms, cell biology of infected cells, innate immunity and anti-pathogen drug design. The biological part of my thesis was performed in collaboration with the group of Professors Ruigrok and Jamin which focuses on replication of negative strand RNA viruses, whereas the major part of the work consisted of electron microscopy and image processing and was conducted in the "Virus Structure and Electron Microscopy Development" headed by Dr. Schoehn.

Helices in Biology

Overview

Helices are everywhere in the biological world, at every scale, in every organism (**figure 1.1**). One of the main structural elements of proteins is alpha helices (**figure 1.1A**), where the alpha carbons of the amino acids are connected to each other forming a helical path and the bases are sticking out of the formed helix (Pauling, Corey, and Branson 1951). As we will see, the work done for understanding the geometry of peptide helices, and in the particular the interpretation of X-ray fiber diffraction pattern in the 50's (Cochran, Crick, and Vand 1952) was crucial for the future upcoming of the first Three-Dimensional (3D) reconstruction from Electron Microscopy (EM) images in the 60's (DeRosier and Klug 1968). The next example of a biological helix, perhaps the most famous, is the double-helical arrangement of bases in DNA (**figure 1.1B**). Again, its structure was solved using fiber X-ray diffraction data obtained quasi simultaneously by several groups in 1953 (Watson and Crick 1953; Franklin and Gosling 1953; Wilkins, Stokes, and Wilson 1953).

At the protein level (**figure 1.1C**), which interests us mainly, helical polymers are also ubiquitous in biology: they are found in bacteriophages, viruses and all eubacterial, archaeal and eukaryotic cells. If it is clearly impossible to exhaustively list all helical protein polymers,

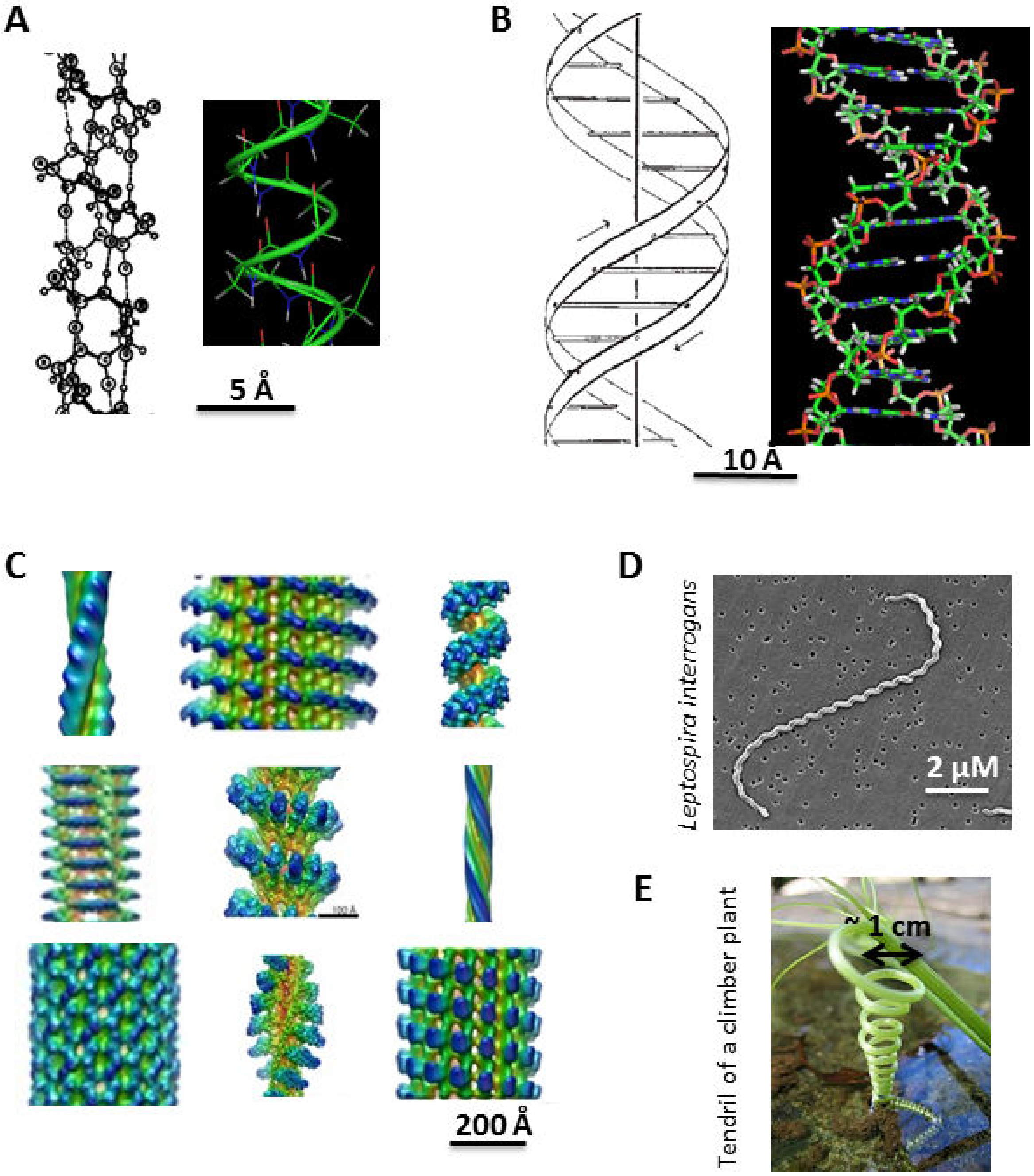


Figure 1.1 : Helices at every scale in Biology

From the panel A to E, biological objects with helical shape of increasing size are shown. The alpha helix (**A**) and the DNA double strand (**B**) structures are shown in their first depiction on the left and in a modern representation on the right. Note that in the representation of the alpha helix by Pauling et al. in 1951 (**A, left**), the handedness was wrong : the helical path is turning anti-clockwise when looking from the helix axis (left-handed helix) whereas in the real structure the helix is right-handed (**A, right**). In **C**, a gallery of 3D reconstructions obtained by Electron Microscopy of various protein polymers forming helices is shown, using a constant scale. Comparing to those protein polymers, at a size scale of ~10^4 higher, the bacterial members of the phylum *Spirochaete* exhibit helical cells (**D**). Again some order of magnitude bigger, animals, or plants, can have parts of their organism which make use of the helical shape to fulfill certain functions (in **E**, a plant tendril used for climbing) we can still mention as examples F-actin, microtubules, myosin thick filaments, phage tails, bacterial pili and flagella, amyloid fibers, viruses capsids and nucleocapsids. The functions that are fulfilled by proteins forming helical polymers are probably even more numerous than their variety, as often one polymer is used for several different cellular processes. Helical protein polymers are for example involved in cytoskeleton (F-actin, microtubules), muscle contractility (actin + myosin), secretion machineries (type III secretion system needle), protection of genetic material from the environment (virus capsids and nucleocapsids), support for long and short range cargo transport (e.g. through interaction of kinesin and dynein with microtubules), whole cell movements (flagella), cell division (microtubules), membrane deformation and scission (BAR domain containing proteins, dynamin), bacterial colony cohesion (type IV pili), force generation for cell crawling (through F-actin polymerisation), and so on.

In addition to the helical protein polymers which form naturally, some proteins can, under certain condition, form helical assemblies, which can be exploited for structural determination using Electron Microscopy. As examples, we can cite the human erythrocyte band 3 membrane domain (Yamaguchi et al. 2010) various ATPases (Pomfret, Rice, and Stokes 2007) or the nicotinic acetylcholine receptor (N Unwin 1993).

At the scale of organisms, it is again not uncommon to find helical motifs, including some bacterial cells (**figure 1.1D**), parts of plants (**figure 1.1E**), the placement of scales on a pine cone or the left-handed helix of the narwhal tusk.

Flexibility of helical protein polymers

Despite the usual terminology used to describe helical bio-polymers, and in contrary to what the gallery of EM reconstructions shown in **figure 1.1C** would tend to suggest, biological protein helices are in reality never truly helices, but always approximate the helical symmetry (which will be defined below) to a certain extent. Helices that respect the helical configuration more are referred to as ‘rigid’ or ‘regular’ helices whereas those that respect helical configuration less are referred to as ‘flexible’ or ‘irregular’. To further clarify the terminology, we must say that these definitions are not equivalent to “homogeneous” and “heterogeneous” which are often used in EM. For example, a sample of protein helices can be heterogeneous (e.g. if it contains two types of polymers formed by different proteins) and be composed of only rigid helices.

As the Thesis title suggests, we will be dealing with rather flexible helices and thus it is worth introducing here how common, or uncommon, flexibility in protein helices is, and what kind of flexibility, and also heterogeneities, can be usually observed (**figure 1.2**). When looking over the literature, finding very rigid examples of protein helices is more an exception than a rule. The most popular example of a regular helix in EM is Tobacco Mosaic Virus (TMV; **figure 1.2A**), which can even be used as a very accurate "ruler" for determining the exact magnification of an EM image. The exceptional regularity of this structure is well illustrated by the fact that it has been used to push the resolution of 3D helical reconstructions from EM images to unprecedented limits, at different times in EM history (Jeng et al. 1989; Ge and Zhou 2011). Moreover, if one looks at the first highest resolution EM structures of helical specimens deposited in the Electron Microscopy Data Bank (EMDB : http://www.ebi.ac.uk/pdbe/emdb/), reconstructions of TMV appear three times (**table 1.1**). Based on the resolution criteria indicated in **table 1.1**, which can be a good indication of very regular structures, we can search for other helices with a comparable rigidity as TMV. Two of these are artificially formed helices which often prove to be exceptionally symmetric, as they result from the "folding" of a perfect 2D crystal on a cylinder, and which are often explicitly named "tubular crystals" (Atsuo Miyazawa, Fujiyoshi, and Unwin 2003; Nigel Unwin and Fujiyoshi 2012). If we restrict ourselves to helices formed naturally, we find two structures of bacterial flagella (Yonekura, Maki-Yonekura, and Namba 2003; Maki-Yonekura, Yonekura, and Namba 2010) and one of F-Actin (Fujii et al. 2010). This is surprising as both of those polymers are known to be rather flexible (**figure 1.2B**), a quality which is required by their functions.

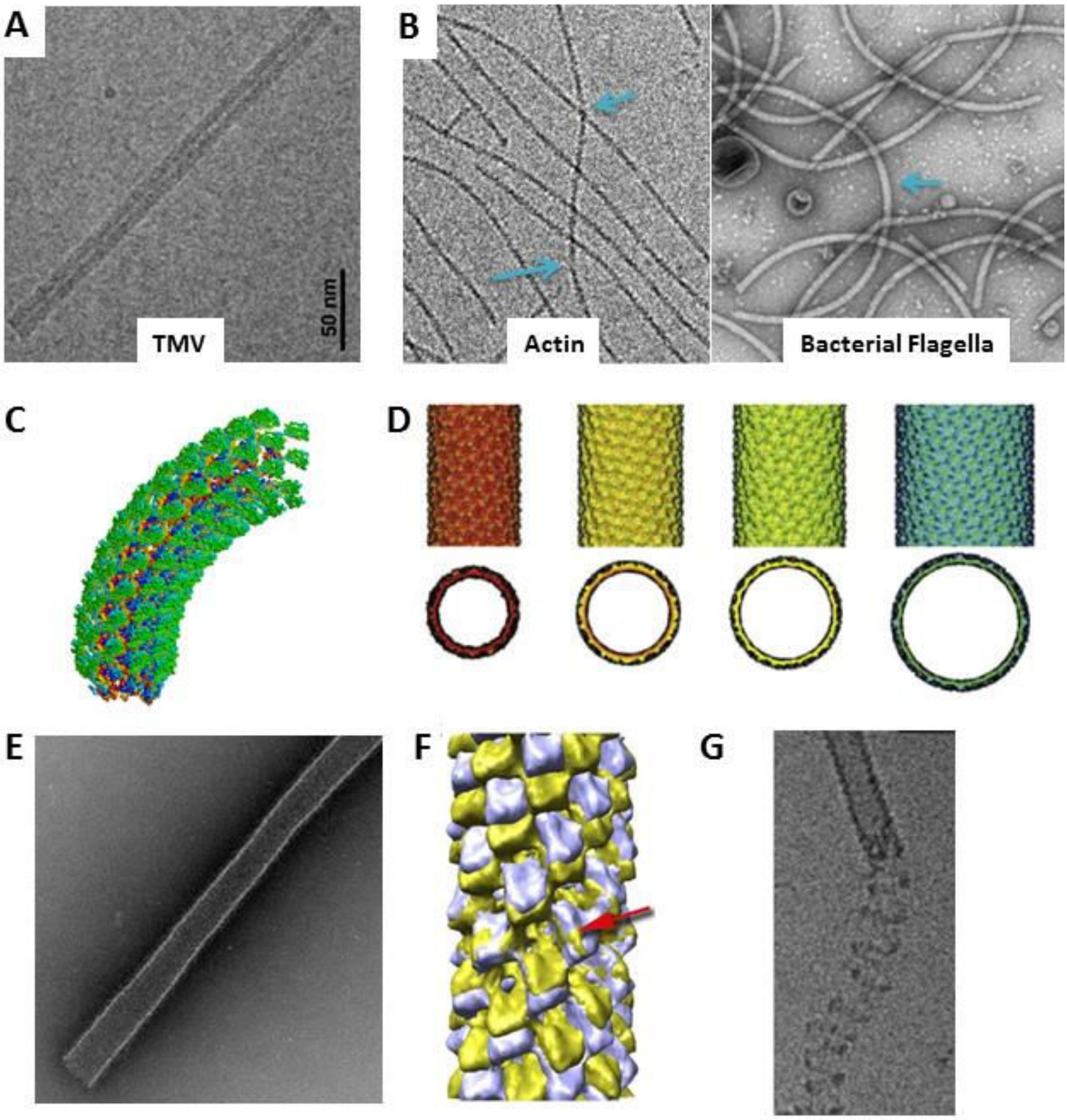


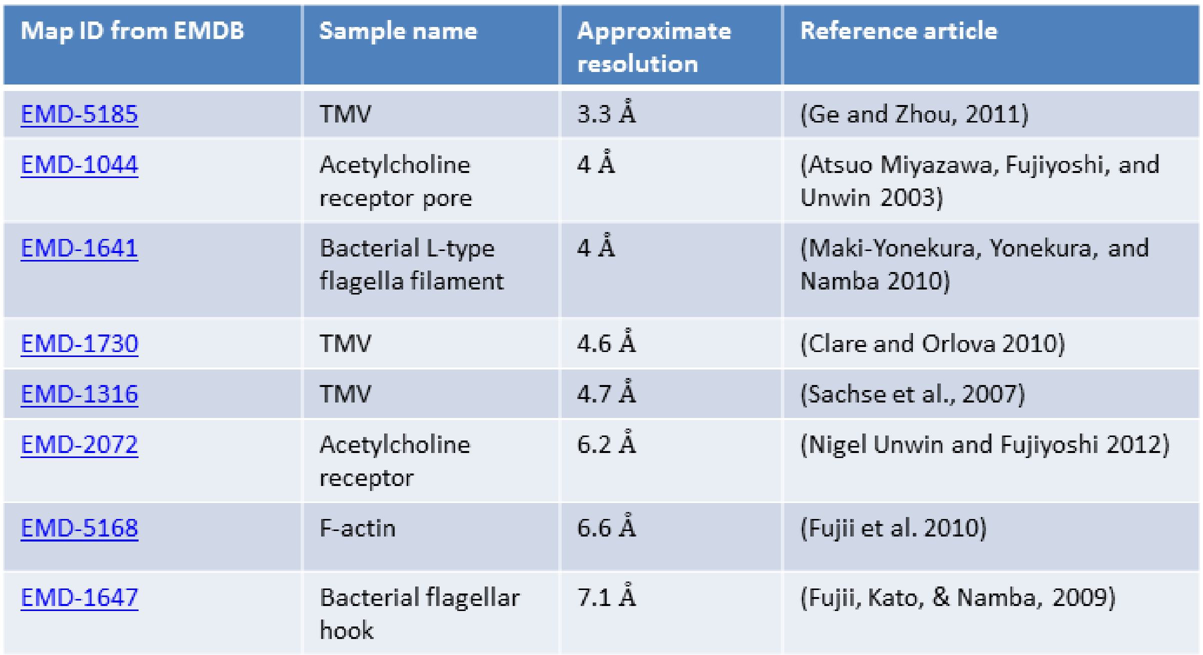
Figure 1.2 : Flexibility, rigidity, and heterogeneities of helical protein polymers illustrated by real examples

Most of the biological helical polymers are not as rigid as TMV (A). Indeed, frequently, bending of the helical axis is observed (**B**), which might be required for the polymer function as for example the model of the flagella suggest (**C**). The blue arrows in **B** indicate zones where this type of flexibility is particularly well observable. Diameter variability is also frequently seen, either among different filaments, as the 3D reconstructions shown in **D** and obtained from different filaments in the same sample illustrate, or within a single helix (**E**). In **F**, two reconstructions of similar diameter but different geometrical relationships between subunits, are superimposed. The red arrow indicate a subunit which has been aligned between the two reconstructions. In **G**, a nucleocapsid of Ebola virus illustrate how helical axis bending, variability of diameter and of subunits arrangement can be found in a single filament. Images were adapted from : **A** (Sachse et al., 2007) ; **B** (Resch et al, , 2002 ; Trachtenberg et al. 2005,) ; **C** (Samatey et al., 2004); **D** (Parent et al., 2012) ; **E** (Lata et al., 2008) ; **F** (Wang et al., 2006) ; **G** (Bharat et al., 2012)

This first type of flexibility, illustrated in **figure 1.2B,** is long or middle range bending of the helical axis, which is very common among filaments (Trachtenberg, Galkin, and Egelman 2005; Resch et al. 2002). To illustrate this type of flexibility more clearly, **figure 1.2C** shows the model explaining how the bacterial flagella is used to generating force for motility, which requires a strong helical axis bending (Samatey et al. 2004). A second type of variability/flexibility which we will encounter is variation of diameter. Variable diameters can be found either within a single helix or among different filaments, as illustrated in **figure 1.2D** and **E** (Lata et al. 2008; Parent et al. 2012). We can remark that variability of diameter was observed frequently for viral capsids like Marburg virus or Ebola (Bharat, Noda, et al. 2012). As a third type of variability, **figure 1.2F** illustrates the fact that, in some cases, even without large changes in the helix diameter, several helical states (relative position of the subunits) can be observed. This is a very common type of flexibility: it has been shown and debated for Actin filaments (E H Egelman and DeRosier 1992; Fujii et al. 2010), viral helical protein polymers (E H Egelman et al. 1989; Bhella, Ralph, and Yeo 2004), and other helical structures (Y. A. Wang et al. 2006). Finally, **figure 1.2G** illustrates the fact that often different types of flexibility coexist in a sample: in this example, we can see on a single filament a far from straight helical axis, a variation of diameter (the bottom portion has a larger width), and clearly the coexistence of several type of interactions between subunits.

To conclude this part, we have seen how common helical symmetry is within biological protein polymers, and how commonly those assemblies exhibit different types of flexibility. Thus we understand how important it is to use and develop methods for reconstructing the structure of these types of biological assemblies, which is the central point of this thesis work.

Before giving a point of view on what could be the reason why helical symmetry is so popular in biology, we need to introduce briefly the terminology which will be used throughout this manuscript to describe helical structures and the convention used for their orientation description.



Map ID from EMDB	Sample name	Approximate resolution	Reference article
EMD-5185	TMV	3.3 Å	(Ge and Zhou, 2011)
EMD-1044	Acetylcholine receptor pore	4 Å	(Atsuo Miyazawa, Fujiyoshi, and Unwin 2003)
EMD-1641	Bacterial L-type flagella filament	4 Å	(Maki-Yonekura, Yonekura, and Namba 2010)
EMD-1730	TMV	4.6 Å	(Clare and Orlova 2010)
EMD-1316	TMV	4.7 Å	(Sachse et al., 2007)
EMD-2072	Acetylcholine receptor	6.2 Å	(Nigel Unwin and Fujiyoshi 2012)
EMD-5168	F-actin	6.6 Å	(Fujii et al. 2010)
EMD-1647	Bacterial flagellar hook	7.1 Å	(Fujii, Kato, & Namba, 2009)

Table 1.1 : List of the highest resolution structures of helical protein polymers solved by EM, as given by the Electron Microscopy Data Bank and sorted by resolution criteria

Note that other structures solved to a better resolution than 7.1 Å (the last of our list) might not be present in this table : this simply would reflect the fact that they were not deposited in the EMDB or that their resolution was not clearly indicated and thus not included in the sorting results made by the EMDB. This table is only indicative, and not supposed to be exhaustive.

Terminology of helix description

A continuous helix is characterized by a radius r and a pitch P (**figure 1.3A,** left). Biological helices are discontinuous, and the simplest case, to which we will restrain ourselves, is similar to a continuous helix on which subunits would be placed at a regular interval (**figure 1.3A**, right). The angle around the helical axis formed by successive subunits is called the angular rotation between subunits (ΔΦ, also written Δphi) and the distance along the helical axis between two subunits is the axial rise (Δz). In many biological helices, the number of subunits per turn (=360/ ΔΦ) is not an integer, which implies that the real repeat of the helix along z (called c), contains several turns (usually noted t) and subunits (noted u).

In the images of filaments what we usually analyze, the helical axis lies near to the plane of the EM support (carbon or ice), so that we see the helices from the "side" (along y in **figure 1.3B**). In this manuscript, we will refer to the rotation out of the plane of the support (and perpendicular to the helix axis) as the "out-of-plane" angle of the helices, the rotation around the helical axis as the "on-axis" rotation, and the rotation around the viewing axis as the "in-plane" rotation (**figure 1.3B**).

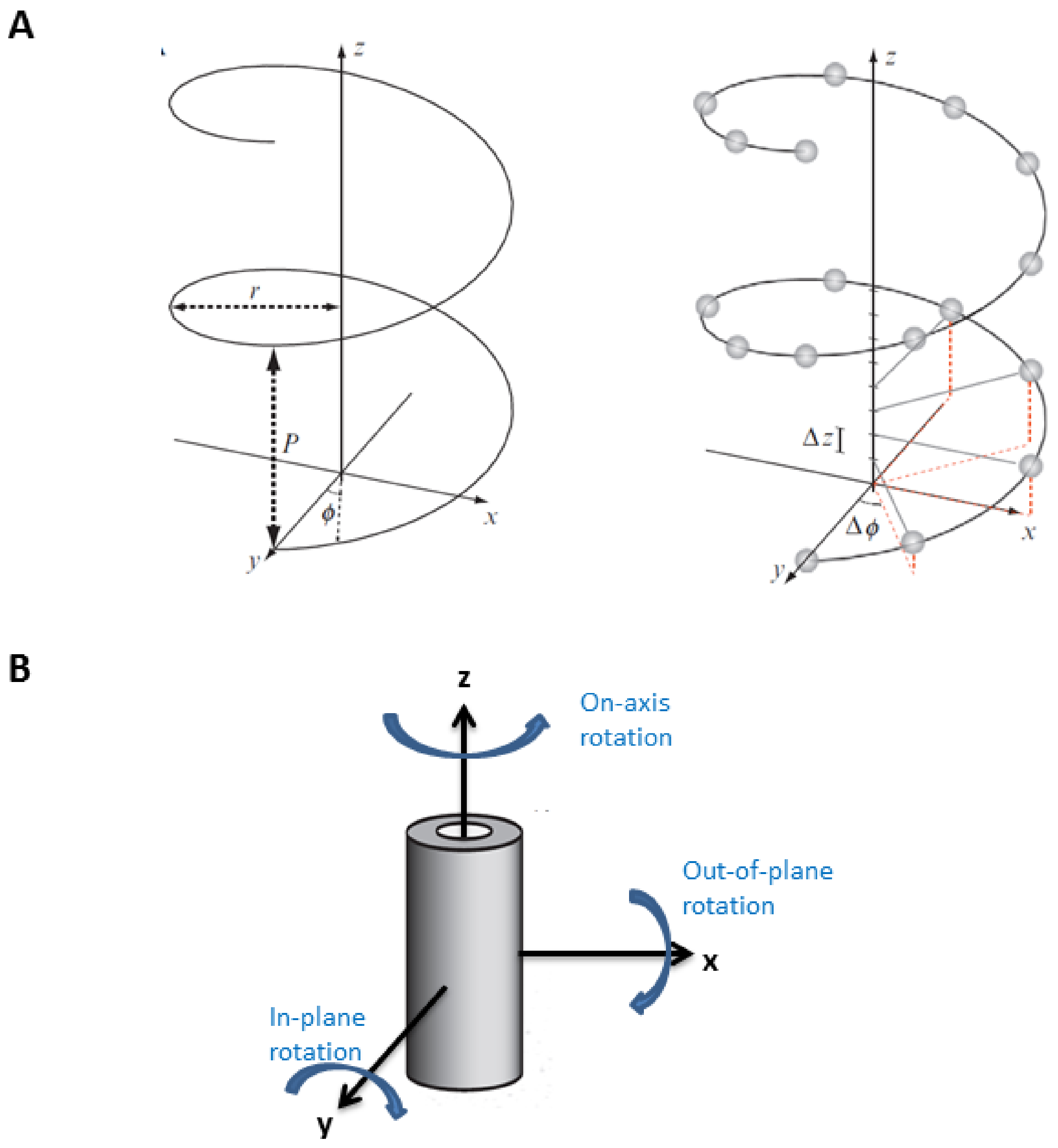


Figure 1.3 : Real-space description of an helix and conventions used for naming rotations

A continuous helix (**A**, left) as well as a helix with subunits represented by small spheres (**A**, right), is schematically represented. Adapted from (Diaz, Rice, & Stokes, 2010). The notations used are : r = radius of helix ; P = pitch of helix ; Δz = axial rise, or incremental translation between subunits ; ΔΦ = angular rotation between subunits. Here, the helix contain an exact repeat of eight subunits in one turn. Thus, here, the repeat of the helix, *c*, is equal to the pitch P. The helix axis is aligned with z (as in **B**). In **B**, schematic representation of the conventions used throughout this manuscript for angles of rotation. In this scheme, y is the viewing direction, so that zx is the plan in which the 2D projection will be generated by the microscope.

An evolutionary point of view?

Helical symmetry is the simplest symmetry to build

Due to their wide and sophisticated range of functions, helical polymers are often seen as resulting from an elegant and complex biological design. Actually, helical symmetry is the simplest form that one can generate from an ensemble of “building-blocks” (here proteins) which are in contact with each other by a defined interaction interface (or “interaction rule”). This is illustrated in **figure 1.4** using the Respiratory Syncytial Virus nucleoprotein crystal structure (Tawar et al. 2009) as the building-block (**figure 1.4A**). A very specific interaction rule, involving two-fold symmetry, is shown in **figure 1.4B**, and this specific rule generates a symmetrical dimer. Another very specific interaction, involving a single rotation between the subunits of a number of degrees that is a divisor of 360, is shown in **figure 1.4C** (left) and will give raise to a closed ring of subunits with C-fold rotational symmetry (**figure 1.4C**, right). By contrast, two completely arbitrary interactions, involving some translation and rotation, are shown in **figure 1.4D left** and **1.4E left**. When this interaction rule is applied to more building blocks, a helical polymer is formed (**1.4D** and **1.4E, right**). In general, the repeated application of an interaction rule involving rotations and translations between two subunits would lead to an infinite helix. So, the structure of helical polymers can be seen to reflect the simplest mode of interaction between identical copies of the same protein.

Thus, one could almost say that evolutionary little “effort” was required to create helical polymers and that on the contrary it potentially had to fight against those. Indeed, any genetic mutation that would lead to the self-assembly of a protein with an interaction rule including a translation and a rotation such as described above could have dramatic consequences at the cellular and organism level. The potential deleterious effects of the formation of non-wanted helical polymers are illustrated by many diseases caused by fibrillar aggregates, including Alzheimer’s, Creutzfeldt–Jakob, and Parkinson’s diseases.

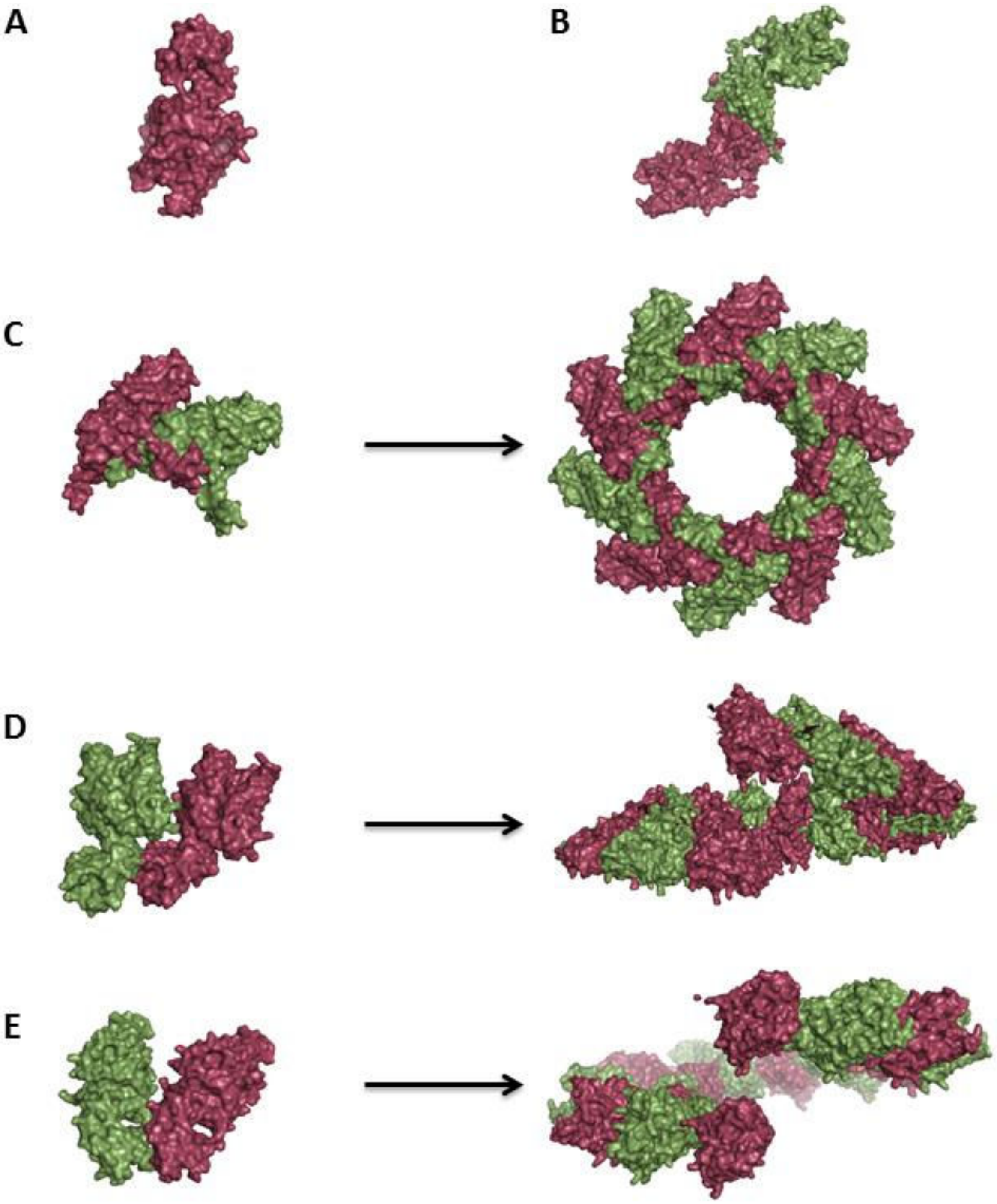


Figure 1.4 : Helical symmetry is the simplest symmetry to build
For the purpose of this fully artificial demonstration, a monomer of RSV nucleoprotein crystal structure (Tawar et al., 2009) was chosen (**A**). In **B**, **C**, **D** and **E**, each time a different interaction rule between two monomers (showed alternatively purple and green) was introduced. When a very specific interaction rule is applied, for example involving a two-fold symmetry (**B**) or a rotation of a divisor of 360 ° (**C ;** 36 °), the assemblies resulting from adding more subunits show a symmetrical dimer (**B**) or a ring with C10 symmetry (**C**). In contrast, when any kind of combination of a translation and a rotation are used to create a contact between subunits, the resulting assembly will form a helix (**D**,**E**).

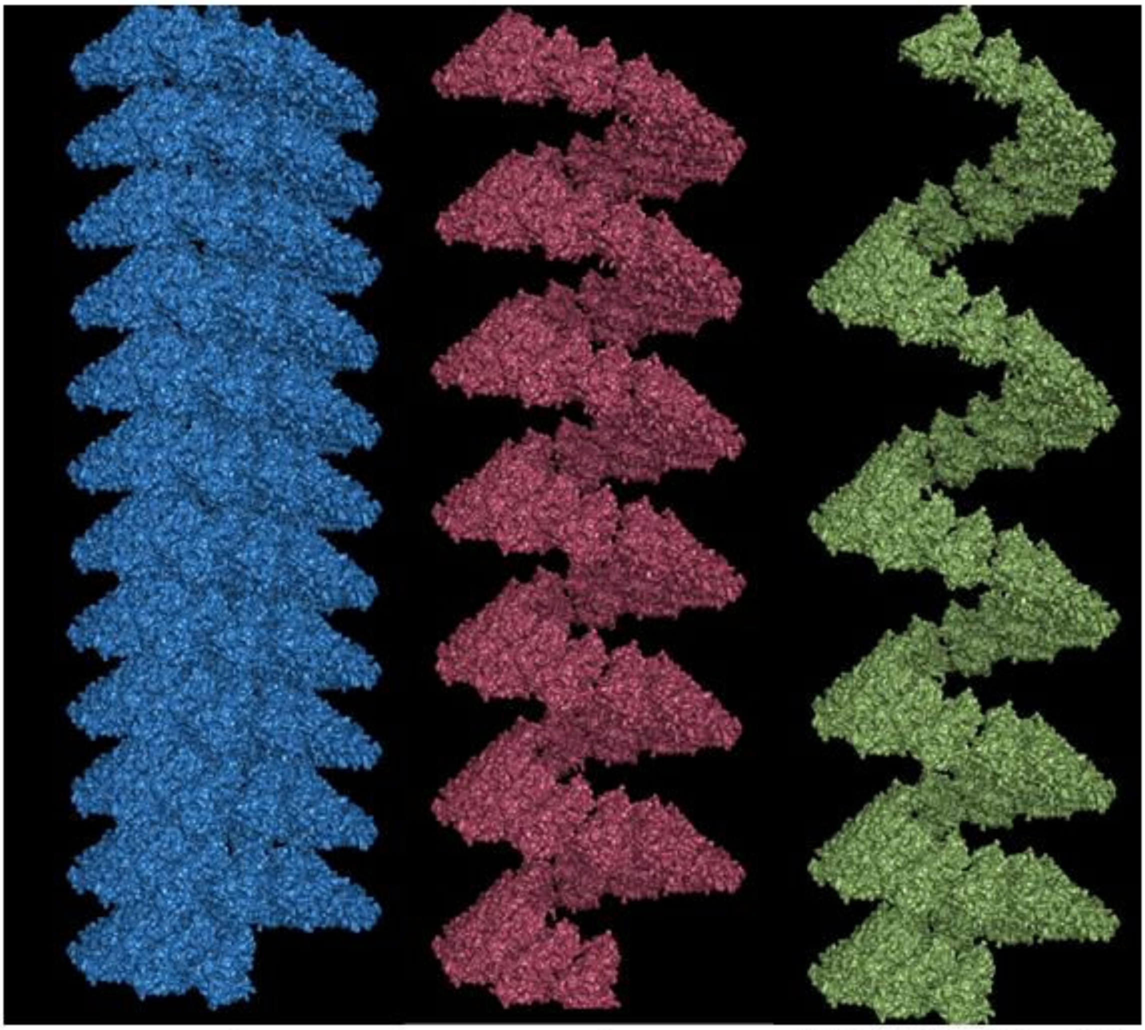
Some advantages of helical symmetry

In addition to its simplicity, helical symmetry offers many advantages: we cannot review all of them here but mention those that are relevant in the context of our biological samples, which will be described below.

Helices can form very long, sometimes huge, assemblies; no other protein assembly has dimensions comparable to naturally occurring helices like microtubules or actin filaments. One advantage is that this permits them to confer particular mechanical properties to entire cells and, thus, eventually tissues, like elasticity or resistance to compressive and tensile forces. Another advantage is that it makes those helices able to interact with substrates of a very large size. For example, Titin, the largest known protein, indirectly interacts with Actin filaments in muscle sarcomeres. In many viruses, such as TMV shown in **Figure 1.2**, the complete viral genome, which can exceed sizes of several micrometers, is covered by a protein, forming very large helical structures.

Another advantage of helical symmetry which we would like to illustrate is that very small variations in the inter-subunit interactions can lead to huge variations in the morphology of the entire assembly, thus potentially conferring very different properties, like the availability of binding sites for interacting partners or its compaction state. **Figure 1.5** illustrates this idea with a completely artificial example (using again the RSV nucleoprotein crystal structure as a subunit), and shows how strongly an only small inter-subunit interaction variation affects the whole morphology. The difference between the helix shown on the left and the helix in the middle is that the translation along the helical axis between successive subunits is increased by 4 Å: at the level of two subunits, this change may be regarded as very subtle, but the effect on the global structure is enormous. If it would be a real object, this change may affect its ability to interact with other proteins or its flexibility for example. On the right part of the figure, we see how even more subtle changes in the interaction interface between subsequent pairs of subunits will dramatically affect the whole assembly if they are "propagated" along the helical axis.

Thus, by acting very locally, and subtly, on a helix, by changing the interaction interface between subunits, for example via the environment of the helix (pH, ionic force, etc..) or through the binding of an interacting protein, one can confer very different properties to the whole object.



Δz = 4.9 Å

Δz = 8.9 Å

Initial (bottom) Δz = 4.9 Å
Then incremented by 0.2 Å
for each new subunit

Figure 1.5 : Subtle changes in subunits interaction can affect the global morphology of a helix dramatically

An **artificial** helix made of the RSV nucleoprotein crystal structure (Tawar et al., 2009) was built with an axial rise of 4.9 Å (**left, blue**). In this helix, the subsequent turn are so close that a large molecule like a protein could not access the inner hollow cavity of the helix. If one increases the translation between subunits of only 4 Å (i.e. less than the pitch of an alpha helix !), the morphology of the resulting helix (**middle, purple**) changes completely and offers new potential interaction interfaces for putative binding partners, with in particular the inner hollow cavity of the helix accessible. The structure shown on the **right (green)** shows that an helix can very rapidly undergo huge morphological rearrangements by "propagating" along the helix a change in the inter-subunit interface. Here, starting from the bottom with an axial rise of 4.9 Å, this parameter is then incremented by 0.2 Å for each new subunit added, so that the structure evolves from the "compact" state shown on the left to an even more extended state that the helix shown in the middle.

As already mentioned, the main focus of this thesis is to study existing image processing methods to obtain 3D reconstructions from EM images of flexible helical polymers, to implement these methods into a useable processing pipeline, and eventually to add new tools to the existing ones. The starting points of this work were biological questions, and biological objects, which fulfilled most of the characteristics of flexible helices that we have described above. So, before going further in the methodology, we would like to introduce those questions, and those structures: the nucleocapsids of negative strand RNA viruses.

Nucleocapsids of negative strand RNA viruses

Generalities on negative strand RNA viruses

Negative strand RNA viruses are enveloped viruses with an RNA genome in the opposite sense of that of mRNA. They possess either a single viral RNA molecule (Mononegavirales order) or a segmented genome. The families of the Rhabdoviridae (Rabies virus, **Vesicular Stomatitis Virus**), Paramyxoviridae (**Measles virus** : subfamily Paramyxovirinae, Respiratory Syncytial Virus : subfamily Pneumovirinae ; Nipah virus; Mumps virus), Filoviridae (Ebola virus) and Bornaviridae (Borna Virus) belong to the order of the Mononegavirales. The genome of the viruses belonging to the family of Arenaviridae (2 RNA segments, Lassa Virus), Bunyaviridae (3 segments, Rift Valley Fever Virus) and Orthomyxoviridae (7-8 segments, Influenza virus) is composed of several single-stranded RNA molecules of negative polarity. The Measles Virus (**MeV**) and the Vesicular Stomatitis Virus (**VSV**) are at the heart of the presented work.

Negative strand RNA viruses have variable morphologies (**figure 1.6**), can infect very different types of host ranging from plants to mammals, and cause many human pathologies.

The Influenza virus and the Respiratory Syncytial Virus (RSV) can cause severe respiratory tract disease. Rabies virus, Nipah virus and some Bunyaviridae are responsible for severe encephalitis. Other viruses from this family, including Ebola or Lassa virus can trigger hemorrhagic fever.

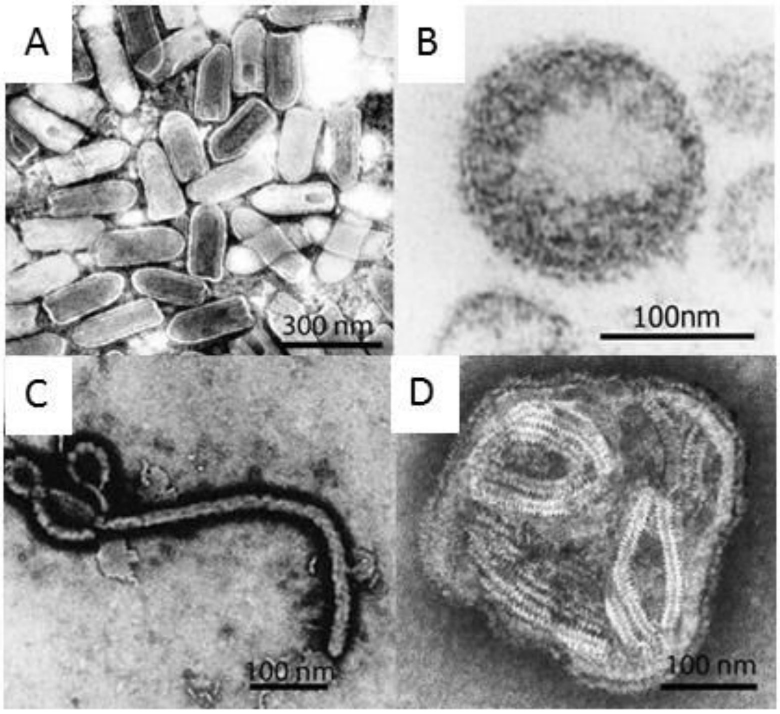


Figure 1.6 : Electron microscopy images of viral particles of *Mononegavirales*
A : Vesicular Stomatitis Virus (*Rhabdoviridae*)
B : Borna virus (*Bornaviridae*)
C : Ebola virus (*Filoviridae*)
D : Parainfluenza virus (*Paramyxoviridae*)

MeV is known to cause Measles, a disease which is characterized by prodromal symptoms of fever, cough, coryza and conjunctivitis followed by the appearance of a generalized maculopapular rash (red plaques on the skin). Measles was estimated to cause more than 400,000 deaths in 2004, almost half of which were in sub-Saharan Africa, and it continues to cause outbreaks in communities with low vaccination coverage (Moss and Griffin 2006). Deaths from measles are mainly due to an increased susceptibility to secondary bacterial and viral infections, which is attributed to a prolonged state of MeV-induced immune suppression.

VSV primarily affects rodents, cattle, swine, and horses and can cause mild symptoms upon infection of humans and other species. In the former, it causes a benign disease characterized by vesicular lesions on the mouth, the tongue, the udder and the hoof of the animals. In contrast to Rabies, which causes fatal disease in humans and animals, VSV is not dangerous to humans. It can thus be easily studied in the laboratory, while retaining the advantage that it shares many common structural and functional characteristics with Rabies.

More generally, it constitutes an excellent model for the replication and transcription of Mononegavirales because it is relatively simple: it carries only 5 genes for which the expression regulation signals are less complex than for the other viruses of this order.

Role of Mononegavirales nucleocapsids

For replication and transcription

During infection, after host cell entry, the first activity of these viruses is the transcription of the negative viral RNA into messenger RNA (mRNA). From a structural point of view, the RNA of negative strand RNA viruses is never naked, neither in the virions nor in the infected cells, but always in a ribonucleoprotein complex. The major protein of this complex is the nucleoprotein (N) which tightly and regularly encapsidates the viral RNA, forming a helical N-RNA nucleocapsid. In all Mononegavirales, two other proteins are also associated with the nucleocapsid : the viral polymerase (RNA-dependent RNA-polymerase; L) and its cofactor, the phosphoprotein (P) (Curran, Pelet, and Kolakofsky 1994).

These N-RNA nucleocapsids provide helical templates for viral transcription and replication (Ruigrok, Crépin, and Kolakofsky 2011). One crucial question concerning this mechanism, schematically represented on **figure 1.7**, is how the polymerase can access the viral RNA for performing its activity.

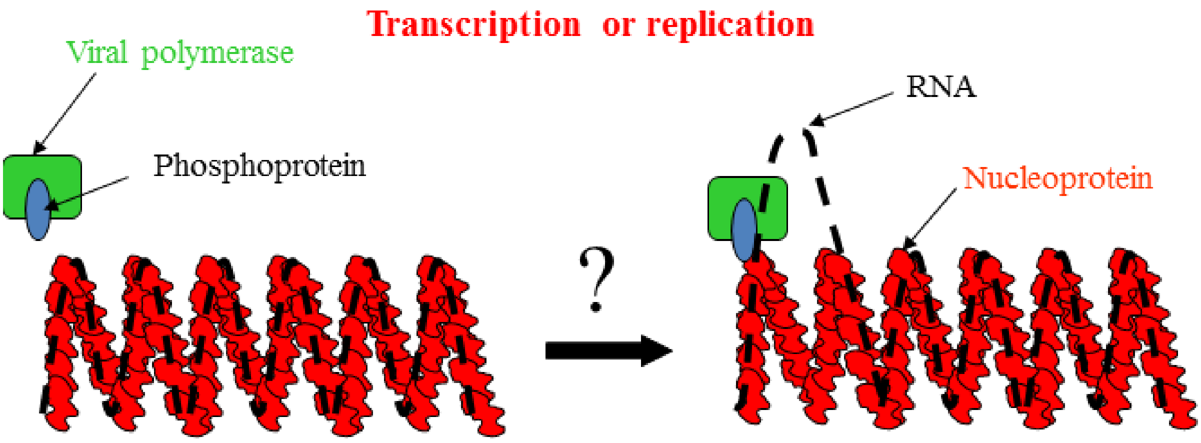


Figure 1.7 : Schematic representation of the interaction between the nucleoprotein-RNA template and the viral polymerase during transcription or replication
Transcription and replication of the viral RNA are initiated by an interaction between N and the polymerase complex, composed of the phosphoprotein (P) and the RNA-dependent RNA-polymerase. The mechanisms by which this interaction frees the RNA molecule at least transiently for polymerase activity are not well understood.

This question appeared when it was shown how strongly the nucleoprotein protects RNA from its environment (Iseni et al. 1998), although part of the bases were shown to be accessible to chemical probes, thus to the solvent (Baudin et al. 1994; Iseni et al. 2000). Moreover, recent crystallographic studies of Mononegavirales N-RNA complexes have shown how the RNA is buried in a nucleoprotein cleft (**figure 1.8**). Although the nucleocapsids are far too big and flexible for crystallization, recombinantly expressed nucleoproteins can also encapsidate short cellular (e.g. bacterial) RNAs that close up into N-RNA rings. In the rings, N-RNA is sterically constrained in a biologically inactive form, but the rings have the advantage of being rigid enough for X-ray crystallography. As an exception, Borna Disease Virus (BDV) nucleoprotein crystallized as a tetramer in the absence of RNA (Rudolph et al. 2003)(not shown on **figure 1.8**). The other three available Mononegavirales nucleoprotein structures, those of rabies virus, VSV, and RSV, crystallized in the form of recombinant N-RNA rings containing 10 or 11 N-protomers (Albertini, Wernimont, et al. 2006; Green et al. 2006; Tawar et al. 2009).

These proteins show two main N-terminal (N-ter) and C-terminal (N-ter) domains, mostly composed of alpha helices (**Figure 1.8,** top row: C-ter is red to yellow and N-ter is green to blue). The subunits in the three N-RNA rings make extensive contacts between their

C-ter-domains, with N-ter extensions (blue "arms" on the figure) reaching to the back of the neighboring N to make an additional domain exchange contact. For RSV and VSV, the C-ter extension also goes to the back of the N-subunit at the other side for additional contacts. The result of those extensive interactions is a very stable N-RNA structure. The C-ter extension of RSV N is very flexible and partially invisible in the atomic structure, and the homologous domains of Measles and Sendai virus are intrinsically disordered and bind to P (Longhi et al. 2003; Houben et al. 2007). In the three structures, the RNA binds in a positively charged (blue) cleft between the N-ter and C-ter domains (**Figure 1.8**, bottom row). In the figure, this channel appears as a hole in the nucleoprotein (RNA has been removed in this representation).

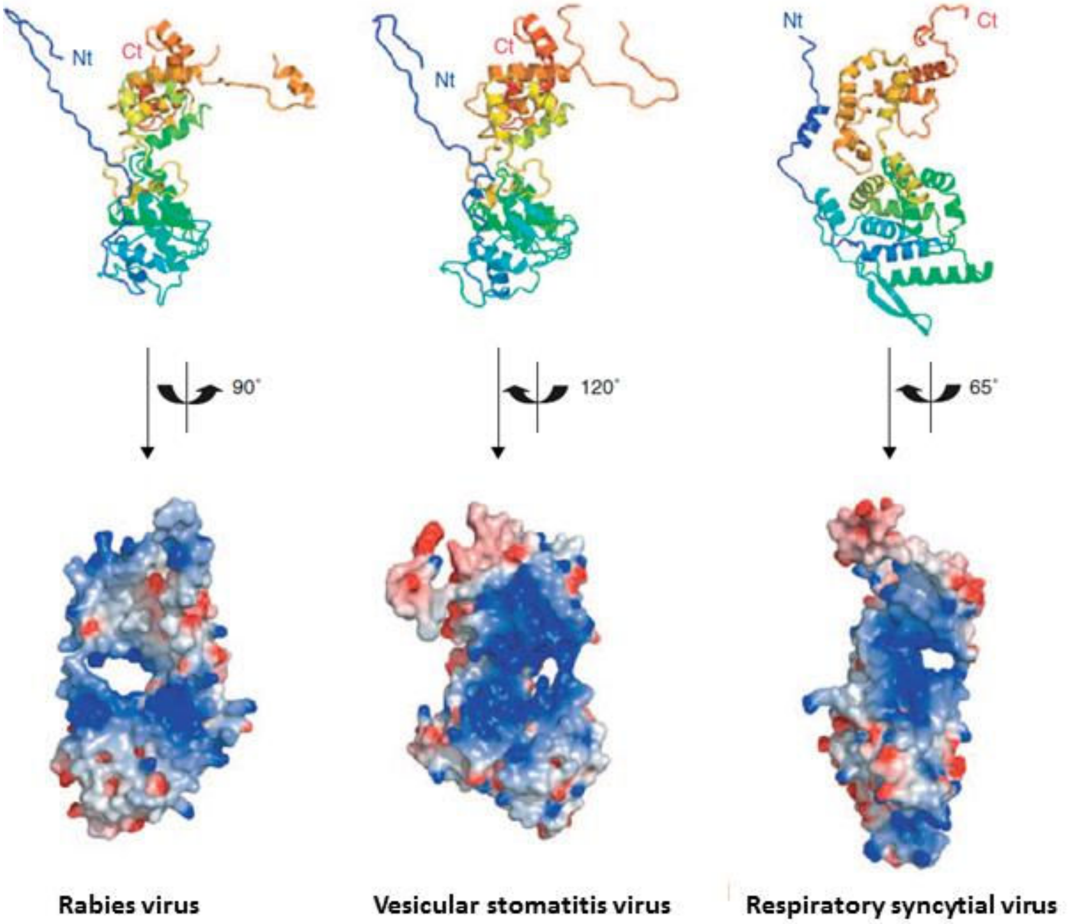


Figure 1.8 : Structures of nucleoproteins of *Mononegavirales*

The cartoon representation in the top raw uses a color coding from purple (N-terminus) to red (C-terminus), whereas the bottom images are turned to show the largest positively charged surface (blue) versus negatively charged (red). The RNA was removed for the Rhabdovirus and RSV structures. The round opening between the N-terminal and C-terminal domains shows the basic channel through which the RNA is threaded. The PDB codes for the structures are 2wj8 for RSV, 2GTT for rabies virus and 2GIC for VSV nucleoprotein.
Figure reproduced from (Ruigrok, Crépin, and Kolakofsky 2011)

As a protective stable scaffold for viral RNA

As mentioned, the nucleocapsids of Rhabdoviruses and Paramyxoviruses are very stable structures. They can support high salt concentrations and high gravity forces during long ultracentrifugation (Blumberg et al. 1984; M. H. Heggeness 1980). The nucleoprotein-RNA interaction is very strong : for VSV, it has been shown that the interaction was resistant to a denaturing treatment with 8 M urea (Iseni et al. 1998). Altogether, this confers to the nucleocapsid a protective role for the viral genome:

- The nucleoprotein protects RNA from digestion by RNases which could be used by the cell as a defense mechanism (Iseni et al. 2000).
- It ensures that the viral RNA will not form secondary structures by interacting with itself (Baudin et al. 1994).
- It is also necessary to avoid the formation of double-stranded RNA between the viral genome and the viral messenger RNA during their synthesis.

Structures of nucleocapsids

Although the isolated nucleocapsids of negative strand RNA viruses are all composed of a viral nucleoprotein (which basic organization does not differ very much between various virus families) and the viral genome, their morphology is very variable. The **figure 1.9** shows EM micrographs of nucleocapsids isolated from virion or formed upon heterologous expression of the nucleoprotein. The recombinant nucleocapsids have a similar morphology as the viral nucleocapsids, and with a similar stoichiometry of nucleoprotein/nucleotides. The **figure 1.9** shows that the nucleocapsids of Paramyxoviridae (Sendai, Mumps and Nipah) shares morphological similarities: they are relatively compact and have a “herringbone” appearance. In comparison, the nucleocapsids of Rhabdoviridae forms a loose coil (Rabies is shown here, but the one of VSV is very similar) and for Filoviridae (Marburg) they appear even less compact. Influenza virus nucleocapsids present a very different morphology, with a supercoiled structure with a terminal loop (Michael H. Heggeness et al. 1982).

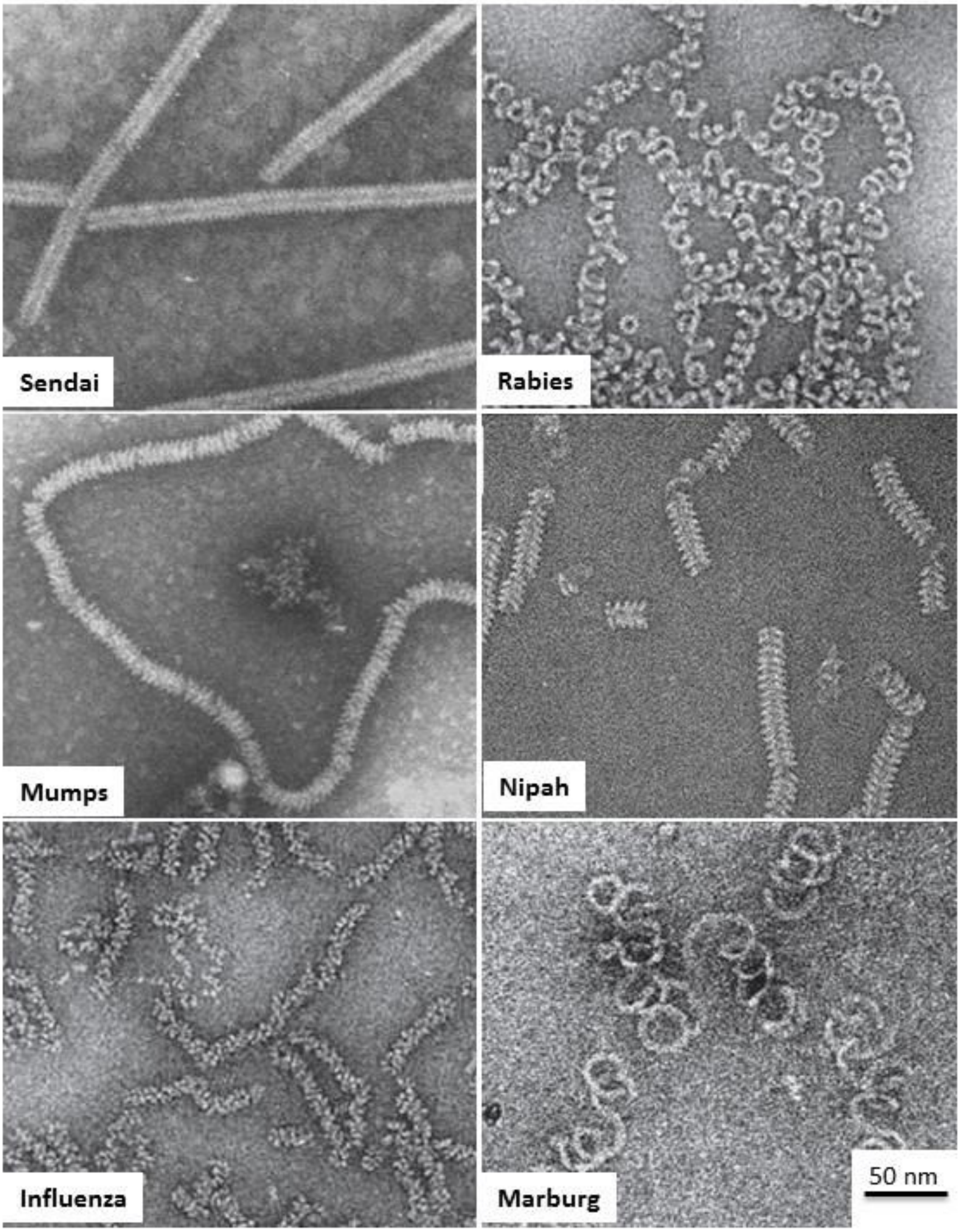
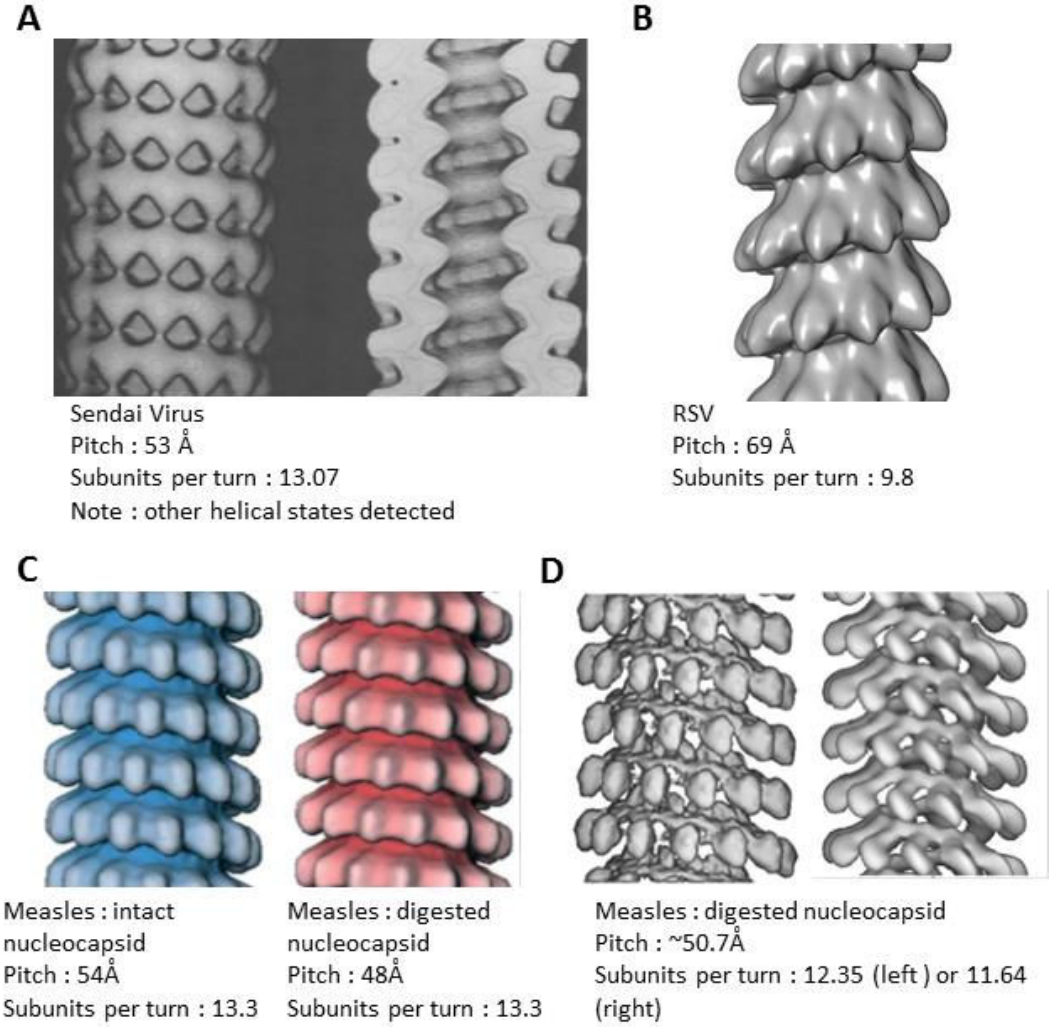


Figure 1.9 : Electron micrographs of negative strand RNA viruses nucleocapsids
All samples were negatively stained. The nucleocapsids of Sendai, Rabies, Mumps, and influenza virus were isolated from virus, whereas the Nipah and Marburg nucleocapsids were produced in recombinant form. All micrographs have the same magnification indicated by the bar underneath the Marburg virus nucleocapsids.
Images were taken from (Finch and Gibbs 1970) and (Ruigrok, Crépin, and Kolakofsky 2011)

Due to the highly flexible nature of these assemblies, there are only a few three-dimensional structures obtained by EM of isolated helical nucleocapsids of negative strand RNA viruses described in the literature. From **figure 1.9**, we can understand that all the structures solved so far are from viruses belonging to the Paramyxoviridae (**figure 1.10**).

A low resolution reconstruction of Sendai nucleocapsids (E H Egelman et al. 1989) shows a arrangement of ~13 subunits per turn with a pitch of 53 Å (**figure 1.10A**). A structure of RSV nucleocapsid (Tawar et al. 2009) (**figure 1.10B**) indicate a relative higher pitch (69 Å) which explains the higher degree of flexibility of the corresponding isolated nucleocapsids. Other than those, only reconstructions of Measles nucleocapsid have been obtained, either in the intact form or digested form (see the next section for more details) (**figure 1.10 C, D**). In (Bhella, Ralph, and Yeo 2004) (**figure 1.10 C**), cryo-negative stain reconstruction showed that there is extensive conformational flexibility within these structures, ranging in pitch from 50 Å to 66 Å, while the number of subunits per turn vary from 13.04 to 13.44 with a greater number of helices comprising around 13.1 subunits per turn. They also showed that in the digested form, the pitch becomes shorter, ranging from 46 Å to 52 Å, while more helices have a twist of approximately 13.3 subunits per turn. In (Schoehn et al. 2004) (**figure 1.10 D**), a 12 Å resolution structure obtained by cryo-EM of the digested form of Measles nucleocapsid marks the highest resolution structure obtained so far for nucleocapsids. This structure show 12.35 subunits per turn, which is different from (Bhella, Ralph, and Yeo 2004), but only a small portion of the total amount of images were used to calculate the reconstruction (< 10%). Another structure, at much lower resolution (25 Å), showed a different symmetry (11.64 subunits per turn), but it is not clear if it is really present in the data or an artifact of the reconstruction method.



A

Sendai Virus
Pitch : 53 Å
Subunits per turn : 13.07
Note : other helical states detected

B

RSV
Pitch : 69 Å
Subunits per turn : 9.8

C

Measles : intact nucleocapsid
Pitch : 54Å
Subunits per turn : 13.3

Measles : digested nucleocapsid
Pitch : 48Å
Subunits per turn : 13.3

D

Measles : digested nucleocapsid
Pitch : ~50.7Å
Subunits per turn : 12.35 (left) or 11.64 (right)

Figure 1.10 : Three-dimensional structures obtained by EM of isolated nucleocapsids of *Mononegavirales*

Surface representation and helical parameters of nucleocapsid reconstruction of :
A : Sendai ; negative staining (E H Egelman et al. 1989). Note that other helical states were also observed.
B : RSV ; negative staining (Tawar et al. 2009).
C : Measles (intact and digested nucleocapsids) ; cryo-negative staining (Bhella, Ralph, and Yeo 2004). Note that a continuum of variable helical parameters was described.
D : Measles (digested nucleocapsids) ; cryo-EM (Schoehn et al. 2004). Two reconstruction, one at 12 Å resolution (left), and one at 25 Å, were obtained.

We have noted the difference of morphology between Paramyxoviridae (except Pneumoviruses like RSV) and other negative strand RNA viruses, especially concerning the relative compactness of the former. Behind this observation, as well as behind the relatively well conserved number of subunits per turn (~ 13) in those viruses, there is a biological reason. SeV and MeV nucleoprotein subunits binds to six bases of RNA. It has also been demonstrated that there is an absolute requirement in both the respiroviruses and the morbilliviruses for the genome to be of a length that is a multiple of six bases (Calain and

Roux 1993). Furthermore, there is evidence that this requirement is more than a simple reflection of the N-RNA stoichiometry. Experiments with minigenomes have shown that both the genomic and anti-genomic promoters are bipartite (Kolakofsky et al. 2005). They consist of a 12 nucleotide region at the extreme 3' end of the nucleocapsid associated with the first two N subunits of the nucleocapsid. A second element consisting of a triplet repeat of hexamers (3' CNNNNN-5') exists downstream between bases 79 and 96, associated with the 14th, 15th and 16th N subunits. The position of this second element is such that in the nucleocapsid, due to their particular symmetry, we would find these elements on successive turns of the helix with the hexamer repeats in-phase with the N subunits. Mutations or deletions that affect the spatial relationship between these elements, or change the phase of the second element hexamers, are highly deleterious to mini-genome replication. It has been suggested, therefore, that these elements may be a "polymerase landing pad".

In contrast, the precise length of rhabdo/pneumo- virus genomes does not appear to be important; they are not subject to a hexamer (or any integer) rule (Pattnaik et al., 1995; Samal & Collins, 1996), and their genomic promoter is not bipartite. However this does not mean that the nucleocapsid (and its end part) must not adopt a particular symmetry in order for the complex polymerase-phosphoprotein to function optimally.

Targets for anti-viral drugs

Nucleocapsids of negative strand RNA viruses are unique structures in the biology of nucleic acids. In our case the nucleoprotein which covers the genome is necessary for the activity of the viral RNA-polymerase. The N-RNA complexes can thus be perfect targets for the development of specific antiviral molecules without toxic side-effects due to their unique belonging to the viral world. The active sites of the viral RNA polymerase, as well as protein-protein interactions like nucleoprotein-phosphoprotein, phosphoprotein- phosphoprotein, phosphoprotein -polymerase are also potentially good target for antiviral drugs. Developing molecules which would modify the helical characteristics of the nucleocapsid, thus hindering the progress of the polymerase on its template, is a promising research area for novel antiviral drugs. For such developments, the 3D structures and the understanding of the molecular mechanisms associated to the different components of the nucleocapsid are primordial.

Measles project : the biological questions

Measles nucleoprotein is composed of two main parts (**figure 1.11 A**), N_{CORE}, which contains the oligomerisation motif and the RNA binding motif (Karlin 2003), and a C-terminal region, $N_{TAIL.}$ The latter contains a short sequence (residues 489 to 506) which binds the viral phosphoprotein carrying the viral RNA polymerase (Longhi et al. 2003), and other cellular factors like hsp70 (Couturier et al. 2010). The molecular recognition element (MoRE) (residues 485–502) of the disordered N_{TAIL} interacts with the C-terminal three-helix bundle domain, XD, of P (residues 459–507) (Johansson et al. 2003) and thereby recruits the polymerase complex onto the nucleocapsid template (Bourhis et al. 2004).

As for other negative strand viruses, when recombinantly expressed, the nucleoprotein of MeV binds non-specifically to cellular RNA and is able to form nucleocapsid-like structures (Fooks et al. 1993; Spehner, Kirn, and Drillien 1991) which can be purified. Nucleocapsids containing the full-length nucleoprotein (MeVND) are very flexible (Bhella et al. 2002). One can however take advantage of the well-known sensitivity of nucleocapsids of negative strand RNA viruses to trypsin digestion (M. H. Heggeness 1980) and the increased rigidity of resulting digested nucleocapsids (**figure 1.11B**). This property enabled Dr Schoehn to provide a 12 Å resolution 3D reconstruction of MeV digested nucleocapsid (MeVD) by cryo-electron microscopy (Schoehn et al. 2004), as shown in **figure 1.10D, left**. Unfortunately, trypsin digestion removes precisely the domain of our main interest, i.e. the N_{TAIL}, which interacts in particular with the polymerase cofactor.

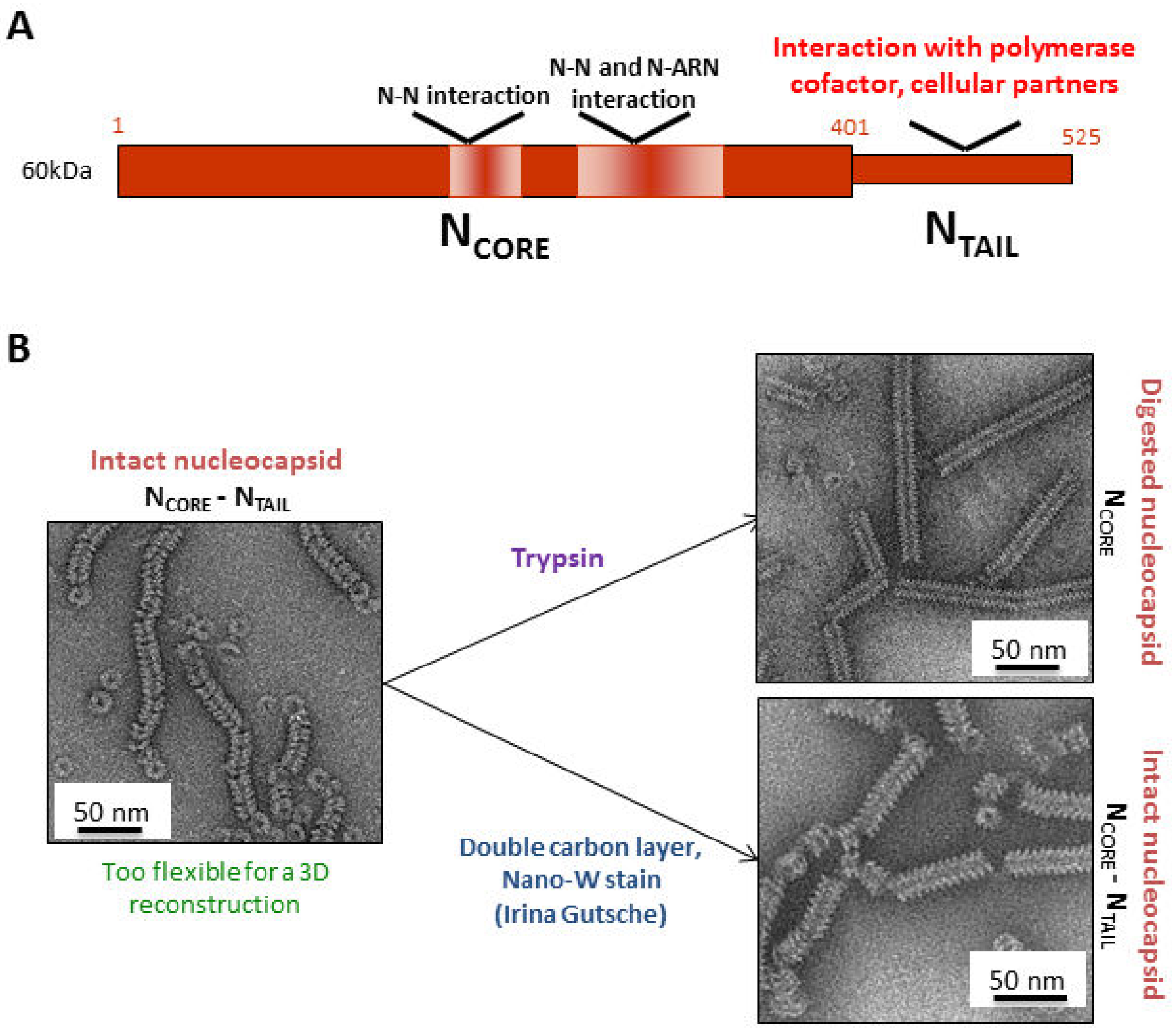


Figure 1.11 : What is the localization of the C-terminal domain of the nucleoprotein of Measles in the nucleocapsid ?
The general organization of Measles virus nucleoprotein N highlights the importance of the C-terminal domain N_{TAIL} (**A**). The flexible intact nucleocapsids (**B, left**) can be rigidified by digestion with trypsin which removes the N_{TAIL} domain or by use of a double carbon layer and Nano-W stain (**B, right**). A comparative structural analysis of both digested and undigested nucleocapsids by EM and 3D reconstruction may thus provide the localization of the NTAIL domain, absent from the 12 Å resolution cryo-EM reconstruction of digested nucleocapsid (Schoehn et al., 2004) .

At the time of my arrival at the UVHCI, Irina Gutsche had just set up optimal conditions for negative staining observation which rigidified the intact nucleocapsids (**figure 1.11B**). This opened up the possibility of getting EM images of intact and digested nucleocapsids under the same conditions and comparing the two corresponding reconstructions. Thus, the idea which constituted the first part of my thesis project was to acquire data sets of intact and digested nucleocapsids by negative staining and to process them to 3D reconstructions, which would thus eventually enable to localize the N_{TAIL} domain.

Concomitantly to this work, a novel important structural information appeared, and raised new questions : the atomic structure of the nucleoprotein of respiratory syncytial virus (RSV) was solved (Tawar et al. 2009). Belonging also to the Paramyxoviridae, RSV thus became the closest species to MeV with a known atomic structure of the nucleoprotein.

Intriguingly, in this structure, the location of the RNA groove is outwards, whereas it is inwards in X-ray crystal structures of N-RNA rings of other negative strand RNA viruses like rabies (Albertini, Wernimont, et al. 2006) and VSV (Green et al. 2006). Indeed, the lateral contacts between N confer to the RSV N-RNA ring an opposite curvature (**figure 1.12**). The internal position of the VSV RNA suggested by the crystal structure was confirmed in the virus particle (Ge et al. 2010) where the helical turns with the smallest diameter (at the tip of the bullet) have a very similar structure to those of the recombinant N-RNA rings. For Measles, although an attempt of RNA localization was done on the digested nucleocapsids (Schoehn et al. 2004) using labeling with cis-platinum and subsequent cryo-EM reconstruction, the obtained result was not completely clear concerning the orientation of the RNA molecule. Furthermore, we cannot completely exclude important rearrangements of the nucleocapsid after digestion.

It was therefore crucial to find out if the difference of RNA localization in the crystal structures (**figure 1.12**) is due to steric constraints in the ring or if it reflects the intrinsic difference in the corresponding nucleocapsids. This would have functional implications, in particular for the access of the polymerase to the RNA molecule, as it was discussed above.

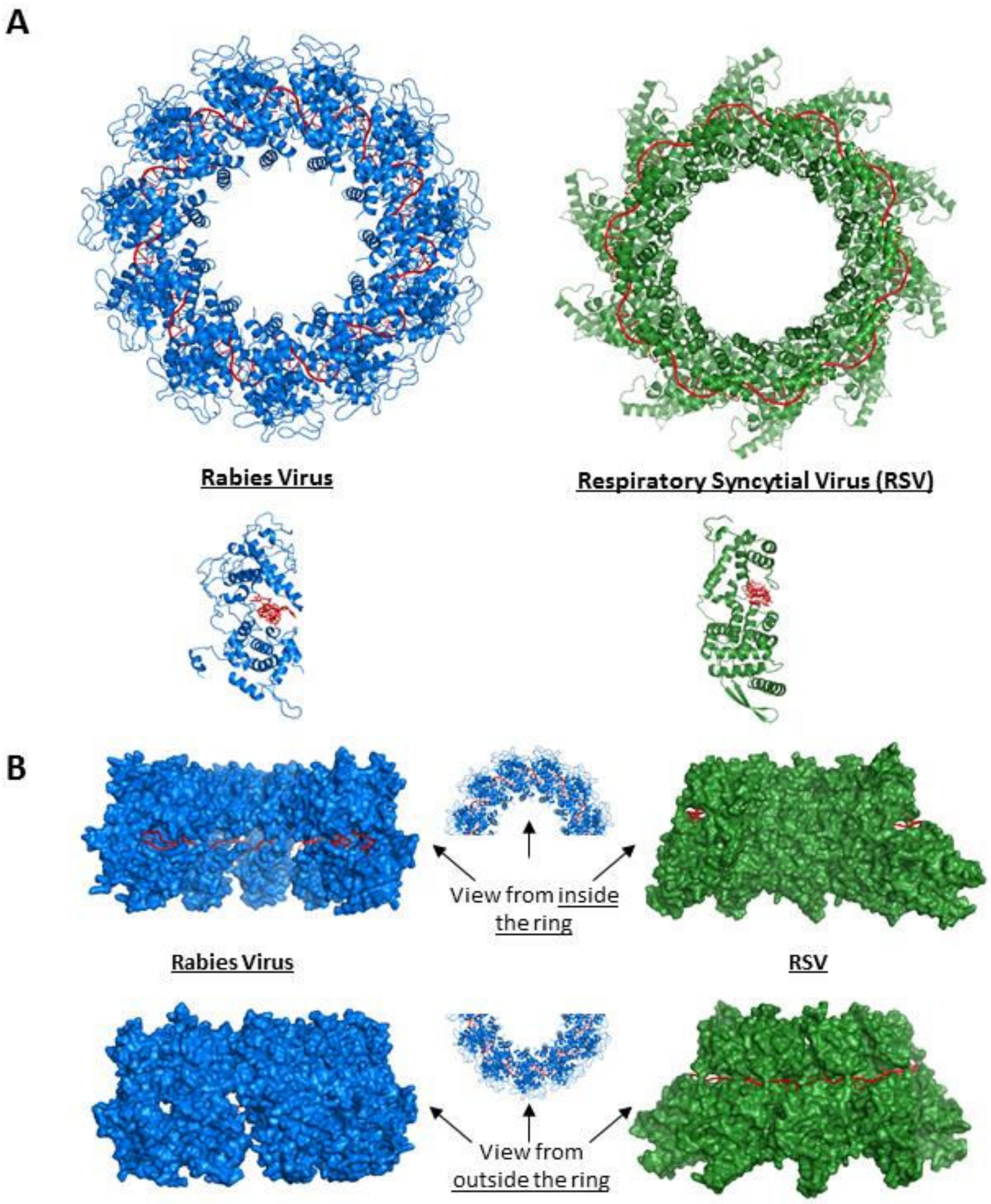

A
Rabies Virus
Respiratory Syncytial Virus (RSV)
B
View from inside the ring
Rabies Virus
RSV
View from outside the ring

Figure 1.12 : RNA groove in Measles nucleocapsids : pointing inwards or outwards ?
A : Comparison of the crystal structure of N-RNA rings of rabies virus (Albertini et al., 2006) and of respiratory syncytial virus (Tawar et al., 2009). The RNA molecule is depicted in red, the nucleoprotein in blue for rabies (left) or green for RSV (right). A closer view of the monomer along the RNA strand direction (bottom) shows the cleft of N in which the RNA is buried. **B** : The surface representation of half of the ring structure, viewed from the inside of the ring (top) or from the outside (bottom) shows the different localizations of the RNA grooves in these structures.

VSV project : the biological questions

Vesicular Stomatitis Virus (VSV), the prototype Rhabdovirus has a lipid envelope enclosing a tightly packed bullet-shaped skeleton. Built by a helical trunk and topped by a conical tip, the skeleton contains the negative-strand viral RNA enwrapped by the viral nucleoprotein N (**figure 1.13A, left**). This N-RNA complex is the template for replication and transcription by the viral polymerase complex consisting of the phosphoprotein (P) and the enzymatic large protein (L). Concerning the structure of the virus particle, the simple wooden model of the 60's based (**figure 1.13A, right**) on 2D negative stain electron microscopy (EM) observations of the skeletons (Nakai and Howatson 1968) was proven to be visionary by the recent 3D cryo-electron microscopy reconstruction of the helical trunk of the skeleton inside the virus particle (Ge et al. 2010) **(figure 1.13B)**. In this reconstruction the viral matrix protein M, which role in the nucleocapsid condensation has been under debate since thirty years (Newcomb and Brown 1981), bridges consecutive turns of the N-RNA helix (**figure 1.13B, right**). Thus the formation of a rigid nucleocapsid core is proposed to be impossible in the absence of M.

However, in the meantime in Grenoble, Irina Gutsche and Euripides Ribeiro were exploring the large conformational rearrangements of purified viral and recombinant N-RNA as a function of pH and salt concentration, and found out that, in the absence of other viral proteins, the nucleocapsid can fold into bullet-shaped structures (**figure 1.13C**). Determining the 3D structure of these reconstituted N-RNA bullets from cryo-EM images constituted my second main project.

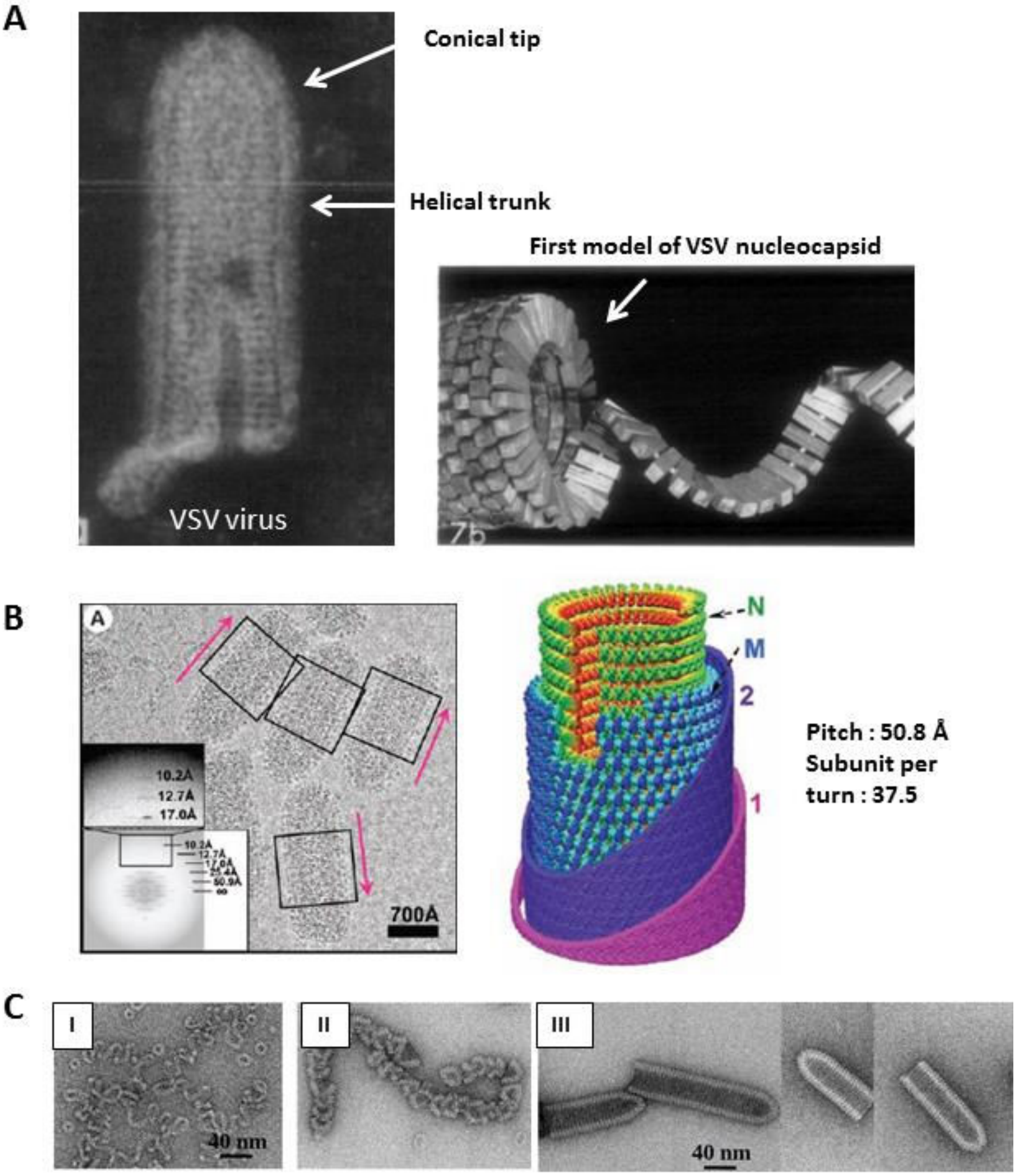


Figure 1.13 : VSV : Structure of the nucleocapsid.

The nucleocapsid of VSV has a bullet shape built by a helical trunk and topped by a conical tip (**A**, Nakai and Howatson, 1968). Since the first model (**A**, right), and beyond the latest reconstruction of the full virion (**B**, left ; Ge et al., 2010) from cryo-EM images (**B**, left), the role of the matrix protein M in the packing of the nucleocapsids into bullet-shaped structures has been under debate. In the virion structure derived from the cryo-EM images, the M clamps adjacent turns of the nucleocapsid, which suggested that it is necessary to form the bullet shape structure. The conformational rearrangements of purified VSV nucleocapsid (**C**), from the loosely coiled ribbons of NC at pH 7.5 and 150 mM NaCl (**C, I**) to string of conical tips at pH 7.5 without salt (**C, II**) and to Bullet-shaped NC at pH 5 (**C, III**). EM images from Dr Irina Gutsche.

Nucleocapsids of negative strand RNA viruses are very large helical structures. Unlike in other viruses like TMV, these helices are rather flexible, at least in their isolated state, making them unsuitable targets for X-ray crystallography. Thus, for these objects, the technique of choice for determination of their 3D structure is Electron Microscopy (EM), associated with appropriate image analysis techniques, which have proven since many years to be suitable for the reconstruction of more or less flexible helical polymers. The following section aims at giving a general introduction to electron microscopy and 3D reconstruction as well as to the specimen preparation techniques that were used in the course of this work, before introducing helical reconstruction.

Introduction to Electron Microscopy

Historical points

Transmission Electron Microscopy (TEM) for the characterization of biological objects is now a well-established method which has necessitated more than a century of developments. In 1878, Ernst Abbe realized that the optic microscopes had reached a fundamental resolution limit and that it was necessary to find new tools to explore objects on a smaller scale (or objects at a higher magnification). TEM originated in 1896, when the phenomenon known as magnetic focusing was discovered by A. A. Campbell-Swinton: he found that a longitudinal magnetic field generated by an axial coil can focus an electron beam. In 1899, Wiechert observed that cathode rays (electrons) can be focused by the action of an electromagnetic field produced by a solenoid. The elements for building an electromagnetic lens are here and, in 1926, Hans Busch presented a complete mathematical interpretation of this effect. Based on these previous works, in 1928, Ruska and Knoll built an optical bench for electrons, under vacuum, which consisted of a small aperture illuminated by an electron beam plus a fluorescent screen to observe the image. A small solenoid was used to create the image of the aperture. In 1931, they managed to further magnify the image created by the first solenoid using a second one placed between the intermediate image and the screen. The magnification at this time was 16x, but rapidly improved over the next years: in 1933, Ruska obtained, for the first time, a resolution better than the best optical microscopes. In 1935, Knoll published the first images obtained from the scanning of a solid sample by an electron beam (signaling the advent of Scanning Electron Microscopy -SEM-). During the next decades, the resolution of EM was constantly improved thanks to the multiple interactions between the needs of the users and scientific progress in domains as various as electronic optics, detectors, informatics and electronics, vacuum science and precision engineering. In biology, the need to visualize macromolecular complexes "in vivo" pushed the microscopists to develop sample preparation techniques adapted to biological objects which are intrinsically very sensitive to electrons. Those included first metal shadowing (Williams and Wyckoff 1944), then negative staining (Hall 1955; Brenner and Horne 1959) before the development of cryo-electron microscopy which made it possible to visualize biological objects at low temperatures in a hydrated state, thus closer to physiological conditions (Dubochet et al. 1982; Adrian et al. 1984; Dubochet et al. 1988).

Basic Principles of Transmission Electron Microscopy

Transmission electron microscopy (TEM) makes use of high energy electrons (accelerated by tensions from ~60 to 400 kV) to create an image of thin specimen (~50-150 nm). The wavelength of electrons, in comparison to visible light, ensures a much better resolution in TEM than in optical microscopy. However, due to imperfections of electromagnetic lenses, the resolution of EM drops off far before what is expected from the De Broglie wavelength of electrons (0.0025 nm for electrons accelerated at 200kV). In practice, the resolution of the best electron microscopes is about 0.5 Å. The general scheme of a transmission electron microscope is depicted in **figure 1.14**. It is principally composed of:

-a system of pumps to maintain high vacuum in the microscope. As the electrons interact strongly with matter, they also interact with molecules in the air. The microscope vacuum must be kept at ~$0.1 \text{x} 10^{-5}$ Pa = $1 \text{x} 10^{-8}$ millibars.

-an electron gun, composed of the source of electrons, a focalization system, and an electron accelerator

-a column containing electromagnetic lenses and diaphragm

-a sample holder (equipped with a liquid nitrogen based cooling system, for cryo)

-a detector (screen, CCD camera, films)

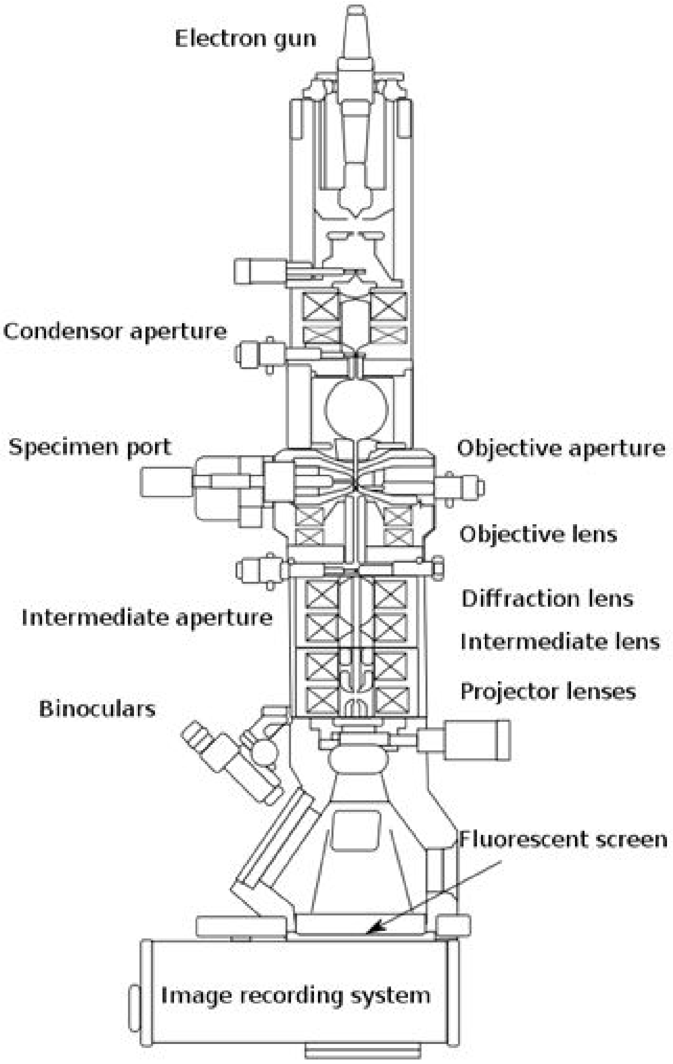


Figure 1.14 : Scheme of a transmission electron microscope

For the image formation, one usually considers only the objective lens, which is the closest to the sample. Indeed, this lens ensures the interaction of the electron beam with the sample and the formation of the first magnified image of the object, and thus will mostly (but not only) determine the quality of the resulting images.

In high-resolution TEM, especially in biology, thin specimens can be considered as "phase" objects, i. e., the interactions between the electron beam and the sample do not significantly change the wave amplitude associated to the electrons but they modulate its phase. By tuning the magnetic lens to produce a defocalization, the exiting wave interferes with itself generating a contrast call phase-contrast. Electrons interacting more strongly (inelastic scattering, see **figure 1.15**) with the sample do not follow the phase approximation above and they generate noise at the image level. For this reason, EM constructors try to minimize the amount of such electrons by using diaphragms and energy filters in order to improve contrast.

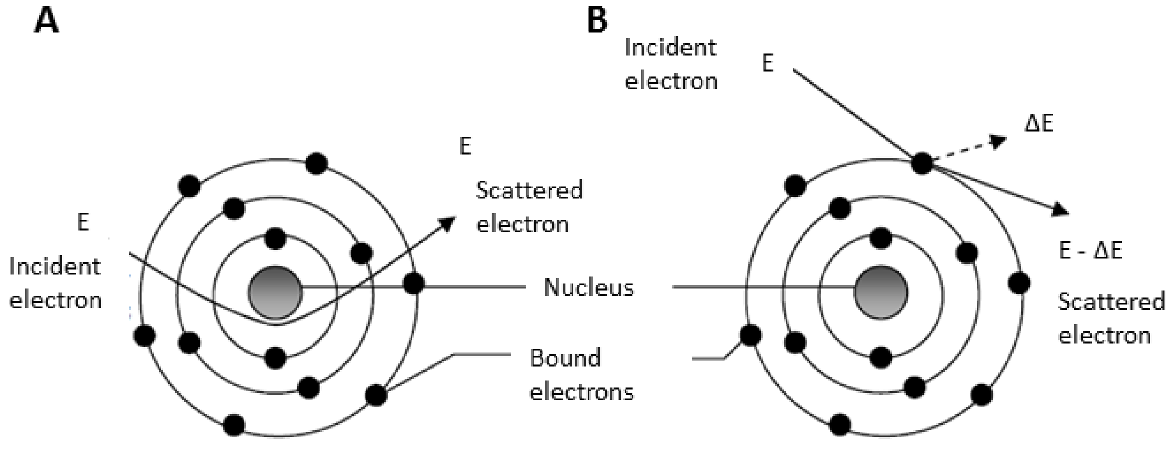


Figure 1.15 : Interactions of electrons with matter
A : elastic scattering, without energy loss
B : inelastic scattering, with a loss of energy ΔE of the scattered electron, and either ionization (loss of an electron) or excitation (movement of an electron into an excited state) of the encountered atom.

A more detailed explanation about the physical principles behind the image formation process is beyond the scope of the presented thesis and references like (Transmission Electron Microscopy: Physics of Image Formation from Reimer and Kohl) should be consulted. What is important for this work is the contrast transfer function (CTF) defined by the optical characteristics of the electron microscope and the amount of defocus used. This function modulates the amount of information transmitted from the specimen to the image depending on the spatial resolution. When represented as a plot, the CTF has the form shown in **figure 1.16**.

As it can be seen, the CTF has negative and positive parts. That means that (i) there are zeros where no information is transmitted. This imposed the combined use of multiple images with varying defocus to fill those gaps and (ii) there are contrast reversals for some resolution ranges that must be compensated before combining images to avoid annihilation of the information.

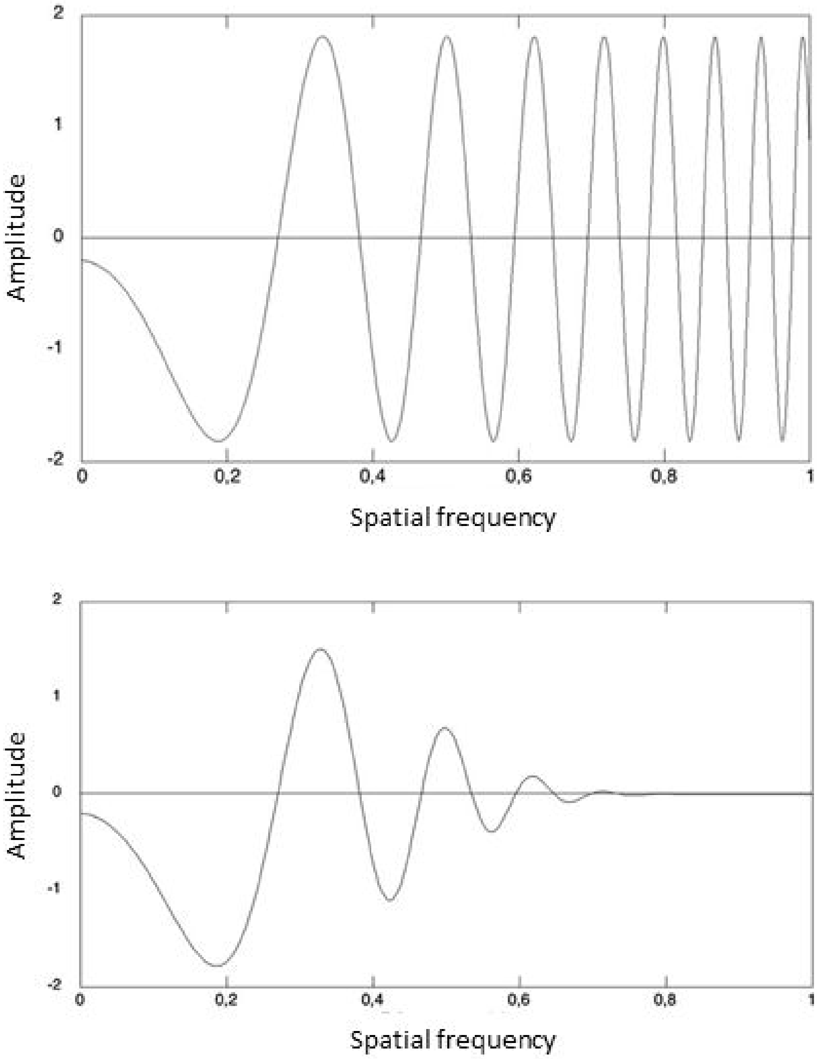


Figure 1.16 : Representation of the Contrast Transfer Function (CTF)

Top : representation of the CTF without decay of amplitudes. Amplitude contrast : 10% ; Cs = 2mm ; Defocus = 3µM

Bottom : representation of the CTF with decay of amplitudes of high frequencies (so, including the envelope function). Amplitude contrast : 10% ; Cs = 2mm ; Defocus = 3µM

Basic Principles of 3D reconstruction of single-particle

Single-particle 3D reconstruction is based on the central section theorem (R. A. Crowther, DeRosier, and Klug 1970; DeRosier and Klug 1968). It says that the Fourier transform of a 2D projection of a 3D object is equal to the slice of the object 3D Fourier transform perpendicular to the projection direction (**figure 1.17**). It allows the recovery of the 3D information from the 2D images produced by an electron microscope.

The numerical realization of the above theorem composes the core of all the single-particle reconstruction softwares available. Some implement it on real space (without using Fourier transforms), others do it in the reciprocal space but the result is (theoretically) the same. The 3D Fourier transform can thus be recovered from its central sections, either by interpolation or by assuming a functional form that depends on a certain number of unknown functions. The main advantage of the methods based on functional forms is the fact that the number of unknowns must always be less than the number of available data points (images). Interpolation schemes allow the reconstruction to be calculated in any situation (even if not enough data is available) but the distortions present in the resulting 3D reconstruction cannot be controlled.

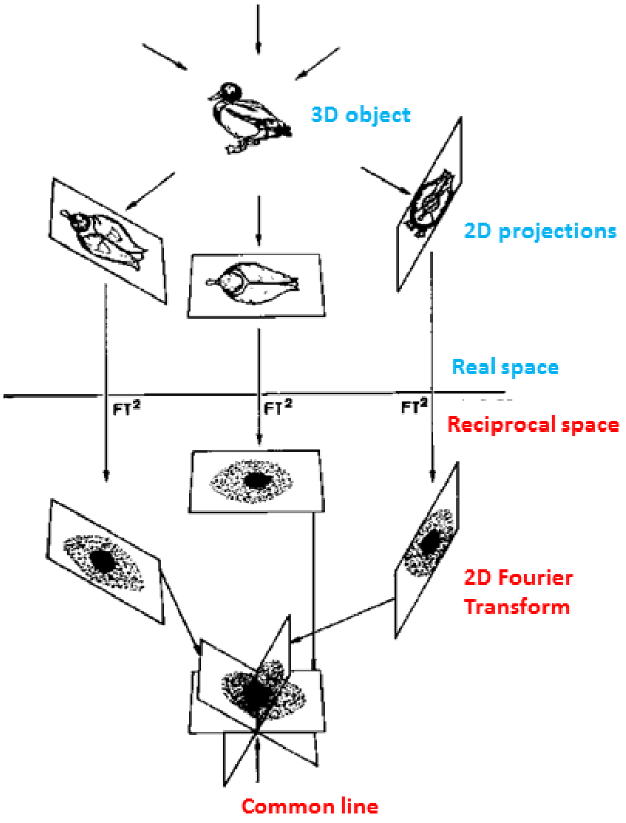


Figure 1.17 : The central section theorem
A 3D object (top, the duck) is projected in 2D. Those projection are meant to represent the images obtained by the Transmission Electron Microscope. The 2D Fourier transform (FT) of the projections are central sections of the 3D FT of the 3D object. Thus each pair of 2D FT of 2D projections shares a common line in reciprocal space. By combining many 2D FT of 2D projection, one can fill the reciprocal space with central sections, and recover ideally the complete 3D FT of the object, back-Fourier transform, and obtain a real-space 3D object.

Projection matching

The projection matching technique is a way to obtain the orientation corresponding to the 2D projections produced by the electron microscope in the form of single-particle images. It consists of the use of a 3D reference from which 2D projections are calculated and systematically compared to the real (experimental) images (**figure 1.18**). If the 3D model is sufficiently close to the 3D structure of the sample, the orientations associated to the synthetic 2D projections can be assigned to the experimental images. This allows a new 3D reconstruction to be calculated and to replace the initial model in an iterative process that follows until convergence. The method above also determines the center of the particles.

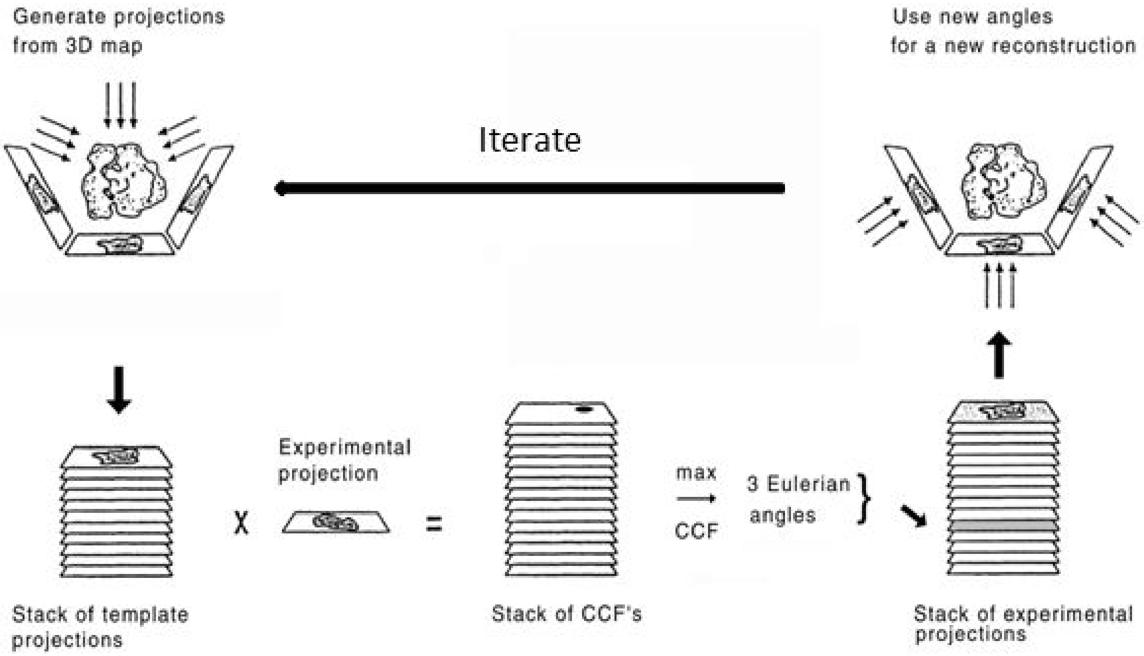

Generate projections from 3D map
Iterate
Use new angles for a new reconstruction
Experimental projection
X
=
max
CCF
3 Eulerian angles
Stack of template projections
Stack of CCF's
Stack of experimental projections

Figure 1.18 : Determination of particles orientation by projection matching
A 3D initial model is used to generate projections at many different orientations (top left). Each of the experimental images is compared to each of the projections by cross-correlation. The maximum of cross-correlation is used to assign the view angles (eulerian angles) of the corresponding projection, to each image. Additionally, the experimental images are usually rotated in the plane and shifted to align with the matching projection. The stack of images (bottom, right) with assigned orientation is then used to calculate a new reconstruction, which is used as an initial model for a new projection matching cycle.

Biological sample preparation for Electron Microscopy

We will now give more details on the sample preparation techniques which were used for the projects presented in this thesis: negative staining and cryo-EM. Negative staining, used for Measles nucleocapsid project, will be comprehensively introduced as I was directly involved in EM grid preparation, whereas cryo-EM (used for VSV reconstituted bullets-shaped N-RNA project) will be more briefly overviewed, as the experimental procedures were performed by Dr Irina Gutsche and Dr Guy Schoehn.

The negative staining technique

Historical background

It has been known since the 1940's that the enhancement of contrast in electron microscopy observations can be achieved through the use of stains, which are dense materials that will associate with the structures of interest. The stains were mainly composed of heavy atoms, like osmium tetroxide or phosphotungstic acid (PTA). The main focus of research at that time was to find conditions to achieve maximum stain absorption with optimum preservation of morphology, in buffer conditions that would not destroy the structures. These stains were used because they directly and covalently interacted with the sample, so this staining technique was what we would call today positive staining. During the 50's, several reports were made on observations of "anomalous" staining pattern. The first was made by (Hall 1955) who studied the effects of staining conditions on the structural aspect and measurable electron density of the well known viruses of tomato bushy stunt (BSV) and tobacco mosaic (TMV). Although the focus of the study was mainly on the effects of stain concentration, buffer composition, washing and fixing conditions, some attention was drawn to some "anomalous" staining patterns and a hypothesis concerning the kind of interaction between stain and sample that would explain these patterns were postulated. Quite visionary, the author makes this remark: "Although the effect shown in Fig. 8 is the opposite to what is usually sought by the use of electron stains, the visibility of particles of low scattering power can be enhanced as well, **if not better**, by surrounding them with dense material rather than impregnating them with dense material". By the "opposite" effect, he meant the fact that the particles, in imperfect washing conditions and low stain concentration, were seen light on a dark background instead of appearing dark on a light background (**figure 1.19**). This effect was also noticed, on TMV again, by (Huxley, 1957), but the first to use and introduce the term of "**negative staining**" were (Brenner and Horne 1959), that observed the same phenomenon with T2 bacteriophage. In the following years, the negative staining technique became the standard for EM observation of viruses (Horne, Hobart, and Ronchetti 1975) or other biological objects, resolving both the preservation and contrasting problem.

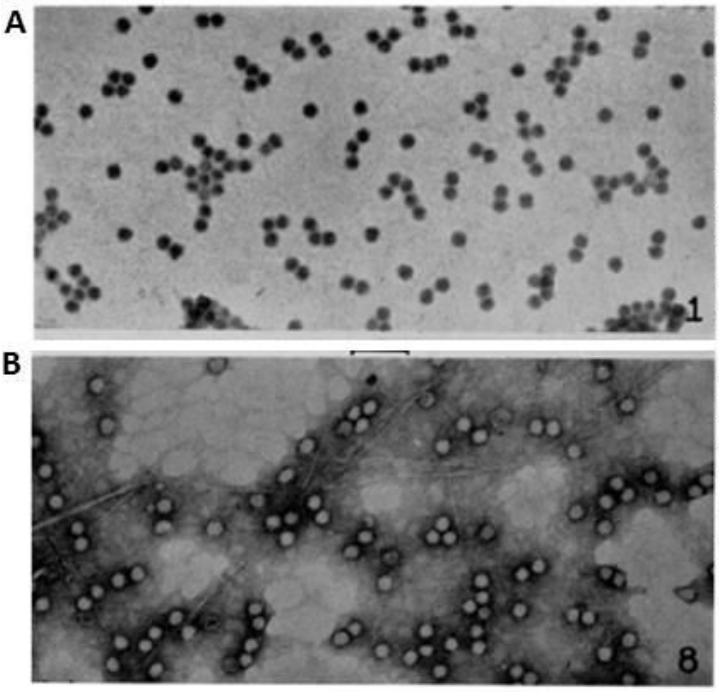


Figure 1.19 : First observation of the negative staining effect : « the anomalous images»

These two images, taken from (Hall, 1955) show two different, inverse aspects that can have tomato bushy stunt viruses under different staining conditions. Whereas the micrograph in A shows the –in 1955- usual positive staining pattern where particles appear dark on a light background, the micrograph B shows white particles on a dark background.
The staining conditions were :
A : Solution at pH 1 containing 40% 12-PTA and 10% $PtCl_4$
B : Solution at pH 4.6 containing only 5% PTA. Insufficient washing.

Principal of the method

The basic principle of the negative staining method is relatively simple: it involves the use of a stain, composed of heavy atoms, that will interact strongly with the electrons of the beam in the microscope (more scattering), and thus give raise to a higher contrast within the images. Whereas for positive staining, the stain directly interacts with the sample, the idea here is that it replaces the hydration shell around the proteins until, ideally, all hydrated volumes are filled with it, forming a cage embedding the protein and protecting it from surface tensions. Next, the ensemble stain-protein is dried as rapidly as possible, and the layer of stain should protect the shape of the protein. In the microscope, this shell of stain around the biological material is much more stable than the material alone, thus preventing rapid specimen degradation due to irradiation by the electrons, a major problem for biological EM.

Although each commonly used stain do not meet all the qualities listed below, one can say that an ideal stain should have these properties:

- High density

- Ability to protect specimen against dehydration effects

- High solubility

- Non-chemically reactive with the specimen

- High melting and boiling points

- Uniform spreading on the support film

- Amorphous structure (i.e. structureless) when dry

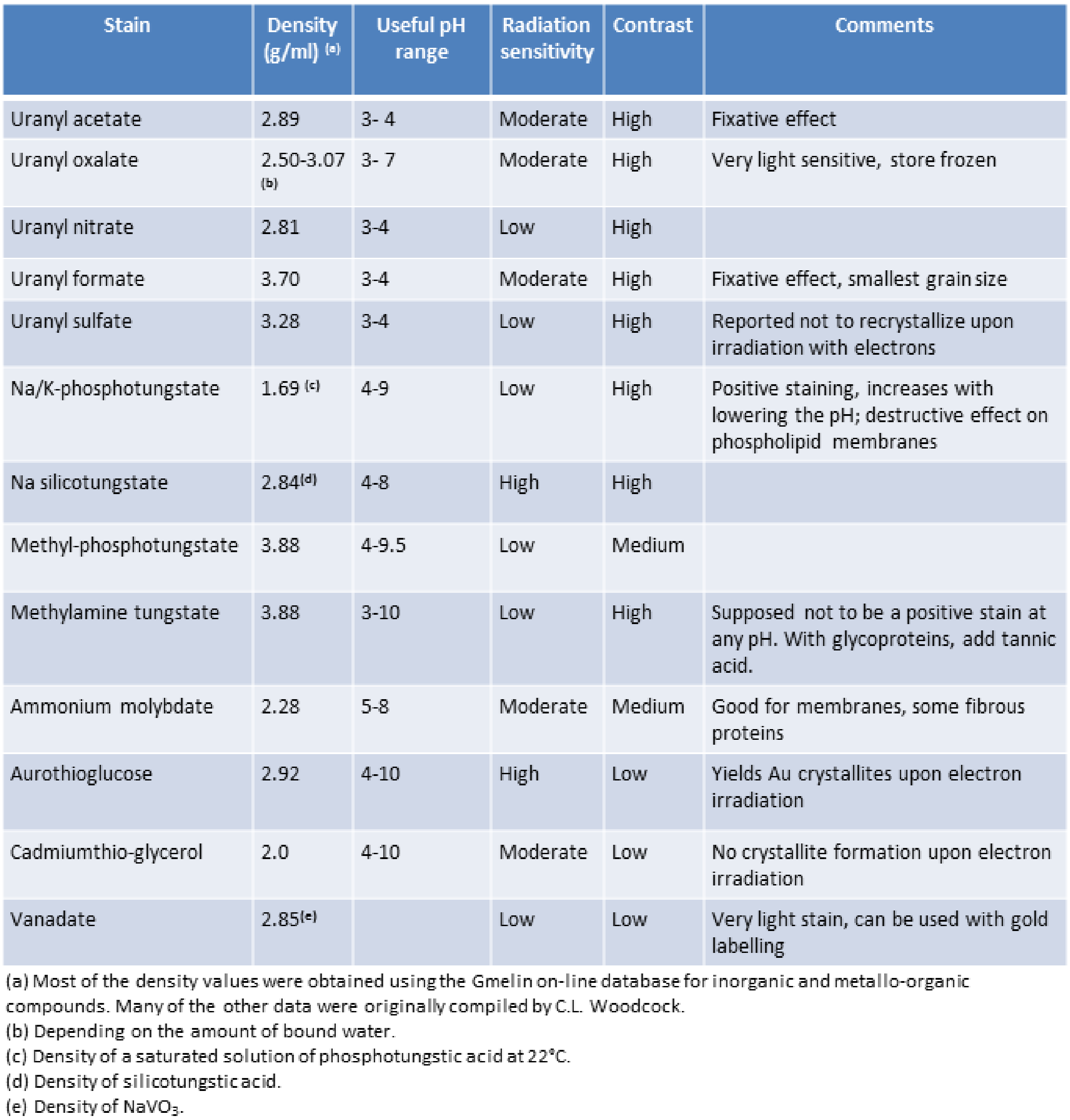
Experimental setup

One can probably find as many negative staining experimental setups as EM laboratories in the world. Furthermore, one could say that any single negative staining experiment is unique, without being far from the truth. Indeed, most of the grid preparation is performed manually, and each step is subject to many parameters. Among these parameters, one can cite :

-environment: temperature, humidity…:

-specimen support: most of the negative staining protocols propose the use of a physical support for the specimen, usually a thin continuous carbon film, although holey carbon film can also be used for particular aims (Hanson and Lowy 1963). The thickness of this film can be variable depending on the requirements of the experiments: use of single or double carbon layer; size of the specimen and buffer composition (some components can be destructive for the carbon film). The time needed for the proteins to adsorb to the carbon film depends on the protein and on the carbon film properties, in particular its hydrophobicity (which can be reduced by glow-discharge), and thus is a parameter to experimentally determine.

-stain: the choice of the negative stain can be crucial, and thus a good negative staining experiment involves assaying as many stains as possible before drawing conclusions on the structural aspect of the specimen. If, for example, a variety of stains are used and similar staining patterns are obtained, then it is likely that the features revealed are consistent with genuine specimen morphology. On the other hand, the property of a particular stain in having an effect on the sample (oligomerisation state, protein conformation/flexibility) can be used in some cases: an illustration of this will be shown in the next section on Measles nucleocapsids. Different stains are commonly used, and for most of them, one can also vary their concentration and the pH of solution, both having potential effects on the staining. A table containing a list of commonly used stain and their properties, reproduced from (Bremer et al. 1992), is shown in **table 1.2**.



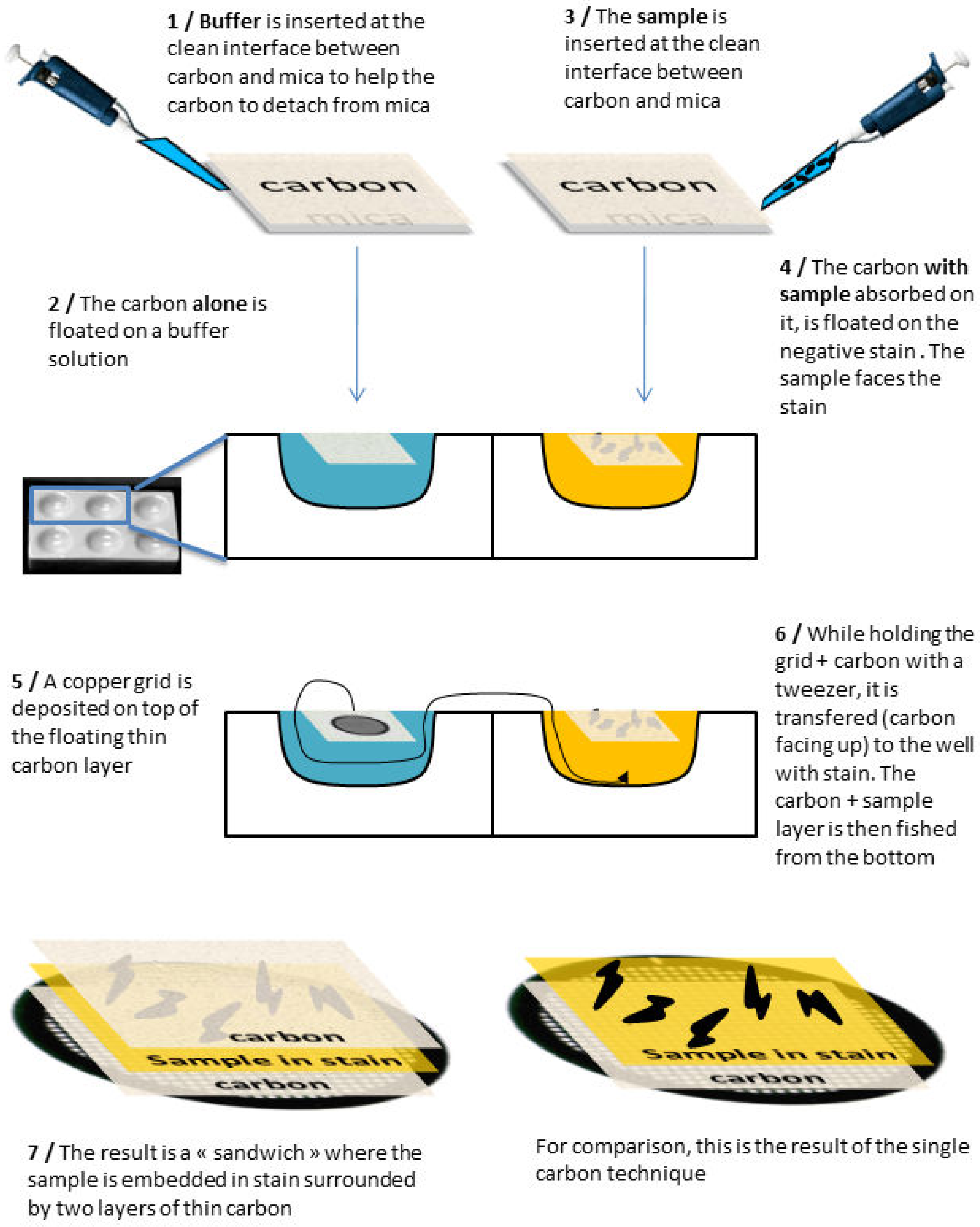
Stain	Density (g/ml) [a]	Useful pH range	Radiation sensitivity	Contrast	Comments
Uranyl acetate	2.89	3- 4	Moderate	High	Fixative effect
Uranyl oxalate	2.50-3.07 [b]	3- 7	Moderate	High	Very light sensitive, store frozen
Uranyl nitrate	2.81	3-4	Low	High	
Uranyl formate	3.70	3-4	Moderate	High	Fixative effect, smallest grain size
Uranyl sulfate	3.28	3-4	Low	High	Reported not to recrystallize upon irradiation with electrons
Na/K-phosphotungstate	1.69 [c]	4-9	Low	High	Positive staining, increases with lowering the pH; destructive effect on phospholipid membranes
Na silicotungstate	2.84[d]	4-8	High	High	
Methyl-phosphotungstate	3.88	4-9.5	Low	Medium	
Methylamine tungstate	3.88	3-10	Low	High	Supposed not to be a positive stain at any pH. With glycoproteins, add tannic acid.
Ammonium molybdate	2.28	5-8	Moderate	Medium	Good for membranes, some fibrous proteins
Aurothioglucose	2.92	4-10	High	Low	Yields Au crystallites upon electron irradiation
Cadmiumthio-glycerol	2.0	4-10	Moderate	Low	No crystallite formation upon electron irradiation
Vanadate	2.85[e]		Low	Low	Very light stain, can be used with gold labelling

(a) Most of the density values were obtained using the Gmelin on-line database for inorganic and metallo-organic compounds. Many of the other data were originally compiled by C.L. Woodcock.
(b) Depending on the amount of bound water.
(c) Density of a saturated solution of phosphotungstic acid at 22°C.
(d) Density of silicotungstic acid.
(e) Density of $NaVO_3$.

Table 1.2 : Properties of various negative stains
From (Bremer, Henn, Engel, Baumeister, & Aebi, 1992).

-specimen concentration: an advantage of negative staining is that the required concentration of sample is relatively low. An appropriate concentration is not only needed to have enough views of particles when recording images or screening a grid, but very importantly to obtain a good staining. Due to the way that the stain deposits around proteins and to the drying process, too low a concentration will cause the stain thickness between each individual particles to be almost zero. Also, due to the tensions at the surface of the stain film, it might be more frequent that the entire particles are not surrounded by stain, especially if their size is big. If the concentration is higher, the proximity of particles will "support" the stain film between particles, and the surface tension in the proximity of every particle will be lower, and thus result in a higher likelihood of obtaining a well embedded specimen. Nevertheless, too high a concentration can lead to protein aggregation during grid drying, to superposition of views of the particles that would make any image analysis impossible and cause the stain thickness to be so high that the particles would be barely visible. Due to the different properties of proteins (electrostatic surface, hydrophobicity, shape …) and of the carbon surface, the adequate concentration has to be experimentally determined. A commonly used starting value for this search is ~0.1 mg/ml.

-**staining protocol**: many possible protocols to apply the negative staining are described. The most widely used is based on the support of specimen on a single continuous carbon layer and the staining is achieved through the "droplet method". In our study, although the droplet method was also tried, the main protocol used for routine observations was based on a simple continuous carbon layer and specimen floating on a large amount of stain. When image analysis was planned, we used a protocol involving two carbon layers to catch the specimen in between. A visual step by step description of the protocol for the double carbon layer technique is poorly documented in the literature, and thus is depicted in **figures 1.20**.



1 / **Buffer** is inserted at the clean interface between carbon and mica to help the carbon to detach from mica

3 / The **sample** is inserted at the clean interface between carbon and mica

2 / The carbon **alone** is floated on a buffer solution

4 / The carbon **with sample** absorbed on it, is floated on the negative stain . The sample faces the stain

5 / A copper grid is deposited on top of the floating thin carbon layer

6 / While holding the grid + carbon with a tweezer, it is transfered (carbon facing up) to the well with stain. The carbon + sample layer is then fished from the bottom

7 / The result is a « sandwich » where the sample is embedded in stain surrounded by two layers of thin carbon

For comparison, this is the result of the single carbon technique

Figure 1.20 : Protocol for negative staining EM grid preparation using the double carbon technique

For clarity, size proportions are not respected

Optimization of observation conditions with negative staining : example of Measles nucleocapsids

Considering the variety of conditions listed above, it seems to always be important, especially for a new project, to try a variety of conditions (staining protocol, pH, temperature, concentration of specimen, stain and buffer, etc.) when preparing specimens for microscopy using negative staining techniques. Quite often, under varied conditions, different features of a specimen will be enhanced, and either complementary or perhaps even contradictory information may be obtained. In the next section we will detail an example of a search for sample preparation condition for EM image acquisition using negative staining, which will in the meantime introduce the type of images we have been working on for the Measles nucleocapsids project.

Effect of salt concentration on nucleocapsid compaction

Striking effects of parameters such as ionic strength, pH or salt concentration on the ultrastructure of viral proteins assembly are widely described (Salunke, Caspar, and Garcea 1989; Lepault et al. 2001). More specifically, the effects of salt concentration on the structure and rigidity of nucleocapsids of several Paramyxoviridae have been reported (M. H. Heggeness 1980) and were already used to obtain conditions where nucleocapsid rigidity was enhanced for facilitating the reconstruction process (E H Egelman et al. 1989). However, similar data in the particular context of Measles nucleocapsids are not available. We thus first tried to find salt conditions which enhanced the rigidity of the nucleocapsids, while trying to conserve homogeneity among the sample (**Figure 1.21**).

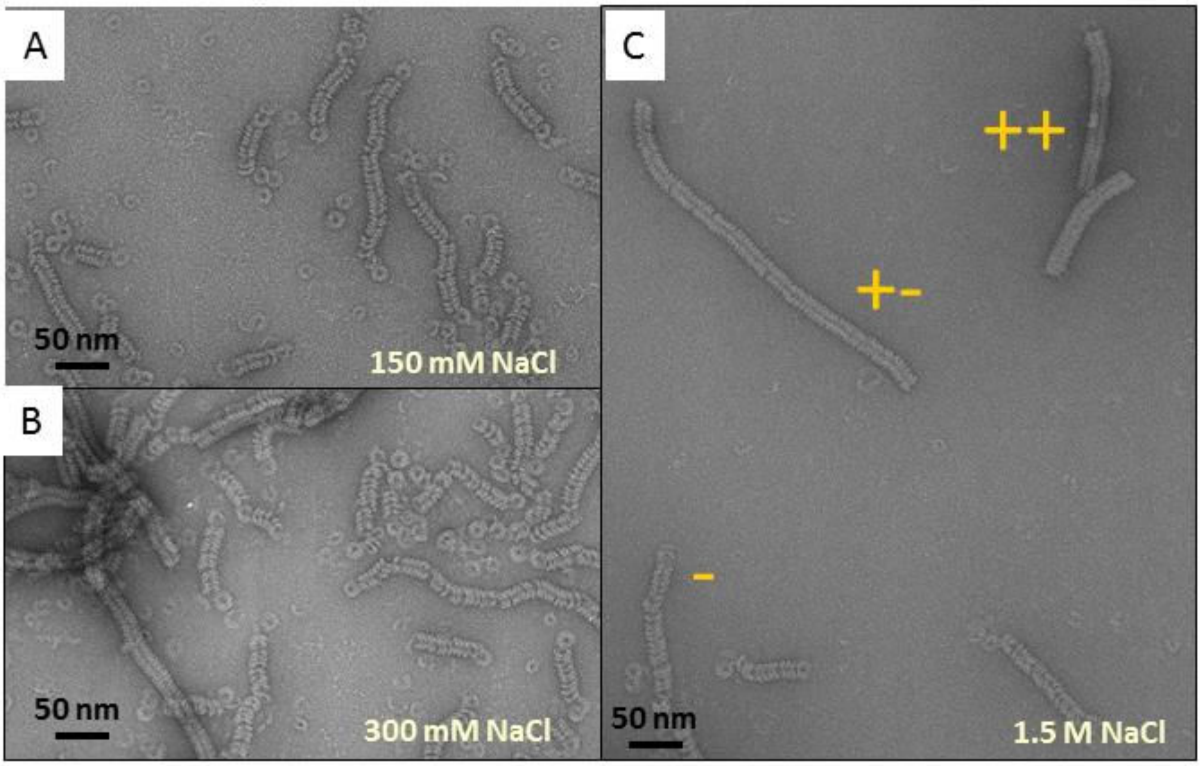


Figure 1.21 : Effect of NaCl concentration on the morphology of MeV non digested nucleocapsids (stain : uranyl acetate)
See text for details

When using uranyl acetate as a stain and a salt concentration of 150 mM (NaCl), the nucleocapsids are very flexible (**Figure 1.21A**). One can also observe ring-shaped structures, either isolated, or attached at the extremities of some of the longer helices. This observation suggests some fragility of the nucleocapsids: the last turn of the helix looses its interaction with the precedent turn and adsorbs then horizontally on the carbon. When a higher salt concentration was used (**Figure 1.21B** : 300 mM NaCl ; **Figure 1.21C** : 1.5 M), the sample aspect is not homogeneous anymore : one can observe structures similar to the one seen at lower salt concentration, and other more rigid helices showing an apparent lower pitch, to a various extent. In the extreme case shown on **figure 1.21C**, one can even see three different compaction states on the same area (noted - , +- and ++ from less to more compact).

The reasons for the lack of homogeneity observed among the samples in the presence of high salt concentration are not clear, notably because of the complexity of the interactions between the sample, the carbon film and the stain, and because of the fast drying. Interestingly, one can remark a certain degree of cooperativity in the nucleocapsids compaction phenomenon. Indeed, we never observed a helix that would be compact only on a part of it (**Figure 1.21C** is a good illustration of this). It is not easy to say, with our experimental setup, if this cooperativity happens in solution or results from interaction with the carbon and the stain. By using metal shadowing, a transition between two very different compaction states occurring within a single helix was observed with the nucleocapsids of Sendai virus (E H Egelman et al. 1989), but we don't know the frequency of such observation. This observation suggests a role of the stain in the phenomenon of cooperativity.

We also observed that the compacted nucleocapsids had a tendency of being grouped on the carbon surface (this can be appreciated on the left part of **Figure 1.21B**). This could be due to conformational rearrangements transmitting from one particle to the other, to heterogeneity at the surface of the carbon film (e.g. variable hydrophobicity), or to local salinity variations caused by dehydration. This last hypothesis is favored by the fact that such areas usually show a thicker stain layer, which is known to be related to the presence of a higher salt concentration.

Considering the lack of homogeneity observed by assaying different salt concentrations, and the difficulties of interpretation of the structures obtained with high salt, we then tested the effects of the choice of the negative stain on nucleocapsid morphology.

Effect of the choice of stain and preparation technique

When using Sodium Silico Tungstate (SST ; pH 7.5) or Methylamine Tungstate (Nano-W; pH7.5) instead of Uranyl Acetate, and a salt concentration of 150 mM NaCl, the nucleocapsids show a more rigid appearance in comparison to UA (150 mM NaCl), with a visibly lower pitch, in a quasi-homogeneous manner (**Figure 1.22**). These important changes could be explained by the difference in pH between the stains, that would modify the charges at the surface of the nucleoprotein and thus the electrostatic interactions, or by direct interactions between the stain and the nucleocapsids. These more compact structures are quite different from the ones obtained by increasing the salt concentration, with a greater measurable pitch, and, more importantly, a visibly better definition of the subunits (clearer stain pattern). In SST, longer structures are observed, but closer examination shows that they consist of several pieces of nucleocapsids interacting with each other via their extremities.

Moreover, in Nano-W, the nucleocapsid appear more rigid, and thus we selected this negative stain for the image acquisition.

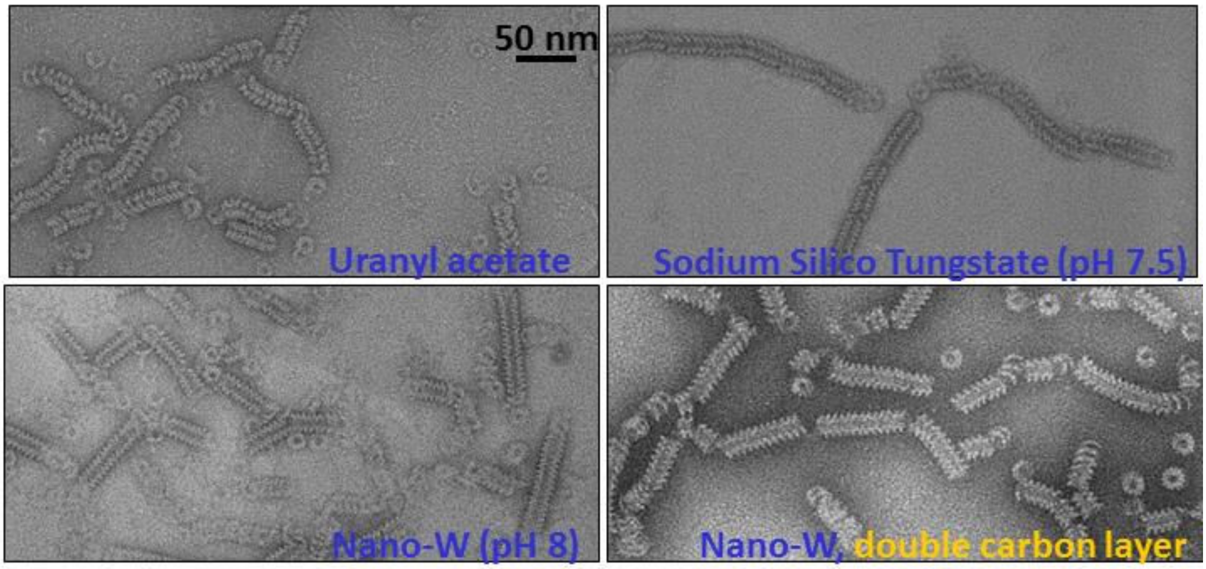


Figure 1.22 : Effect of stain and preparation method on the morphology of MeV non digested nucleocapsids
See text for details

In addition to the stains, we also varied the protocol for grid preparation. As described above (**figure 1.20**), the double carbon technique may ensure a better embedding of the specimen into the stain, especially for high molecular weight samples (Deckert et al. 2006). For helices, this is particularly important to preserve the symmetry of the particles. When applied to the Measles nucleocapsid, this technique proved to be efficient, with a better quality of staining, and has even slightly further enhanced the rigidity of the structures (**figure 1.22, bottom right**). A more uniform distribution of the stain and the eventual interaction of the nucleocapsids with both carbon films could be reasons for this unexpected effect.

To be consistent in the grid preparation technique between our two Measles nucleocapsid samples (trypsin-digested and intact), we applied the same protocol to both samples, and acquired images for further processing. **Figure 1.23** shows a typical micrograph of Measles non-digested nucleocapsids (MeVND) and digested nucleocapsids (MeVD). The entire image processing which will be described later was performed starting from such images.

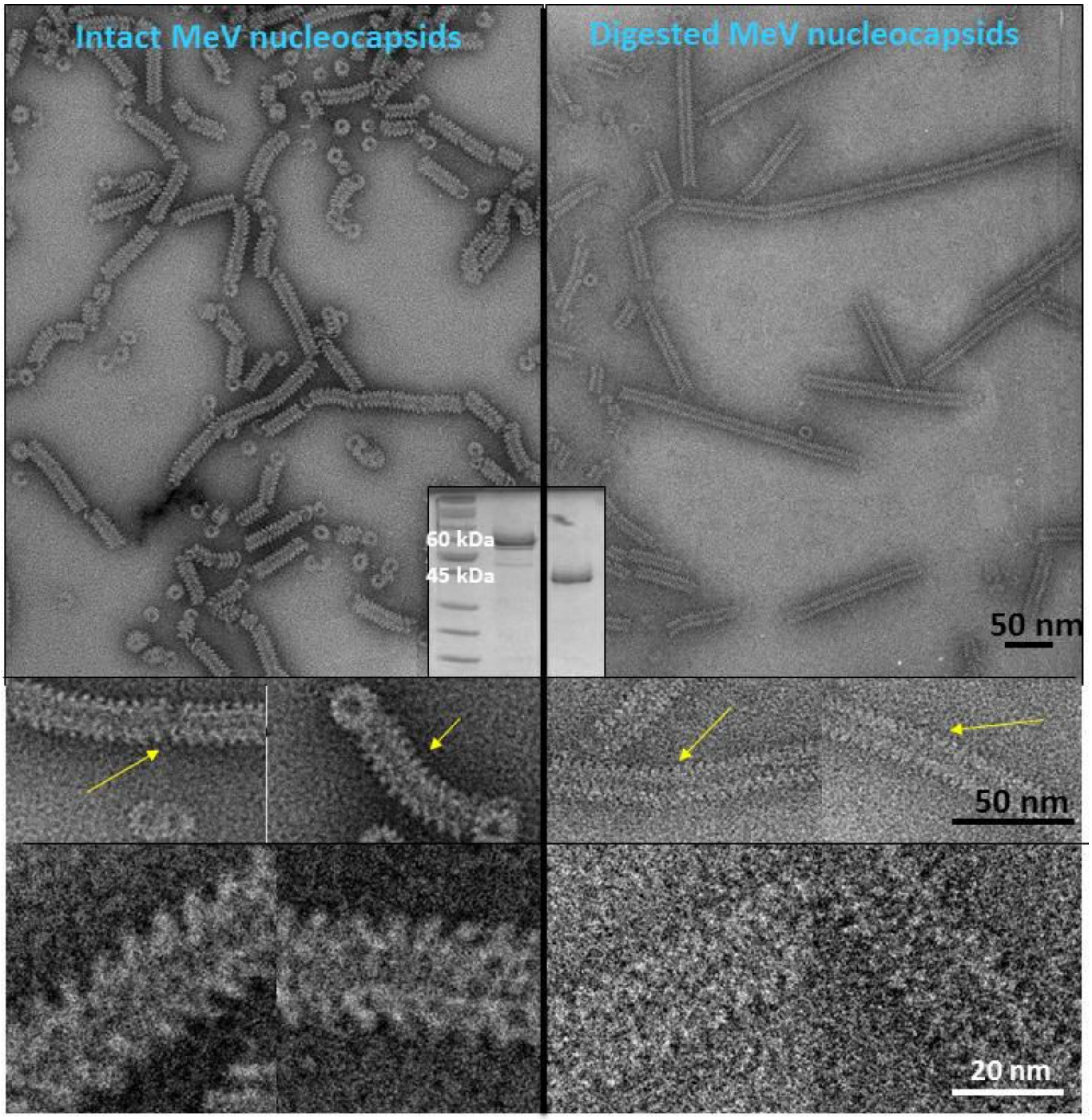


Figure 1.23 : Typical micrographs of recombinant Measles intact and digested nucleocapsids

The top part show a large field of view illustrating the typical aspect of micrographs of MeV intact nucleocapsid (left) and digested nucleocapsids (right). Inset : SDS polyacrylamide gel of the two samples. In the middle, examples showing discontinuities (or broken nucleocapsids) and bending of the helical axis. The bottom show typical boxed images from filaments, which were further used for processing.

Cryo-Electron Microscopy

Historical notes on cryo-EM

Although the negative staining technique had brought so much to the biological EM field since the 50's, providing a simple and easy high contrast imaging technique and making possible the first 3D reconstructions, the fact that the preservation of specimen was far from ideal pushed the community to find new ways of better preserving the specimens. Indeed, once thin protein crystals were examined, it was clear that conventional negative stains fell far short of the ability to preserve the crystalline order at the resolution needed to visualize and trace the polypeptide chain (P. N. T. Unwin and Henderson 1975). On the other hand, the quality of the microscope was such that the theoretical resolution limit of the instrument was enough to resolve such fine structures. What was needed was a way to preserve the hydration of the crystals (or other samples) to avoid deformations due to dehydration, to adsorption on the supporting film and to the stain itself.

The idea of keeping hydrated samples at a cold stage (now referred as **cryo-electron microscopy**) to preserve native structures was not new (with mainly contributions from Fernandez-Moran in the 50's), but it's value for high-resolution was shown only much later when the same protein crystals that showed insufficient electron diffraction with negative staining (usually less than 8 Å) were observed in a frozen-hydrated state. Taylor and Glaeser used liquid nitrogen to freeze thin catalase crystals (Taylor and Glaeser 1974), without cryo protectants, and obtained electron diffraction patterns with a resolution of 3.4 Å, marking the beginning of high-resolution biological EM. However another problem had to be solved : although the temperatures used for cooling down the water were extremely low (-180), the cooling **rate** was such that ice crystals were formed, introducing another artifact and possible deformation source in the observations. The theoretical possibility of cooling water without formation of ice crystal (water remains amorphous = **vitreous water**) was known since a long time, but a practical setup for achieving such results was only first described in the Nature journal by (Brüggeller and Mayer 1980), and consisted in a violent projection of a small droplet of liquid material into the cryogen. In the meantime, a group of electron microscopist in the European Molecular Biology Laboratory (EMBL) in Heidelberg, and especially Jacques Dubochet, showed how to vitrify thin water layers by immersion in liquid ethane, giving raise to the nowadays worldwide used method for cryo-electron microscopy of biological specimens (Dubochet and McDowall 1981). In both cases, the key of success relied in increased cooling rates due to small sample sizes, which size was fortunately exactly compatible with observation by transmission electron microscopy. Amazingly, the major inherent difficulties to this new technique (mainly, how to form a thin enough and uniform aqueous layer, how to avoid any carbon support, how to surmount the very low contrast, reduce beam damage, etc…) found elegant solutions in the next very few years (the 1980-1984 period ; to cite among others : (Dubochet et al. 1988; Adrian et al. 1984; Dubochet et al. 1982).

Subsequently, a number of three-dimensional reconstructions appeared, in particular of helical objects (Mandelkow and Schultheiss 1986; Trinick et al. 1986; Lepault and Leonard 1985). The first single-particle reconstructions, done on highly symmetrical specimens preserved unstained in vitrified ice, included an icosahedral virus reconstruction (Vogel et al. 1986) and clathrin coated vesicles (Vigers, Crowther, and Pearse 1986). The first 3-D reconstruction of an asymmetric structure by cryo-EM appeared 5 years later (J Frank et al. 1991). By today's standards, the resolution of these early reconstructions was modest, but they showed that preservation was greatly improved when the specimen was not dried and they opened the way to the extensive list of 3-D structures of frozen-hydrated specimens that were to come.

Experimental aspects of cryo-EM

The key to of obtaining of good cryo-EM grid is the speed of freezing, in order to obtain vitreous water and to avoid ice crystal formation (opaque to electrons). This very high speed freezing is achievable through :

-the low mass of the grid (and thus its low calorific capacity)

-the speed to plunge the grid in the cryogen

-the choice of the cryogen. As we mentioned, liquid nitrogen, although cold enough, has a too low calorific capacity. Instead, liquid ethane (or a mix propane/ethane) is a good cryogen.

The mechanical support for plunge-freezing can be called "guillotine" (Dubochet et al. 1982), and consist of a system to trigger the plunging (**figure 1.24**), a tank with liquid nitrogen containing a smaller tank filled with liquid ethane, a support for storage of freshly prepared grids.

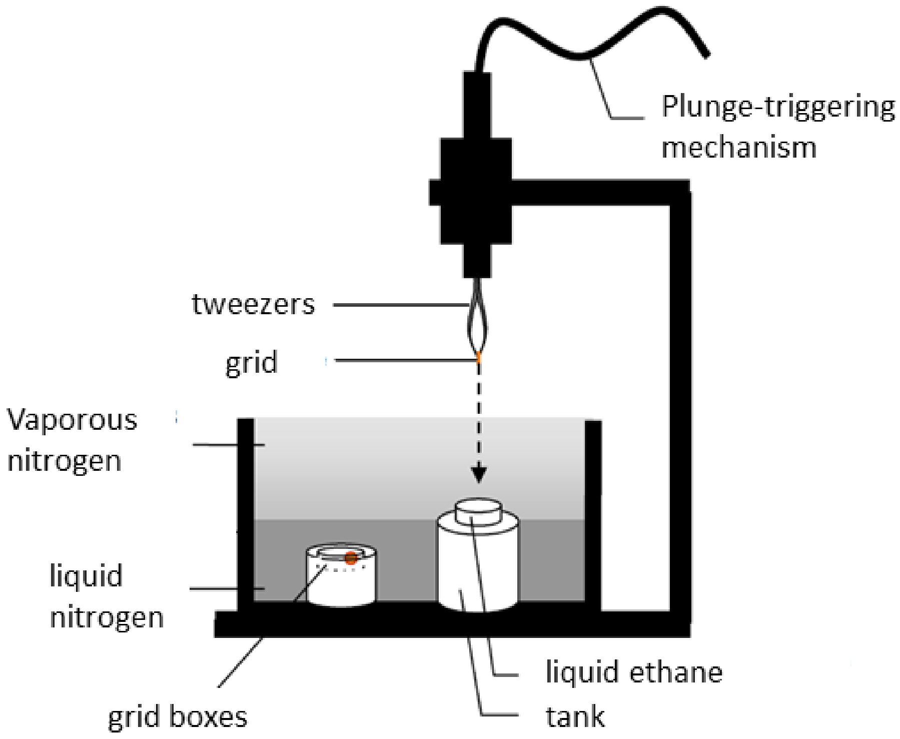


Figure 1.24 : Experimental design of a plunge-freezing device for cryo-EM grid preparation

Example of application of cryo-EM : the reconstituted bullets of VSV

N-RNA was vitrified (Irina Gutsche and Guy Schoehn) on carbon-coated quantifoil 3.5/1 grids (Quantifoil Micro Tools GmbH, Germany). The grids were observed with a Phillips CM200 transmission electron microscope with a LaB_6 filament at 200 kV. Images were recorded under low electron dose conditions at 27,500x magnification on Kodak SO-163 films and negatives were digitized with a Zeiss scanner (Photoscan TD) to a pixel size of 2.55 Å at the specimen level. The **figure 1.25** illustrates two typical micrographs of reconstituted bullets without M (left) and with M (right). One can note the presence on both pictures of views close to the helical axis. All further image processing for VSV was done starting from images similar to those.

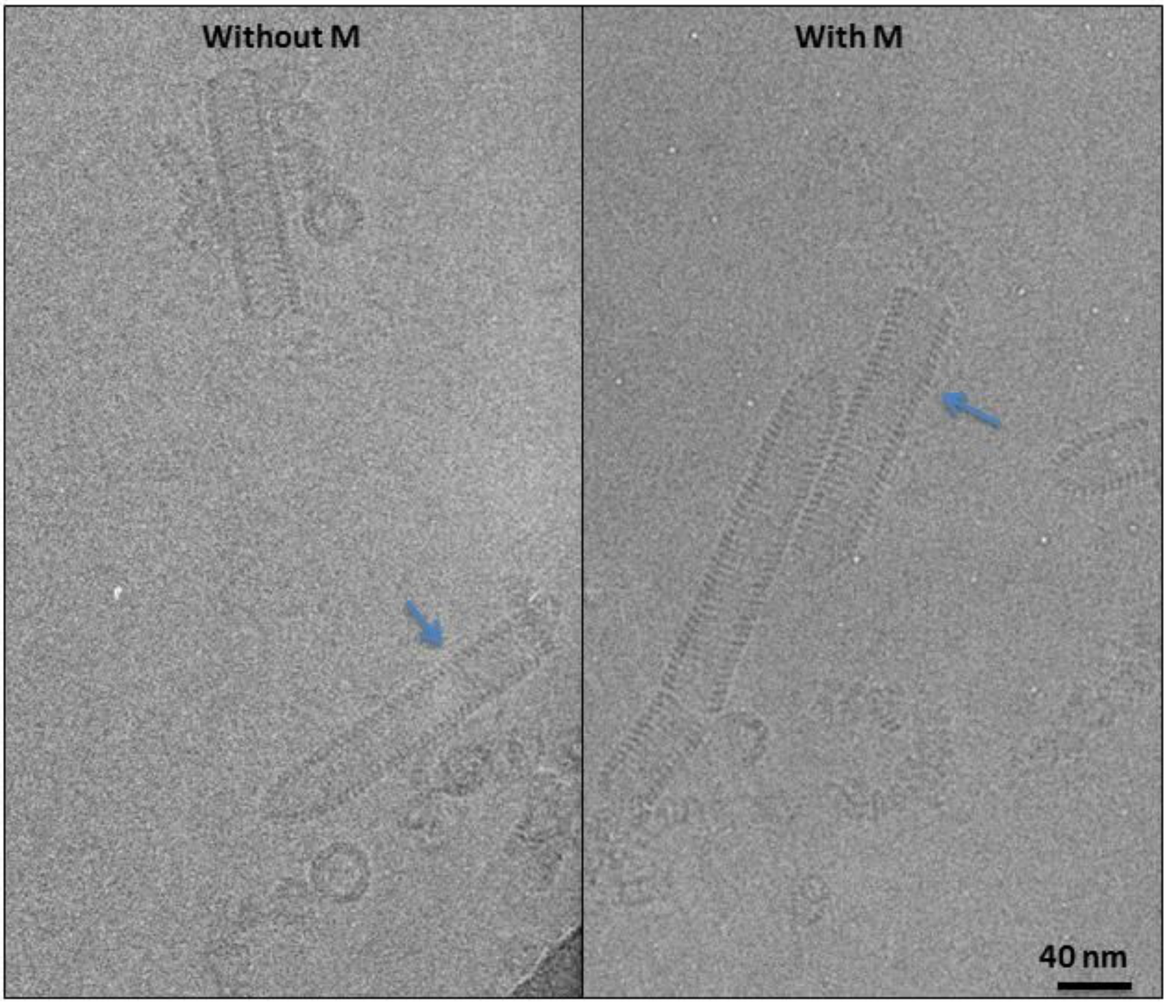


Figure 1.25 : Typical cryo-EM micrographs of VSV N-RNA reconstituted bullets with or without matrix protein added

The blue arrows indicate a switch in visible diameter of the structure. See text for details.

Introduction to helical reconstruction

Methods evolution: from classical methods to single-particle approaches

Historical points

The first ever published reconstruction of a three dimensional object from a set of electron microscopy images was one of a helical object, the bacteriophage T4 tail (DeRosier and Klug 1968). This work presents a general method of 3D reconstruction from EM images of any type of object, with or without symmetry. The method relies on the fact that the Fourier transform of a two-dimensional projection of a three-dimensional object is identical to the corresponding central section of the three-dimensional Fourier transform of the object. The choice of its first application to a helical object is however not a pure coincidence. Ten years earlier, one of the authors (Klug) was indeed already implicated in fiber diffraction studies (Klug, Crick, and Wyckoff 1958), which followed the helical diffraction theory initially developed by (Cochran, Crick, and Vand 1952). This reciprocal space formulation was necessary to solve the structure of DNA (Watson and Crick 1953) and to understand the geometry of polypeptides (Bamford, Hanby, and Happey 1951) .

In the years following the structure of the bacteriophage T4 tail, the general theory of structure determination from projections was continuingly enriched, and almost all the introduced concepts are still used nowadays (R. A. Crowther, DeRosier, and Klug 1970). The theory of reconstruction of structures with helical symmetry (Fourier-Bessel reconstruction method) became also more advanced, and different steps of its practical application were extensively described (DeRosier and Moore 1970). Thus, it is not a surprise that many of the three-dimensional structures published in the following years were of helical objects (Moore, Huxley, and DeRosier 1970; Wakabayashi et al. 1975; Amos and Klug 1975; P. N. Unwin and Klug 1974; R A Crowther and Klug 1975).

Brief description of the classical method

What made, and still makes, helices such an appealing target for three-dimensional reconstruction, except the fact that the mathematical background is known since the first hours of EM? As noted in the earliest paper, the reason for this attractiveness is the following: a projection of a helix contains many different views of the structure (the subunit), and in theory "a single view may often provide sufficient information to derive the three-dimensional structure" (R. A. Crowther, DeRosier, and Klug 1970).

We will not detail here the theoretical background of the Fourier-Bessel helical reconstruction method, which we will now refer to as the "classical method", but a simple way to understand how one can combine the helical diffraction theory from (Cochran, Crick, and Vand 1952) and the reciprocal space formulation for three-dimensional reconstruction from projections (DeRosier and Klug 1968) is the following. In 3D, the Fourier transform of a discontinuous helix is 0 everywhere except on the so called "layer planes", which positions are determined by the descriptors of helical parameters, expressed as pitch P and axial rise, or number of turns and subunits in the repeat, t and u (**figure 1.26A,B**). The relationship between layer planes position and helical parameters is called the "selection rule" (Klug, Crick, and Wyckoff 1958), and also involves an integer n, which can take multiple values for each layer plane (the solutions of the equation given by the selection rule) and defines the order of Bessel functions contributing to this layer plane (**figure 1.26B**). What is not obvious at a first glance is that the involvement of the Bessel functions in the theory in only due to the use of cylindrical co-ordinates for this reciprocal-space formulation, and not to the helical symmetry in itself (i.e., in these co-ordinates, Bessel functions would be used whatever the symmetry of the object studied). The particularity for helices is that Bessel terms will be systematically zero, unless their order n satisfies the selection rule, which is thus the "true characteristic of a helical structure" (Klug, Crick, and Wyckoff 1958). The **figure 1.26C** shows the characteristics that the Bessel functions will confer to the 3D Fourier transform: its amplitude is cylindrically symmetric about the meridian, but the phase (colors) oscillates azimuthally, depending on the Bessel order of each layer plane, with n oscillation in one full revolution. If we now consider the projection of a helix as generated by the electron microscope, we understand from (DeRosier and Klug 1968) that the transform of the projection will be a central section in the described transform of the helix as schematized on **figure 1.26D**. Therefore, the central section will cross the layer planes, which will give rise to the so-called layer lines in the observed transform. This crossing will be perpendicular to the layer planes if the helix was imaged exactly perpendicularly to the helix axis (**figure 1.26D, left**), or slightly inclined if the helix has an out-of-plane tilt (**figure 1.26D, right**). Thus, if we start from the Fourier transform of a single projection of a helix, which is at the beginning only one section in 3D transform of the original object (thus not enough to reconstruct), we understand that by determining the order of the contributing Bessel functions to each layer line (a process called indexing), and defining the out-of-plane tilt of the particle, one will be able to "reconstruct" the information originally present on each full layer plane, by using the above-mentioned properties of the Bessel functions. An inverse transformation of this reciprocal information, now three-dimensional, will then enable to recover the desired real-space density information. To determine the order of the Bessel function from the initial transform of the 2D projection, one just needs to understand the argument of the Bessel function which characterizes the distribution of amplitudes, and which depends on the reciprocal radius and the real-space radius of the particle (**figure 1.26E**), while knowing the behavior of a Bessel function as a function of its order and its argument X (**figure 1.26F**) : it will be given by a measure of the reciprocal distance of the first maximum seen on each layer line, knowing the particle radius.

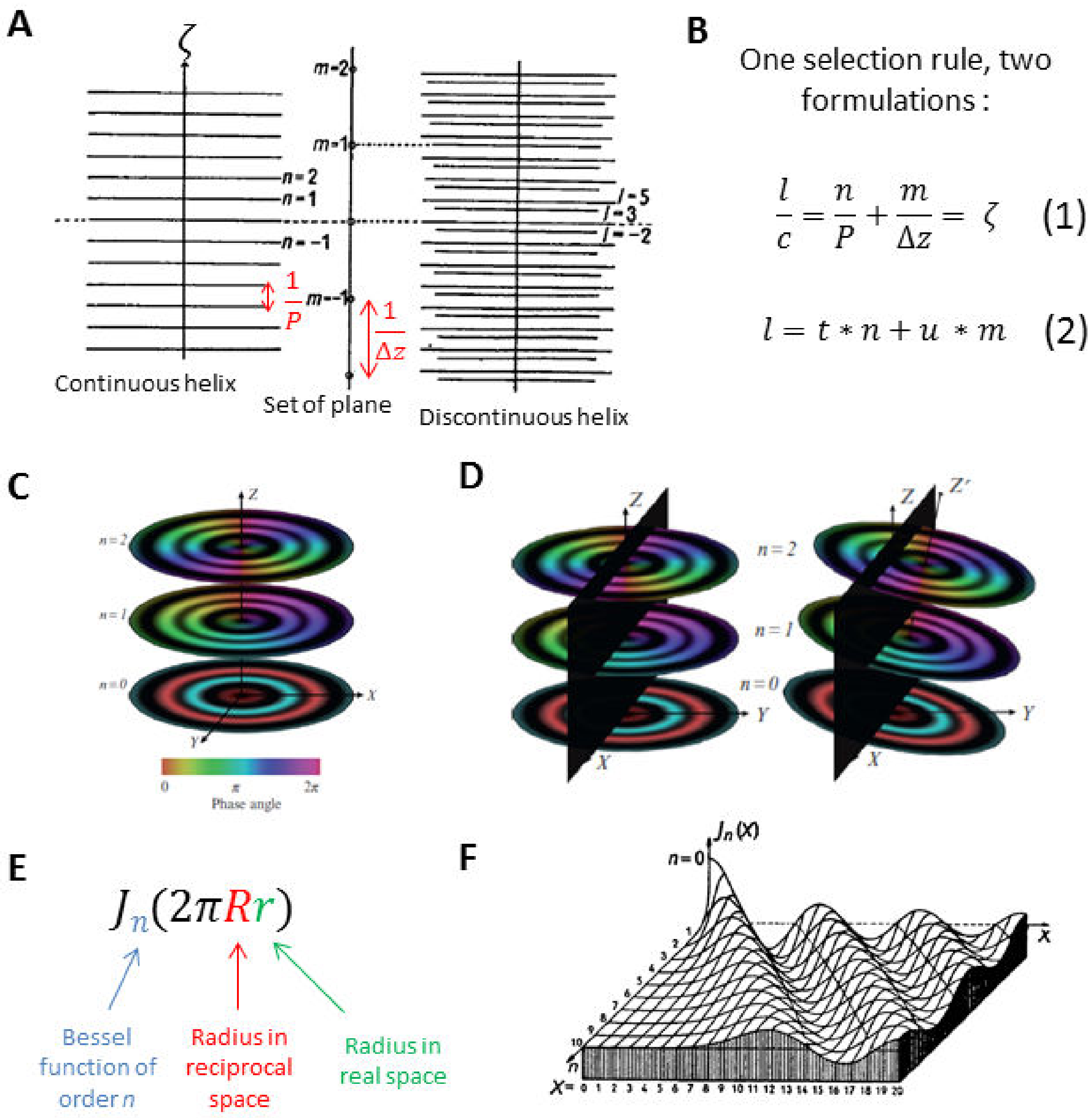


Figure 1.26 : Classical method : a few notions
A discontinuous helix can be seen as a multiplication of a continuous helix with pitch P and a set of planes separated by a distance Δz. The transform of a continuous helix is finite only at planes at height $\zeta = \frac{n}{P}$ (**A**, left), the transform of a set of planes only at points at heights $\zeta = \frac{m}{\Delta z}$ (**A**, middle). The transform of a discontinuous helix is the convolution of the transform of each of its component, and thus is not zero only at planes at heights ζ defined by the selection rule (**B**). A representation of the layer planes in 3D Fourier space shows the behavior of the Bessel functions (**C**). The Fourier transform of 2D projections of an helix are central sections in this Fourier space filled by layer planes, and will thus exhibit layer lines (**D** ; left : no out-of-plane tilt ; right : with out-of-plane tilt). **E :** The arguments of Bessel functions contributing to the layer lines. **F** : representation of Bessel functions of different orders n as a function of their argument X. Figures on **A** and **F** were reproduced from (Cochran, Crick, and Vand 1952), and from panels **C** and **D** from (Diaz, Rice, and Stokes 2010)

Limitations of the classical method

The classical method as described above can suffer from several following limitations.

First, if one uses the method on a single projection as described in the earliest paper, the resolution that can be achieved will be limited by the relatively poor signal of a single image and the limited number of views of the subunit. A way to overcome this, would obviously be to combine information from several projections, if they correspond to objects of the same helical symmetry parameters (Wakabayashi et al. 1975). As a clear counter-argument to the one that the classical method is limited in resolution, one must cite recent studies making elegant use of the classical methods and culminating in reconstructions at resolutions below 5 Angstrom (A Miyazawa et al. 1999; Atsuo Miyazawa, Fujiyoshi, and Unwin 2003), which could in some cases be used even for ab initio building of an atomic model (Yonekura, Maki-Yonekura, and Namba 2005) !

Another limitation comes from the fact that, for each layer line, there are systematically several solutions n of the selection rule, so that several Bessel functions of different orders contribute to them simultaneously (DeRosier and Moore 1970). Depending on the different order n on the same layer line, Bessel functions can overlap even at low to middle resolution, in which case it will be very difficult to index the transform and extract the layer line information properly. Fortunately, many structures do not suffer from the Bessel-overlap problem until high resolution (sometimes even until resolution which was anyway impossible to achieve for other reasons), thanks to the behavior of Bessel functions of high order n (**figure 1.26E,F**) which are effectively 0 until a certain radius in reciprocal space, i.e. until a certain resolution. If only these higher order terms overlap with a lower order term on a particular layer line, the Bessel-overlap is not a problem anymore. Another type of Bessel-overlap occurs when two Bessel functions of relatively close order contribute to two different layer lines that are at a very close reciprocal height, so that they cannot be distinguished from the Fourier-transform of the original projection. A number of methods were designed to solve or at least limit this problem, including decomposition algorithm combining data from different views (R A Crowther, Padrón, and Craig 1985) or tilting the specimen (Stewart and Kensler 1986). One should also note recent developments of the classical method, which seem to efficiently overcome this problem (H. Wang and Nogales 2005).

A last but not least restriction of the classical methods is that they are limited by the requirement of high helical order in the sample to be studied. Below certain regularity, the indexing of the Fourier transform which is a prerequisite of the method will become impossible.

Overcoming limitations of classical methods by single particle approaches

To overcome this problem, new computational methods based mainly on single particle image processing techniques (eventually combined with classical helical reconstruction), have thus appeared, first described in (Bluemke, Carragher, and Josephs 1988). Most of the developments of these methods had actually only emerged since the beginning of the 2000's, and were successfully implemented to address a number of problems in helical assemblies (E H Egelman 2000; E. H. Egelman 2007; Holmes et al. 2003; Sachse et al. 2007; Li et al. 2002; Pomfret, Rice, and Stokes 2007). The relative novelty of such approaches leaves room for constant developments of many "adds" of sub-steps in the processing and significant improvements (Ramey, Wang, and Nogales 2009; Ge and Zhou 2011; V. Korkhov and Sachse 2010), as well as discussions about the best way to use this or that approach. During my thesis, I tried, as far as possible, to follow some of those improvements and to take them into account for the image processing, although this was not always possible due to the high rate of introduced changes. We also tried to take part in the methods evolution, by providing our own "adds", as we will see later.

The single particle-like algorithms cited above share some common points. First, for the use of these algorithms, images of several helical filaments are chopped into small fragments, each containing typically a few to tens of turns of the helix. Then, similarly to single particle processing, fragments can be sorted based on their features to reduce heterogeneity, which we will further detail in the result section of this manuscript. A reconstruction of the filament can be then calculated by placing each segment in 3D space according to its relative orientation through iterative alignment and sorting.

The principle difference between currently existing methods is the way to take the helical symmetry into account.

In the most widely used method, the so-called iterative helical real space reconstruction, IHRSR (E H Egelman 2000), the algorithm determines the local helical symmetry present in a reconstructed volume, imposes this symmetry, and then uses this new volume as a reference for a subsequent cycle of projection matching (**figure 1.27**).

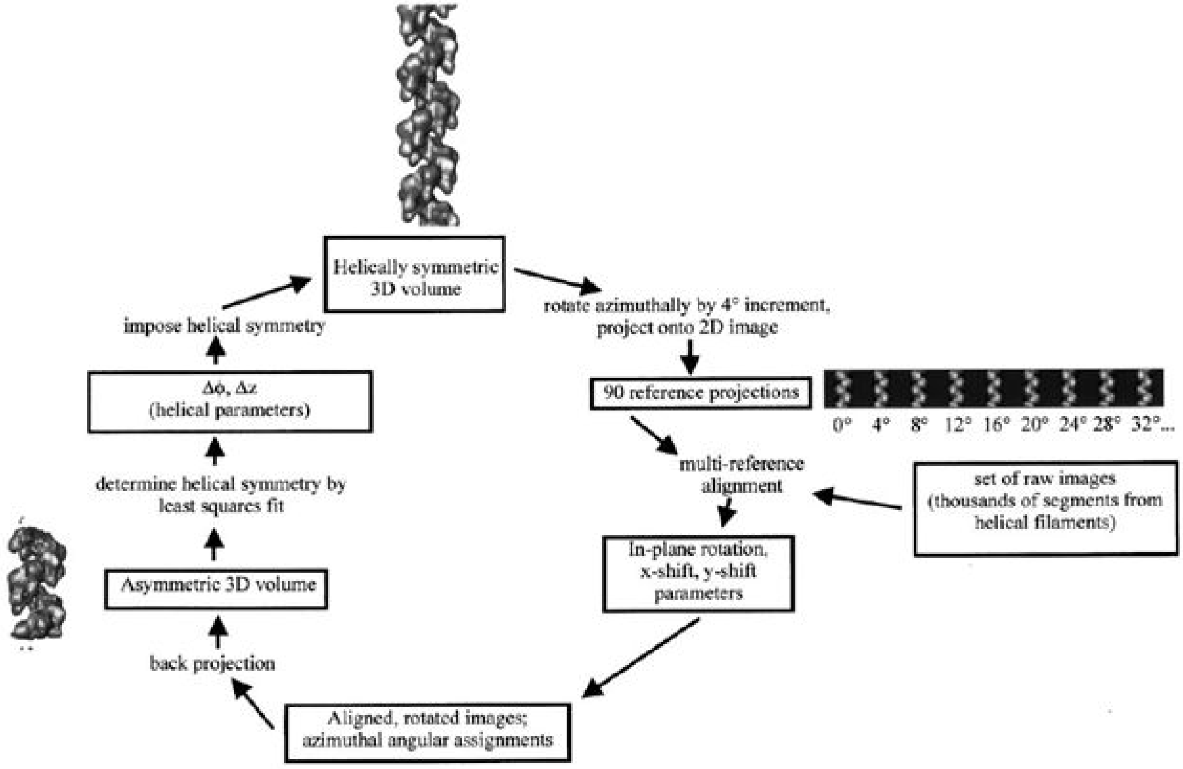

Helically symmetric 3D volume
rotate azimuthally by 4° increment, project onto 2D image
90 reference projections
0° 4° 8° 12° 16° 20° 24° 28° 32°...
multi-reference alignment
set of raw images (thousands of segments from helical filaments)
In-plane rotation, x-shift, y-shift parameters
Aligned, rotated images; azimuthal angular assignments
back projection
Asymmetric 3D volume
determine helical symmetry by least squares fit
Δφ, Δz (helical parameters)
impose helical symmetry

Figure 1.27 : The IHRSR method

The cycle used for the iterative algorithm. The procedure begins with a helically symmetric reference structure (shown at the top), which is then rotated about the filament axis to generate reference projections. These projections are used in a multi-reference alignment procedure with the raw images to determine the five parameters associated with each raw image: an azimuthal angle, an in-plane rotation angle, an x-shift, a y-shift, and a cross-correlation coefficient against the reference. The in-plane rotation and the shifts are applied to each image, and these ``aligned'' images are then used with the known azimuthal orientations to generate a 3D reconstruction by back projection. The resulting volume has had no symmetry imposed upon it, but is clearly a segment of a helical filament. A least-squares procedure is used to determine the helical symmetry of this segment, and these parameters are then imposed to generate a new helically symmetric reference volume. The entire procedure is then iterated until a stable solution is obtained, with no further changes in helical symmetry.

Figure and legend were adapted from (E H Egelman 2000).

The determination of the helical parameters on the non-perfectly symmetric volume, which is done by the program called hsearch, takes the following steps. First, a starting guess for the axial rise Δz is imposed for a refinement of the angular rotation ΔΦ. The best solution for ΔΦ is determined by calculating the mean square deviation between voxels of density at different symmetry-related positions in the volume, and varying ΔΦ. The minimum in the mean square deviation defines the best fitting ΔΦ. The found value of ΔΦ is then fixed for a refinement of Δz using the same approach, and the two steps are iterated once. In practice, all these calculation are done in cylindrical coordinates (E H Egelman 2000). Once the "best" pair of ΔΦ and Δz are obtained, the helical symmetry is imposed on the volume (using the program himpose), and this newly symmetrized volume is used for a new iteration of projection matching and reconstruction, and the whole process is repeated until stabilization of the helical parameters and of the reconstruction features. One important advantage of the method is its easy automation which allows including it into an automatic reconstruction procedure as used for truly single particle objects, and does not require as much manual intervention as in the classical method. In our hands, and also noted by others (Edward H Egelman 2010; Y. A. Wang et al. 2006), the main disadvantage is the requirement of quite precise initial guesses for the algorithm convergence, and the possible failure to find the correct solution (see the reconstruction part later in this manuscript). Another negative point might be the relatively "inaccurate" way of imposing the symmetry (using himpose program), because it implies interpolation in 3D which may affect the very high resolution terms of the structure. For this reason, a very recent improvement of the symmetrisation algorithm was implemented to minimize interpolations errors (Ge and Zhou 2011).

This last point makes a good transition to another method that was published the first year of my thesis and on which we focused (Sachse et al. 2007), because it has been developed partly to compensate the "inaccurate" way of taking into account the helical symmetry in IHRSR. This method also relies on an iterative projection-matching based algorithm (**figure 1.28**).

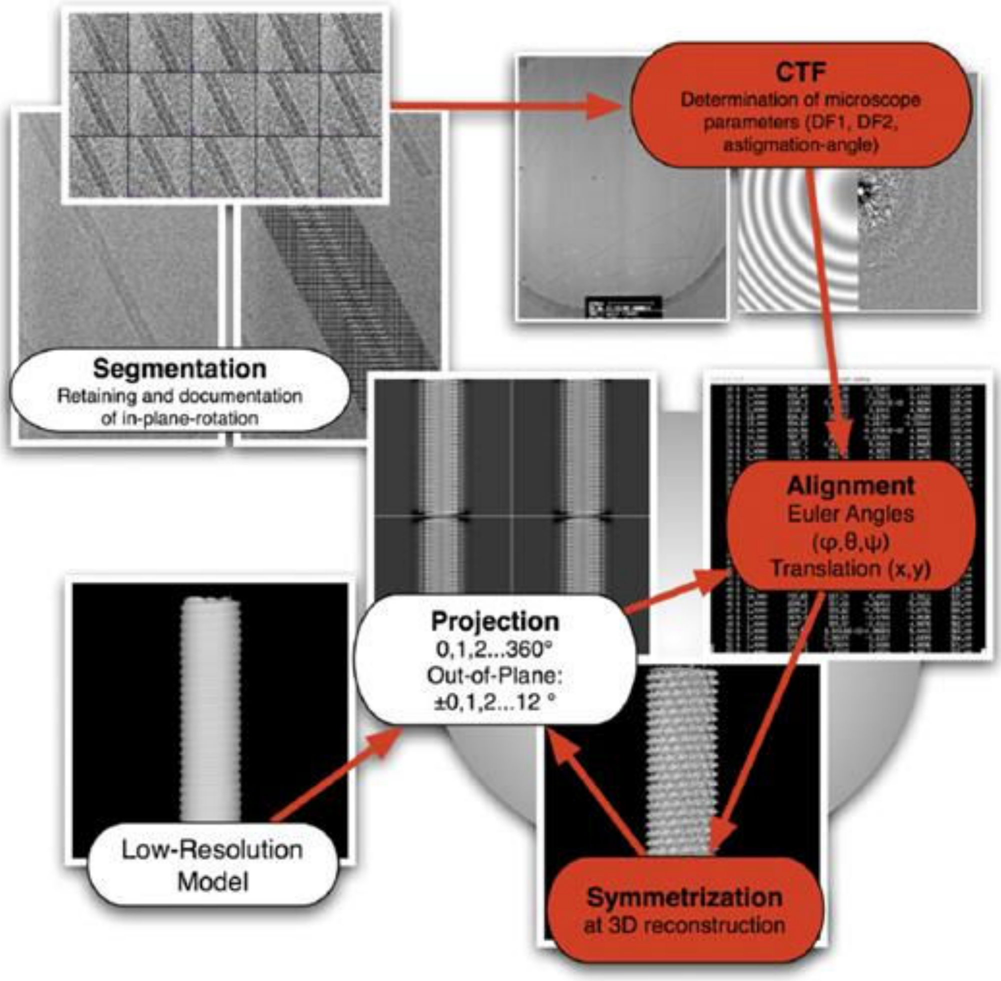


Figure 1.28 : Flow chart of data processing including alignment restraints and symmetrisation based multiple inclusion of the 2D segments (Sachse et. al, 2007)
Major adaptations of the IRSHR procedure are highlighted in red. Segments are processed in an iterative reconstruction cycle based on projection matching. **Alignment:** projections are matched with the overlapping segments extracted from the micrographs. Restraints on alignment are imposed derived from the continuity of the virus particle. **3D reconstruction:** the orientational parameters are used to merge the segments into a 3D volume. Each image is inserted multiple times according to its symmetry-equivalent views
Figure and legend were adapted from (Sachse et. al, 2007)

One major difference however, is exactly the way of imposing the symmetry. The authors note that during the segmentation procedure (which is usually done using the "90% overlap" rule (E H Egelman 2007)), many symmetry-related views are not taken into account, because the distance between successive cropped segments along the filaments is larger than the axial rise Δz. In order to recover these views and in the same time impose the helical symmetry, they propose to include each segment in the reconstruction as many times as the number of missing views. To do this, additional copies of each segment are generated during the alignment, each one shifted along the helical axis by a multiple of the axial rise and included in the reconstruction with a multiple of the rotation angle between the subunits (according to the imposed translation). To avoid including empty areas near the edges of the segments, after translation along the helical axis, the segments are windowed in a smaller image. In addition to exploiting all possible views initially present in the images, this symmetrisation procedure is also more correct than himpose when using a weighted back-projection algorithm (Radermacher 1988) or iterative algebraic reconstruction methods (ART ; (P Penczek, Radermacher, and Frank 1992)) for reconstruction. This is due to the fact that the weights (or in the second case the “correction factor”) that are calculated from the distribution of data in Fourier space are affected by symmetrisation in 3D (as done by himpose) in an “input images independent manner”, whereas this is not the case when multiple version of each image are included in the reconstruction. Additionally to this new symmetrisation procedure, the method of (Sachse et al. 2007) introduces an alignment parameter validation scheme that exploits the geometry of filaments : we will come back to those validations later in the manuscript (part “Introduction into the developed scripts”).

We mostly investigated, and thus detailed, the two methods cited above, but we can refer to different ways of applying single particle approaches for helical reconstruction (Holmes et al. 2003; Ramey, Wang, and Nogales 2009), which should be further explored in the future.

Plan of the manuscript

The manuscript is basically articulated around the successive steps of image processing. After the images of several helical filaments were chopped into small fragments, we applied classification procedures to sort images upon helical parameters, diameter, and other structural features. Although the classification of truly single-particle is widely described in the literature, this is not the case for helical samples. Thus, the first part of the thesis provides a detailed discussion of our results for each project, after an introduction on classification methods. In addition to the classification of real images, this part includes our method for classifying power spectra (amplitudes of Fourier Transform), which we used mainly as a way to detect and sort symmetry heterogeneities.

The second part of the manuscript focuses on the symmetry determination step, which is a prerequisite for 3D reconstruction. In these regards, I introduce our efforts to develop a new method that works on a single 2D real-space projection of a helix. The details of the method, as well as its application to ideal cases and real data set are presented. This work raised interesting considerations concerning ambiguities of helical parameter determination, which is extensively discussed based on the results of our method.

The next step of the processing, the 3D reconstruction by single-particle approaches, constitutes a small part of the manuscript (part 3), in which we discuss some encountered difficulties and possible solutions, and offer perspective for improvement of this part of the processing. The reasons why this part is only briefly developed are twofold. First most of the methods are already described in our article (Desfosses et al. 2011) and in the manuscript in preparation (Desfosses, Ribeiro, Schoehn, Blondel, Guilligay, Jamin, Ruigrok and Gutsche, in preparation). Second the ways that we used to combine processing steps described in the literature and eventual new approaches are mostly detailed in the next part : “Introduction into the developed scripts”.

Indeed, an important part of my work consisted in setting up a processing pipeline that can be used by others, facilitates the use of single-particle approaches described in the literature and add some new possibilities. The last part of the manuscript presents this pipeline and gives guidelines to use it, similarly to a software manual. It also describes two other scripts, the one that applies the symmetry parameter determination on 2D projection described in the second part, and the other one that determines symmetry at the 3D volume level.

Results and Methods

PART 1 : Two-dimensional classification and Introduction to Multivariate Statistical Analysis

Purpose and history

The amount of images that can be produced by an EM experiment is easily very large. Naturally, no human eye can reasonably analyze several thousand of images and gain any useful information out of them. Furthermore, restricting the analysis to randomly chosen small subset of those images, is not only a statistical nonsense, but also impossible due to the superposition of the expected ideal image with random noise, imaging artifacts (especially for negative staining), and low signal-to-noise ratio (especially for cryoEM).

Firstly however, one common initial step when starting a new project and after having recorded micrographs and selected particles, is to try to gain understanding of the structural characteristics of the sample from the 2D images, which classically implies a human visual inspection. Therefore it appears necessary to find a way to compress the total amount of data in order to be able to extract useful information from a large set. What do we mean here by "useful information"? In particular, one may want to answer the questions such as: is the sample homogeneous? What is the global shape –characteristic dimensions- of the object(s)? Is(are) the object(s) symmetrical? Is there conformational flexibility?

Secondly, in the case of a data set that presents heterogeneity, whether it arises from conformational variability or from composition (e.g. in the case of protein complexes) of the observed particles, one absolutely needs to distinguish between various states and separate them into more homogeneous subsets if one wishes to use averaging of images for 3-dimensional reconstruction purposes (Klaholz, Myasnikov, and Heel 2004).

Finally, compressing the total amount of information in order to separate the data set into classes of similar images makes possible to compute class-averages with increased SNR. This is not only useful to visually characterize the sample as already mentioned, but is often needed in order to construct a first initial model in the case of an unknown structure, should one wish to use common lines techniques (Serysheva et al. 1995) or random-conical tilt

(Radermacher et al. 1987). As we will see in the next part of this manuscript, creating those higher SNR class-averages can also be useful to determine the symmetry of the object of interest, for example in the case of helical sample.

The first attempts to reduce the high complexity of large experimental datasets used alignment and averaging of original images (Markham, Frey, and Hills 1963; Joachim Frank 1975). Several methods were employed, ranging from the use of a complex physical apparatus in the first descriptions of EM image averaging (Markham et al. 1964), to more modern computing methods, in particular the largely and currently used auto- and cross-correlation functions for this purpose (Joachim Frank et al. 1978; Joachim Frank, Verschoor, and Boublik 1981). An important problem that should be taken into account prior to averaging is however that single molecules can lie in different orientations on the support film. Moreover, one also has to consider possible genuine structural variations in the sample as well as possible systematic variations in stain distribution when using staining techniques. A significant step towards the solution of those problems was the introduction of multivariate statistical analysis (MSA) methods, usually in the form of correspondence analysis (Marin van Heel and Frank 1981; Joachim Frank and van Heel 1982), which leads to a large reduction of the total data volume and thus facilitates the understanding and the classification of the data set for averaging purpose. Being probably the most widely used method, it will be described in more detail.

General description of the existing methods

It is not the point here to give a detailed mathematical background of the method, that can be found for example in (Joachim Frank and van Heel 1982), but since it is of primary importance to have at least an “intuitive” or a “visual” understanding of the classification method to be able to interpret our analysis and results, we will introduce the basics of commonly employed classification methods.

Principles of correspondence analysis illustrated on a model data set

Data set description

The **figure 2.1** illustrates the successive steps of this classification method based on the principal component analysis, using an example chosen by Marin Van Heel for his PhD thesis: a purely artificial data set of 64*64 pixel images representing human-shaped heads. The **figure 2.1A** shows the sources of variability that was introduced in this artificial data set, to simulate the variability that can be found among real images. Three parameters were varied on the heads: the shape of the head (round or long), the size of the mouth (large or small) and the direction of the look (left or right). Thus, eight combinations of those parameters are possible, and the aim of the classification will be to identify the three different sources of variability and then to separate images according to their characteristics. It must be noted that although in this example the different images correspond to different objects (except the variation of look direction that could correspond to 180 degrees flipped view), the same general reasoning and method outline would apply to different views of the same object. To be more realistic, a random noise is then added to the images, to create 10 different copies of each of the possible head-characteristic combination (**figure 2.1B**).

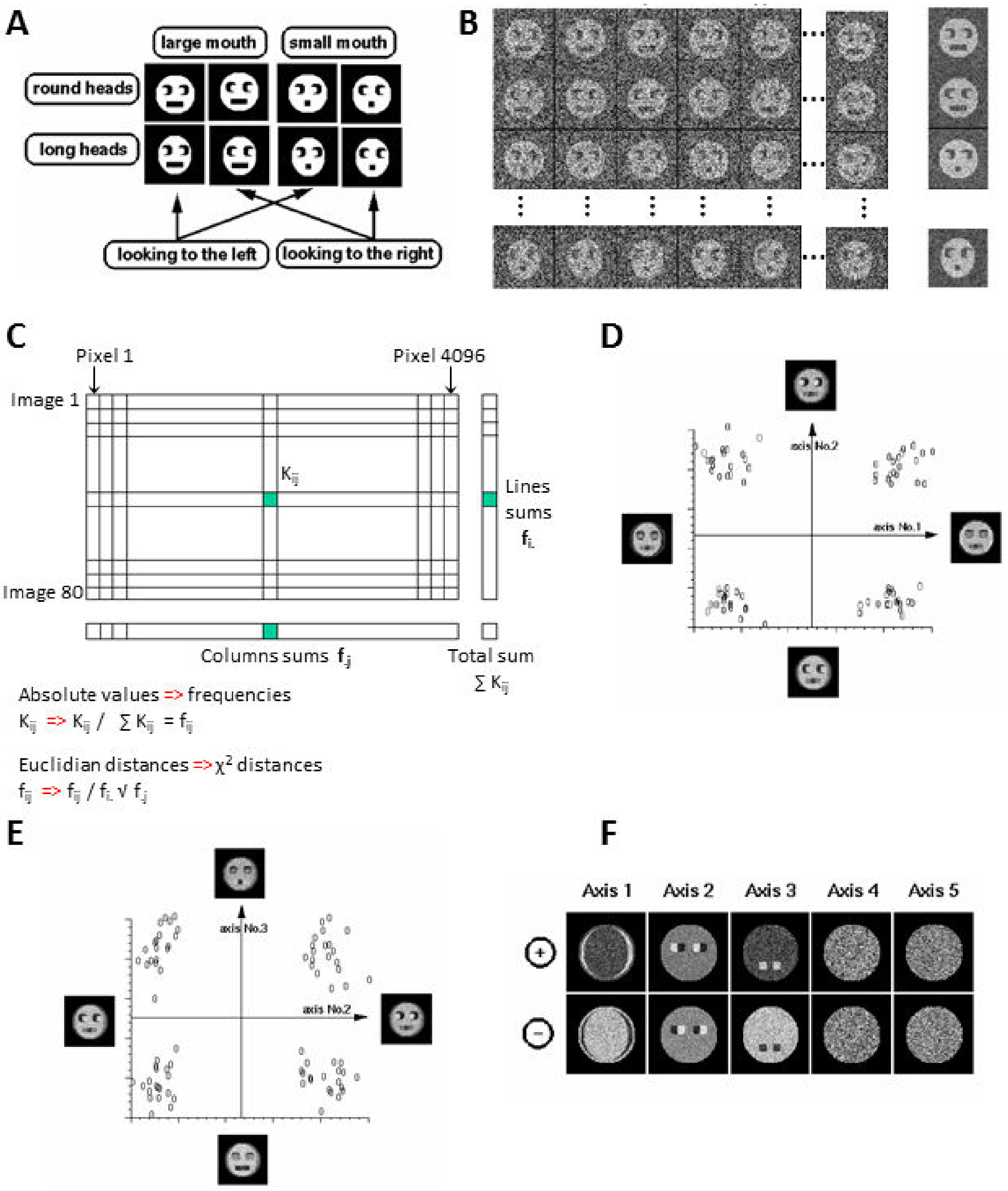


Figure 2.1 : Illustration of the Multivariate Statistical Analysis Method

All possible combinations of three different characteristics with two possible feature each were used to create eight versions of human-shaped heads 64*64 pixels images (**A**). For each version, 10 noisy images were created leading to a data set of 80 images to classify (**B**). The data set is then represented on a 2D matrix containing all pixels values (**C**) which is used to calculate the χ^2 distance matrix. From this matrix will be calculated the eigenvectors, in the order of their decreasing relative importance. The map of the images relative to eigenvectors 1and 2 (**D**) and 2 and 3 (**E**) are shown. The first five eigenimages calculated from this data set are shown in their "positive" and "negative" versions (**F**).

Data representation in the eigenvector space

We have thus the “experimental observations”, that are the 64*64 elements of an array that represent the images, as realized in 80 independent sets of measurement, from which one wishes to identify common trends and clusters. A way to do this is to measure the “inter-images variance direction”. First, all the measurements are arranged in a 4096 * 80 (4096=64*64) matrix, as shown on the **figure 2.1C** : each of the measured intensities of the 4096 pixels of each image are distributed in the columns, whereas each image now constitutes an individual raw. From this matrix one can then calculate the $\chi 2$ distances between any two rows or two columns, to give a new symmetrical matrix containing those distances, which represent thus the variances among the initial images. This new matrix is finally used to calculate the eigenvectors and corresponding eigenvalues that characterize the inter-images variance directions, according to the following steps :

-The strongest variance direction (= biggest direction of extension of the data) is represented by the first eigenvector (i.e. also factorial axis or eigenimage), and the coordinate system of the data cloud is rotated/ shifted such as the first unit vector of the new coordinate system points in the direction of the maximum inter-image variance (first eigenvector). In our concrete example, the first factorial axis corresponds to the variation of the head shape (**figure 2.1D**).

-One then determines the next maximum inter-image variance (second eigenvector), orthogonal to the first. The fact that each new eigenvector is orthogonal to all the precedent indicates that they characterize independent variations (non-correlated). Once again, one re-orients the coordinate system in regard to the new factorial axis. In Van Heel’s artificial data set, the second biggest direction of extension of the data is related to the variations in the direction of the look (**figure 2.1D**).

-This step can be repeated to calculate eigenvector 3 (corresponding to the mouth size in our example, **figure 2.1E**), then eigenvector 4, and so on. Therefore, one can express in decreasing order all the independent variances of our data.

In this way, each image can be represented by an expansion of eigenvectors or factors, which are ordered by their relative importance.

An alternative explanation of the method, equivalent but maybe intuitively more accessible, is the following: each image can be represented by a point in a multidimensional space where each axis is used to represent the intensity of one pixel of the image. Thus, the number of dimension of this space is equal to the number of pixels in the original image. According to their characteristic, the points corresponding to the different images in the data set may form separated clusters in this space, and/or show common trends illustrated then by grouping the images when looking along certain direction of the multidimensional space. Determining the directions of extension of the data cloud in the multidimensional space is then similar to determining the eigenvectors/factorial axis mentioned above. One would try to recover the preponderant vectors that, once combined, would make possible to place each image into the space. The representation in **figures 2.1D** and **2.1E** can be related with the current explanation, each point would represent an image and one would look at the multidimensional space along the direction perpendicular to the eigenvectors 1 and 2 (for **figure 2.1D**) or 2 and 3 (**figure 2.1E**). In this explanation, it becomes clear that the coordinate system representing the extension of the data in the multidimensional space defines an intensity value for each pixel, meaning that each vector can actually be represented as a pair of "positive" and "negative" eigenimages.

The meaning of eigenimages

The **figure 2.1F** illustrates for our artificial example the eigenimages corresponding to the first 5 factorial axes. Because one can read the direction of variation in two opposite ways, there is always a "negative" and a "positive" version of the eigenimages. It must however be noted that in the usual software performing correspondence analysis only one of the two versions of the eigenimages are shown, so only one will be shown in our real-case examples. Having introduced only three different sources of variability in the artificial data set, we can remark that only the first three of those show relevant information, the last two only show a noise pattern. Thus, although each artificial image of the data set was the combination of various sources of variability, the method was able to identify each of them separately. In a real case, it means that even if different sources of variability contribute to each image, like when structural heterogeneity is present in addition to the distribution of views, one can in theory separately distinguish each of those using MSA.

As the method of calculation of the factorial axis suggests, the successive eigenimages should reflect the decreasing order of the variances of the data: the first eigenimages correspond to the greatest degree of variability, the second to the second highest, and so on. In this artificial example however, the three different characteristics of the head were introduced exactly in the same relative amount, and exactly the same number of noisy images were created for each characteristics combination. Thus in this particular case the order of the eigenimages does not arise from the prevalence of one source of variability, but most probably from the fact that the images were more "affected", in term of number of pixels concerned, by the changes in the width of the head, less by the look direction, and even less by the mouth size. If only 5 % of the images represented a thinner head, the corresponding pair of eigenimages would not be the first in the order.

The benefits of data reduction

We should also note the large reduction of data that was achieved: instead of the 64*64 = 4096 density values per image, each image can now be described as a combination of only a few eigenimages associated with relative weights. In the chosen example only three eigenimages were sufficient, and it is remarkable that for a real data, usually less than 10 are already sufficient (Marin van Heel and Frank 1981) ! This huge compression of the total amount of data will then make the next step, the classification of the images, much simpler. As will be described in our experimental examples of classification, one can also benefit from the reduction of the data into a few eigenimages in order to classify images only according to some of the eigenimages. For example, in Van Heel's example, one could have first classified images according to the look direction using the weight of the corresponding eigenimages.

Classification strategies

The literature on classification of large image data sets is rich (Ohi et al. 2004; White et al. 2004), and we will not go into the details of different methods used as it is not necessary for the understanding of our results. The general classification strategies all rely on distance measurements to evaluate similarity between images studied. From those similarities measurements, the classification algorithm will then aim at partitioning of images into classes. Classification strategies can be globally divided into two main categories:

-Direct methods (**figure 2.2A**), like the widely used k-means clustering (PA Penczek, Zhu, and Frank 1996), where the images are grouped based on their distance from a set of predefined classification centers. To refine the classification and make it more “data-based”, one can take the centers of mass of the newly determined classes as new classification centers, and repeat the procedure until stabilization of the classes’ content. In this method, the number of classes chosen is simply related to the number of classification centers used for data clustering.

-Hierarchical procedures (**figure 2.2B**), in particular the Hierarchical Ascendant Classification (HAC) procedure implemented in the IMAGIC software (M. van Heel et al. 1996) used in this work for classification purposes. In this method, one starts with as many classes as there are images, and then merges two most similar classes at a time to form bigger classes until one ends up with one class containing all the images. The “stop signal” of this merging procedure is thus just given by the final number of classes wanted. The algorithm can be represented by a tree in which the merging of classes can be followed.

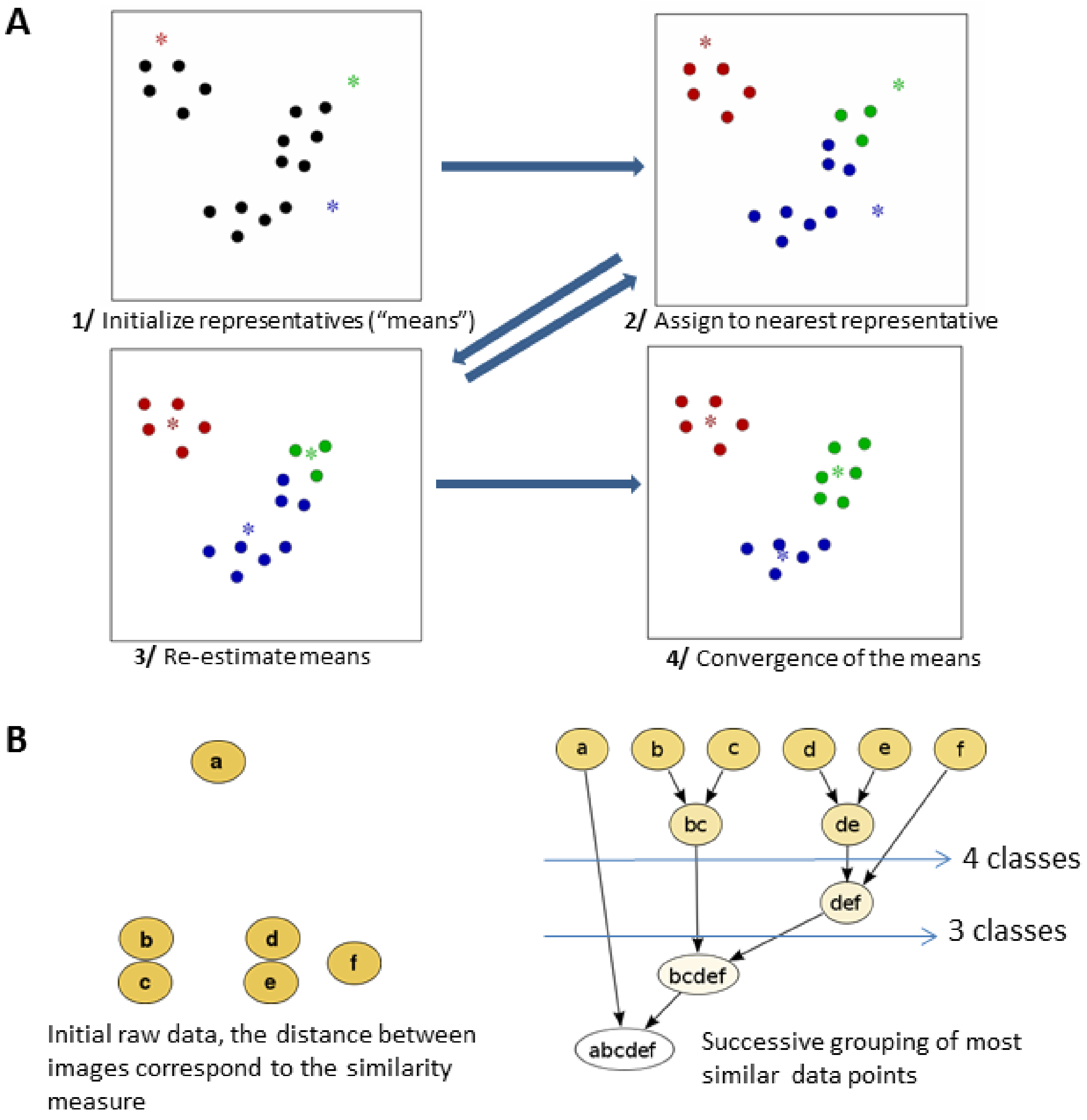

A
1/ Initialize representatives ("means")
2/ Assign to nearest representative
3/ Re-estimate means
4/ Convergence of the means
B
a
b
c
d
e
f
Initial raw data, the distance between images correspond to the similarity measure
a
b
c
d
e
f
bc
de
4 classes
def
3 classes
bcdef
abcdef
Successive grouping of most similar data points

Figure 2.2 : Two different classification strategies : k-means clustering and HAC

(A) The k-means clustering algorithm partitions *n* data points (the black dots in 1) into k clusters in this way : seeded with k initial cluster centers, e.g. randomly placed (1 : colored asterisks), it assigns every data point to its closest center (2) and then re-computes the new centers (3) as the means (or center of mass) of their assigned points. This process of assigning data points and readjusting centers is repeated until it stabilizes (4).

(B) In the HAC method, one also start from the similarity measures between images (represented left) and first consider each image as a class. The algorithm will then merge two most similar classes at a time to form larger classes (right). Therefore, at each merging event, there will be less classes left. The algorithm is stopped when the wished number of classes is obtained (the blue arrows on the right part represent the stop signal).

Classification and image alignment

Until now, for the sake of clarity we omitted the fact that in practice MSA/classification is usually coupled to a step of alignment of raw images. The aim of the alignment is to get rid of three different sources of variability among the images: rotation in the plane, translations in x, and translations in y. In this way, one can avoid creating classes which differ only in translation and in-plane rotation of the object while being redundant in terms of the actual viewing direction. Additionally, alignment enables grouping more images in fewer classes, which both improves the SNR in the classes and reduces the data size, thus making the analysis easier. Moreover, the class-averages will show less blurring due to averaging of misaligned images. To keep the classification procedure as unbiased as possible, one will try to use as alignment references only data that was produced without a priori considerations: the class-averages themselves. The **figure 2.3** illustrates the iterative procedure that is often used to couple MSA, classification and multi-reference alignment. As shown in the **figure 2.3**, often not all the classes are used as references for the alignment of the data set, firstly because one wants to keep for each characteristic out-of-plane view only one in-plane rotation class, and secondly because one might wish to discard some classes, for example if they show a poorer quality than others (less detailed), eventually that correspond to particles other than the actual object of study (for example individual components of a complex of interest), or show a bad centering. However, this selection step makes the alignment procedure more biased, and one needs to be as careful as possible to always select the most representative classes, while paying attention to avoid redundancy of references. For the above-mentioned reasons, it is clear that if one wishes to extract detailed information from a classification of the data, one need to carefully execute several manual tasks and perform a lot of visual inspection; these time-consuming but clearly crucial steps are described in many application examples, like in (Burgess et al. 2004). Moreover, it should be specified that the above example of the means of integrating the classification into an iterative procedure is, although common, not the only way of processing, and additional steps may be added to this general pipeline. Just as an example, it has been proposed to refine the classification of the data according to more subtle changes than the viewing direction, for example the presence or absence of a small ligand, by re-classifying images that were already grouped in an orientation class (Klaholz, Myasnikov, and Heel 2004; Elad et al. 2008).

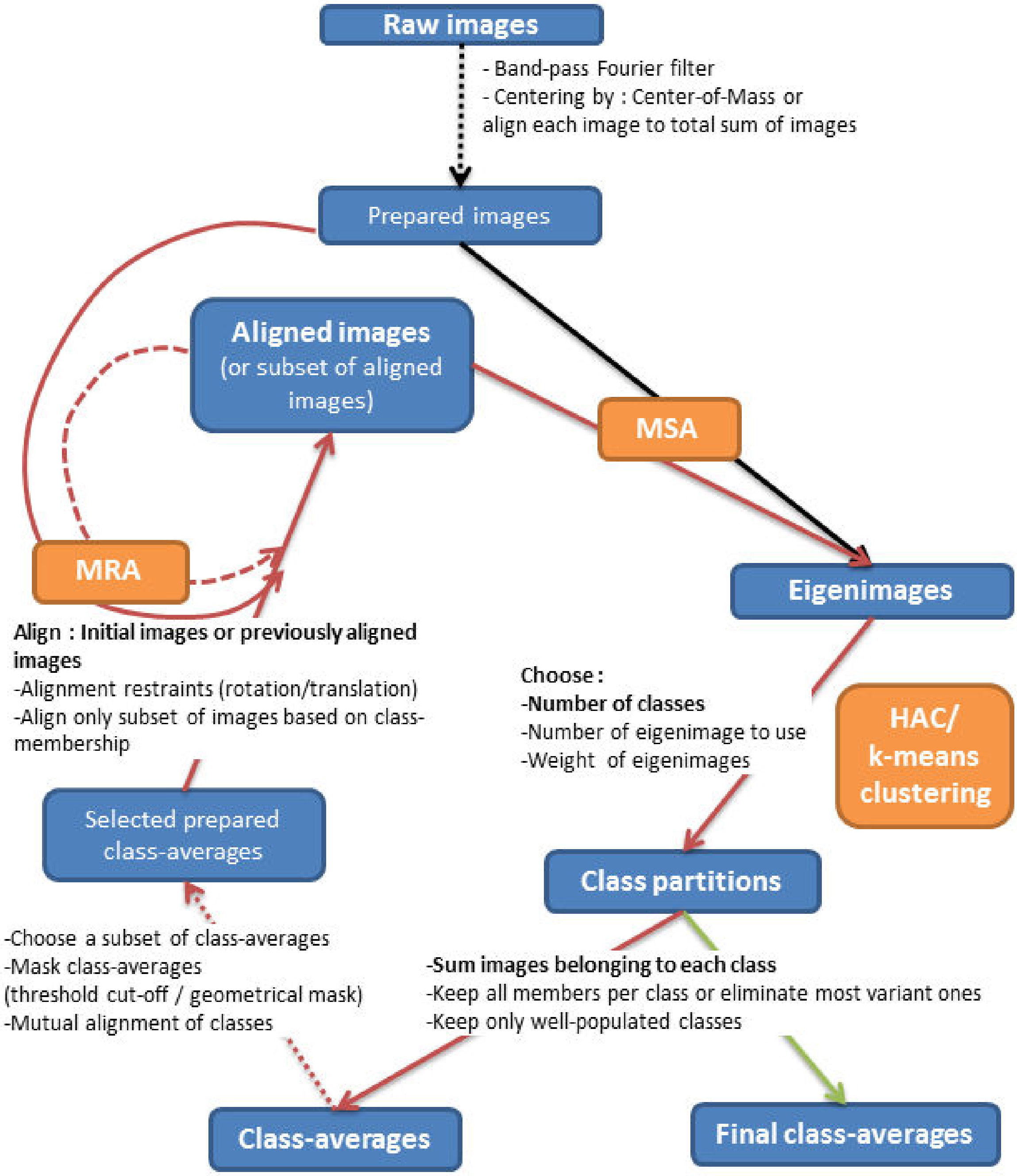

Raw images
- Band-pass Fourier filter
- Centering by : Center-of-Mass or
align each image to total sum of images
Prepared images
Aligned images
(or subset of aligned images)
MSA
MRA
Eigenimages
Align : Initial images or previously aligned images
-Alignment restraints (rotation/translation)
-Align only subset of images based on class-membership
Choose :
-Number of classes
-Number of eigenimage to use
-Weight of eigenimages
HAC/
k-means
clustering
Selected prepared class-averages
Class partitions
-Choose a subset of class-averages
-Mask class-averages
(threshold cut-off / geometrical mask)
-Mutual alignment of classes
-Sum images belonging to each class
-Keep all members per class or eliminate most variant ones
-Keep only well-populated classes
Class-averages
Final class-averages

Figure 2.3 : Scheme of the steps for image classification

The main obligatory steps are noted in bold, other facultative steps are noted in normal characters. In the case of the separation of the data set in multiple subset of images according to class-membership, all steps of the refinement are then done for each subset. Abbreviations used : MSA Multivariate Statistical Analysis ; MRA Multi-Reference-Alignment ; HAC Hierarchical Ascendant Classification

→ : First iteration

→ : Iterative refinement cycles

→ : Last iteration

Dashed arrows indicate optional steps that can be skipped (then one would directly go to the next step).

Further considerations and our examples of applications

We have seen so far how MSA and HAC can be used to classify noisy images, heterogeneous in viewing direction and/or presenting genuine structural variability. It should be noted that very different methods were proposed, and applied, for classification of noisy heterogeneous EM images, including the use of self-organizing maps (SOM ; (Marabini and Carazo 1994; Radermacher et al. 2001)), or topology representing networks (Ogura, Iwasaki, and Sato 2003), the latter claiming in one paper much better performance over other methods. However, it is not fully clear if the relative non-success (in term of broad use by the EM community) of alternative methods is due to non-clearly announced disadvantages or if it is just a result of a loss in the unavoidable competition between software packages available for similar tasks, where the outcome depends not only on efficiency and precision, but also on factors like user-friendliness, accompanying advertisement, etc.

Various applications of the MSA/HAC methods are very widely described in the literature, thus it might seem unnecessary to dedicate one part of this manuscript on the application of classification techniques to our data. However, we will see now three application examples, each one for a particular reason.

The first example concerns classification of top-views, i.e. along the helical axis, of very short helical segments, or rings/pseudo-rings, of recombinant Measles N-RNA and of reconstituted VSV N-RNA bullets. These single-particle-like examples are on one hand a good illustration of how one can sort a heterogeneous data set with different particle sizes/characteristics, and on the other hand how symmetry (or here "pseudo-symmetry") information can be extracted from the analysis of the eigenimages and classes. Although not explicitly described in our articles (Desfosses et al. 2011) (Desfosses, Ribeiro, Schoehn, Blondel, Guilligay, Jamin, Ruigrok and Gutsche, in preparation), this rough information on symmetry was useful to further gain confidence in values independently obtained from the analysis of side-views of the helical particles.

The second example focuses on the classification of side-views of helical segments for both Measles and VSV nucleocapsid projects, and gives new insights into application of classification methods to helical specimens. Indeed, the literature mainly describes classification of isolated single particles, and it is very rare to find information about classification of helical particles. In particular, the eigenimages accompanying classification are almost never shown, and no explanations concerning their interpretation and the way they can be used are provided. One notable exception is the work of Nogales group (Ramey, Wang, and Nogales 2009), which shows the use of classification to address heterogeneity issues in a relatively didactic manner, although the eigenimage analysis is only briefly described, and which was published while we were in process of performing our analysis.

As a third example of classification, we will introduce our method that uses classification of power spectra of helical segments in order to sort them according to symmetry parameters.

Classification of real images

Rings / Pseudo rings classification

Ring-shaped nucleocapsid structures

The atomic scale structural information available for the nucleoproteins of negative -strand RNA viruses arisen from the natural ability of recombinant N to bind short RNA segments and induce their circularization into rings of various sizes (Albertini, Clapier, et al. 2006; Chen et al. 2004). In 2006, two groups were able to isolate and crystallize rings of N-RNA of rabies virus (Albertini, Wernimont, et al. 2006) and of VSV (Green et al. 2006), thanks to the efforts spent in the biochemical separation of the samples according to the rings' size. Since then however, only one new nucleoprotein structure of a negative-strand RNA virus, namely the one of RSV (Tawar et al. 2009), was solved based on such a circular arrangement. When looking at recombinant Measles N-RNA by negative stain EM, we can also see rings, or ring-like structures, in variable relative amount depending on the expression batch, on the purification conditions and on the enrichment of the sample in only very long helices by glycerol cushion pelleting. However, nobody has so far succeeded in isolating (not even to mention in crystallizing) homogeneously sized Measles N-RNA ring. Furthermore, and there is still no proof that the observed rings are indeed truly circular , and that what we observe on the EM micrographs are not opened rings or very short helices viewed along the helix axis. In particular, we could not reproducibly and convincingly show clear side-views of ring structures, which would definitely address this question, although the existence of true rings does not necessarily imply that we can easily observe side-views by negative staining, due to potential strong preferential orientation on the carbon film, as it was already described for other negative strand RNA viruses, like rabies virus (Iseni et al. 1998; Schoehn et al. 2001). During this work, although we were mainly interested in the structure of the actual helical nucleocapsids, a significant amount of time was spent in attempts to isolate and obtain homogeneous rings samples, by combined use of CsCl gradients, glycerol gradients and native gels, without much success (results not shown).

The number of subunits in the ring-shaped structures as an indicator of the helical symmetry

However, even not purified to great homogeneity, the rings or “pseudo-rings” structures of Measles N-RNA are not uninteresting in the light of our structure determination pipeline. Indeed, their advantage is that the subunits can appear clear viewed from the top, and that one may count their number in one turn of a ring-like structure. Therefore, we may appreciate the variability of number of subunits that can be seen in one turn and further relate the diameter of the observed ring-like structures to the number of subunits that they contain in order to compare this information with the diameter of our helical nucleocapsids. Moreover, determining the major pseudo-symmetry population of the rings can give an indication of the most stable lateral arrangement of the nucleoprotein on the RNA string. Altogether, even if these insights may not allow to directly infer the possible symmetries of the long nucleocapsids due to potential rearrangements of the nucleoproteins between short and long N-RNA structures, they can improve our confidence in the symmetry parameters of our structures.

As far as the reconstituted VSV N-RNA bullets observed in amorphous ice are concerned, that some bullet-like structures were systematically found to be very tilted out of the plane of the ice, and sometimes until the structures appear as a large ring. These structures most likely contain more than one helical turn, therefore we have to keep in mind the superposition of the contribution of those turns in the final projection image, and thus carefully interpret our observations.

We performed 2D-classification of ring-like structures of Measles and VSV N-RNA, by following the method described above, using the IMAGIC software. The raw filtered images were subjected to MSA and classified using HAC to produce high SNR class-averages. A subset of representative class-averages was chosen as references for multi-reference alignment of the initial images (using translations only, except for the last alignment cycle for VSV). These steps were repeated until the alignment parameters stabilized and the appearance of class-averages did not further improve. The reason to use only translational search for the alignment of the images is that a rotational alignment could make the main symmetry contribution less clear from the observation of the eigenimages because the total sum of images (first eigenimage) would already contain this information, thereby eliminating the source of variation among the images that we actually aim to observe. As the IMAGIC software manual specifies: "Looking at the eigenimages of the (rotational) unaligned data-set is a powerful method for an unbiased finding of the particle's symmetry" (GmbH Image Science Software, 2010). Furthermore, for ring-like structures, the in-plane rotation variability information should be clear from the eigenimages. However, for VSV, due to the very low number of initial images to classify (~200), a last rotational alignment was performed in order to merge images that only differed by an in-plane rotation into common class to be able to have a sufficient SNR to visualize the subunits. The principal results of the classification of the ring-like structures are shown for Measles on **figure 2.4** and for VSV on **figure 2.5**.

Classification of Measles virus N-RNA pseudo-rings

Eigenimages as indicators of the circular symmetry

For Measles, we have a good example of how we can by apply the classification procedures to a raw data set where the noise makes difficult to extract the information that we are interested in (**figure 2.4A**), and which represent a highly heterogeneous data set. When looking at the first eigenimages of the translationally aligned data set (**figure 2.4B**), we can see that the main contributions to the image are circularly symmetric, ranging from C12 to C14. From eigenimages 2 and 3 (the first being the total average of images), we see that the eigenvectors can be grouped two by two, regarding the symmetry they represent. The difference between the two is an in-plane rotation of 360/2n degrees for a Cn symmetry, thus a rotation of half the angular distance between subunit, or, to employ the terms from (M. van

Heel et al. 1996) on a similar case, ‘‘90°’’ out of phase in a rotational sense. Although usually not represented, each eigenimage that we see has an inverted version that implicitly exists, where a white pixel would be black and inversely. Thus, the eigenimages 2 and 3, for example, both have a contrast inverted version so that there are in total four positions represented by the eigenfactors for each “subunit” (a bright white spot) : at 0 degree rotation (arbitrary), at 1 * (360/13)/4, at 2 * (360/13)/4, and at 3 * (360/13)/4 degrees rotation. Thus those eigenvectors sample regularly the different possible position of the subunit in one “asymmetric unit” of the C13 symmetry, and we can logically conclude that their combination makes possible to represent the variability of in-plane rotation for rings of this pseudo-symmetry. The same reasoning can be applied for the eigenimages pairs 4 and 5 (C14 symmetry), 6 and 7 (C12 symmetry). The order of the eigenimages reflecting the relative predominance of a particular contribution to the images, we can draw the conclusion from this analysis that the preferred pseudo-symmetry of those short helices/rings is 13 subunits per turn. Thus, the lateral contacts between subunits may be energetically more favorable for this pseudo-symmetry, providing an additional argument for our later trials for symmetry determination from side view of longer helical segments. It can be noted that the 3D reconstructions actually showed later that this a priori favored 13 subunits per turn symmetry is very close to the one of the majority of the nucleocapsids segments (~12.9 subunits per turn). From the eigenimages alone, it seems difficult to conclude anything regarding the fact that N-RNA of Measles could form truly closed rings, as would be needed for crystallization, or not. Indeed, although we see circular symmetries on the first eigenimages, we first have to keep in mind that a projection perpendicular to a ring plane would look almost identical to a projection along the helical axis of a very short helix, e.g. of only one turn. Secondly, if those symmetrical contributions are the most preponderant, the images are actually the expansion (combination) of several eigenimages, and the next ones usually represent asymmetrical contributions, like the eigenimage 9 in **figure 2.4B**.

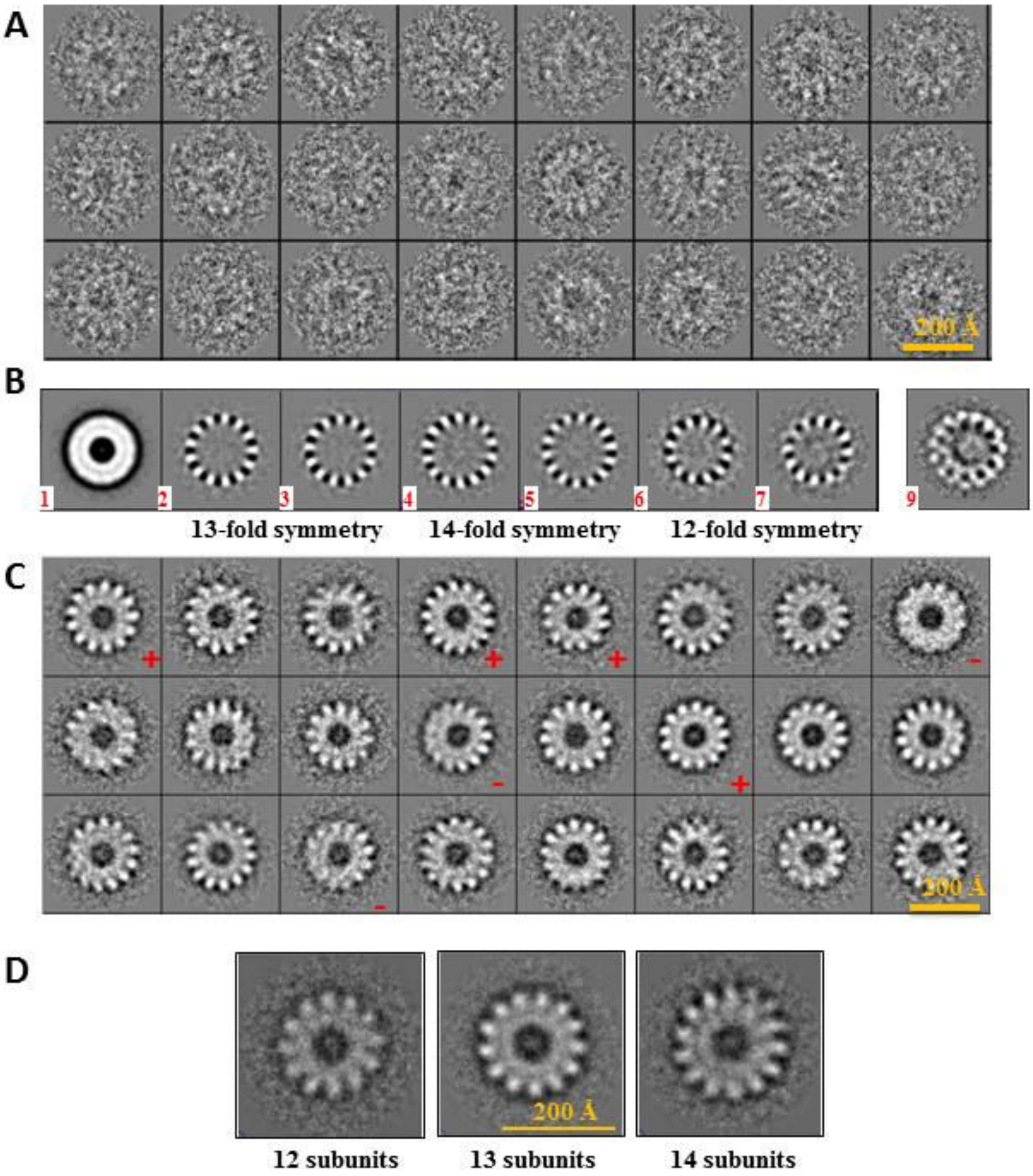


Figure 2.4 : classification of ring-shaped structures of MeVNC (negative staining)

A gallery of representative raw images of ring-shaped structures of MeVNC used for classification is shown (**A**). Eigenimages (**B**) arising from the translationally aligned data set illustrate the variability of pseudo-rotational symmetries found among the sample and the preponderance of the pseudo 13-fold symmetry, as well as asymmetrical contribution. The numbering of the eigenimages is indicated in red. A subset of the obtained class-averages (**C**) show various class quality, from well-resolved and symmetric classes (noted with **+**) to poorly-resolved and/or asymmetric classes (noted with **-**). The badly-resolved side of the asymmetric class-averages suggest that it corresponds to the begin of a second turn of a very short helix that superposes with the first turn. Some selected class-averages corresponding to various number of subunits are shown in (**D**).

Interpretation of the class-averages

The **figure 2.4C** shows some the final class-averages obtained by our classification procedure. The high level of signal makes now possible to manually count on the class-averages the number of subunits that they represent, that vary, as predictable from the eigenimages, from 12 to 14. We can also see that some classes show a less clear definition of subunits (some are noted with a '-" on **figure 2.4C**) on one side of the ring, that would be an indication that those parts actually correspond to the superposition with the begin of a second turn of a very short helix. On the other side, some class-averages look much more symmetric and regular in appearance (some are noted with a "+" on **figure 2.4C**). Again, this is not a sufficient proof that such truly symmetric structures exist, as this regular aspect might simply result from averaging of several non-perfectly closed rings together. Upon a more thorough examination of the well-defined class-averages (**figure 2.4D**) for each pseudo symmetry, we can see that the apparent diameter ranges from ~190 Å for the 12-rings to ~210 Å for the 14-rings and is of ~200 Å for the 13-rings. These values will again be useful to compare with the observed diameter of the longer helical segments and to gain more confidence in the results of independent methods of symmetry determination.

Classification of VSV N-RNA pseudo-rings

Meaning of eigenimages and pattern in class-averages

For VSV, we could isolate from our set of 88 cryoelectron micrographs around 200 ring-like structures (a subset is shown in **figure 2.5A**). As already mentioned, these rings most likely represent projections of (close to) 90 degrees tilted short bullet-like structures. Due to the low number of images to classify, both the clarity of eigenimages and the quality of class-averages that can be expected is significantly lower. The eigenimages of the non-rotationally aligned data set (**figure 2.5B**) are indeed difficult to interpret except maybe the two first after the total average (i.e. number 2 and 3). The eigenimage number 2 shows that the preponderant contribution to images (source of variance) points either toward the interior of the ring (as seen by the white circle) or toward the exterior (the strong black circle would be white in the inverted version of this eigenimage). This contribution being also relatively constant over the periphery, one can presume that this eigenimage points to a diameter variability of the ring-like structures. The later analysis of the helical segments would tend to confirm this interpretation. The eigenimage number 3 shows a contribution to the images that is actually not circular, but slightly ellipsoid. Associated to this pattern, there is a strong contribution at the exterior of the ellipsoid in the direction parallel to the small axis of the ellipsoid (**figure 2.5B**, orange arrows). Together, these observations suggest that this eigenimage represents the trunk of short VSV N-RNA bullets that are less than 90 degrees tilted out-of-plane of the ice: the end of the "hollow cylinder" structure would then show an ellipsoid projection, with densities corresponding to the projection of the rest of the trunk going out of this ellipsoid in the direction where the rest of the trunk is tilted, so parallel to the small axis of the ellipsoid. After iteration of the classification procedure, we could obtain a subset of class-averages (**figure 2.5C**) from which a subunit pattern could sometimes be identified. The latter became clearer when adding the possibility of rotational search during the alignment (**figure 2.5D**, for example classes surrounded by a blue circle). This pattern (**figure 2.5E**, orange arrows) consists of two distinguishable stronger densities (**figure 2.5E**, green arrows), connected by a lower density region, in agreement with the bi-lobed appearance of nucleoproteins of negative strand RNA viruses at low resolution (Schoehn et al. 2001).

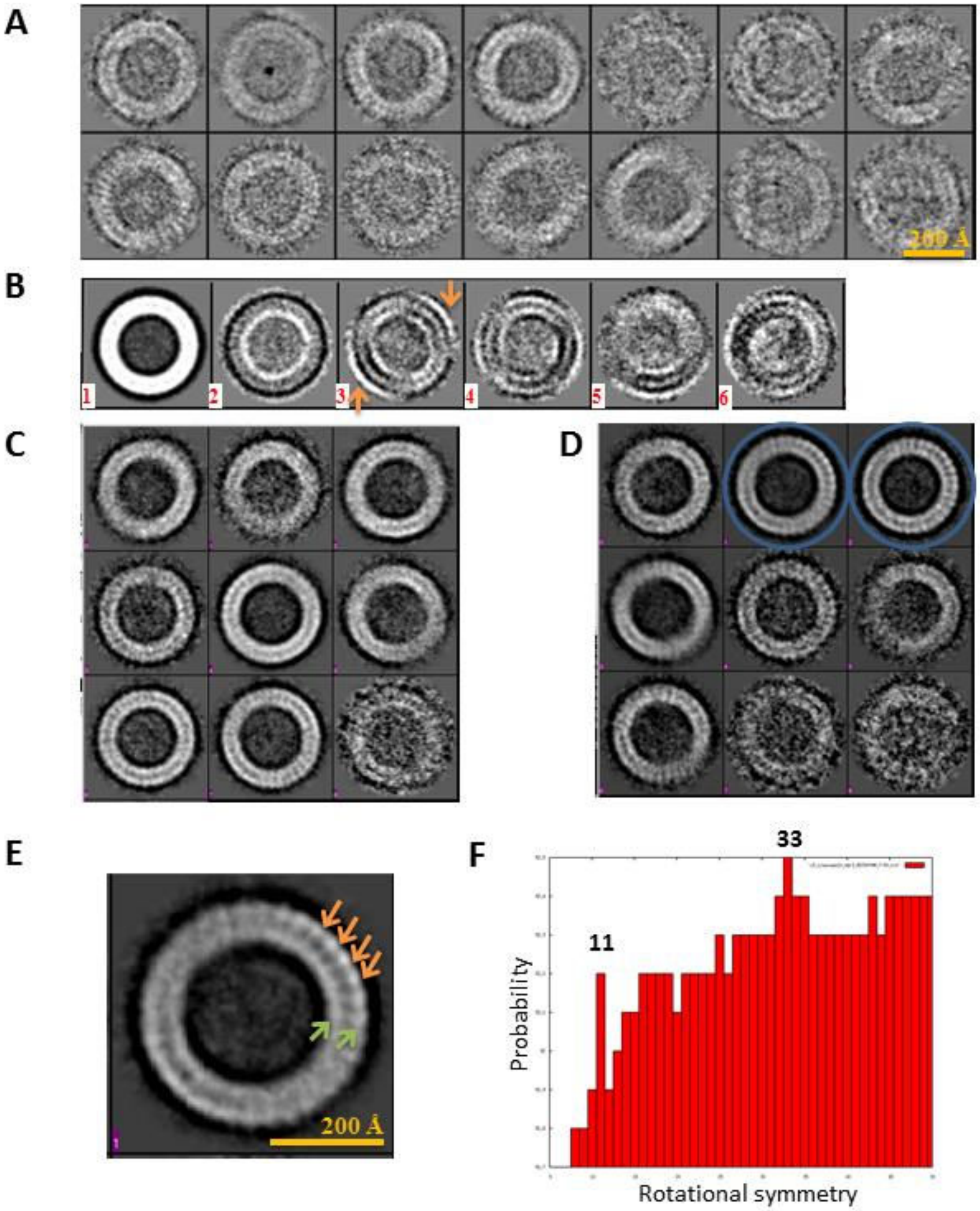


Figure 2.5 : classification of top-views of VSV N-RNA bullets (cryo-EM)

A gallery of raw images of ring-shaped structures of VSV N-RNA bullets used for classification is shown (**A**). The first 6 eigenimages arising from the translationally aligned data set are shown (**B**), and numbered in red. Some class-averages before (**C**) or after (**D**) rotational alignment makes in the best cases appear a subunit pattern (blue circles), which is better visualized on an enlarge version of one class-average (**E**, orange arrows). This subunit pattern is composed of two stronger densities (**E**, green arrows) connected by a lower density region. The panel **F** illustrate the determination of the rotational symmetry of the class-average shown in **E** by the IMAGIC software.

Determination of the circular symmetry

The fact that we observe this subunit pattern is interesting in itself. Indeed, this pattern should be smeared out unless we either have bullets portion with only less than two turns (so that at least on a portion of the ring-like structure there is no interference of the projections of N from successive turns), or the N densities almost superpose in subsequent turns (which would mean that the number of subunits per turn is very close to an integer). We can reasonably exclude the first option, as such extremely short bullets trunks were never observed in any other orientation within the ice and thus can be supposed to be inexistent. Thus we can hypothesize that a non-negligible set of the bullets trunk that were averaged into the classes showing the subunit pattern indeed contain a nearly or exactly integer number of subunits per turn. On some class-averages the number of subunits per turn can be straightforwardly counted by hand (e.g. for the one surrounded by blue circles on **figure 2.5D**). A better and more precise estimation can be made by using for instance routine implemented in IMAGIC that takes an image and a rotational symmetry as input arguments and gives a "probability" (no further details are given in the software documentation) that this symmetry is actually correct. By iterating over all the symmetries that one wishes to test, one can obtain a plot of symmetry probability as a function of the symmetry tested. We made such plots on several class-averages that gave a clear subunit pattern, and we show on **figure 2.5F** one representative plot, calculated from the class-average shown on **figure 2.5E**. This plot shows a maximum of the symmetry probability at 33 subunits per turn, and another clear peak at 11 subunits per turn (i.e. a divisor of 33). Despite these indications, we cannot simply assume this value of 33 as being the number of subunits per turn for two main reasons. First because of the small number of N-RNA bullet top views and the consequently poor statistics, and second because this value was determined only from the classes showing the subunit pattern. For the remaining classes, we do not know if a pattern is not recognizable because the symmetry is too far away from an integer, because the corresponding trunks are too long (so that even an almost integer symmetry would become blurred by the number of turns), because the tilt is too far from 90 degrees, because of a low SNR, or for any other reason. However, we can a posteriori note that the 33 subunits symmetry was independently determined during the analysis of the side-views of bullets trunks for 3D reconstruction purposes as being a major symmetry of the N-RNA bullet trunks.

Further considerations and more examples of applications of classification methods

Up to now, we described the usage of the ring-like structures we observed in the case of the Measles and VSV nucleocapsids as an example of how one can take advantage of classification in order to extract information of a data heterogeneous in various aspects. Although useful, this ring-like data set was actually just a by-product of our sample preparation. We will see now the trials that were made to use classification to gain more knowledge about the long helical particles and to find a way to obtain more homogeneous data sets. This processing step is only rarely described in the literature for helically symmetrical objects.

Examples will be shown both for the negative stain images Measles non-digested and digested nucleocapsids and for the cryo-EM images of VSV bullets, but not necessarily in an equal manner. One of the reasons for this differential presentation lies in the genuine differences between the preparations of the two types of nucleocapsid in terms of structure or heterogeneity for example. In addition, during the course of this thesis we gained more and more insights into the interpretation of different steps of classification. In these lines, the interpretation and the use of the eigenvectors will be presented in more detailed for VSV and the information extracted from the class-averages will be deeper explored for Measles.

Classification of verticalized helical segments

For the helical filament, we used the classification procedure described on **figure 2.3** using the IMAGIC software, with some particular additional steps adapted to our objects. As the helical segments were pre-verticalized using the coordinates of the extremities of the respective long filament, we could restrain the rotational alignment search to only roughly -10/+10 and 170/190 degrees. Also, for the selection of class-averages to use as alignment references, we could use the particular geometry of the helices. For example, 1D projection of the 2D classes along the helical axis can be used to judge about the correct centering of the classes. Other particularities of the classification procedure that we can use will be more detailed later.

We saw in the previous part how classification can be used to sort heterogeneous data set on the 2D level. For the sake of the 3D reconstruction process however, we do not want to necessary reconstruct each different "state" of the particles. First because we might be interested in only reconstructing the major population to have a chance of actually getting an acceptable result, because some states may be too much underrepresented, or because some heterogeneities just correspond to damaged particles or other "bad" particles like the ones suffering from staining artifacts, flattening, or (in our case of helical objects) any deformation that would break the helical symmetry. Thus, the first step of classification is often used to discard images.

As a way to get rid of "bad" images

The **figure 2.6** shows examples of "good" (**figure 2.6A**) versus "bad" (**figure 2.6B,C,D**) class-averages for the three samples (VSV N-RNA bullets, digested and native MEVNC) we are interested in. Several criteria that are easily identifiable were retained for assigning a class to the "bad particles". In the case of Measles non-digested nucleocapsids, as could be already be judged from the raw micrographs, many filaments show a long range bending of the helical axis in the plane of the carbon filament. When the bending degree is too high, it can even be detected on the smaller windowed segments, and several class-averages representing such segments (**figure 2.6B**) can thus be used to discard corresponding images. Sometimes also a discontinuity in the projection pattern can be a hint of a helical symmetry break or can result from an accidental boxing of junction between ends of two different filaments. As we already mentioned, the HAC can produce classes containing various numbers of members. In the case of a truly single particle project, due to possible non uniform distribution of views, classes with very few members may represent underrepresented orientation on the grid. In our case however, due to the helical symmetry and our overlapping segments boxing scheme, we cannot have underrepresented views. Thus, classes containing only very few members (**figure 2.6C**) can indicate that they represent rare features of the segments like unusual symmetry or particle distortions and thus corresponding images can be reasonably discarded for the next steps of the processing. Due to the averaging of fewer particles, those classes often show fewer details. Other types of classes sometimes also show fewer details, with a blurred aspect, even if they sometimes contain more images (**figure**

2.6D). This indicates a failure of the classification algorithm to regroup images with truly similar features. The reason of the failure is usually not known: it could lie in a low SNR, in images showing unique or rare features, in imaging artifacts, etc… Thus getting rid of corresponding images at this step can at least not be a bad choice.

After discarding images, a classical aim of the classification is to identify heterogeneities, and naturally distinguish heterogeneities due to view angle from the ones due to structural differences, to separate images into more structurally homogeneous classes.

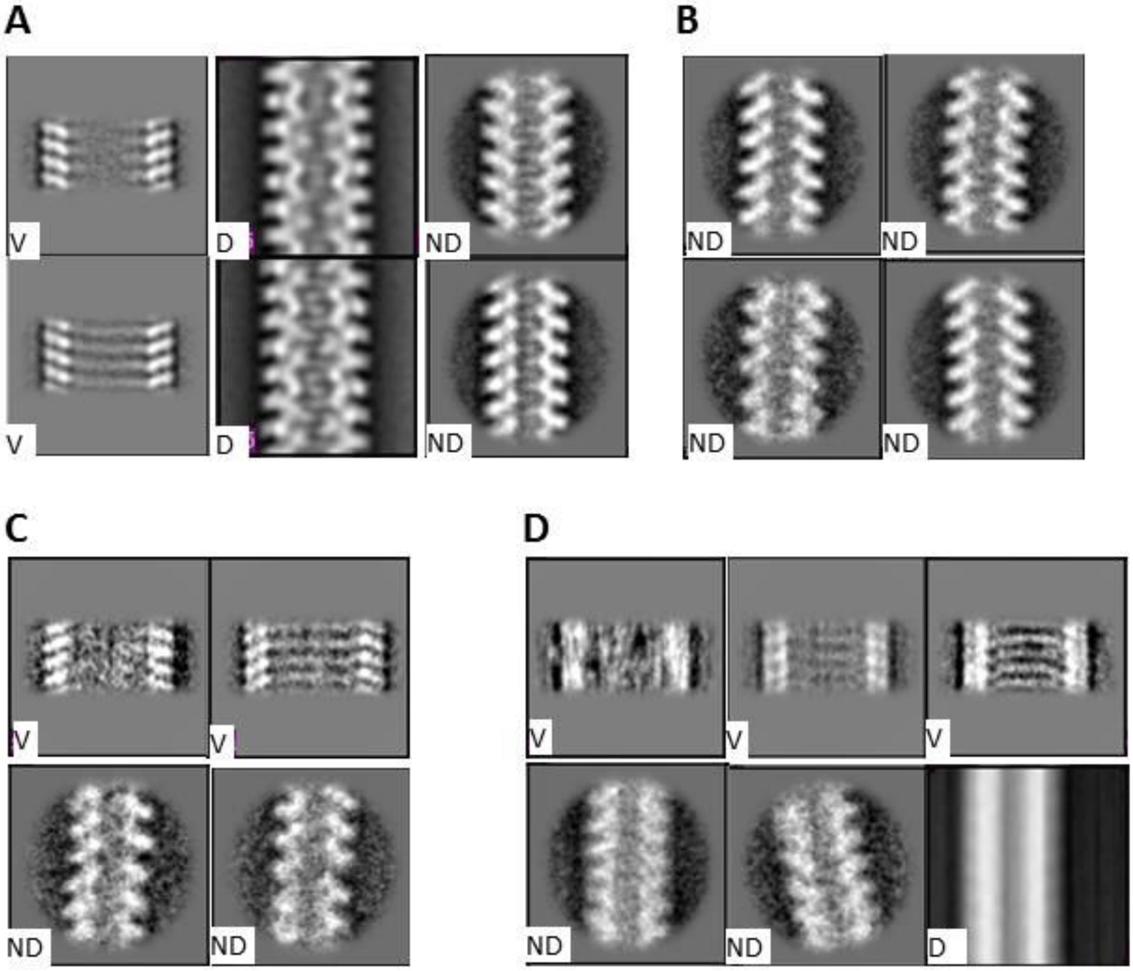


Figure 2.6 : Classification as a way to eliminate bad segments
The sample is indicated on the class-average as VSV (for VSV N-RNA bullets), ND (Measles non digested nucleocapsid) and D (Measles digested nucleocapsid). For those qualitative comparisons within same sample, the relative scale between different samples is not respected. The size of the box for VSV is 510 Å , for MeaslesND 448 Å and for MeaslesD 350 Å. The panel A represent class-averages judged as "good", the panel B shows bending, the panel C underrepresented class-averages and D classes with few details, or blurred aspect.

As a way to identify heterogeneities

For the non-digested Measles nucleocapsid, a clear source of heterogeneity appeared to be the distance between the densities pointing outward the helix (**figure 2.7A**). This variability could arise either from differences in out-of-plane tilt, as an out-of-plane tilted helix would have a 2D projection were those densities would appear closer to each other, or due to variability in the helical pitch. We cannot completely discard the first possibility, but we judge unlikely that it is the main explanation for several reasons. First we are dealing with a negative stain data set, which means that the particles are absorbed to a carbon film and should not present a very high out-of-plane tilt, whereas the differences in measurable inter-turn distances would require a very high out-of-plane tilt to be explained. Secondly, images belonging to the same filament were found to sometimes belong to various inter-turn distances classes (results not shown ; it can be noted that this was also observed by (Bhella, Ralph, and Yeo 2004), Figure 5). Thus we have to consider that those helices can present a certain degree of variability in their pitch and take this into account for the 3D processing. However at this point, the data was not separated into pitch classes as we had at that time no easy and automatic way of dealing with pitch variability (like a simple automatic pitch measure) but was instead done on the power spectra level as will be shown later (reference to the part with PS classification).

For Measles digested nucleocapsids that show much more rigid and straight helices we might expect less heterogeneity, and indeed, contrary to the non-digested nucleocapsids, no strong pitch variability was observed. However, still some striking differences could be observed between certain classes. In particular, whereas ~95 % of the images belong to classes which show a very characteristic projection pattern, around 5 % are grouped in classes showing a very different pattern (**figure 2.7B**). It can be noted that the characteristic pattern of the majority of the segments is amazingly similar to what is shown in (Schoehn et al. 2004), Figure 3, from cryo-images of a similar specimen, suggesting a relatively good preservation of the specimen in our negative-stain preparation. As mentioned, we can rule out the possibility that the low-represented classes represent particular view angles because we must have an even distribution of the views. Thus, they could represent a rare symmetry of the digested nucleocapsids or an artifact like flattening. Another hypothesis, tempting to formulate because of the sample preparation needed, would be that those classes represent partially digested nucleocapsids or alternatively more digested nucleocapsids (the SDS gels often showed subtle bands at slightly different sizes that the expected one after digestion, results not shown). If one compares the projection pattern of this population with the one of the non-digested nucleocapsids (for example the lowest pitch on **figure 2.7A**), we can indeed note similarities which would tend to confirm this hypothesis and suggest an incomplete digestion. Whatever the explanation, this is a good example of how a heterogeneity problem, which was not necessarily expected based on the raw micrographs, can be detected by classification and taken in account to clean our data set from such too different images. However, the question: "Do all the other classes really represent projections of identical objects, for example in term of symmetry?" does not have any clear answer. If detecting one source of heterogeneity actually shows that it exists, not detecting heterogeneity doesn't mean that it doesn't exist. This shows a limitation of this human inspection-based method, and highlights the need of associating automatic methods to the classification step to for example identify symmetry corresponding to the classes, as will be proposed later in this manuscript (part on helical symmetry determination on 2D projection).

For the VSV bullets, we can also appreciate even larger variability of the aspect of the class-averages (**figure 2.7C**). Not only the pattern of the inner part of the projection seems to be very variable (**figure 2.7C**, left), but the first inspection of the classes also suggest diameter variability (**figure 2.7C**, right). The understanding and the ways to deal with those variability has been more deeply explored while trying to make use of the eigenimages, and thus we will look at it in more details in the next part ("going further…").

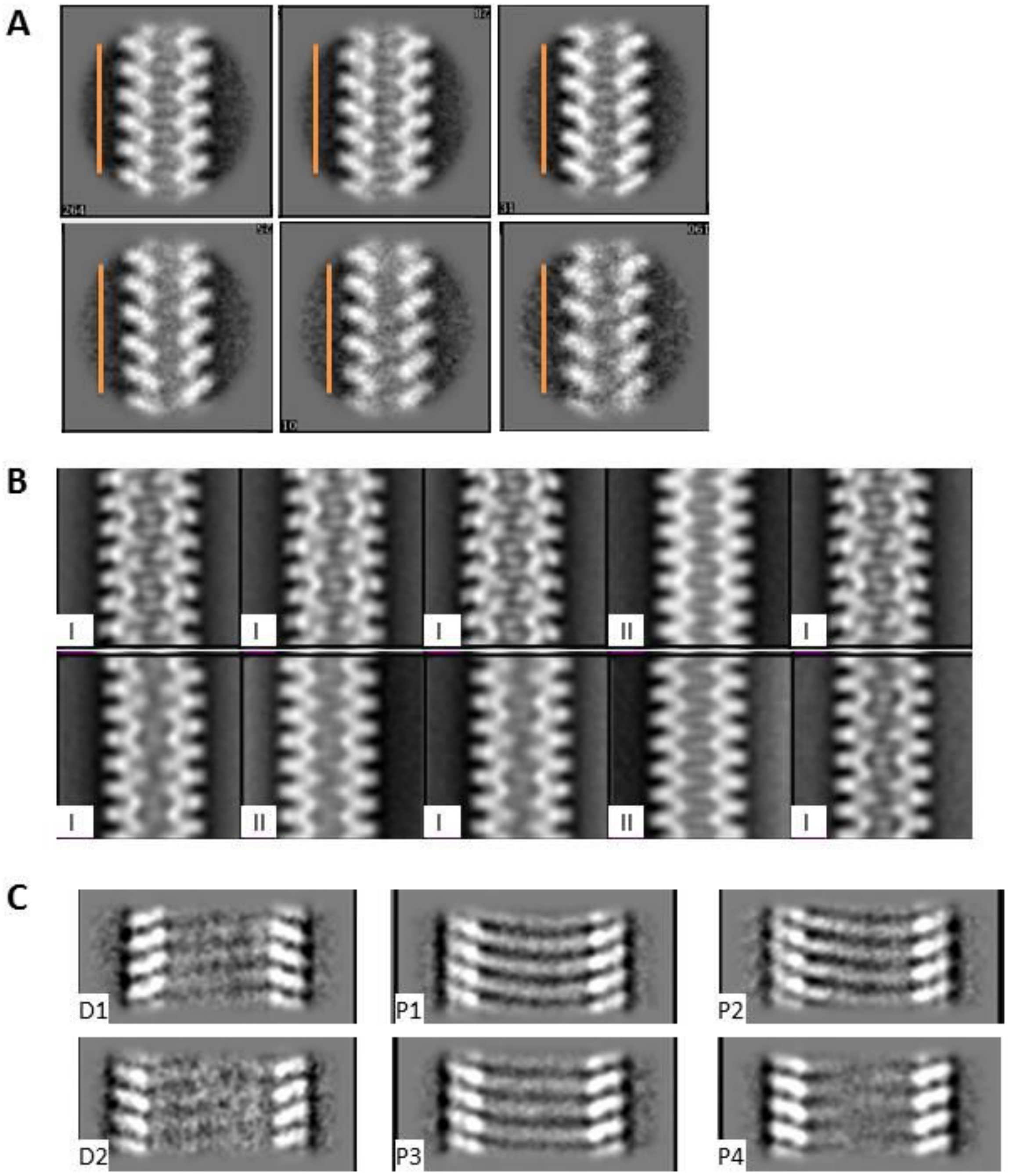


Figure 2.7 : Classification as a way to detect heterogeneity
The classification of Measles non-digested (A) and digested (B) recombinant nucleocapsids, as well as of VSV N-RNA bullets (C) revealed heterogeneous appearance of the class-averages. For MeVND, the distance between turn appears variable (the orange bar in each class in A is of constant length ; classes are manually sorted from left to right and top to bottom by increasing pitch). MeVD sample shows a minority of classes with a different aspect (noted II in B versus the majority noted I). VSV sample shows important diameter variability (example noted D1 and D2 in C), as well as various projection patterns, especially considering the center of the structure (noted P1 to P4 in C). The width of the boxes are the same than indicated on Figure 6.

After having discarded “bad” images, identified sources of heterogeneities and refined the classification, we are able to get “good” class-averages with a high SNR, from which we would like to extract information at 2D level. The **figure 2.8** shows such good class-averages, using the goodness criteria of Imagic, called “Overall Quality” which, although again obscure, takes for sure at least into account the mean variance among images contained in a class, thus reflecting the homogeneity of the members of this class. For VSV, we also show two classes with not such a good quality score (**figure 2.8D**) because these still provide valuable information. We can mention that for helical objects, looking at the power spectrum of each class will also give important indications, in particular regarding the regularity, i.e the straightness / the symmetry preservation, of the classes. As a chapter will be dedicated to the PS analysis, this issue will not be discussed here.

When looking at the projection pattern of the digested nucleocapsid (**figure 2.8A**), we remark that, within a single class-average, and whatever the class chosen (**see figure 2.7B and 2.6A**), there is a repetition of motifs along the helical axis every three helix turns (highlighted by circles on the bottom part of the **figure 2.8A**). A true repetition of a motif in the projection is in theory only attained after a translation of the repeat distance c along the helical axis, after u turns. Here, the visual assessment, although very convincing, is not quantitative and thus we cannot conclude that the exact repeat is attained, but we can say that the helix contains in three turns a number of subunits at least very close to an integer. Thus, we expect the number of subunits per turn to be close to X.33 subunits per turn if the number of units in the repeat u is odd and X.67 if u is even. The work of (Schoehn et al. 2004) on cryo-EM images of Measles digested nucleocapsid resulted in two reconstructions obtained by a single particle-based approach. The reconstruction with the higher resolution shows 12.35 subunits per turn, the one with the lower resolution has 11.64 subunits per turn, whereas a combined Fourier-Bessel/single particle approach gave a symmetry of 12.33 subunits per turn. Remarkably, these values are compatible with our observations.

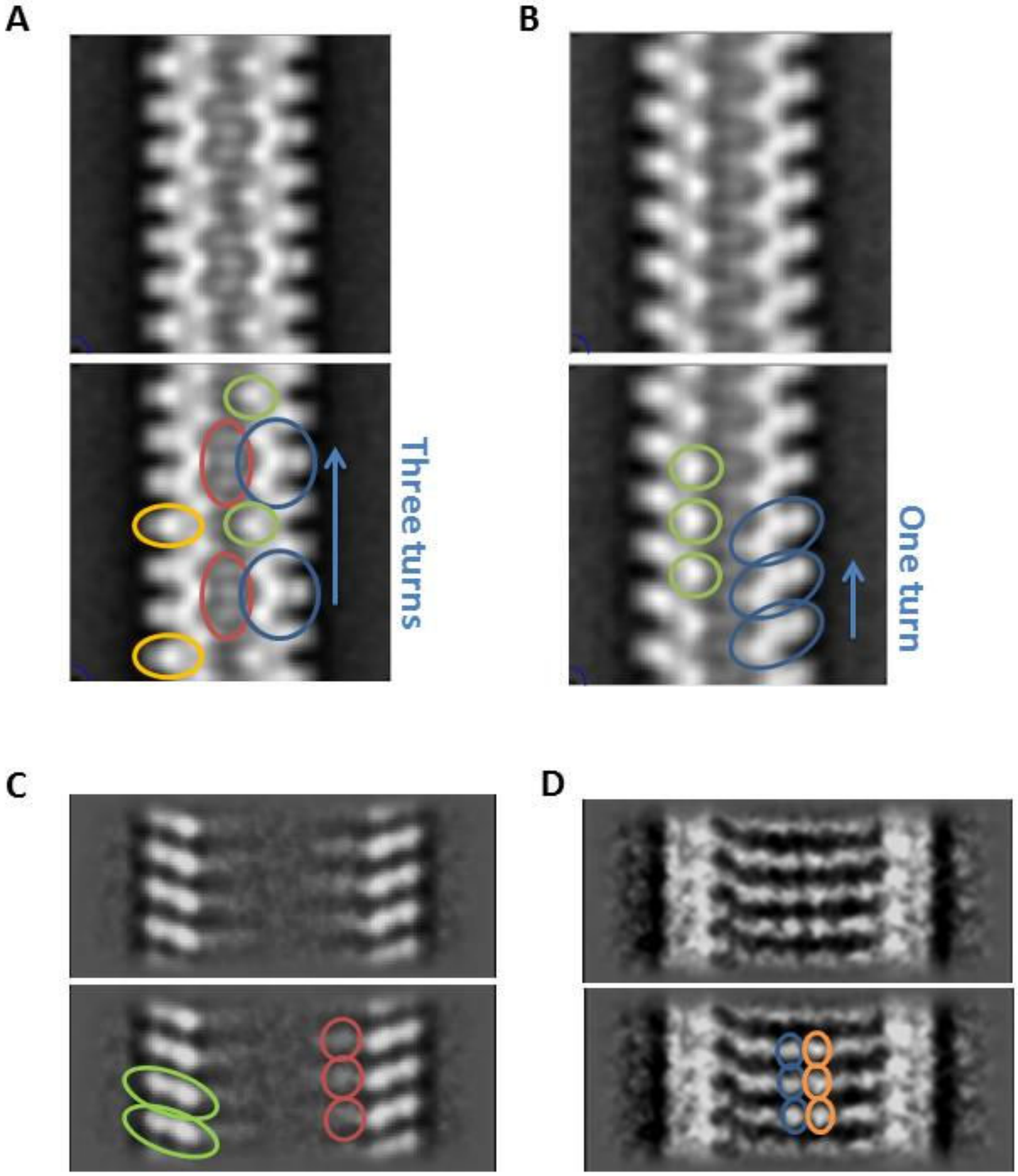


Figure 2.8 : Observation of class-averages : detection of motifs in the projection pattern
One representative majority-type class-averages of MeVD (A), MeVND (B) and VSV N-RNA reconstituted bullets (C) are shown. For VSV, another less representative type of class-average is shown (D). On each panel, the class-averages is shown twice, with a superimposition of colored ellipsoids which highlight patterns in the projection (motifs) that seems to repeat along the filament, as visually determined.

For the non-digested nucleocapsids, the situation is very different (**figure 2.8B**). First the classes usually show fewer details, in particular for the inner part of the helix projection, making the analysis more difficult/ less precise. This could be an effect of the symmetry, but most likely it mainly comes from the fact that, as already shown above, these helices are less ordered. Contrary to the digested nucleocapsids, for the native ones the “three-turn” repeat is not observed, and instead the classes show a density pattern that seems to repeat each turn (**figure 2.8B**, bottom), which would suggest that the number of subunits per turn might be close to, or exactly, an integer.

For VSV, the higher diversity of projections type makes the interpretation more uncertain. A good class as defined by IMAGIC shows almost no inner pattern (**figure 2.8C**). Given the high number of subunits per turn expected for these objects, the inner part of the projection is a complex superposition of many subunits from the near side and the far side of the helix, potentially explaining this observation. On the exterior part of the projection, we can recognize the global bi-lobed nucleoprotein shape, which appearance in subsequent turns seems almost constant (**figure 2.8C**, bottom, green marks). Thus, as for Measles non digested nucleocapsid, we may expect a close to, or exactly, integer number of subunits per turn. As shown on **figure 2.8C** (bottom, red circles) the slight inner densities seen close to the edge of the projection appeared quasi stacked in the direction of the helix axis, but not exactly. Interestingly, some less represented classes (and with a worst IMAGIC goodness score), but showing a different and “discrete” inner pattern, also show this kind of quasi stacking of densities (**figure 2.8D**). The structure of the full VSV virion exhibits exactly a half number of subunits per turn (37.5), which gave rise to a particular pattern in the 2D projection where motifs were found to be identical after a translation along the helical axis corresponding to two helix turns (Ge et al. 2010). Although we cannot be very precise from the rough observation of the classes, we can at least exclude such a half integer value of symmetry, and rather suppose based on the several observations that we have, that the symmetry is almost an integer number of subunits per turn.

Before closing this part on the conclusions drawn from the class-averages observation for our three projects, an important point has to be made. This part of the manuscript was written after the 3D reconstructions were calculated and symmetries determined: thus our look on these data is now strongly biased by our a posteriori knowledge, and it is much easier to do those interpretations, although they all seem reasonable and justified. A part of the classification work (on Measles) was done at the very beginning of the thesis, and at that time our way of analyzing the results was yet less advanced. However, it also shows that using a posteriori information to re-interpret older data makes possible to gain knowledge, in our case in the potential classification outputs, in order to have more tools and insights for a new project.

In the **figure 2.8**, we have seen “good” class-averages and the information we can draw from them. These results are however not quite easy to get; indeed one needs to do careful multi-reference alignment by choosing “representative and good” classes, understand the kind of variability represented from the classes, separate the data set into more homogeneous subsets to improve the classes, etc. It can be sometimes difficult to sort out these issues based on the classes themselves, and thus a better understanding of the eigenimages accompanying the classification and their possible use is important.

Going further: trying to investigate the meaning of eigenimages

When looking at the eigenimages, one should keep in mind that their order is crucial: the first eigenimages represent the most important contributions in our images. In a relatively similar manner, one can say that the more a dataset is homogeneous (less variations), the less eigenimages are required for its description. As an extreme example, Marin van Heel's artificial data set necessitated only three eigenimages to be described (**figure 2.1F**). One should also remember that the eigenimages are associated with a weight, specific to each image of a data set, that represents their contribution to the formation of this particular image.

Meaning of eigenimages for Measles projects

The first 16 eigenimages of an aligned data set for each Measles project are shown on **figure 2.9 (A** : digested, **D** : non-digested). After the total sum of images (eigenimage 1), both samples exhibit contributions (**figure 2.9A,D** eigenimages 2 and 3), that has a constant pattern for each turn, that visibly reflects the global aspect of the images. This means density blobs pointing outward the helix, each spaced along the helical axis by a distance corresponding to the pitch, and which radial position reflects the helix diameter. These two eigenimages are already sufficient to notice significant differences between the native and the digested nucleocapsids. The next contribution (**figure 2.9A,D** eigenimage 4) can be attributed to a remaining non perfect centering of the segments, as will be explained on the VSV example. For the digested sample (**figure 2.9A**), after the unclear eigenimage 5, we see contributions (eigenimages 6 and 7) that also show a pattern that is identical in each turn but with a different aspect. Interestingly, the same kind of pattern is also observed for the non-digested sample, but slightly further away in term of contribution importance (**figure 2.9D** eigenimages 8 and 9). We can only hypothesize here that these contributions could represent out-of-plane tilt of the filaments, which in the case of Measles makes appearing on the projection the bi-lobed shape of the nucleoprotein (as simulated from low-resolution filtered version of reconstruction of (Schoehn et al. 2004) ; results not shown). This could be verified a posteriori, for example by plotting the contribution of those eigenimages to individual raw images as a function of the out-of-plane tilt as found by projection matching during 3D refinement. Interestingly, before this putative out-of-plane contribution, we can observe for Measles non-digested nucleocapsids three eigenimages (**figure 2.9D** number 5,6,7) that are not present for the digested sample, presenting a “discontinuous aspect”. We can exclude that they are used to represent broken filaments, given the attention with which the micrographs were boxed, and strengthened by the fact that even if some broken segments were still included in the data set, they would represent a minority of the images and thus would not require 3 eigenimages, with such a high degree of importance as inferred from their position. We saw from the analysis of real-space classes that one main source of variability observed for this sample was the helical pitch. These eigenimages are actually compatible with this, and we see two reasons for this. First, they show density contributions that are differently spaced (the orange bars of same size drawn on **figure 2.9D,** eigenimages 2 and 5,6,7 illustrate this “stretching” or “compression” of densities). Second, when one computationally superposes two eigenimages of the type describing the global major pitch of helices of two different pitches (for example the eigenimage number 2 from the digested and the non-digested data set), the appearing pattern (**figure 2.9E**) is very similar to the one in the eigenimages 5,6 and 7 (**figure 2.9D**). However, the reason why three eigenimages are needed to explain this variability (and not two as usually required to reflect the variability in the translation along the y axis) is not clear. One can envision a potential role of the bending of helices that would also result in a stretching/compressing of density in the helix direction. Naturally, this could be verified by several ways. One possibility would be to classify the images only according these eigenimages (eventually in addition to eigenimages number 2 and 3), create a few classes and see if the difference appearing is bending or pitch. Another way would be to use an independent method to separate images according to pitch (for example reference-based, as in (Bhella, Ralph, and Yeo 2004), or based on power spectrum), perform the classification on these subsets and see if this type of eigenimages disappear.

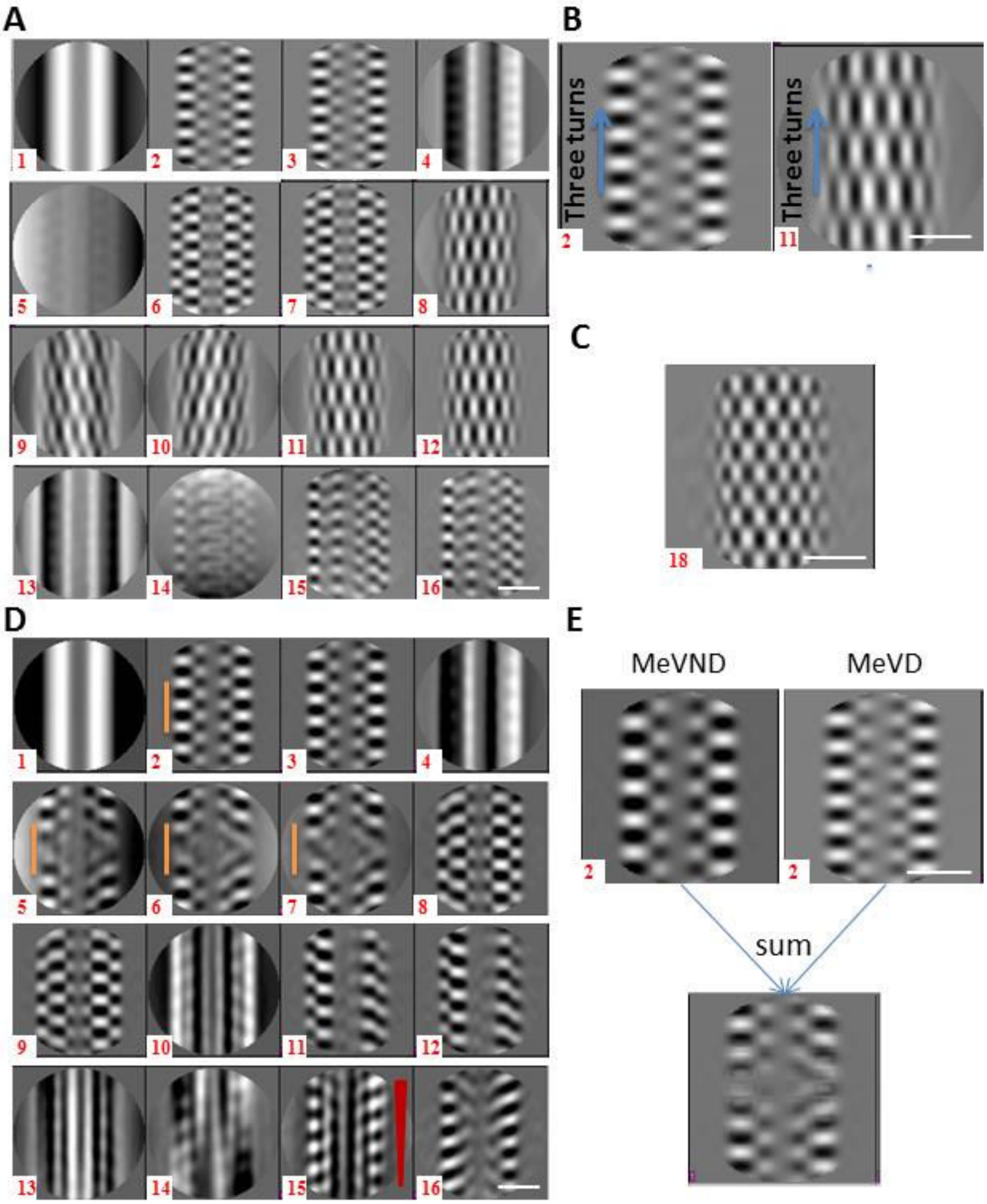


Figure 2.9 : Eigenimages of MeVD and MeVND and their interpretation

The 16 first eigenimages of MeVD (**A**) and MeVND (**D**) are shown, as well as eigenimage number 18 for MeVD (**C**). The panel **B** shows for MeVD the periodicity of one pitch (left) that can be seen on a number of eigenimages, versus the periodicity of 3 turns seen in others. The panel **E** shows for MeVND that one can obtain an eigenimage pattern that is relatively similar to the one observed on eigenimages 5,6 and 7 (**D**), by summing up two eigenimages reflecting different pitch (eigenimages number 2 of MeVD and MeVND). Drawings on the eigenimages are described in the text. The ranking of each eigenimage is indicted by the number in red. The scale bar drawn in white in each panel indicates 100 Å.

For the digested sample, although the class-averages seemed to show a repeat of the motifs every three pitches, the first 7 eigenimages we have described so far have a periodic pattern repeated every pitch along the helical axis (**figure 2.9A**). Indeed, the three turn-repeat is represented by the next 5 eigenimages (**figure 2.9A** ; number 8 to 13). A closer view on eigenimages 2 and 8 illustrating this fact is depicted on **figure 2.9B**. Thus the combination of these two eigenimages for instance allows creating a projection pattern that reflects both the pitch and the repeat. For the non-digested sample, we do not see such a pattern. A departure from an exact motif repetition every turn is depicted only later in the eigenimage order (e.g. **figure 2.9D** eigenimage 16), and does not show any short distance repeat, but a progressive variation of the pattern in each turn. This fact supports the previously ventured that native nucleocapsids may have a close to integer number of subunits per turn. This is supported by eigenimages 13 and 15 (**figure 2.9D**), which show quasi vertical arrangement of densities (eigenimage 10 most probably corresponds to diameter variability as will be shown for the VSV data set). The particularly well defined eigenimage 15 (**figure 2.9D**) even shows how the quasi vertical arrangement of density motifs slightly varies from turn to turn (the red symbol depicts the gradual increase of the density at the right side of the eigenimage, that is correlated with a decrease at the other side). This eigenimage shows even more interesting details, for example it makes apparent the lateral and vertical arrangement of smaller density motifs. For the digested sample, others eigenimages also contain such kind of finer information, as for example the eigenimage shown on **figure 2.9C**. One is tempted to interpret these patterns in terms of subunit arrangement along the helical path and among successive turns, but we found it too hazardous to further interpret this because of the fact that the near and the far side of the helix are merged in the projection. Interestingly, the number and arrangement of density motifs seen perpendicularly to the helix axis in the eigenimages was found to agree with the final number of subunits per turn in the reconstruction of the non-digested MeVNC – indeed, there are slightly more than 5 density stripes in the eigenimage (**figure 2.9C**), i.e. just about a half of the number of subunits per turn, so that these motifs could reflect the front or the back side of the helical lattice. This should be further examined and proven, for example by creating artificial data set with pseudo-helices that would actually be made of two halves with two different helical lattices, projecting them so that the two lattices are superposing, and performing a classification of such simulated 2D projections. The ability of the eigenimages to discriminate the two different lattices, or not, should then appear clearly.

Meaning of eigenimages for the VSV project

For the VSV bullets (**figure 2.10**), the eigenimage number 2 of a non-aligned vertical data set (**figure 2.10A**) shows vertical stripes that are not symmetric in respect to the helix axis (contrary to the eigenimage 9), but that are relatively constant along the tube. We can note the similar eigenimages number 4 observed for both Measles project (**figure 2.9A,D**). As each eigenimage has an implicit inverse version, it shows that the data set has as a strongest variance source towards "the right" or towards "the left" of the helical axis. This suggests that this eigenimage could represent the centering variability among the segments. However for a non-trained eye, as it was our case, it can be not obvious. A way to understand to what this eigenimage corresponds is to classify the images only as a function of the variation it describes. In other words, we can give the HAC algorithm only the relative weight with which this particular eigenimage contributes to the formation of each image as a basis for partitioning. Doing this and generating only a few classes (10 of which are shown on **figure 2.11A**), one can actually see that the difference among classes is the centering. From those class-averages, one can now choose the best centered ones by looking at their 1D projection along the helical axis, and use them to do a multi-reference alignment of the badly centered data set allowing only for translation search perpendicularly to the helix axis. By classifying the obtained aligned data set, we can observe that the eigenimage thought to correspond to centering discrepancies disappears, confirming the original hypothesis. This is an important point: to confirm that one has correctly identified a variability source among the images (orientation/centering/heterogeneities), one should be able to classify according to this variability source and observe the disappearance of the corresponding eigenimage upon either alignment of the images or separation of the data set into subsets. In our case (and more generally for helical samples), one can thus use the presence, or the relative numbering of such an eigenimage to judge about the correct centering of the images. This may even be done during the reconstruction process on the images aligned to the projection matching references to assess the outcome of the alignment.

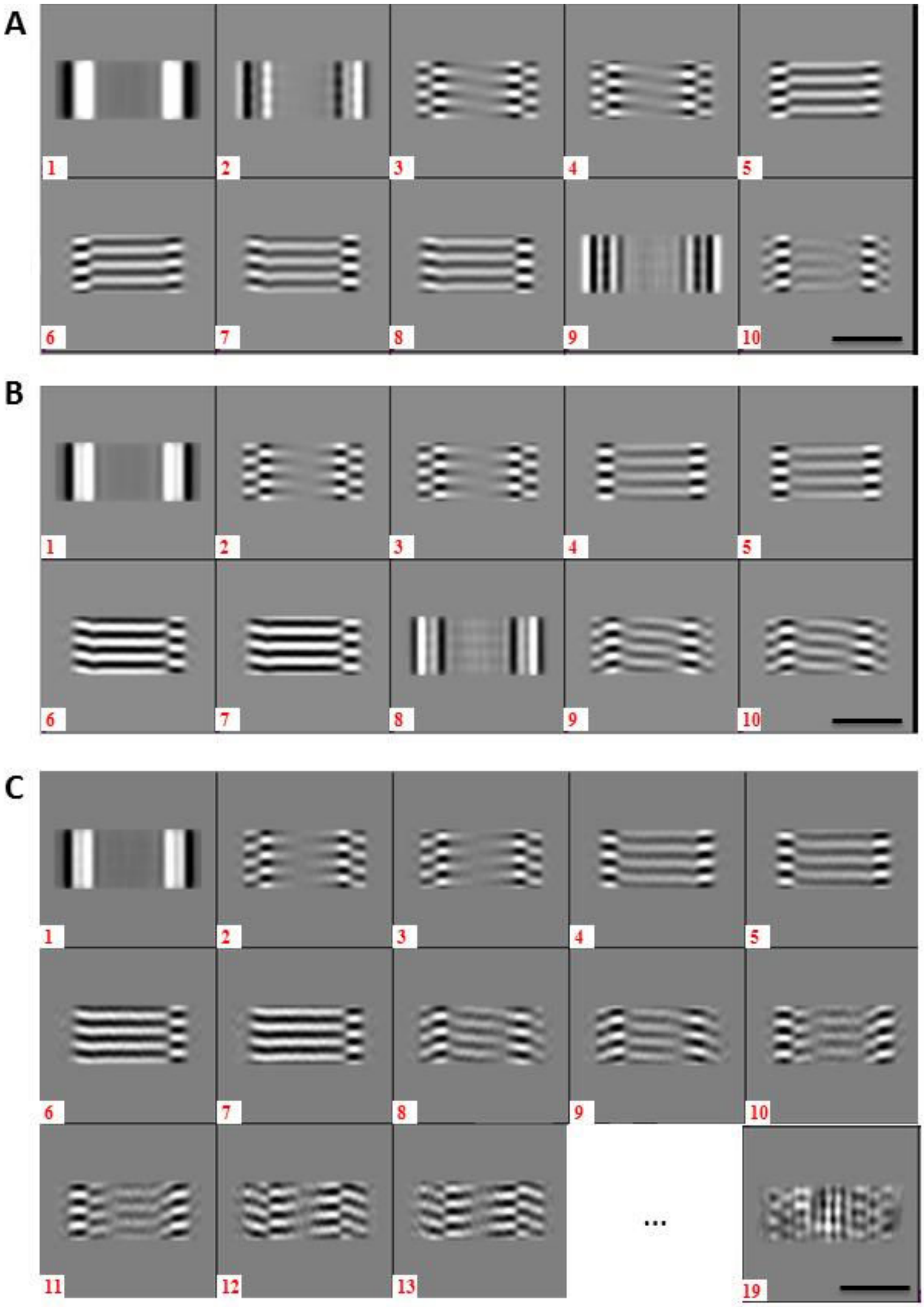


Figure 2.10 : Eigenimages of VSV N-RNA reconstituted bullets

The first eigenimages of VSV N-RNA reconstituted bullets are shown for the unaligned images (**A**), after centering (**B**) and after classification according to diameter (**C**, for the diameter class of ~390 Å), both based on eigenimage interpretation. Black scale bar indicates 200 Å.

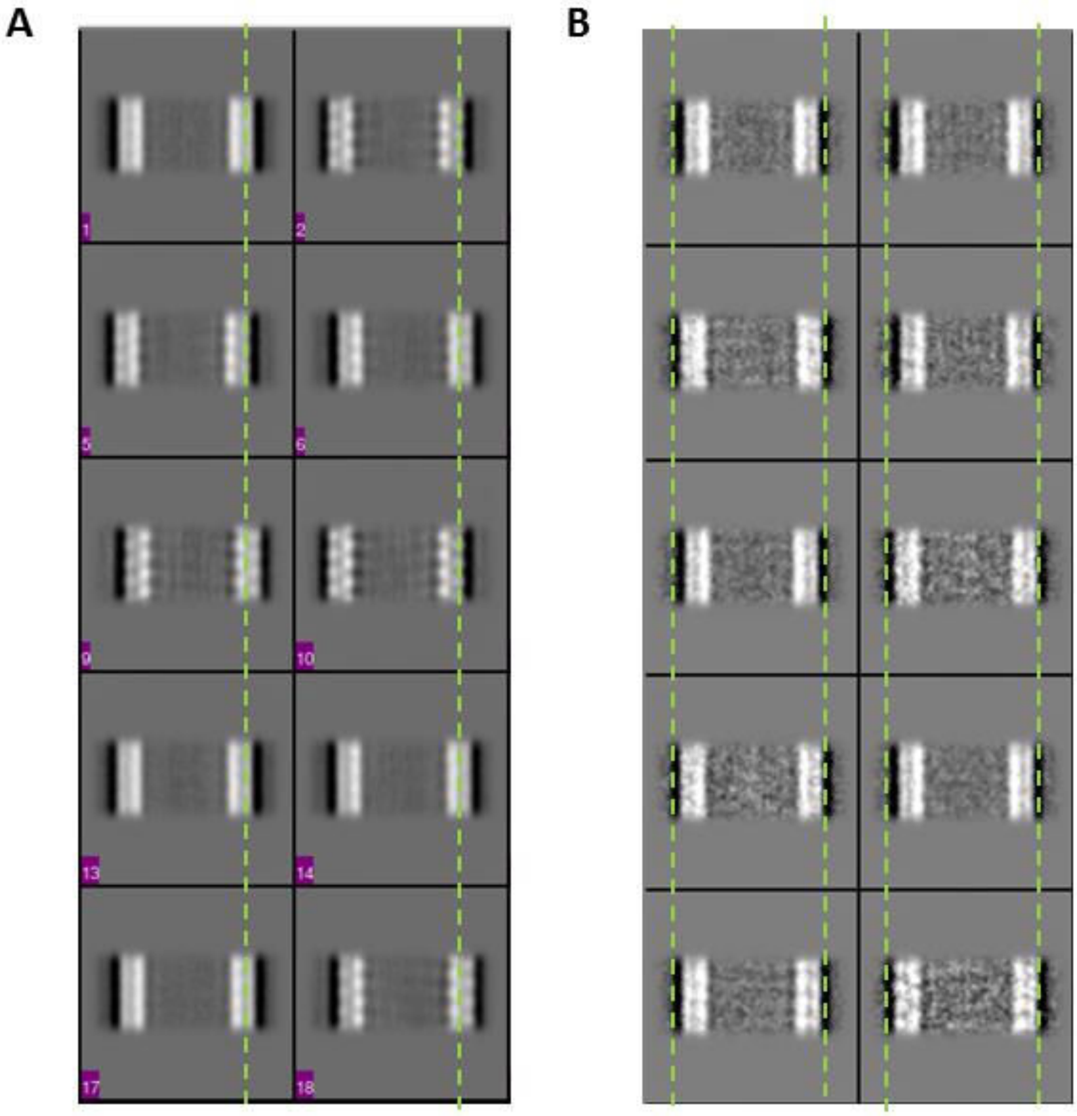


Figure 2.11 : Class-averages obtained by classification according to eigenimage attributed to centering variability and to diameter variability
A subset of class-averages obtained by classification according to centering (A) or diameter (B) variability are shown. The vertical dashed lines help to highlight both types of variability.

Another eigenimage that differs from the majority of the others that describe the global structure of the filament is the eigenimage number 9 at the **figure 2.10A**, that became number 8 on **figure 2.10B** in its inverse version. We see again stripes parallel to the helix axis at the edge of the structure, but this time the pattern is mirror-symmetric around the helical axis. Thus we have a source of variability that is independent of the position along the helix axis and that points symmetrically towards the exterior of the structure: this brought us to think about diameter variations, that could be in a certain extent already seen from the micrographs, the top views (**figure 2.5B**, eigenimage number 2) and the class-averages of the side-views (**figure 2.7C,** left). Again, we verified this by classifying the images only using this eigenvector, this time not in order to obtain references for multi-reference alignment as we did for the centering, but in order to split the data set. A small number of classes (100) was generated to clearly see the differences (**figure 2.11B** shows some of those) and to split the data set, but a higher number (400) was used to obtain a finer plot of diameters as shown in the VSV manuscript (Desfosses, Ribeiro, Schoehn, Blondel, Guilligay, Jamin, Ruigrok and Gutsche, in preparation). As expected, the main difference between the generated class-averages is their apparent diameter. The **figure 2.12A** shows class averages of two extreme diameters and their corresponding 1D density profiles of the projection along the helix axis. Based on this, we can see that we can use this classification step for an automatic separation of the images. From the measurable diameter of their corresponding class average (which can be for example assessed as the distance between corresponding zero-crossings in the 1D density profile), the images can be separated according to their diameter.

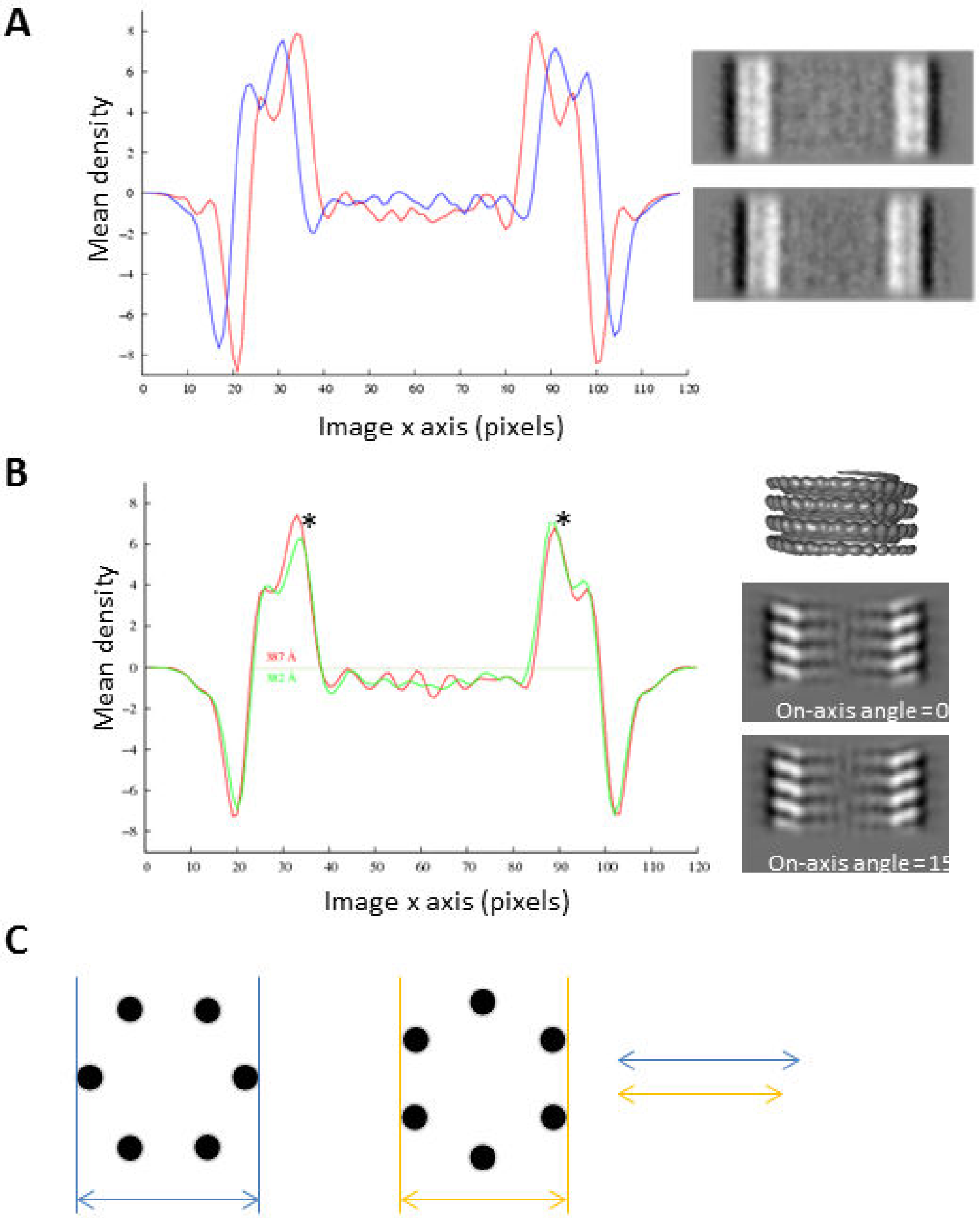


Figure 2.12 : Sorting a data set according to diameter using 1D density profiles and effects of the on-axis view angle on measurable diameter

The panel **A** shows the 1D density profiles (left) of two extreme diameter class-averages (right). The blue profile corresponds to the class shown on the top and the red one to the class on the bottom. To simulate the effect of on-axis view angle on the measurable diameter on the 1D density profile, an artificial model similar in size to VSV reconstituted bullets was generated (**B**, top right) projected at all on-axis views every 1°. The density profiles of the views showing the most difference in diameter (5 Å) as judged from the crossing of the density profile with 0 are shown (left, red and green profiles) as well as corresponding projection (right middle and bottom). The asterisks highlight the density profile maxima, were the difference is easier to visualize. The panel C schematically illustrates the fact that the apparent diameter can be more variable in some cases. The blue and yellow double arrows have same size, respectively.

An automatic splitting of the data was proposed to be done directly on the raw images based on their 1D projection for example in the IHRSR++ modified version of original Egelman's scripts (e.g. in (Parent et al. 2010)). One should however be aware of two issues. If raw images are used directly, the SNR is very low (especially for cryo) and thus the measure of diameter can become quite imprecise. However this is not the case when measuring on a class-average, which is one advantage of our procedure. A second problem, and this is true when working on raw images as well than on class-averages, is that one has to take in account the fact that the angle of view of the particle may influence the 1D profile in such a way that it could lead to a different measure of the diameter. Of course this influence will also depend on the symmetry of the helix and the shape of the subunit. In order to get an idea of the extent to which the on-axis angle view can influence the measured diameter in our case, we simulated a 3D model with dimensions and number of subunits per turn in the same range of those expected for VSV. This model was projected at each direction around the helical axis every 2 degrees and an automatic measure of diameter using the proposed procedure was done. Here the difference between extreme values was very small, around 5 Angstrom (**figure 2.12B**), and thus images were split only if they corresponded to classes with measurable diameter difference superior to this value. However, as mentioned, this difference depends on the symmetry and we are here in an extreme case, because of the very high number of subunits per turn, i.e. low angle between subunits, so the profile will change only a little when rotating around axis. A very schematic drawing on **figure 2.12C** makes clearer what we could expect for an opposite case, i.e. with a relative strong influence of the view angle on the measured diameter.

Once the data was properly centered and separated in diameter classes, we can use MSA on each subset to see the remaining eigenimages, and of course further refine the classification outcomes. As expected, we can see, for one diameter class on **figure 2.10C** (and the same was also observed for the other diameter subsets), that the eigenimage identified as corresponding to diameter variability disappeared.

The remaining eigenimages (on **figure 2.10C**, from a data set which is already aligned and homogeneous in diameter, eigenimage number 2 to 13) are present as couples with a relative shift of a quarter of the pitch in the y direction, and show strong differences in the inner pattern. Although they should describe the general structure of the helices, we can ask ourselves what the differences among the pairs actually represent. To try to answer this question, one can rationally think of what kind of features in the sample, or what kind of variability can explain these eigenimages. Visibly they do not show enough details to represent the organization of the subunit nets and the possible variability in it. Notably some eigenimages further in term of importance –**figure 2.10C number 19**, are showing more details which one might indeed be tempted to interpret as subunits network, even if we didn't hypothesize more about it at that time. Could the difference between remaining eigenimages be due to variations of the pitch? If one manually measures the distance between two turns (easily discernible at the left and right edges of the images), one can eventually see some small differences between these eigenimages. But these differences are so small, that they seem insufficient to explain why the inner pattern of the images is so different. Another remaining source of variability that would also explain the small variations in measurable distance between turns at the 2D projection level would be a variation of out-of-plane angle of the particles in the ice. Again, a way to understand better the meanings of these eigenimage is to compare them to what one would obtain using a synthetic reconstruction, as it was done for variability of diameter depending of view angles. This time the reconstruction shown in **figure 2.12B** was used to create projections, with a variability of out-of-plane angle in addition to on-axis variation. Either no out-of-plane was allowed, or +/- 4 degrees (2 degree step), or +/- 12 degrees (2 degree step). The **figure 2.13** shows the eigenimages obtain from the MSA performed on each of those data sets. As expected, the more different out-of-plane versions of the images were created, the more information-containing eigenimages are required to describe the data set. We can note that even in the case with the highest extent of out-of-plane variability (**figure 2.13C**), we need less eigenimages to represent the data set than in the real case. One reason for this observation would be the discrete out-of-plane variability of the simulated data contrary to a continuous variability in the real case, the latter thus presenting more fine differences that need to be described. Interestingly, although the eigenimages for the simulated data with the highest extent of out-of-plane are not really the same than in the real data, we can make some similar observations. The first eigenimages (number 2 and 3 both for real data set on **figure 2.10C** and artificial data sets on **figure 2.13A,B,C**) show a "discontinuous" projection pattern, with the highest contribution at the edges of the structure, whereas the next ones show "lines" of density from left to right (eigenimages 4 and 5 on **figure 2.13B,C** and reminiscent eigenimages 4 to7 on **figure 2.10C**), that accounts for low out-of-plane tilt (because they are present already when only 4 degrees of tilt was simulated (**figure 2.13B**)). Next eigenimages for the simulated data set (**figure 2.13C**, number 6 to 9) corresponding to higher out-of-plane tilt show a more complex pattern comparable to the real data set (number 8 to 13 on **figure 2.10C**).

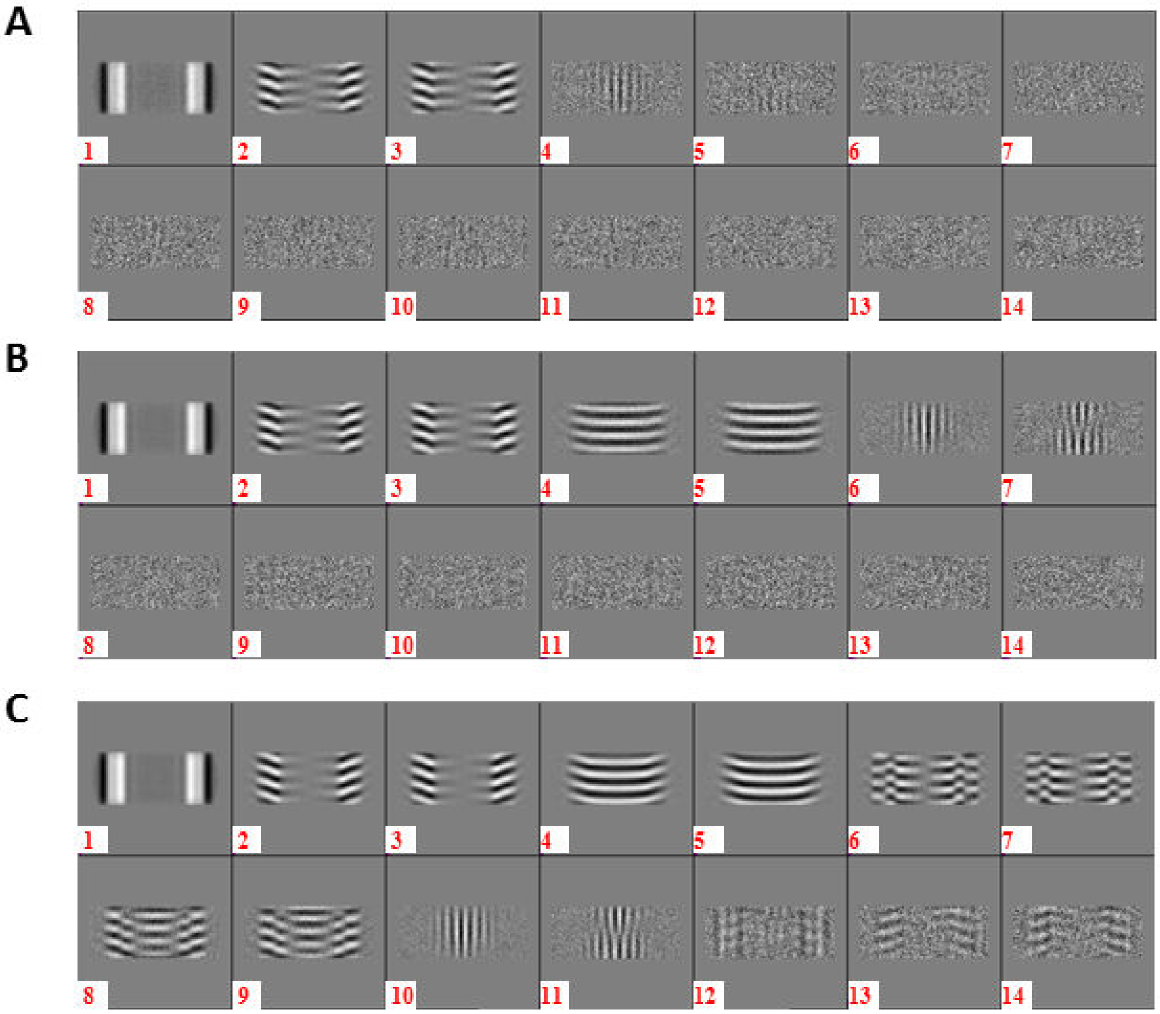


Figure 2.13 : Evolution of eigenimages as a function of out-of-plane angles imposed on a simulated data set with similar characteristics to VSV N-RNA reconstituted bullets

A noisy data set was generated from the 3D model shown on figure 2.12, by varying the on-axis angles for projection from 0 to 360 ° every 2 °, and generating either no out-of-plane views, or out-of-plane views until 4 °, or until 12 ° , with a step of 2 °. The 14 first eigenimages generated from those data set are shown respectively in **A**, **B**, and **C**.

An interesting independent validation of the fact that these kinds of eigenimages explain the out-of-plane variability of the filament was made a posteriori, once we had a reconstruction of the reconstituted VSV bullet trunks. Two sets of images were created during the classification step, one with images corresponding to classes that we would identify as "in-plane" (with the "discontinuous" pattern represented by first eigenimages) and another set of images corresponding to images that showed the "out-of-plane" pattern. As these two sets of images were compared by projection matching to the reprojection of our current reconstruction of the bullets, one could indeed see a good match between our predictions based on visual inspection and the outcome of the projection matching (**figure 2.14**).

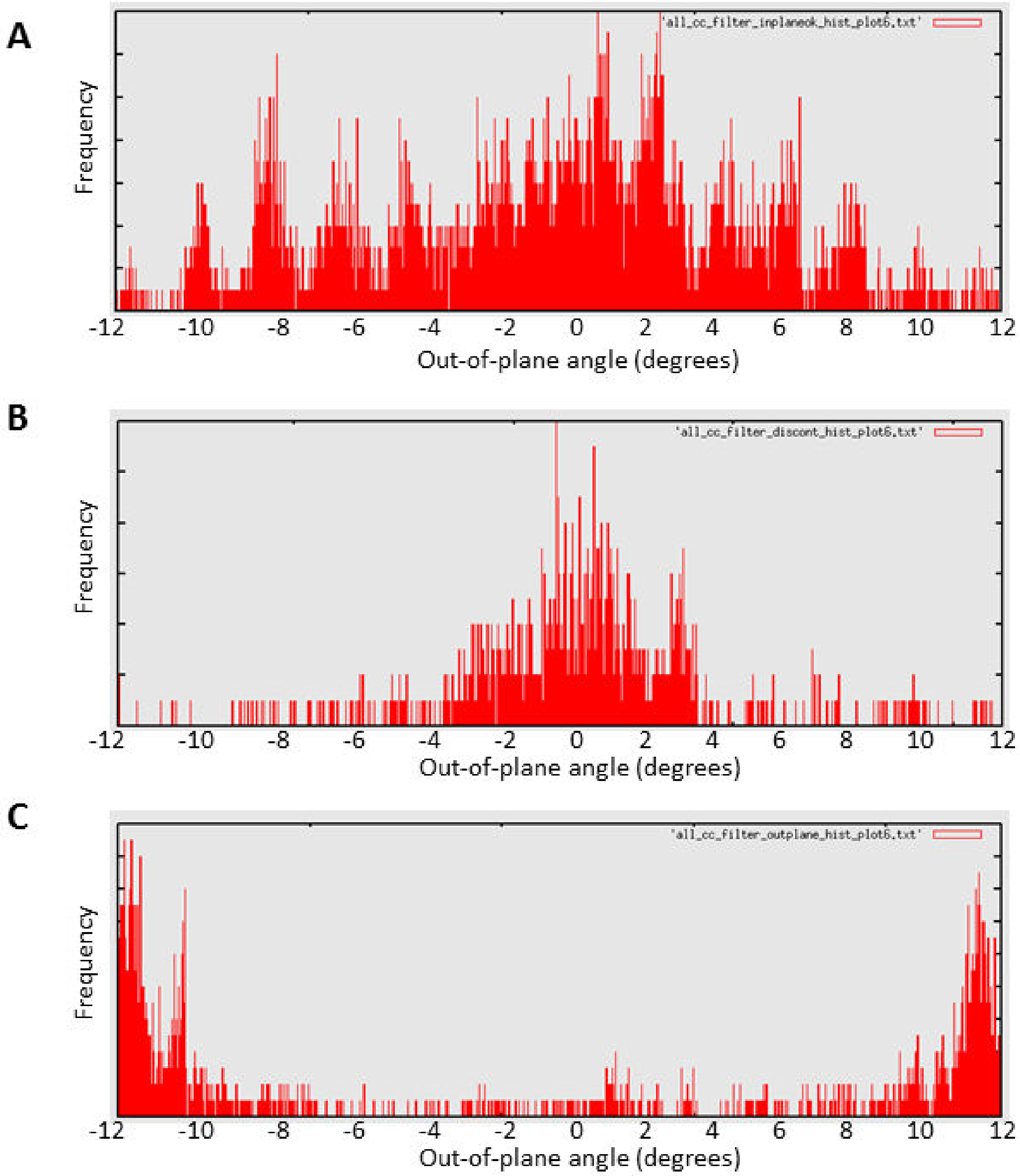


Figure 2.14 : Projection-matching based out-of-plane distribution, after sorting of images according to out-of-plane tilt using classification and interpretation of eigenimages

The distribution of images as a function of out-of-plane as determined by projection matching against a VSV trunk reconstruction is shown for :

A : a selection of images that we currently had for reconstruction (selection done by other means during the reconstruction procedure)

B : images belonging to class-averages that we attributed to views with no out-of-plane tilt, after eigenimages interpretation

C : images belonging to class-averages that we attributed to views with a high out-of-plane tilt, after eigenimages interpretation

So, we have seen how one can take useful information from the classification step and from the analysis of the eigenimages to extract what kind of variability is present among our sample, and to improve the homogeneity of the images. Some perspectives to improve this step of the processing, and how to couple it better to the other steps, will be presented now.

Criticism and perspectives for classification of real images

One important criticism of the analysis presented above is its unquantitative nature. Indeed, we show neither how and how many particular steps of the classification helped to get a better final result on the 3D level, nor exactly which measures or values extracted from the classification can be used in an automatic manner to decide which image should be included in the next processing steps. We can also say that the relative use of the real-space classification step for the Measles projects is poor in contrast to what could have been done. It was mostly used to get rid of images corresponding to clearly (as visually assessed) badly-resolved, curved, or discontinuous classes, but not in order to separate different helical states, as could for example have been done for Measles non-digested nucleocapsid that showed clear pitch variability. Instead, reference-based methods were used (Desfosses et al. 2011), which can clearly suffer from more bias than reference-free classification. For VSV, it was crucial to use the presented diameter classification in order to be able to obtain a 3D reconstruction without symmetry imposition that showed distinguishable subunits densities (it was not possible before). However fine effects on the 3D reconstruction quality of including images belonging to particular classes were not further examined.

The main reason of this lack of quantitative assessments is the fact that the classification and the reconstruction steps were insufficiently coupled. Testing different parameters of classification, creating pools of images corresponding to various combinations of classes (for each classification trial/parameters), requires a lot of manual steps and the outcomes are currently not easily associable with objective quantitative measures (e.g. reconstruction resolution, convergence ability of the IHRSR procedure, symmetry parameters after refinement). A few scripts were written during the VSV project to include the class membership as given by IMAGIC into the selection files for SPIDER, but this was not sufficient as it still required a lot of manual steps and did not give easy ways of judging the outcome.

We think that it could be useful to associate classification and reconstruction not only in the forward way, but also in the reverse direction, i.e. improve classification and the way we can use it based on the reconstruction outcomes. For example, there are many simple questions for which we have no answer, like : To which type of classes the images included in the best reconstruction belong? Are these classes associated to a particular combination of eigenimages weights? Can we associate symmetry parameters to the use of particular eigenimages in the case of heterogeneous sample? Regarding the reconstruction outcome, should we separate images when corresponding classes show diameter differences of 1, 5, 10, 15 Angstrom? What is the best balance between better resolution thanks to a finer pitch sorting and worst resolution due to the decrease of the number of included images?…

To answer those questions, one needs tools to plot/visualize/analyze in a parallel manner the data available from the classification and from the reconstruction steps. But upstream to plotting, above all, one needs tools to store all those data, even in the most complicated cases: for example when each image was used in several classification attempts with different parameters, when one would like to test various combinations of images belonging to particular classes, and start a 3D refinement, for each combination, from a set of different helical symmetry parameters…

One promising way to deal with such complexity may be the use of databases, where all possible information on each segmented image, each corresponding filament, each class of any classification trial, each reconstruction test, would be stored. Then an image can be virtually associated with as many parameters as wanted. For example, it would be straightforward to apply image selection filters like "keep only images that were found to belong to classes with intra-class variance below than X in each of my N tests, that have an associated weight of Y for this eigenimage type, that were always found to correlate well with reconstructions refinement which converged to a symmetry of Z subunits per turn, etc….". The use of databases has recently been added in the world of EM software via the EMAN2/SPARX packages (Hohn et al. 2007; Tang et al. 2007), and one should thus explore this new data storage system and apply it if suitable to the requirements of processing of helical structures and in particular to for a smart coupling of classification and reconstruction.

Classification of power spectra

Rationale

We have seen in the introduction that a remarkable property of projection of helices was their particular signal in Fourier space, with discretization of the information on layer lines on which again the information is not distributed evenly. Roughly speaking, whereas in the case of a projection of an asymmetrical structure useful coefficients are distributed everywhere in the Fourier space, for helices we have a “condensation” of the signal, making it also much more visible, even when looking at the Fourier Transform of a single projection image. The precise position of the amplitude peaks of information is dependent on the helical symmetry, as well as on the out-of-plane tilt angle of the particles which has a precise and predictable effect. Thus, if one could apply classification techniques to FT (or preferably first to the Power Spectra -PS-, to have pixels containing real values instead of the complex ones that contains the phase information in the FT) of images of helical particles, one should be able, thanks to the relatively strong signal, to separate images according to the helical symmetry (and eventually out-of-plane tilt). Moreover, the process of class-averaging should make possible to significantly enhance signal for each symmetry class, thus providing useful for retrieval of the symmetry parameters from the PS.

The idea of using MSA techniques to classify power spectra of images (or modified power spectra) instead of their real-space version is not new. It was for example applied on rotationally averaged power spectra of micrograph pieces to assay local quality of cryo-EM images taken on carbon-coated grids with thin carbon film (Gao et al. 2002). It has also been used to sort power spectra of picked particles according to similar CTF parameters (Sander, Golas, and Stark 2003), after high-pass filtering of power spectra to raise the relative signal of the frequency range with fast sign changes of the CTF.

However, using the power spectra of images as a basis to classify images of helical objects, in the same way as one would do with the real original images, is to our knowledge not described in the literature.

We can ask ourselves about the advantages, if any, of classifying the PS of images instead of their real-space version, because they are in the end representing almost the same information, with even a loss of the phase information in the case of the PS. There are actually multiple advantages:

-precisely because we only look at the square of the amplitudes and thus get rid of the phase information, we do not have the translation variability information between similar images. For classification of the real-space images we had to consider that two different classes might be created only by taking in account differences in translation. We also had to calculate/interpret eigenimages that just reflected translational variability. For the MRA of the images against the classes, we had to give a non-negligible search range for translations to account for badly centered images. Altogether, this represented an important waste of calculation time and made the analysis more complicated. Classification of PS would therefore in principle allow to create fewer classes (much quicker analysis / easier to split data), and restrain the alignment during MRA to rotational search only.

-the on-axis variability is also not considered when looking at the PS: only the phases vary for various on-axis angles whereas the amplitude of the FT is constant (see introduction, **figure 1.26C**). In contrary, the on-axis view can have a very strong effect on the motifs in the real-space projection, due to the varying superimposition of the near and far side of the helix. This fact is for example well illustrated by the moiré pattern varying along a microtubule projection (Chrétien et al. 1996). For our classification purpose, we are not particularly interested in splitting data corresponding to various on-axis views, as we mainly want to separate helical symmetries. We see consequently the advantage of classifying the PS instead of the real-space images.

-Finally, although the interpretation of PS is not straightforward, we expect that their classification can visually show symmetry heterogeneities better than classification of the real-images (at least when looking at the class-averages on the real-space level). This point will be widely illustrated using the example of Measles Digested nucleocapsids.

Method description

We propose here one method for classification of Power Spectra, but it should be mentioned that it is still in progress and that the current description only reflects our first attempts. More ideas of the method improvement will be given in the perspective part. The general steps of the iterative procedure are illustrated on **figure 2.15**.

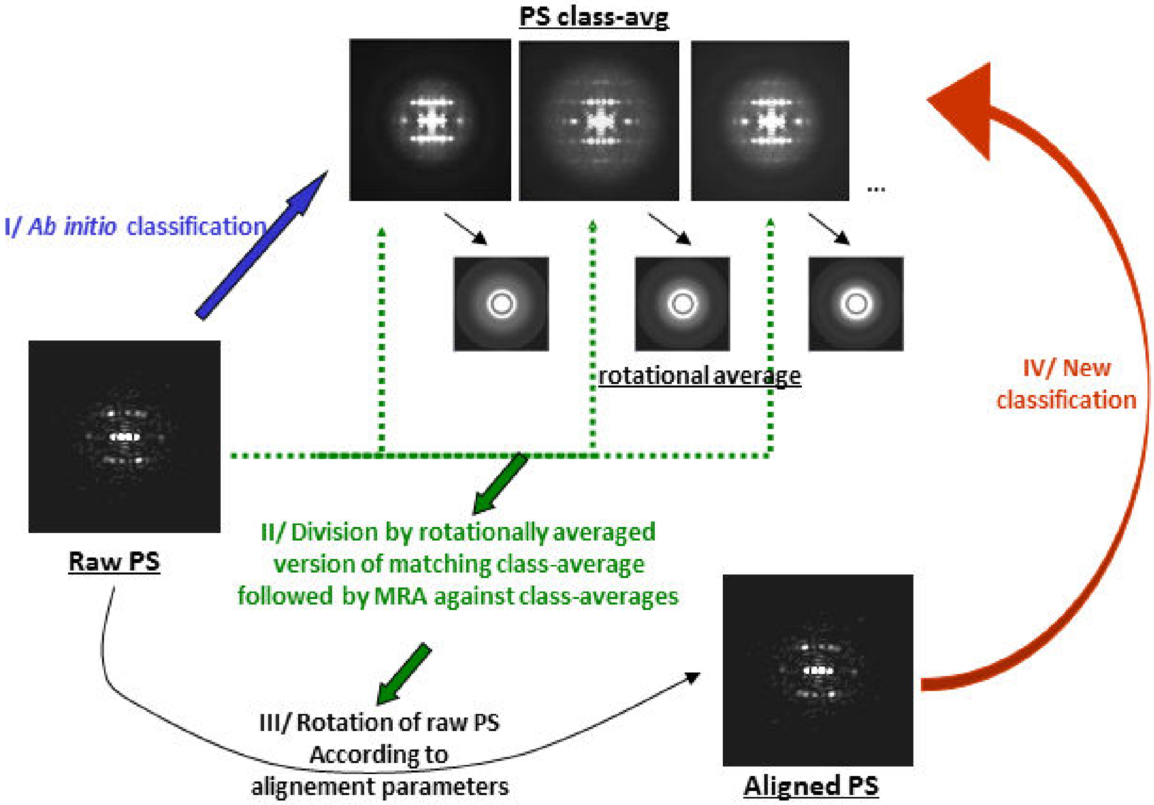

PS class-avg
...
I/ *Ab initio* classification
rotational average
IV/ New classification
Raw PS
II/ Division by rotationally averaged version of matching class-average followed by MRA against class-averages
III/ Rotation of raw PS According to alignement parameters
Aligned PS

Figure 2.15 : Scheme of the proposed method for the classification of images based on their Power Spectra

See text for details

The PS of the pre-verticalized (using boxing) initial images, which can be low-pass filtered, and preferentially padded into a larger image before PS calculation to decrease the frequency sampling step, are classified by MSA-HAC to produce class-averages with higher SNR (**step I**). At this step, one may try various classification parameters, like the mask used for MSA, the number of classes to compute, etc., in an object-dependent manner. After classes were calculated, each raw image is divided by the rotational average of the corresponding class average (**step II**): the aim of this step is to increase the signal at higher frequencies in comparison to the typically strong low frequency contribution. For helices, this is particularly useful to lower the contribution of the strong reflection at the equator. As one would then also do for real images, we choose a subset of representative and “well-defined” class-averages for a subsequent MRA of the modified raw images (**step II**). Here, representative and “well-defined” class-averages can be obviously more clearly assessed than for the real images classes. If we identify several symmetries by observing variation in the position of the layer lines and/or in the intensity maximum on the layer lines, then one should keep at least one class of each detected type. The “well-defined” criteria can be directly judged from the highest resolution reflections in the PS classes. This criteria must be applied for each symmetry class (if several) detected and not beforehand, because different symmetries can potentially be associated by variation of the regularity. Because one variability source that we want to get rid of during the classification is the remaining in-plane rotation (the segments are not perfectly vertical before PS calculation), we should also keep for MRA only classes that do not show in-plane rotation. In the first rounds of classification, if one chooses only a low number of classes, the expected random distribution of the remaining small in-plane rotations of the segments around 0 degrees should ensure that the classes will actually show a vertical pattern (due to the averaging procedure): in our hands, we found it to be the case. As mentioned, the selected class-averages will then be used as references for a MRA of the individual PS that were divided by the rotational average of their corresponding class average (**step II**). Due to the property of PS, no translational search is necessary for this alignment step. Once an in-plane rotation is found for each PS, it is applied to the non-transformed PS (**step III**) that will be used for a new classification round. Steps **II**, **III** and **IV** are then iterated until the classes and the determined angles of rotation stabilize. The fact that we currently use the non-transformed PS for the new classification rounds can be arguable. Indeed, we mentioned that we wanted to relatively enhance the signal at higher resolution, and this is now only done for refinement of in-plane rotation search. Actually, the reason why we did not use the transformed PS for each new classification cycle is related to the current design of the procedure: each PS is divided by the rotational average of the corresponding class, so that all the PS from one class are modified similarly. As this modification is strong, we can then expect that all images modified similarly will be then again classified together, despite of genuine but slighter differences that they could show. Naturally this should be quantitatively verified, and other possible better procedures will be proposed in the perspective section.

We will now see, based on MeVND and MeVD examples what type of results can be expected from this method and what are the difficulties encountered.

Some results

As a way to better visualize helical diffraction

A usual way of increasing the SNR in the PS of images is to compute the total sum of the PS of the segments (E H Egelman 2007; Narita et al. 2001; Y. A. Wang et al. 2006). To do this while not losing signal, the segments must not only be well rotationally aligned, but also possess the same symmetry, which is often not really the case.

The **figure 2.16** show for MeVD (A) and MeVND (B) the total sum of power spectra (left) of the verticalized segments (from boxing) and for each, two chosen PS class-averages as calculated using the procedure described above (right). As expected from our first observations, we can see that the total sum of PS from MeVD images shows higher resolution details than the total sum of PS from MeVND images. In particular, whereas the total sum of MeVND PS shows only one clear layer line at 1/60 Å^{-1} (and eventually a very faint second layer line at 1/30 A^{-1} indicated by a green arrow), the PS sum for MeVD shows a clear last layer line at 1/25 A^{-1} (green arrow).

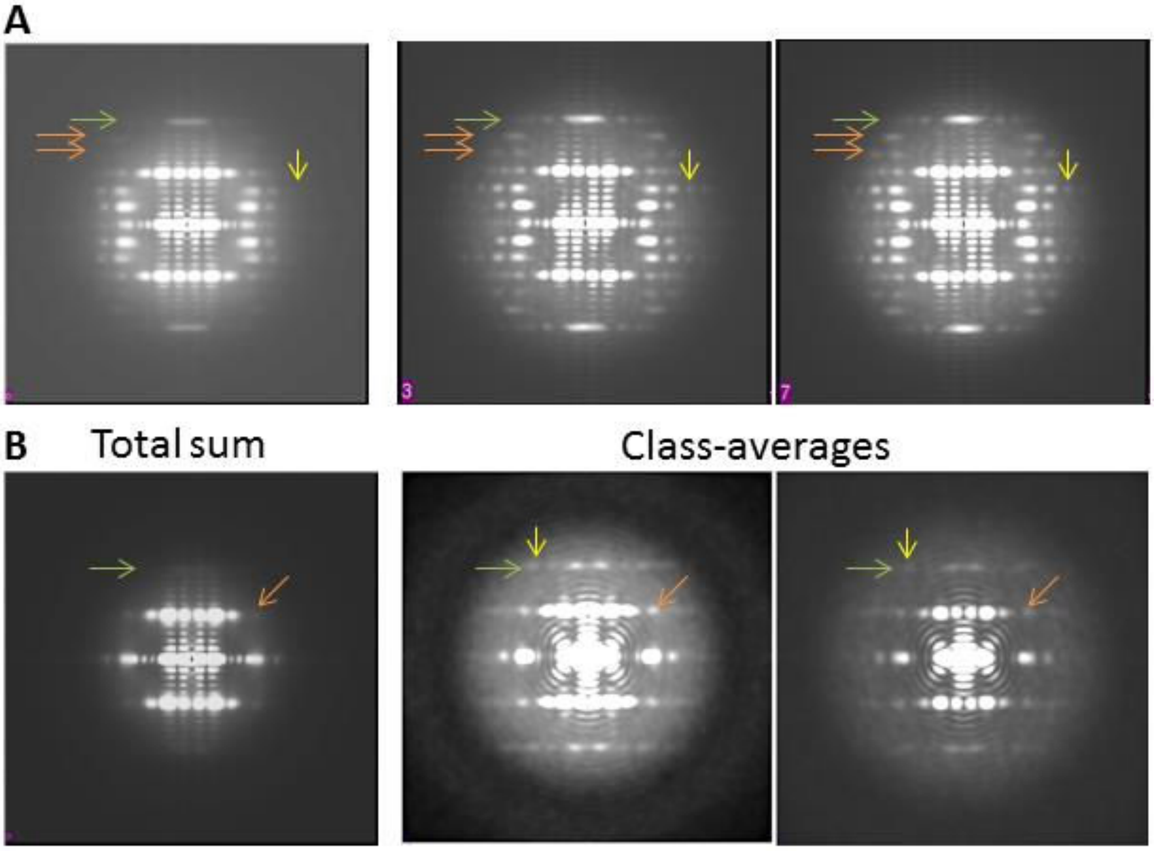


Figure 2.16 : Classification of power spectra as a way to better visualize helical diffraction

The total sum of the power spectra of verticalized segments of MeVD (**A**, left) and MeVND (**B**, left) is compared to two chosen power spectra class-averages as calculated using the presented method (right)

Compared to the total sum, we can observe from the two class-averages shown on the right for each sample that the classification procedure indeed significantly improved the quality of the PS, in term of higher resolution signal. For MeVND, it is striking: the second layer line at ~1/30 Å^{-1} not really observable from the total sum of PS is now well defined (green arrows). We can even observe on this layer line a peak further away from the meridian (yellow arrows) that could indicate the contribution of a Bessel term of higher order on this same layer line (because its intensity is higher than the precedent intensity peak on the same layer line). Similarly, on the first layer line at ~1/60 Å^{-1} more peaks are visible further away from the meridian (orange arrows), either due to repulsion of Bessel functions or to the contribution of another Bessel term. So we see that on this example, simply looking at the total sum of the PS would have led to the wrong conclusion of a very poor ordering of the structure in the range of resolution below 60 Å. In contrary, the classification of the PS show that at least a part of the segments can give raise to diffraction signal up to 1/30 Å^{-1}.

For MeVD, we cannot see on the class-averages additional layer lines at higher resolution than the already observed one at ~1/25 Å^{-1} (green arrows), suggesting that we were already close to the resolution limit. However, two layer lines between the ~1/25 Å^{-1} and the ~1/50 Å^{-1} layer lines are now much better recognizable (orange arrows). In principle, these layer lines might help to index unambiguously the diffraction pattern. In addition, similarly to the MeVND example, we can appreciate the improvement of the quality of the PS class average in comparison to the total sum of PS via the appearance of more repulsion of the Bessel functions (as exemplified by the yellow arrows).

To summarize, we have seen as a first advantage of our method of PS classification how we can improve the quality of PS. Now, as we have done for real images, we will see how one can use this classification to discard images.

As a way to discard images

Poorly diffracting images

For any type of protein or protein complex studied by EM, a limit for obtaining good resolution structures is the regularity/homogeneity of the objects studied. For helices, a simple way of appreciating the regularity of the helical assembly is to measure at which resolution the PS of images still shows layer lines with intensity peaks. The presence or absence of a layer line being a relatively strong change, we can expect that our classification procedure would make possible to detect and group such different PS. Indeed, when creating many classes, we can observe for MeVND that some classes show the second layer line at 1/30 Å^{-1} (**figure 2.17A**) whereas other don't (**figure 2.17B**). So we can see that in this case, PS classification may provide a way to select images corresponding to more regular objects. However, and this is a general remark not only applicable to MeVND, this must be done cautiously : if different classes actually represent different symmetries, some symmetries might be are associated with less order, and one might still wish to get a 3D structure for each symmetry class, even if a poor quality.

For MeVD, the layer line at 1/25 Å^{-1} is always seen, illustrating the higher ordering of those structures, and no effort was done to sort classes upon more finer changes, like the presence or absence of more or less repulsion of Bessel functions. One of the reasons for this, and this is also valuable for the presence of the second layer lines in MeVND example and more generally for PS classification, is the following: we have to think about the possibility that the position of the peaks that we use as an indicator of the regularity coincide with a region close to a zero of the CTF, where the amplitude of the signal is damped anyway.

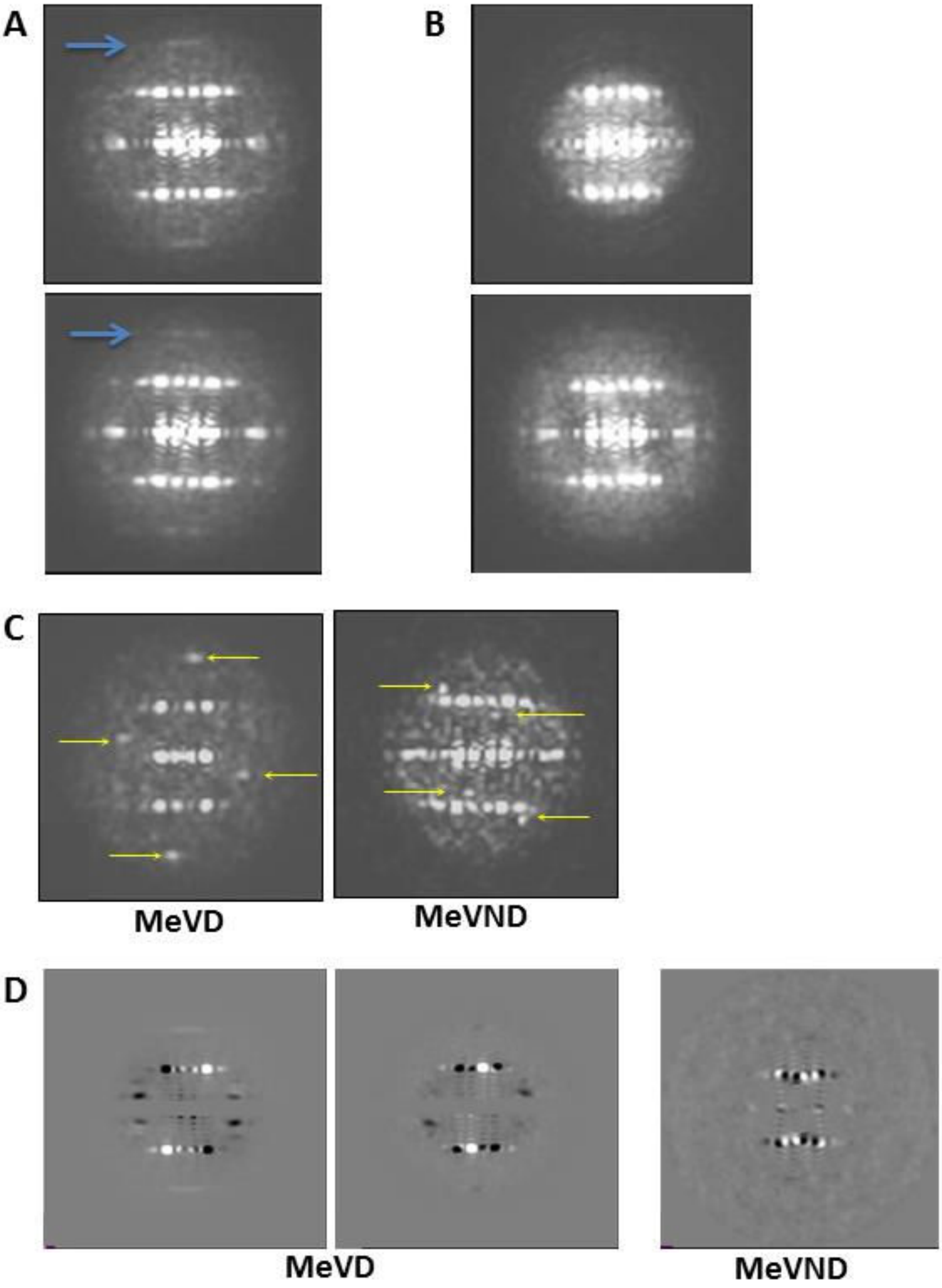


Figure 2.17 : Classification of power spectra as a way to get rid of poorly diffracting images and one-sided stained segments

For MeVND, when generating many class-averages, some showed the second layer line at 1/30 Å^{-1} (**A**, blue arrows) whereas other didn't (**B**). The panel **C** shows class-averages, for each sample, which exhibit an asymmetric pattern in the PS (yellow arrows), indicating one-sided staining. This feature is already detected from the analysis of some eigenimages (**D**). The pixel size in A, B and C is different than in D.

One-sided staining

An unpredicted result that came out of our procedure of PS classification was that several classes, both for MeVD and MeVND, were showing an asymmetrical PS pattern (**figure 2.17C**) in respect to the meridian. Reflecting this, several eigenimages representing a relative important contribution given their rank (usually in the first 10) showed asymmetric patterns (**figure 2.17D**). This type of asymmetry is due to the fact the only one side (the near or the far side) of the helix contribute to the formation of the image (Klug and DeRosier 1966). Thus, despite our efforts to ensure a complete embedding of the particles in stain using a double carbon sandwich technique (Frank, 1996), it seems that a part of our images actually represent filaments that were only partially embedded in stain. Although such a partial embedding can be useful if one wishes to assess the hand of a helical reconstruction (V. Korkhov and Sachse 2010), we must avoid it for 3D reconstruction purpose because it disrupts the true helical symmetry. PS classification is therefore a valuable way to discard such images, which were not detected in the classification of real images.

As a way to separate different symmetries

The variability of pitch of MeVND

As already seen from the classification of real-space images, MeVND sample seemed to exhibit large pitch variability, which we expect to even more clearly appear in the PS classification. The **figure 2.18** shows eight PS class-averages on which the position of the first layer line relatively to the equator varies. To represent this important source of variability, a high ranked eigenimage (number 3) shows very clearly the variable position of the first layer line (**Figure 2.18**, bottom right corner). It is clear that a variation of pitch can explain this behavior: the closer the layer line is to the equator, the higher the pitch of the corresponding structure. However we should keep in mind that another source of variability can account for variation of position of the layer lines: the out-of-plane tilt (as seen in introduction, **figure 1.26D**). When there is out-of-plane tilt, the FT of the 2D projection corresponds to a central section that is tilted relatively to the helix axis (**figure 1.26D,** right), which means that it crosses the layer planes at a different reciprocal spacing; the higher the out-of-plane tilt, the larger the spacing. The difference in layer line position is simply given by

$$L(\alpha)=L * \cos(\alpha)$$

with L(α) being the position of the layer-line when an out-of-plane tilt of α degrees is applied, and L the position of the layer-line when no out-of-plane tilt is present. The out-of-plane tilt alone is however not sufficient to account for the difference in the layer-line positions that we observe for MeVND. For example, the effect of a 12 degrees out-of-plane tilt (which is already fairly high) at the layer line positioned at ~27 pixels from the origin and the closest to the equator (**figure 2.18,** top left class), would be only an approximately 0.6 pixel shift ($27/\cos 12 = 27.6$), whereas we observed up to 5 pixel displacement (if one compares for example class 1 and class 8 on **figure 2.18**).

We have a second indirect proof that the effect that we see is not due to out-of-plane tilt but lies on the pitch variability. A visual comparison between the classes shown at **figure 2.18** shows that the further away the first layer line is (thus presumably the smaller the pitch), the more pronounced higher resolution signal can be observed (second layer line appearing). The eigenimage corresponding to the layer line position variation (**figure 2.18**, bottom right corner) highlights this effect: the "white signal" corresponding to higher pitch is associated with less order, in comparison to the "black signal" (low pitch). In the light of the "out-of-plane tilt hypothesis" no logical explanation for such a behavior can be found, whereas in the light of the "pitch variability hypothesis" the explanation is straightforward. The helices with a higher pitch have less contact between turns; they are less "packed", and thus more flexible. Therefore if we want to restrict our 3D reconstruction attempts to the most regular structures, we should go for the smaller pitch classes.

In our PS classification procedure, the sorting according to pitch as exemplified on **figure 2.18** was done manually, but it would be easy to set up an automatic procedure. One could choose a well-defined layer line (like in MeVND the first layer line), and calculate for each PS class-average the 1D projection of the class average along the layer lines direction: its maximum value could then be used to automatically sort images based on their pitch.

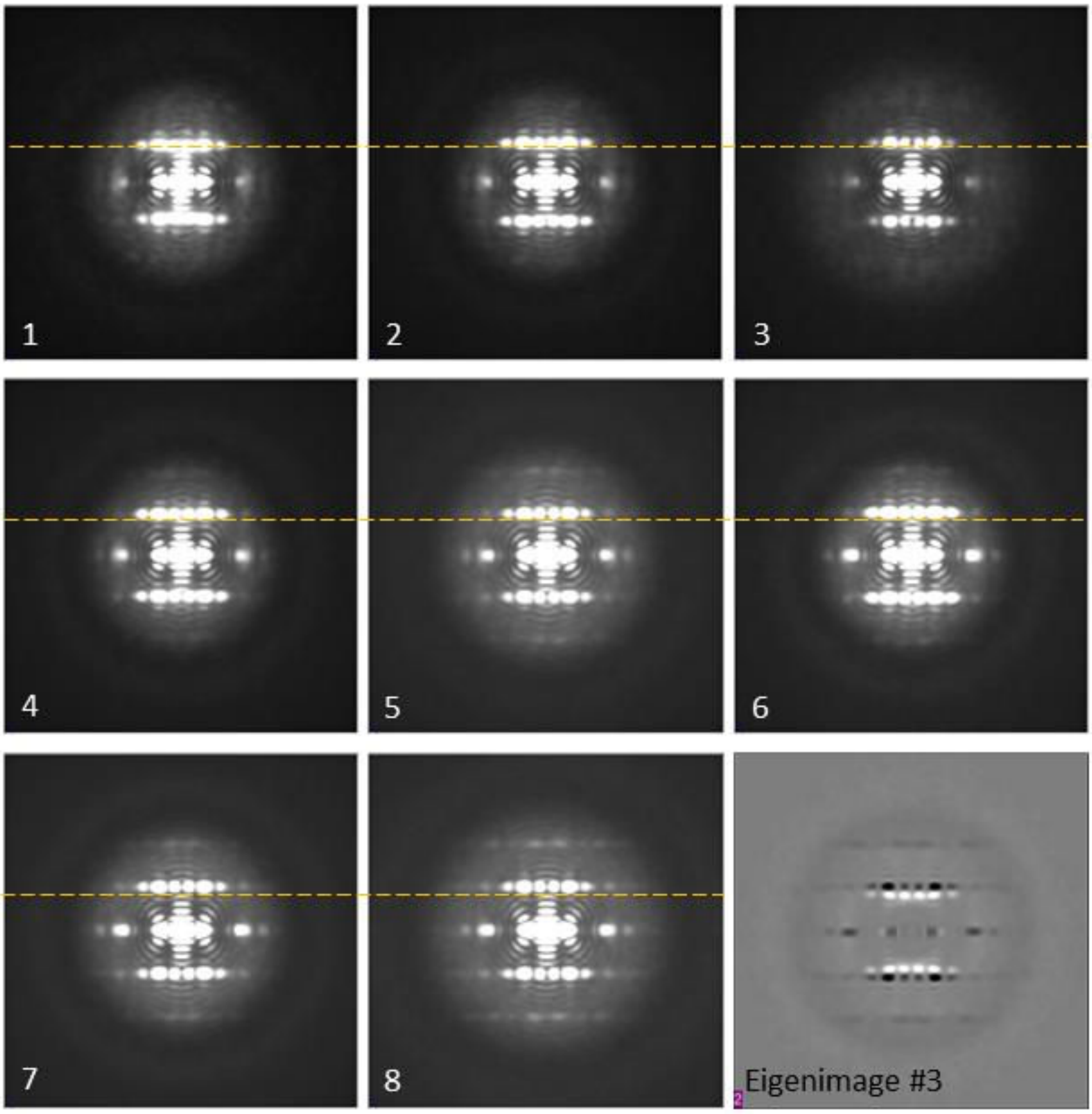


Figure 2.18 : The classification of power spectra of MeVND reveal pitch variability and higher order of lower pitch segments

Power spectra class-averages of MeVND, numbered 1 to 8, show pitch variability as illustrated by the variable position of the first layer line regarding to the equator. The dashed orange line is crosses the power spectra class-averages at exactly the same height (at ~1/70 Å^{-1}). As already seen from the class-averages, but highlighted by looking at the eigenimage describing pitch variability (bottom right), the lower pitch structures (diffraction in black on the eigenimage) are more regular than the high pitch structures (diffraction in white).

Two helical states of MeVD

For MeVD, the classification of real images showed that a small proportion of the images (~ 5%) belonged to classes with a very different pattern than the majority of the segments and to what was already described (Schoehn et al. 2004). Those images were removed from the data set and therefore we expected the remaining images to represent a relatively homogeneous sample. For example no pitch variability was suspected from the first analysis. However, in our attempts of 3D reconstruction using the IHRSR method (E H Egelman, 2000) with those images, we were systematically observing problems of convergence of symmetry parameters, suggesting a heterogeneity issue. The classification of PS and subsequent inspection of the class-averages gave an explanation to this problem: there were clearly at least two different type of helical symmetry in the sample, recognizable by two types of pattern in the PS class-averages, which we will name "Type 1" (**figure 2.19A**) and "Type 2" (**figure 2.19B**), the later representing ~ 30 % of the images. The main differences clearly appear by comparing the sum of each type of class-averages (**figure 2.19C**) : the two layer lines between the equator and the layer-line at 1/50 Å are closer to each other in the Type 2 classes than in the Type 1 (**figure 2.19C**, blue arrows). Moreover, the first intensity maximum on the layer line at 1/25 Å appears further away from the meridian in the Type 2, in comparison with the Type 1 were it seems that only one peak is present (**figure 2.19C**, orange arrows). We tried to identify what could be the source of this heterogeneity. We first looked at the distribution of the two types of PS patterns among the micrographs and found out that it was not an even distribution: one micrograph typically contained only one type of images, and furthermore, the micrographs containing "Type 2" images were grouped in one image acquisition session at the microscope (which corresponded to one grid, and one batch of sample). As a possible source of heterogeneity, we thought about flattening, which could have happened more extensively on this particular grid, due to the condition of grid preparation, that are never 100% reproducible (e.g. dehydration rate). However, we tried to simulate the effects of flattening on the power spectra of images by creating a flattened 3D "pseudo-helical" model, and looking at the PS of its projection, and we could never reproduce the differences that we see between classes of Type 1 and 2. A second option might be that the sample itself was different, for example if incomplete digestion by trypsin occurred, which could be barely detected on the regular SDS gels if only a few more amino acids are present.

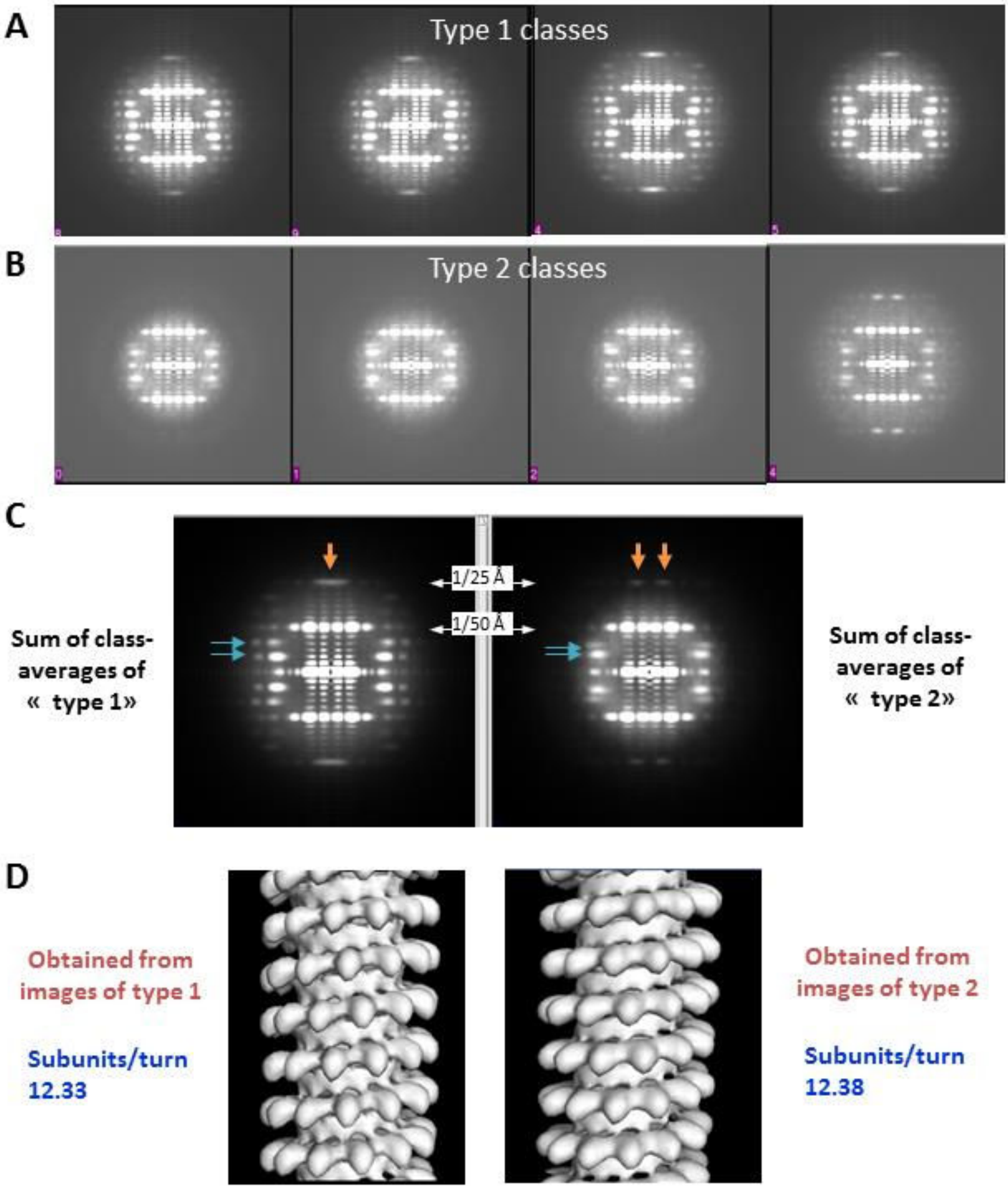

A
Type 1 classes
B
Type 2 classes
C
1/25 Å
1/50 Å
Sum of class-averages of « type 1»
Sum of class-averages of « type 2»
D
Obtained from images of type 1
Subunits/turn 12.33
Obtained from images of type 2
Subunits/turn 12.38

Figure 2.19 : Detection of two helical symmetries on MeVD by classification of power spectra

The examination of power spectra class-averages of MeVD revealed two types of patterns, noted « type 1 » (**A**) and « type 2 » (**B**). The sum of class-averages of each type highlight the differences (**C**) between the two types. The sorting of images based on this classification outcome helped the IHRSR procedure to converge, and we obtained the two reconstructions shown in **D**, with slightly different symmetry parameters.

After sorting the images based on their PS type, we could easily obtain convergence of symmetry parameters using IHRSR procedure, and interestingly we found two different numbers of subunits per turn, respectively 12.33 and 12.38 for types 1 and types 2 images (**figure 2.19D**), which indeed can give rise to the differences we were observing in the PS. We double checked these final values by taking images of type 1, making a reconstruction and searching for the helical symmetry starting from final symmetry parameters found from type 2 images, and vice versa. The parameters again converged to those initially found. Whether this slight change in the symmetry could be explained by the presence of a few more amino acids, or if we have to think about other explanations is still unknown.

Anyway, this example showed that a heterogeneity that was not detected in our first approaches and that prevented the reconstruction process could be detected by classification of the PS. Furthermore, if we can trust the final values found for the symmetry of type 1 and type 2 images, it shows that a very small variability can be resolved by this way. In order to assess if the real-space classification also "detected" this heterogeneity, it would be interesting to check if the images corresponding to the different type of PS classes were systematically partitioned into different real-space classes, or not. Based on our knowledge from the PS classification, we should also examine the PS of real-space class-averages more carefully and try to find the same patterns.

Encountered problems and some perspectives

From the first results that we obtained, it seems that for helical samples the classification of PS can successfully complement real-space classification. However, the details of the method and the use of its outcome were not optimally pushed. The validation of the improvements it allowed were also not systematically done, except for MeVD were the sorting of the two different symmetries was crucial for correct structure refinement. The same remarks that were made regarding classification of real images, concerning perspectives for a better validation and use of the classification outputs (e.g. use of database), hold true for the PS classification as well. We will now see the principal difficulty encountered, and discuss perspectives to circumvent this problem and more generally to improve the method of PS classification.

CTF effects

We already noted above that the use of classification of power spectra of individual images using MSA had to our knowledge only been done in order to classify images according to their CTF, thanks to the strong contribution of the Thon rings to the signal. In (Sander, Golas, and Stark 2003), the authors note the following: “It may therefore be taken that the subsequent classification is really based upon CTF parameters and not upon structure factors”. Here we have shown that at least in the case of a helical sample, the structure factors can be an important property of the images based on which the individual PS will be classified. Indeed, we have seen both for MeVND and MeVD that we could distinguish several helical states based on PS classification. This suggest that the CTF determination and correction approach proposed in (Gao et al. 2002; Sander, Golas, and Stark 2003) could be difficult to apply when many helical filaments are present in the images, especially if different symmetries coexist, eventually even in the case where the rotational average of the power spectra and not simply the original power spectra are classified (Gao et al. 2002).

However, it is true that an evident source of variation and of similarities among images that should be particularly well visible in reciprocal space is indeed the CTF. For the sake of simplicity, this issue was not mentioned above. It is however true that the results of our classifications were also strongly influenced by the CTFs, which can be seen as a potential drawback of our method. Considering the PS classification of MeVND, we can see already in the eigenimage number 2 (the first after the total average) that concentric circles which look quite similar to Thon rings appear (**figure 2.20A**, left). Actually, in this eigenimage, the signal corresponding to the helical component (equator + one layer line), appears only in one “direction” (here strong black signal, but the inverted version is implicit), with no variability source, as it is for example for the next eigenimage, that we showed at **figure 2.18,** bottom right corner. On the contrary, the contrast of the concentric circles alternates between black and white, which means that that the variability explained by this eigenimage is actually the position of the Thon rings. Other eigenimages further away in the ranking also clearly show the CTF contribution (**figure 2.20A**, number 4 and 9). Moreover, if one looks carefully at several class-averages, one can see classes that exhibit a similar helical pattern (at least at the level of our inspection), but that seem to only differ in the position of the Thon rings (**figure 2.20B**). Thus, due to the CTF effects, we have to consider when looking at different PS class-averages that the only difference between them may in some cases arise only from the position of the Thon rings.

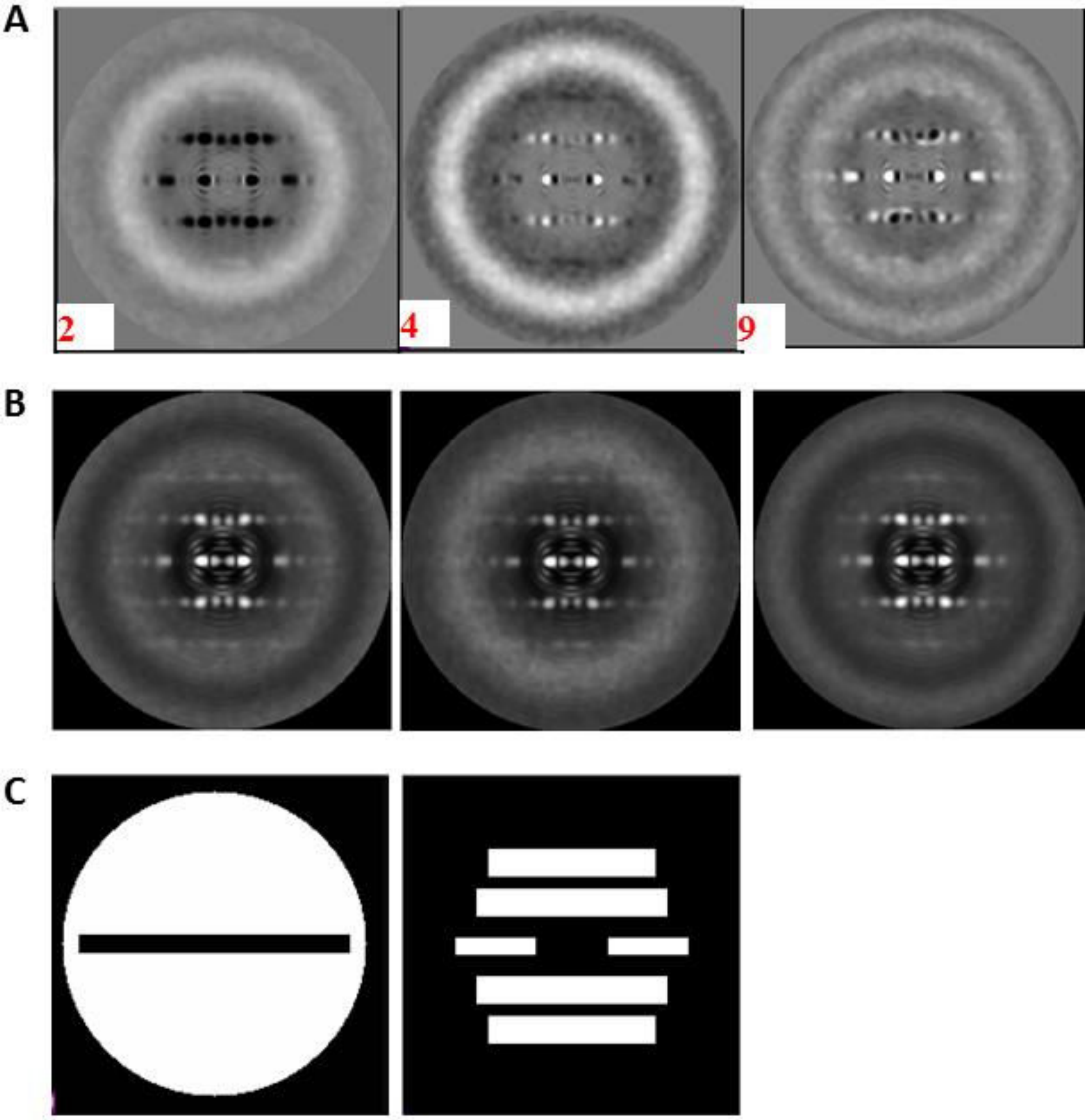


Figure 2.20 : Influence of the CTF on the classification of power spectra of individual segments (example of MeVND)

The first eigenimage after the total-average obtained from the classication of power spectra for MeVND (**A**,number 2) showed that the first source of variability among the PS is the position of the Thon rings. Other eigenimages further away in term of importance also show strong Thon rings signal (**A**, number 4 and 9). In accordance to this observation, several class-averages seem to only differ by the position of the Thon rings, and show the same helical signal (**B**). The panel **C** shows different types of masks used for MSA. See text for details.

Reduction of CTF effects

To try to restrict the classification to only the structure factors of the sample and not to take the CTF variations into account, we made preliminary trials of the use of different masks for MSA, which only encompass the visible helical signal (**figure 2.20C**, right), rather than just using a mask that excludes the equator which showed to improve the classification (**figure 2.20C**, left). However, this could not completely counter-act the problem of classification according to CTF and requires more manual intervention. We can understand this failure by just considering that some intensity maximum that should be in the PS (e.g. a first maximum of a Bessel function along a layer line, thus a strong signal for classification) can be just canceled out in some images, due to a minimum of the CTF at the same resolution shell. The use of masks for PS classification should be further tested and evaluated.

One way to circumvent the problem of classification according to CTF, would be to classify images that were acquired at similar defocus. First a separation of the data set according to defocus would be done (e.g. using software like CTFTILT or CTFFIND3), and the described PS classification would be done on images from same defocus groups. However, depending on the defocus range used to group images, this would result in much less images to classify in each group, and thus a lower expected SNR in the class-averages. Furthermore, a subset of images, particularly if it is small, is not necessarily representative of the ensemble of the data, especially when lots of heterogeneity exists. Finally, this method would add an additional difficulty, which would be to identify in the PS class-averages coming from different defocus groups which ones actually represent the same symmetry and which ones don't.

Another way to diminish the influence of the CTF, would be to apply to the raw PS a function that compensates for the CTF oscillations after modeling those oscillations (i.e. mainly the defocus and the decay of amplitudes), using a regular CTF determination software. This could be done for example by using an appropriate CTF correction with a Wiener filter. Of course this would never be able to compensate for the variability in positions of the zeros of the CTF for which we have no signal from the sample, and which could matter for the classification.

Improvement of individual PS alignment

We mentioned during the method description that one should keep as references for alignment of the PS only class-averages that do not exhibit in-plane rotation (i.e. that are perfectly vertical), as this is a source of variation that we want to get rid of. In practice, we barely eliminated classes for this reason, and hoped that by generating only a few classes at the beginning, one would average the uncertainty of in-plane and thus have a vertical class-average and that the alignment/classification iterations would have a tendency to correctly verticalise the PS. However, this would not work if there is a systematic bias in the in-plane orientation that is likely to happen due to manual boxing. Thus, instead of the manual step of class-average selection, one could easily imagine a way to verify that the PS class-average is actually vertical, and correct for eventual in-plane rotation. To do this, one could use the vertical mirror symmetrical property of the PS of helices: one would compute the difference between the left and the right side of the PS as a function of in-plane angle applied, and find the minimum of this function. Of course this would not work in the case of PS of one-sided stained filaments. Therefore they should be removed beforehand. On the other side, it can be envisioned that during this procedure of finding the minimum of the difference left/right of the PS, the classes corresponding to one-sided segments would be automatically detected (because the minimum would not be as pronounced as for good classes).

A better way to enhance higher resolution signal

In order to classify images upon finer variation in the PS, in particular at higher resolution, we think that a good way of higher frequency signal enhancement would be useful. What we did up to now was simple: we often used a mask for MSA that removed the strong equator contribution, and additionally sometimes the lowest frequencies, but not tested systematically the effects of using various masks. For the alignment of the PS using the PS class-averages as reference, we used a division of the raw PS by the rotational average of corresponding class-average for higher resolution signal enhancement purpose. One possibility to enhance high resolution signal for the classification itself, would be to divide each PS by its rotational average and classify those modified PS, and not divide by the rotational average of corresponding class-average that introduce too strong features. One could also play with different ways of correcting for CTF decay (as proposed above for reduction of CTF effects). To summarize, there is a lot of space for improvement of this part of the method and different possibilities should be systematically tested and validated.

Gain more benefit from eigenimages

For the classification of real images, we saw how we could improve the classification outcomes by putting more efforts into the understanding of the eigenimages and use of their weighting for classification (e.g. classification according to diameter for VSV). The same approaches should be used for the classification of PS to obtain separation of images according to precisely defined criteria like one-sided staining, pitch variability, etc. For example, once the images are separated into finer pitch classes, one could eventually detect finer variability in the structure of these subsets, by using a classification workflow similar to the one described in (Elad et al. 2008).

Better differentiate pitch variation versus out-of-plane tilt

As we could see during the analysis of classification of PS of MeVND, it can be tricky to differentiate if the differences in layer line heights that can be observed are due to pitch variability or to various out-of-plane angles of the corresponding segments. Especially if the pitch variations are small, and with higher resolution data, this difficulty could be limiting. A way to overcome this problem would be to not only consider the PS, but keep the phase information available at some points of the procedure. Indeed, calculating the difference of the phases on the left and right side of the PS (at the same meridional and equatorial position), is a way to detect out-of-plane angle (Wakabayashi et al. 1975). A potential way of including the phase information in our classification procedure would be to first classify the left/right phase difference map of the images in order to split the data set into out-of-plane angle groups that would then be further separately classified. This may be however not so simple, for example due to the influence of centering of filaments on phases distribution, as well as to the in-plane angle that is not perfectly defined at the beginning. The possible ways of taking into account the phases for the PS classification procedure should be further explored and tested.

PART 2 : Ab initio symmetry guess and the ambiguities in helical symmetry

General remarks on the existing methods

Except for a few methods of reconstruction of helical filaments that do not require prior information on the symmetry (angular reconstitution: (Paul et al. 2004); (Hodgkinson et al. 2005)), virtually all the currently used methods critically rely on the precise knowledge of the helical symmetry parameters. In the oldest, classical method the symmetry parameters are obtained by the analysis of the Fourier transform from a filament projection ((Klug, Crick, and Wyckoff 1958) ; (DeRosier and Klug 1968) ;(DeRosier and Moore 1970)). This analysis is not always straightforward, and can be hampered both by problems due to the symmetry itself, when the indexing of the diffraction pattern is impossible (Bessel-overlap), or by the irregularity of the filaments resulting in an interpretable diffraction pattern. Although analysis of FT can be helped by specific programs ((Ward et al. 2003) ; (Whittaker, Carragher, and Milligan 1995) ; (Toyoshima 2000); (Metlagel, Kikkawa, and Kikkawa 2007) ; (Owen, Morgan, and DeRosier 1996) ; (Beroukhim and Unwin 1997); (Yasunaga and Wakabayashi 1996)), it requires good knowledge of the underlying theory as well as significant human intervention. Nevertheless, this method has the advantage of being fully ab initio, and does not necessitate additional information except of relatively easily obtainable ones, like the radius of the particle. The "real-space methods" described in the literature are also based on the exact knowledge of symmetry. In the case of IHRSR (E. H. Egelman 2007), the procedure is able to refine to correct symmetry parameters only when starting from an initial guess very close to the true values (See above, see (Edward H Egelman 2010)). A more recently described method (Sachse et al. 2007) that uses constraints on the alignment derived from the helical symmetry, also requires a very precise knowledge of the symmetry, even if a recent extension of this method (Low et al. 2009) makes it possible to refine the symmetry parameters starting from roughly determined ones. Still, both available real space methods do indeed rely on the initial helical symmetry estimations which are either classically derived from the FT (V. M. Korkhov et al. 2010) or imported from previous studies. In a recent paper (Ramey, Wang, and Nogales 2009), propose a method for an "ab initio reconstruction of helical samples with heterogeneity, disorder and coexisting symmetries". However, in this paper, the “ab initio” term refers to the use of a 2D-reference free classification step, and what they further call “ab initio symmetry estimation” is based on the classical way of the FT indexing.

In the scope of the present work, we therefore searched for alternative means to obtain an initial guess of the symmetry, that do not rely on manual analysis of FT (since it seemed to be not feasible in our case, and since it is not always an easy task), and that require as little human intervention as possible. Two ways were explored: the first, based on 2D real images and the second, based on their power spectra. Most of the effort was invested in the first method which is described in detail below.

Proposed method of ab initio symmetry determination on 2D real images

Method summary

The basic idea at the heart of the method we propose is relatively simple. It consists in inspecting nearly all conceivable symmetries and identifying which is most likely to be true. To do that, we cut out of a 2D real image successive segments along the helix axis, and assign view angles to each of them, in order to reconstruct a 3D model. Via the angle assignment and the shifts between the segments, we can impose any symmetry we want on the 3D model. This 3D model is then reprojected, and the average cross-correlation between the segments and the corresponding reprojections is recorded for each tested symmetry. We anticipate that an inspection of the profile of the average cross-correlations as a function of the imposed symmetry parameters, will allow to determine which parameters are true.

Although simple, this method raised several questions and problems, that will be first detailed, and pointed out a central problem in helical reconstruction on which an original point of view will be given : the ambiguities in symmetry parameters determination.

Theoretical considerations

The proposed method is based on the hypothesis that assigning the correct view angles to a set of 2D images of a given 3D object, would allow to reconstruct a 3D model which projections would have a higher correlation with the initial images than a 3D model reconstructed from the same images, but with wrongly assigned view angles. Despite the fact that this assumption is at the basis of the projection matching method, one has to consider the possibility that two different 3D models can share several identical projections and thus that there might be no unique solution for assigning the "correct" angles to a set of images, in terms of correlation level between the images and the reprojections of the reconstructed 3D model. In the frame of this hypothesis one can propose that starting from an ensemble of 2D images of any 3D object (with or without symmetry), one could determine the correct 3D model/models by testing all different combinations of view angles for the input images while calculating 3D models and recording the CC between the reprojections of these models and the original images. In practice, such approach is computationally too demanding for a common single-particle. Indeed, if one assumes that the images of individual particles are centered, the number of combinations of the angles (one in-plane and 2 out-of plane) to test would be:

$$C = \left(\frac{360}{dpsi}\right)\left(\frac{360}{dtheta}\right)\left(\frac{360}{dphi}\right)^{N}$$

where dpsi, dtheta and dphi are the angular sampling in degrees, and N the number of assessed images. Even if one has only 10 images and considers an angular sampling of 4°, the number of combination to test would be 7290000. In the case of a helix, there exists a symmetry-dependent relation between the translation along the helical axis and the on-axis angle view of the helix. It means that knowing the helical parameters, if one constructs a set of segments regularly placed along a straight helix, all view angles will depend on the assignment (out-of-plane, in-plane, on-axis rotation) of the first segment. In line with the above, if one wishes to exhaustively test every combination of angles in order to compute a reconstruction from a set of images regularly placed along the helix axis, considering that the on-axis view angle of the first image is arbitrary (not a variable), the number of combination would be :

$$C = \left(\frac{360}{dphi}\right)\left(\frac{360}{dtheta}\right)$$

This number does not depend on the amount of images, but only on the angular sampling of the in-plane and the out-of-plane angles to be assigned to the first segment, on which values will depend all the angles of the next segments. If one now consider a helix that does not have out-of-plane tilt (or a known one), and that the in-plane rotation is known, there is no more combination to test: the view angles of the ensemble of segments will only depend on the helical symmetry. Accordingly, it is possible to reconstruct a 3D volume from a helix projection by segmenting the projection and assigning view angles corresponding to the symmetry, to each segment. Inversely, we propose that in a case where we don't know the helical parameters, one can "exhaustively" investigate every possible symmetry by assigning different sets of view angles to segments along a projection. Since for each symmetry the view angles of all segments on the projection are related each other, this approach is not as computationally demanding as a similar exhaustive approach would be for other single particles, and the number of reconstructions to compute is equal to the number of different helical parameters to test. An overview of the method is presented in **Figure 3.1**.

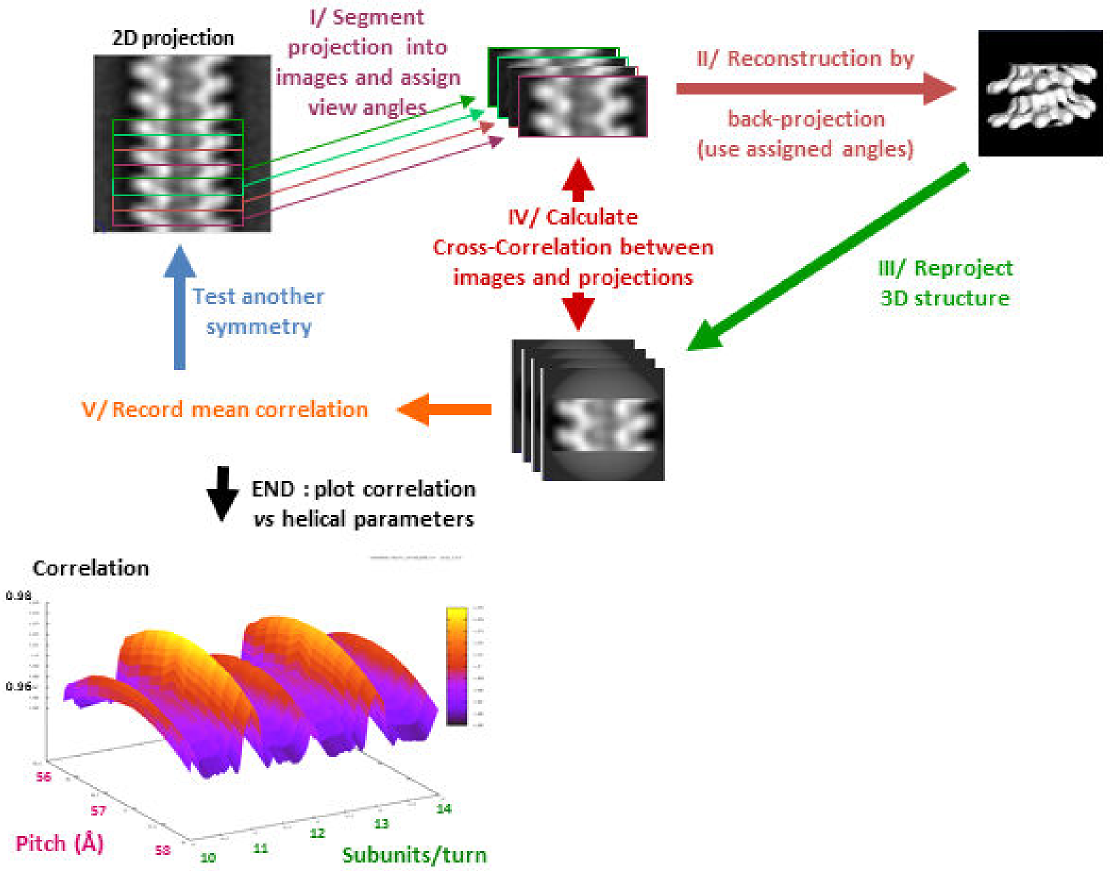

2D projection
I/ Segment projection into images and assign view angles
II/ Reconstruction by
back-projection (use assigned angles)
III/ Reproject 3D structure
IV/ Calculate Cross-Correlation between images and projections
V/ Record mean correlation
Test another symmetry
END : plot correlation *vs* helical parameters
Correlation
0.98
0.96
56
57
58
Pitch (Å)
10
11
12
13
14
Subunits/turn

Figure 3.1 : Scheme of the method of estimation of helical parameters from a single projection
The input image shown here as an example is a class-average of images of intact nucleocapsids of Measles virus. See text for details.

Step by step method description (Figure 3.1)

(I) The input projection is segmented into successive images along the helical axis, and view angles are assigned to each segment according to the symmetry (starting from an arbitrary on-axis angle for the first image). In theory, any kind of projection can be used as input, as long as the helical axis of the image is correctly centered and the in-plane and out-of-plane angles are known. In practice, our current version of the method script is designed for a projection with the helical axis vertically aligned and with the out-of-plane angle of zero. In addition, for an ab initio estimation of the symmetry (e.g. for a new project), it appears to be easier to run this procedure on class-averages (as obtained by reference-free classification) rather than on much noisier raw images, although this is not a rule. Not only should the extracted segments be regularly placed along the helix axis, but also the distance between each segment must be a multiple of the tested axial rise. Indeed, if it is not the case, then we would be looking at views of different objects (see **Figure 3.2**).

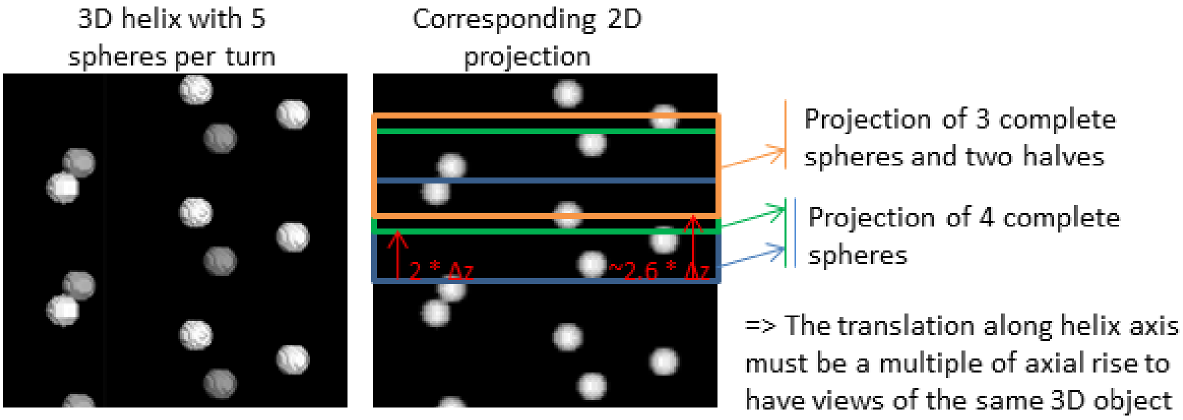

3D helix with 5 spheres per turn
Corresponding 2D projection
Projection of 3 complete spheres and two halves
Projection of 4 complete spheres
2 * Δz
~2.6 * Δz
=> The translation along helix axis must be a multiple of axial rise to have views of the same 3D object

Figure 3.2 : How to segment a projection of an helix to have different views of the same object

This restriction signifies that interpolation must occur when segmenting the projection, except in the cases where the tested axial rise is a multiple of the pixel size. It also means that, given the fact that our input projection is generally limited in length, the number of segments that can be extracted from it will depend on the symmetry tested (the smaller axial rise, the more segments can be extracted). The symmetry tested will also affect the homogeneity of the filling of the angular space. As will be shown below, all these points could have effects on later steps of the procedure.

(II) Using the segmented images and the corresponding view angles, a 3D volume is calculated by back-projecting the segments using interpolation in Fourier space (SPIDER command BP 3F). As the number of images that will be used for reconstructing the volume is generally very limited, we have to keep in mind the potential effects of the reconstruction algorithm, in particular the effects of interpolation when the 3D Fourier space is so sparsely filled.

(III) The computed 3D volume is then reprojected using the directions defined by the assigned view angles of the segments, and **(IV)** a normalized CC between each input image and the reprojections of the volume is calculated, inside the area defined by the usable density on the input images. As the number of images included in the reconstruction is small, we expect to have a high correlation for each individual image and its corresponding reprojection, as the input image itself will highly contribute to CC (conservation of information during back-projection and reprojection). We expect that when the wrong symmetry is imposed (as in a case of a wrong angles assignment in an asymmetric situation) this correlation will decrease due to the influence of the other images included in the volume.

(V) From these individual CCs, an average correlation is calculated and associated with the current tested symmetry parameters. For clarity reasons, this average CC between segmented images and corresponding reprojections of the calculated reconstruction will be later referred to as **ACC** (for **A**verage **C**ross **C**orrelation). Other parameters such as the number of images included in the reconstruction and the standard deviation of the CCs between images and reprojections are also recorded.

The five steps described above are repeated for every symmetry tested, and the ACC is plotted as a function of the imposed helical parameters.

Critical points. Illustrations by a case study of RSV nucleocapsids

The very limited number of subsequent segments

As mentioned, in our method the symmetry-dependent number of segments that can be extracted from a single input projection is usually very low. Firstly, this number cannot exceed $\mathbf{int(L/\Delta z)*N}$, where **L** is the length of the projection, $\mathbf{\Delta z}$ the axial rise and **N** the number of starts for N-start helices. Secondly, in order to include more than one subunit in the reconstructed volume and while paying attention to avoiding the image borders, the number of views is even lower because at both extremities, the distance between the center of the extracted segment and the border should be less than the half of the length of the final reconstructed volume.

Let us consider an example of a 360 Angstroms long class-average of the intact measles virus nucleocapsid. Such a class-average contains 6 helical turns with a 60 Angstrom pitch (**Figure 3.1**), and if we wish to include two-helical-turn segments (i.e. 120 Angstroms) in the reconstructions, then the distance between the center of the first and the last segments that can be extracted will be 240 Angstrom (360 – 2 x 60). If we test imposition of symmetries between 8 and 16 subunits per turn (i.e. axial rise varying between 7.5 and 3.75 Angstrom), the number of segments that can be extracted will vary from 32 (= 240 / 7.5) to 64 (= 240 / 3.75). Despite the fact that each segment contains several views of the subunit, such low numbers of images are usually far too low for attempting a meaningful reconstruction, and may be considered as a problem for the method. The angular space (here we will consider only the on-axis angular space) will indeed be very sparsely filled, and, even more embarrassing, this filling will strongly depend on the symmetry. For example, while testing a symmetry with an angular rotation **Δφ** between subunits being is a divisor of 360, the number of different available views will only be equal to **360/Δφ**.

Effects of number of segments on the cross-correlation and considerations on interpolation

Some effects of number of images and filling of angular space on CC are shown in **Figure 3.3**. The **panel A** shows a 3D model constructed based on the Xray crystal structure of the RSV nucleocapsid ((Tawar et al. 2009) EMD-1622), such that it contains at least one full repeat (dashed yellow line, attained after 23 turns -225 subunits-), to have the optimum sampling of angular views. The **panel B** shows its projection, from which portions of variable length were used as input images to impose the true symmetry using the described method. The **panel C** shows the plot (red curve) of ACC as a function of the number of views used (i.e. of the length of the input image), and the standard deviation (green curve) of CC calculated between segmented images and reprojections. **Panels D** and **E** show plots of the correlation as a function of the view angle, for different number of views used for reconstruction (67, 165 and 225 views in **panel D** ; 224,225 and 226 views in **panel E**).

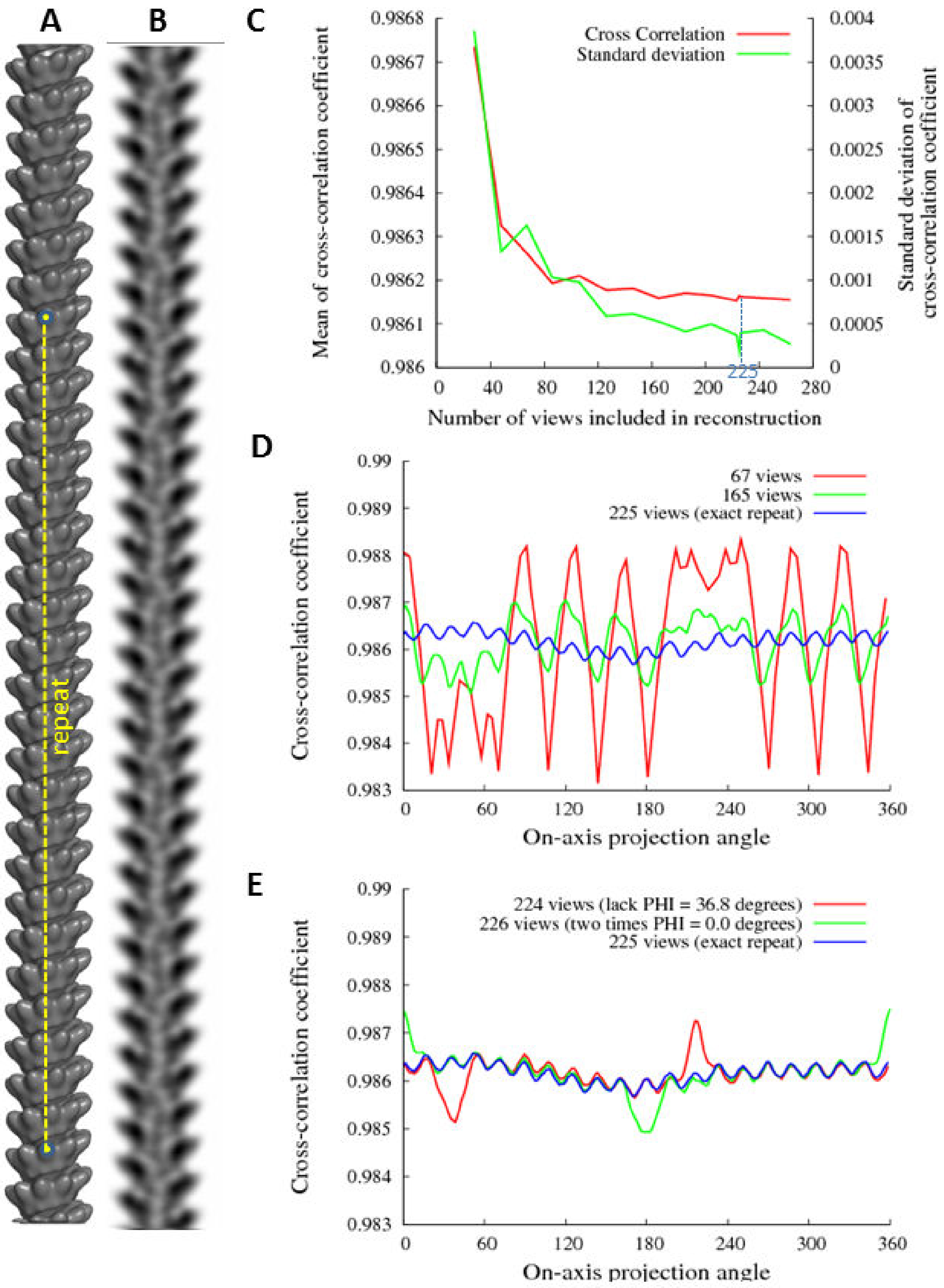


Figure 3.3 : Case study on RSV nucleocapsid : effects of number of segments and angular sampling on CC and ACC

Admittedly, the differences in correlation presented in this example might seem to be insignificant, and thus not worth an examination. However, this test case is on the contrary very revealing, because the 3D object of study was created artificially, and is therefore perfectly symmetric and ideally centered, which can never be a prerequisite if analyzing the real data. Moreover, it happens that due to the helical symmetry of this test object, the angular space tends to be correctly filled. Indeed, the angular rotation between subunits is 36.8 °, so that after one turn there is a shift of angular views of 8° in regard to previous turn => after 5 full turns, the lack of view between 0° and 36.8°, 36.8° and 73.6° etc.. is almost regularly filled (with 8, 16, 24 and 32 degrees views for the empty space between 0 and 36.8 °, and so on). On the whole, the example of the RSV nucleocapsid presented in **Figure 3.3** not only allows to globally illustrate our method, but also to better understand the following results on a variety of different test cases that will be presented later. We will see that the effects in term of CC variations can actually be much stronger (up to 15-20%), due in particular to shorter initial images and/or to more unfavorable symmetry.

The **panel 3C** shows the decrease in correlation between the segmented images and reprojections of reconstructed model as a function of the number of views used. At the first glance, this result can look surprising, because we are all used to an improvement of the correlation when more and more images are included in the reconstruction. However, if one considers a "reconstruction" built from one image only (say with the on-axis angle = 0°), we would expect to have 100% correlation with the original image when reprojecting this reconstruction in the appropriate direction, simply due to the conservation of information during the steps of back-projection and reprojection. When adding a second image to the reconstruction according to the symmetry (with an assigned on-axis angle of 36.8° in this example), the fact that this angle is not 0, 90, 180 or 270 degrees implies a necessity for an interpolation. In fact, the interpolation needed to include this image into the reconstruction will slightly deteriorate not only the reprojection in the direction of this image, but also in the direction of the first one. The same holds true as more and more images are added, which might be a reason for such an observation even despite the true symmetry being imposed.

We can also note that the interpolation needed to add this second image cannot be performed in an optimal way because it is not possible to benefit from information contained in neighboring planes in the 3D Fourier space (we will have only zeros, except at the common line between the transforms of the first and the second image; therefore the effects of such a "bad" interpolation might be greater).

When a wrong symmetry is imposed, then the negative effect of the number of images becomes even more pronounced, since the influence of the addition of each new image, although affecting only the area around one single line in reciprocal space of each of the other images placed, will be inconsistent. Altogether, these considerations on the influence of the image number on CC enable to apprehend the profile of ACC as a function of the symmetry imposed, i.e. for a fixed pitch as a function of the number of subunits per turn. They imply that the observed decrease of ACC with the increasing imposed number of subunits per turn is normal and does not necessarily mean a movement away from the true symmetry value.

Effects of angular view on the cross-correlation and more on interpolation

Our first tests of the method, either when imposing a true or a false symmetry, frequently showed differences of up to 15% in correlation between the segmented images and the corresponding reprojections of the reconstructed volume, depending on the segmented image. The first and the last segmented images had much lower CC with reprojections, and visually, one could observe that there was a small shift between these images and corresponding reprojections (difficult to show here by a figure), and that these reprojections had a "blurred" aspect, comparing to the other ones. A way to understand this, is to imagine how the view angles are filled during the segmentation of the input image. Let us take a very simple case of segmentation, say of a projection of a helix with 3.8 subunits per turn ($\Delta\varphi \approx$ 94.7°), from which we extract 12 views. The **Figure 3.4** shows the assignment of the view angles for each of the 12 extracted segments, numbered 1 to 12. The distinctive characteristics of the segments extracted from the middle of the projection (images 5, 6, 7 and 8) is that they are placed closer between two images in term of angular views. This means that, when filling the Fourier space for reconstruction, the needed interpolations can be done much better than for the first (1,2,3,4) and last four (9, 10, 11, 12) images. Here again, these effects and their amplitude are symmetry-dependent.

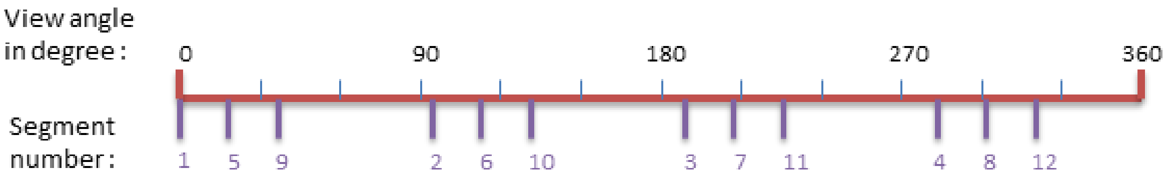


Figure 3.4 : View angle assignment of 12 segments extracted from a projection of an helix with 3.8 subunits per turn

The **figure 3.3D**, shows the correlation as a function of the view angle for several given numbers of images used for a reconstruction from the projection of the artificially made RSV model. Again, stronger variations were observed in less favorable cases, but with globally the same characteristics. We can see that the more images are used, the less variations in CC are observed. This can be explained by a better filling of view angles, closer to the "ideal" profile where the number of views equals the number of subunits (225) in the whole repeat. The same can also be appreciated in **figure 3.3C**, where the standard deviation of cross-correlation between segmented images and reprojections is plotted in green.

Effects of views at 180° and implications for particular symmetries

In **figure 3.3E**, we see what happens if either one less or one more than the optimal number of 225 views, are used for reconstruction. The lack of the 36.8° on-axis view causes a decrease of CC between segmented images which assigned view angle is close to this value and the reprojections corresponding to these angles. This is in accordance with the remarks made above about the effects of filling of the Fourier space: images with assigned angles near 36.8° will lack information for a proper interpolation comparing to the other views. Inversely, the effect of including two times the 0° view causes an increase of CC for this view and its neighborhood. Interestingly, we observe in both cases inverse effects on the CC of images with assigned on-axis angles at ≈180° away from these views (that is, at almost the same plane in Fourier space). Thus, the contribution of a view at 180° has a negative effect on CC, even when the true symmetry is imposed, which can be due to interpolations effects and/or to a non-perfect centering (that would cause a shift of the 180° view in respect to the 0° view).

These 180° effects are interesting, since when we wish to test symmetries that have an even (or nearly even) number of subunits per turn, we will only have views that are 180° apart each other. If it is the true symmetry, then we will eventually observe a weaker increase of ACC comparing to false symmetries imposed, both due to these effects and to a poor angular sampling. If it is not the true symmetry, then the opposite views will affect each other strongly, in an incoherent manner, and thus a strong decrease of ACC for these symmetries is expected, more than for other false symmetries. For an odd (or nearly odd) number of subunits per turn, there is one view that is repeated at each turn, and thus a weaker negative effect is expected.

Intermediate conclusions

In conclusion, these preliminary remarks are important to keep in mind in order to correctly analyze the CC profiles that will be shown in the next parts: Whatever the true symmetry is, the individual and mean correlations that will be measured between segmented images and corresponding reprojections will be affected by the described effects of the number of images, the angular sampling and the filling of Fourier space in a symmetry-dependent manner, and in an “input image dependent manner” (e.g as far as the available length is concerned). This will interfere with an “ideal” profile, depending only on the difference between the imposed symmetries and the true one.

Applications to different helical structures

Different types of data plots to facilitate analysis

A typical 3D plot of ACC as a function of the imposed helical parameters is depicted in **figure 3.5A** (3D view) and **3.5B** (top view of the 3D plot). We can observe a global variation of ACC according to the imposed pitch, centered on the nearest value to the true pitch (here 23 Å). This bell behavior of the ACC as a function of the pitch, observed for all projections **of one-start helices** tested up to now (both theoretical or experimental projections), is quasi-independent of the number of subunits per turn imposed (**figure 3.5C**).

The plot of ACC as a function of the number of subunits per turn shows more “high-frequency” oscillations, and also shows similarities within different pitch imposed (**figure 3.5D**. In this case however, one can observe more variability in sharpness, relative height and precise position of the peaks (in some cases even more than in presented one). Due to the behavior of ACC as a function of pitch, one can reduce the calculation time by first refining the pitch, and then imposing the pitch found and refine the number of subunits per turn. Moreover, the value of pitch can be often relatively easily obtained by other ab initio methods (measure on PS, direct measure on real images…). Thus, for the sake of clarity and simplicity, in the following part we will consider only the 2D plot of ACC as a function of the subunit number, while the pitch will be kept fixed.

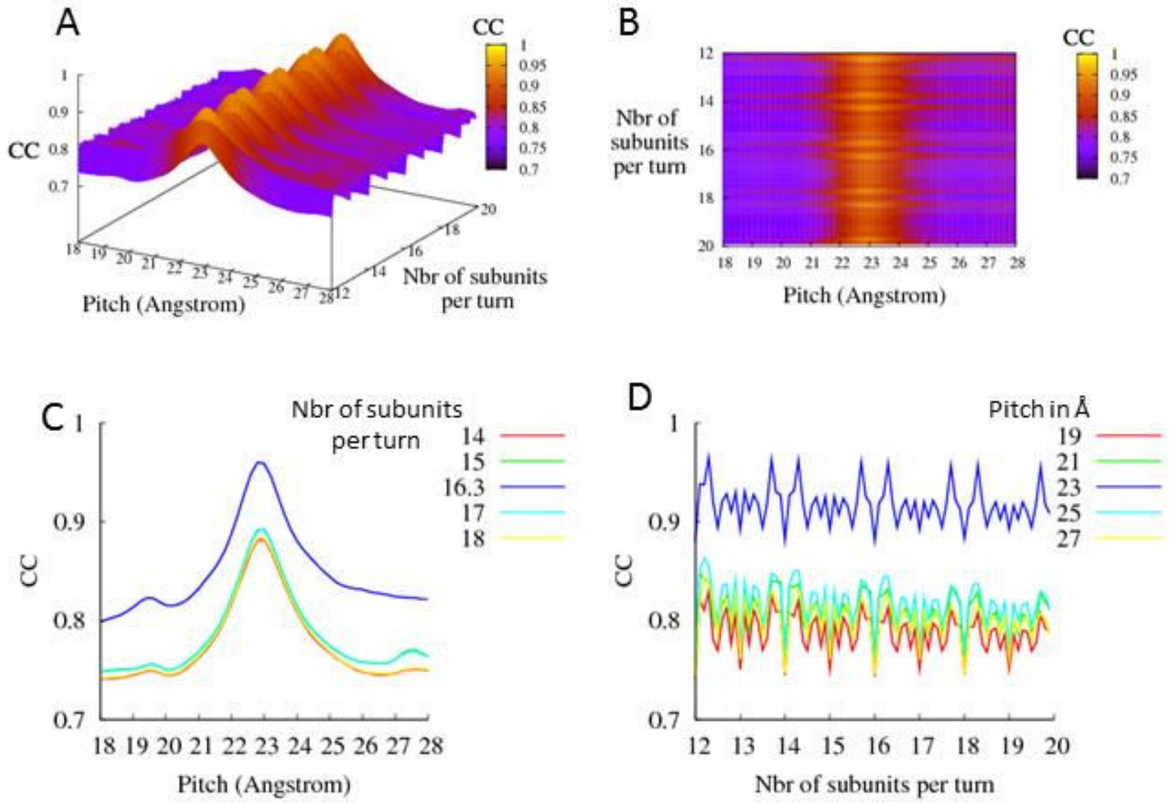
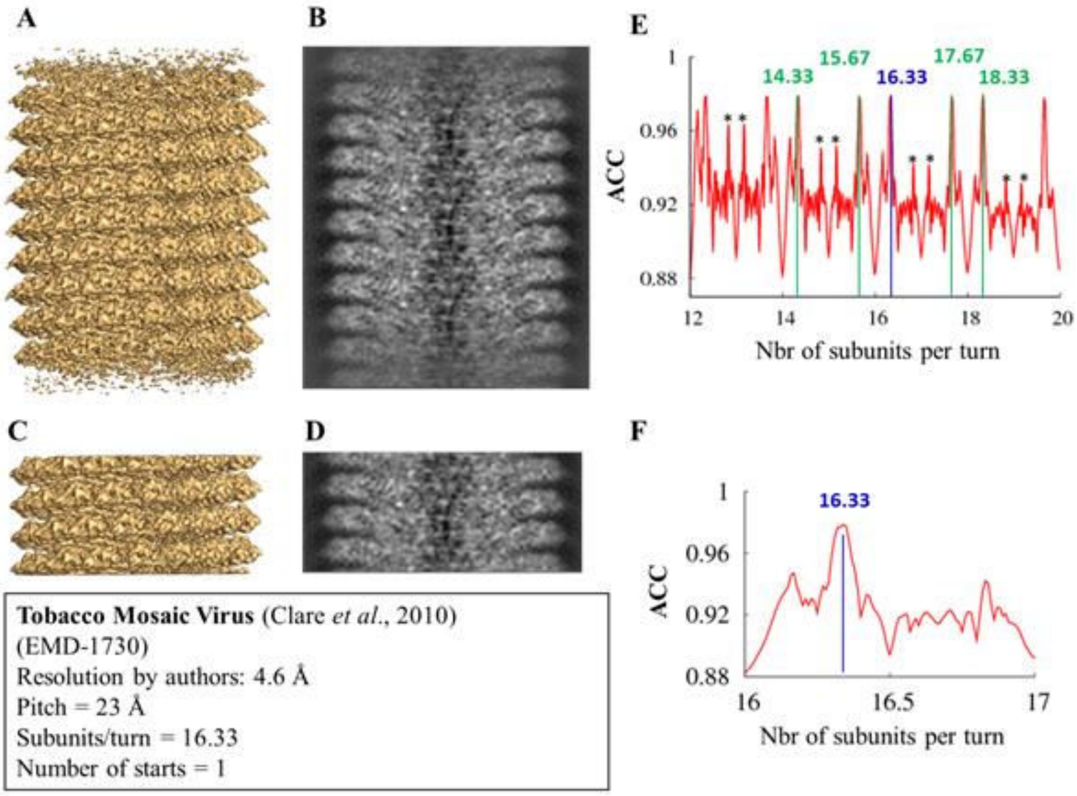


Figure 3.5 : Type of plots of results of symmetry determination on 2D projection

A : 3D plot of ACC as a function of pitch and number of subunits per turn
B : Top view of the 3D plot of ACC as a function of pitch and number of subunits per turn
C : Plot of ACC as a function of pitch for different number of subunits imposed
D : Plot of ACC as a function of number of subunits for different pitch imposed

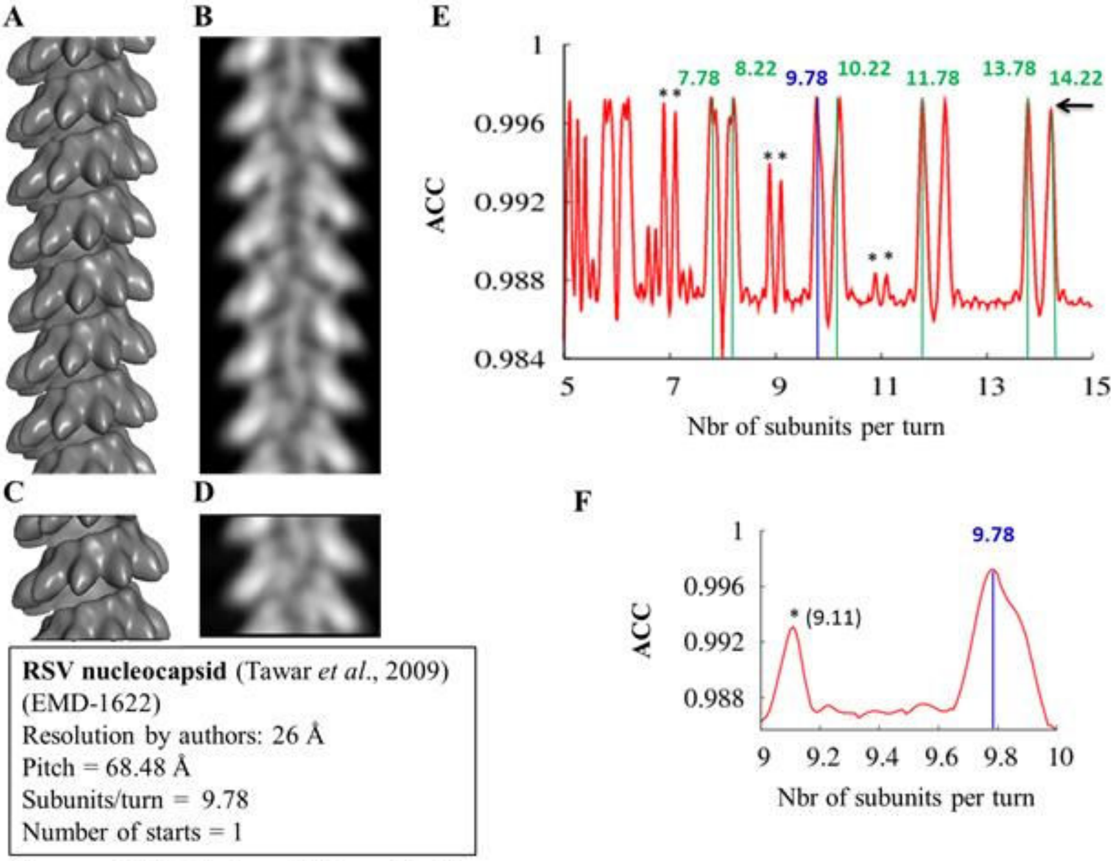
Description of the test methodology

Before applying the method to experimental data, we tested it for several "ideal" test cases, i.e. which are perfectly symmetric and for which we know precisely the symmetry parameters (**figures 3.6 to 3.10**, legends in the following text). Thus, several helical EM maps, at various resolution, were downloaded from the EMDB. Some representative results will be shown here : the structure of TMV (**Figure 3.6A** ; (Clare and Orlova 2010)), of RSV nucleocapsid (**Figure 3.7A** ; (Tawar et al. 2009)), of Flagellar Hook (**Figure 3.8A** ; (Fujii, Kato, and Namba 2009a)), of the Bacteriophage fd (**Figure 3.9A** ; (Y. a Wang et al. 2006)) and of the **Nitrilase** (**Figure 3.10A** ; (Thuku et al. 2007)). From these maps, a projection was calculated with in-plane and out-of-plane angles set to 0 (**Figure 3.6 to 3.10, panels B**). Features of the maps like their resolution and the symmetry parameters are indicated in the text box included in the figures. The method described in **Figure 3.1** was applied to each of these projections. The length of the segments to cut out from the projection and to include in the reconstruction was chosen in order for each segment to contain at least 2 turns of helix, except for the 5-start helix (**Figure 3.9**), for which the pitch of the one-start helices was too high for such segmentation. The pitch information was fixed to the known value, and the number of subunits to test was incremented every 0.01 subunits per turn. The choice of this step is such that the very exact value of number of subunit per turn in the structures will not be tested, as the precision of the real values is usually higher than 0.01, but values reasonably close will be included in the test. The range of symmetry tested here is huge to better appreciate the global behaviors of the CC profiles but this is not necessarily what one would always do in a real case, as we often have some knowledge to restrict the search range. The plot of ACC as a function of the number of subunits per turn imposed is shown in **Figure 3.6 to 3.10, panels E.** The **panels F** on these figures provide a magnified view of this plot around the true value.



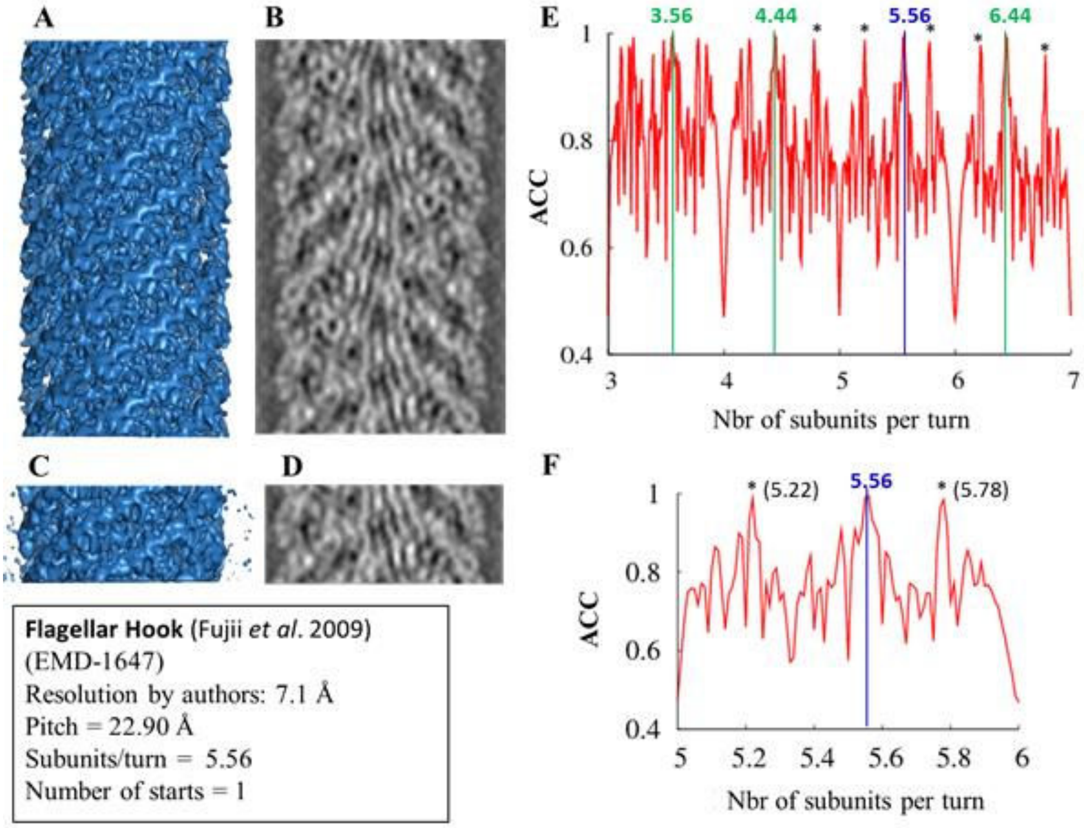
Tobacco Mosaic Virus (Clare *et al*., 2010)
(EMD-1730)
Resolution by authors: 4.6 Å
Pitch = 23 Å
Subunits/turn = 16.33
Number of starts = 1

Figure 3.6 : Test on TMV projection



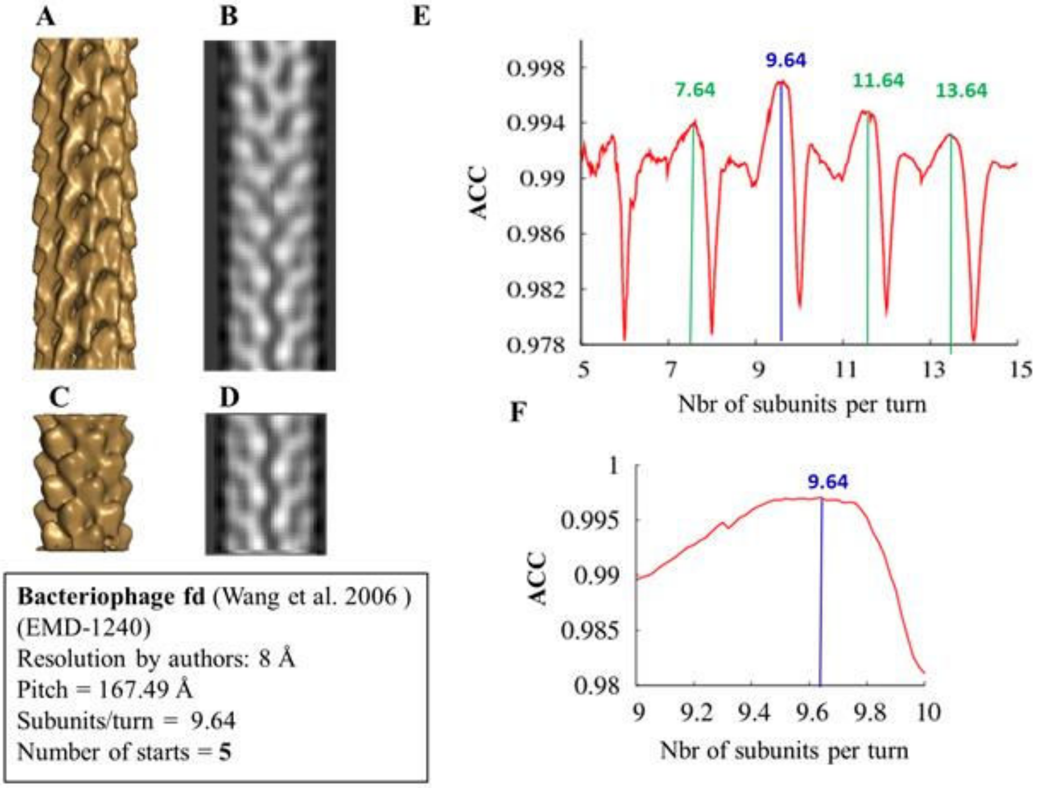
RSV nucleocapsid (Tawar *et al*., 2009)
(EMD-1622)
Resolution by authors: 26 Å
Pitch = 68.48 Å
Subunits/turn = 9.78
Number of starts = 1

Figure 3.7 : Test on RSV projection



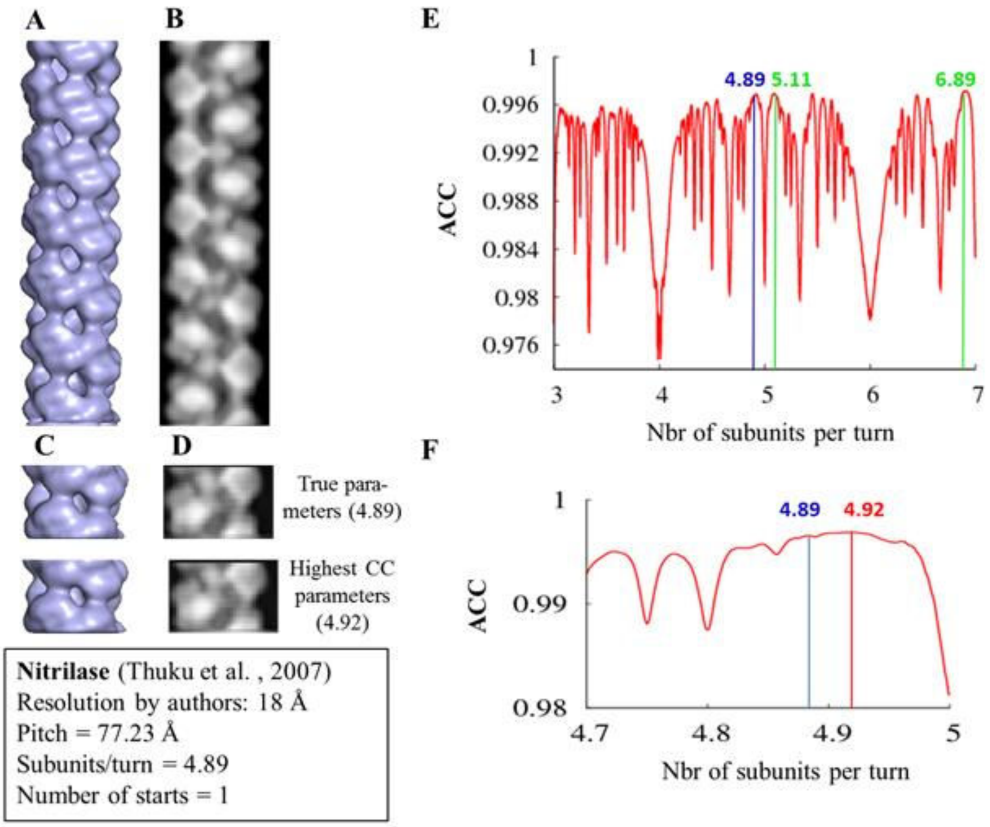
Flagellar Hook (Fujii *et al*. 2009)
(EMD-1647)
Resolution by authors: 7.1 Å
Pitch = 22.90 Å
Subunits/turn = 5.56
Number of starts = 1

Figure 3.8 : Test on Flagellar Hook projection



Bacteriophage fd (Wang et al. 2006)
(EMD-1240)
Resolution by authors: 8 Å
Pitch = 167.49 Å
Subunits/turn = 9.64
Number of starts = **5**

Figure 3.9 : Test on Bacteriophage fd projection



Nitrilase (Thuku et al. , 2007)
Resolution by authors: 18 Å
Pitch = 77.23 Å
Subunits/turn = 4.89
Number of starts = 1

Figure 3.10 : Test on Nitrilase projection

Results on the test cases -Overview

What first appears when looking at the ACC profiles on a wide range of tested symmetries (**Figures 3.6 to 3.10, panels E**), is that on the one hand, the true symmetry, indicated by a vertical blue line, does not appear as a unique solution in term of ACC peak, except for the case of the several-start helix (**Figure 3.9E**). On the other hand, a peak of ACC corresponding to the real helical parameters is observed in most of the cases in a close vicinity to the true solution (**Panels F**). However, it is not always the case (**Figure 3.10D**), and tests with other structures downloaded from the EMDB were not systematically successful. Visually, the reconstructed volume corresponding to the peak of CC (**Panels C**), and its reprojection (**Panels D**) are very similar to the original structure (**Panels A**) and projection (**Panels B**), despite the low number of images included in the reconstructions.

As expected, the plots show a strong decrease of ACC for integer number of subunits per turn, especially for even number of subunits per turn. This is more likely due to the mentioned problems of angular sampling and “180° effects” than to departure from the true symmetry. We see that even of the case of the helix which symmetry is the closest (among our examples) to an integer number of subunits per turn (Nitrilase, **Figure 3.10** ;4.89 subunits per turn) , the ACC drops very rapidly to reach one minimum for the nearest integer value.

In order to get a better feeling of the difficulties in finding the true helical symmetry and of the non-uniqueness of the solution, a visual comparison of projections of 3D structures with different symmetries corresponding to different peaks of CC appears informative (see **Figure 3.11** for an illustration of these ambiguities using RSV test). The volumes obtained by applying different ambiguous symmetries to the input projection are very different in term of the shape of the subunits (**Figure 3.11B**), but their projections (**Figure 3.11A**), as well as the PS of the projections (here not shown) are similar, even if one considers a projection of a high-resolution case such as TMV. The current example on RSV illustrates that some peaks of correlation can be easily discarded as false solution when the 3D structure has no biological sense (e.g : on **Figure 3.11**, the 8.89 symmetry). Furthermore, at this point, any prior knowledge on the subunit, like its global shape, the number of domains or contacts between subunits, would help to decide which solutions are more likely to be true.

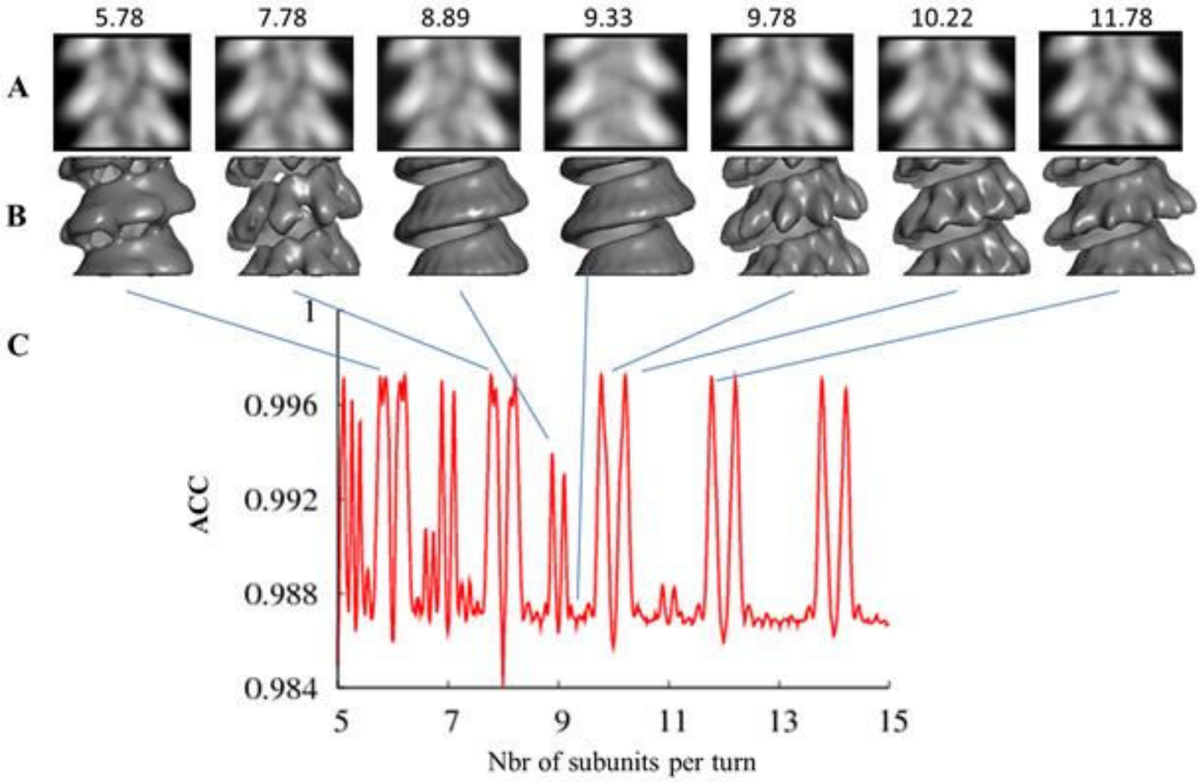


Figure 3.11 : The ambiguous determination of symmetry parameters on RSV image
From the RSV projection shown in Figure 7, an ACC plot was calculated (**C**), and reconstructions corresponding to various symmetry parameters are shown in **B**, and corresponding reprojections in **A**

The overview on the results of these tests on ideal cases gives us an idea of what can be expected from the method. On the one hand, an unambiguous symmetry determination seems impossible except may be of several-start helices (which we plan to analyse more extensively in the near future). On the other hand, even if spurious solutions appear unavoidable, the true solution also appears in most of the cases as a maximum of ACC. Thus, the uncertainty in the symmetry determination is reduced to a restricted number of possibilities given by the ACC maxima. Since one normally possesses additional information on the symmetry and/or on the subunit assembly, our method allows to restrict the uncertainty even more and leads to the true symmetry determination.

Finally, we have to keep in mind that the present examples were done with perfect images, in the absence of any noise, so that real cases with noise could introduce further ambiguities. Having gained all the presented knowledge from artificial test cases, we are now ready to analyze some real examples and compare them with the artificial ones.

Results on the real cases –Overview

Three different cases will be shown here, covering both negative staining and cryo images as well as one-start and several-start helices. In all the cases, the input image used to test the method is a class-average, resulting either from an ab initio classification using MSA, or from projection matching, as will be specified on the figures. The **Figure 3.12** shows the results on a non-digested MeVNC class-average (**Figure 3.12A**) obtained by MSA. The 3D ACC plot (**Figure 3.12B**) shows the bell shape behavior of ACC as a function of the imposed pitch, independent from the number of subunits per turn imposed, and enable to determine the precise pitch for this class-average. The ACC profile according to the number of subunits per turn shows two major peaks around 11 and 13 subunits per turn. Despite the fact that the absolute values are slightly greater for 11 than for 13, the experience we have acquired on artificial data (as far as possible effects of number of images and angular sampling are concerned) inspire caution and teach us to take such small differences of ACC with care, and to rely on the profiles of ACC rather than on the exact values. Indeed, the values of ACC for peaks corresponding to lower number of subunits (7 and 9, not shown on the figure) are even greater than for 11 and 13, and ACC tends to decrease with an increase of the number of subunits per turn. This tendency is in accordance with a decrease of ACC as a function of the number of images included in reconstruction (more subunits per turn => smaller axial rise). Another important point in this example is that, despite the fact that integer numbers of subunits per turn are usually strongly disadvantaged by the method, they appear here as ACC peaks. The method was repeated over several different class-averages, and the results always showed an increase of ACC for –or close to (less than 0.1 subunits per turn away)- odd number of subunits per turn. Together with the final reconstructions that were obtained (~odd number subunits per turn), this suggests that the method is able, at least in the present case, to overcome the problems posed by such symmetry.

The **Figure 3.12C** shows the reconstructions and their reprojections corresponding to the two peaks at 11 and 13 subunits per turn. The hand of reconstructions is arbitrary imposed, as it is not determinable from the projection. As expected, although the shape and the assembly of the subunits are fully different, the projections of both reconstructions are similar. Without any other information, we would probably have to consider both solutions as possible. In this case however, we can take an advantage of the presence of ring-shaped top views (which could be short segments of helices) on the electron micrographs. The corresponding class-averages are represented at the **Figure 3.12D**. The major “symmetry” is the 13-fold, even if pseudo-rings with less or more subunits are also present, whereas the 11-fold symmetry was almost never observed. Together with the comparison of the 11 and 13 subunits/turn reconstructions with the 12 Å cryo-EM reconstruction of the digested nucleocapsid (Schoehn et al. 2004), this provides a strong evidence for the 13 subunits/turn symmetry as being the true one.

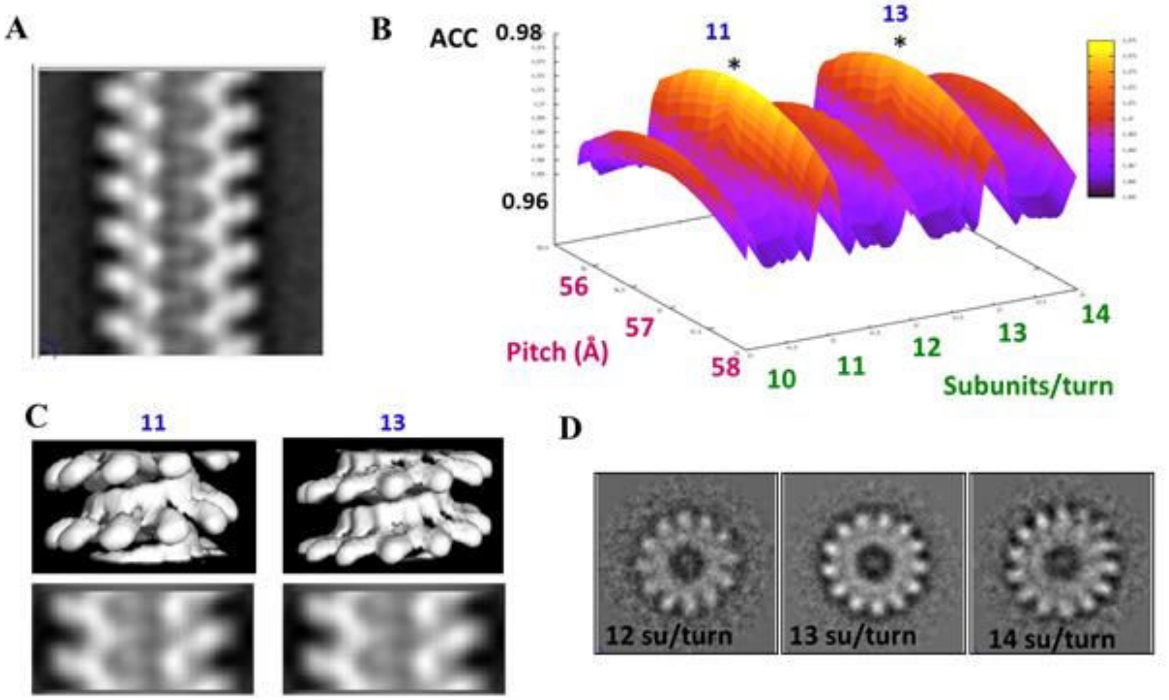


Figure 3.12 : Estimation of symmetry parameters on non-digested MeVNC image class-average

A : Class-average (MSA) of negative-stained non-digested MeVNC images

B : ACC plot from class-average shown in A

C : Reconstructed volumes corresponding to the peaks indicated by asterisks in B

D : Class-averages of top-views of ring-shaped MevNC

The **Figure 3.13** shows the results on a cryoEM class-average (obtained by projection matching) of the digested measles nucleocapsid (**Figure 3.13A**) taken from the work of (Schoehn et al. 2004). The **Figure 3.13B** shows the corresponding 3D ACC profile, viewed perpendicular to the number of subunits per turn axis. Again, several solutions (11.67, 12.33

and 13.67 subunits per turn) gave a comparable ACC, and a survey of a wider symmetry range shows supplementary ambiguous solutions. Two of these solutions, the ones that make sense regarding the diameter of the helix and comparison with the top views, are shown at **Figure 3.13C**. Interestingly, almost exactly these two symmetries were found by (Schoehn et al. 2004) using the IHRSR method (**Figure 3.13D**), the 12.35 solution leading to better resolution and being more consistent with the metal shadowing experiment.

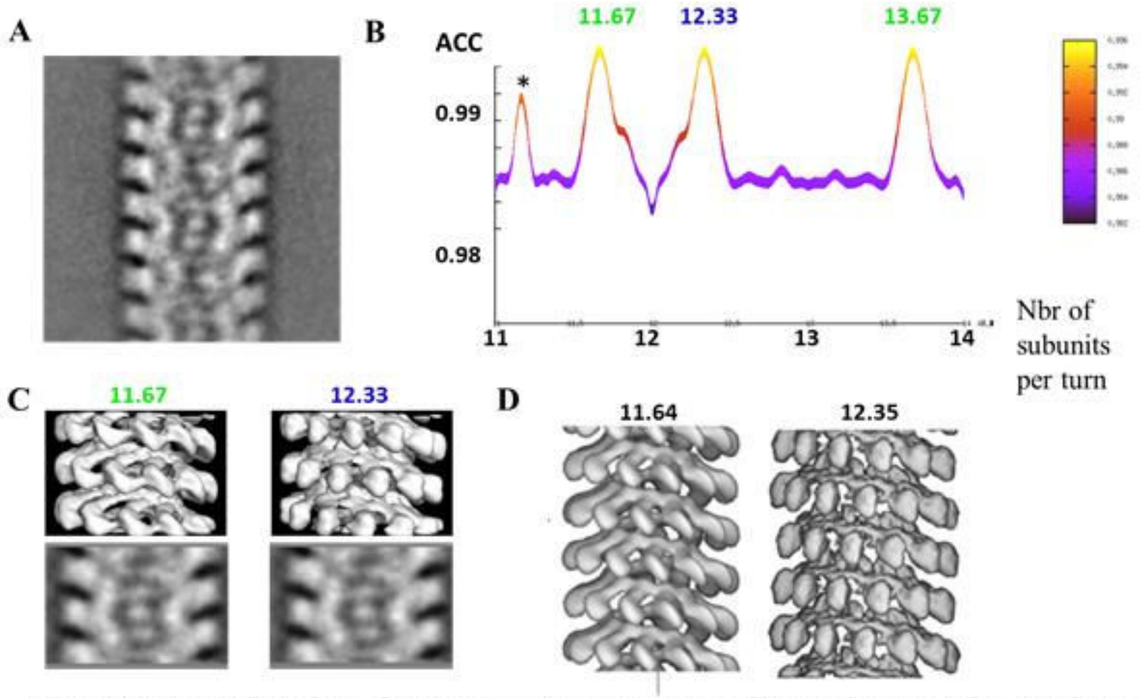


Figure 3.13 : Estimation of symmetry parameters on digested MeVNC image class-average
A : Class-average of cryo-EM digested MeVNC images (Schoehn et al., 2004 ; projection matching)
B : ACC plot from class-average shown in A
C : Reconstructed volumes corresponding to the peaks indicated by asterisks in B
D : Two of the reconstructions obtained by (Schoehn et al., 2004)

When applied to **our** class-averages of our images of the same sample by negative staining, we found the same two solutions (or very close to), the only difference between these tests being a poorer quality of the reconstructed volume from negative stain class-averages. Similarly as for the non-digested sample, the additional information that we dispose suggests that the ~12.3 subunits/turn solution was the true one, which enabled us to use it as a starting point for refinement of the structure of digested MeVNC by negative staining EM.

As a last example of the application of the method in an experimental case, we chose a class-average, obtained by MSA, of images of TspO in a helical form (V. M. Korkhov et al. 2010) (**Figure 3.14A**). This example is revealing because the final reconstruction obtained by these authors was a several-start helix, and the artificial several-start projection tested above showed a particular behavior of the ACC (**Figure 3.7**). However, to begin with, we will not consider the projection as arising from a several-start helix – first, we are not supposed to know it, and second, there is no apparent reason that would hinder the determination of the helical parameters of the one-start helix that is repeated C-fold symmetrically in the whole assembly.

Considering that we have absolutely no prior information on this sample, we tested a huge range of helical parameters, with a pitch varying from 30 to 1400 Å every 5 Å and a number of subunits per turn from 5 to 40 every 0.2 Å, making in total ~50000 different symmetries tested. Despite the high complexity of this test, the ACC profile as a function of pitch and number of subunits per turn (**Figure 3.14B**) is relatively simple and shows a very different profile comparing with what we usually observe. There is no bell shape behavior of the ACC according to the pitch independent to the number of subunits per turn imposed, and in contrary we observe a dependency of ACC **both** on the pitch and on the number of subunits per turn. To understand the nature of this dependency, we can look at this same ACC plot, but as a function of angular rotation between subunits **Δφ** and of axial rise **Δz** (**Figure 3.14C**). The ACC profile then clearly shows that the multiples of **Δz** ≈ 32 Å give globally higher ACC, whatever **Δφ** is imposed. This distance corresponds to the spacing between the apparent horizontal striations on the class-average, suggesting that the one-start helix that we try to detect has one subunit per stack, implying that the whole assembly consists of several one-start helices related by rotational symmetry (otherwise one subunit would correspond to an entire stack ring, that is not compatible with the known MW of the protein -18kDa-). To further refine the parameters of the one-start helix, we then did a search on both **Δφ** (on a wide range) and **Δz** (around 32 Å) but with a finer step of search (0.01 Å on **Δz** and 0.01 ° on **Δφ**). Figure 14D shows a slice through the generated 3D plot for the **Δz** giving the highest correlations (**Δz** = 32.32 Å). Angular rotation of **Δφ** = 9.49 ° gives unambiguously the highest correlation. These values are very close to the one corresponding to the final reconstruction published in (V. M. Korkhov et al. 2010) ; **Δz** = 32.67 Å and **Δφ** = 9.53. Based on this data alone, we are unable to discriminate if these small differences are due to imprecision of our method, to a lack of very precise information on the class-average, or to a real difference between the helical parameters corresponding to this particular class-average and the one corresponding to the whole set of images included in their final reconstruction.

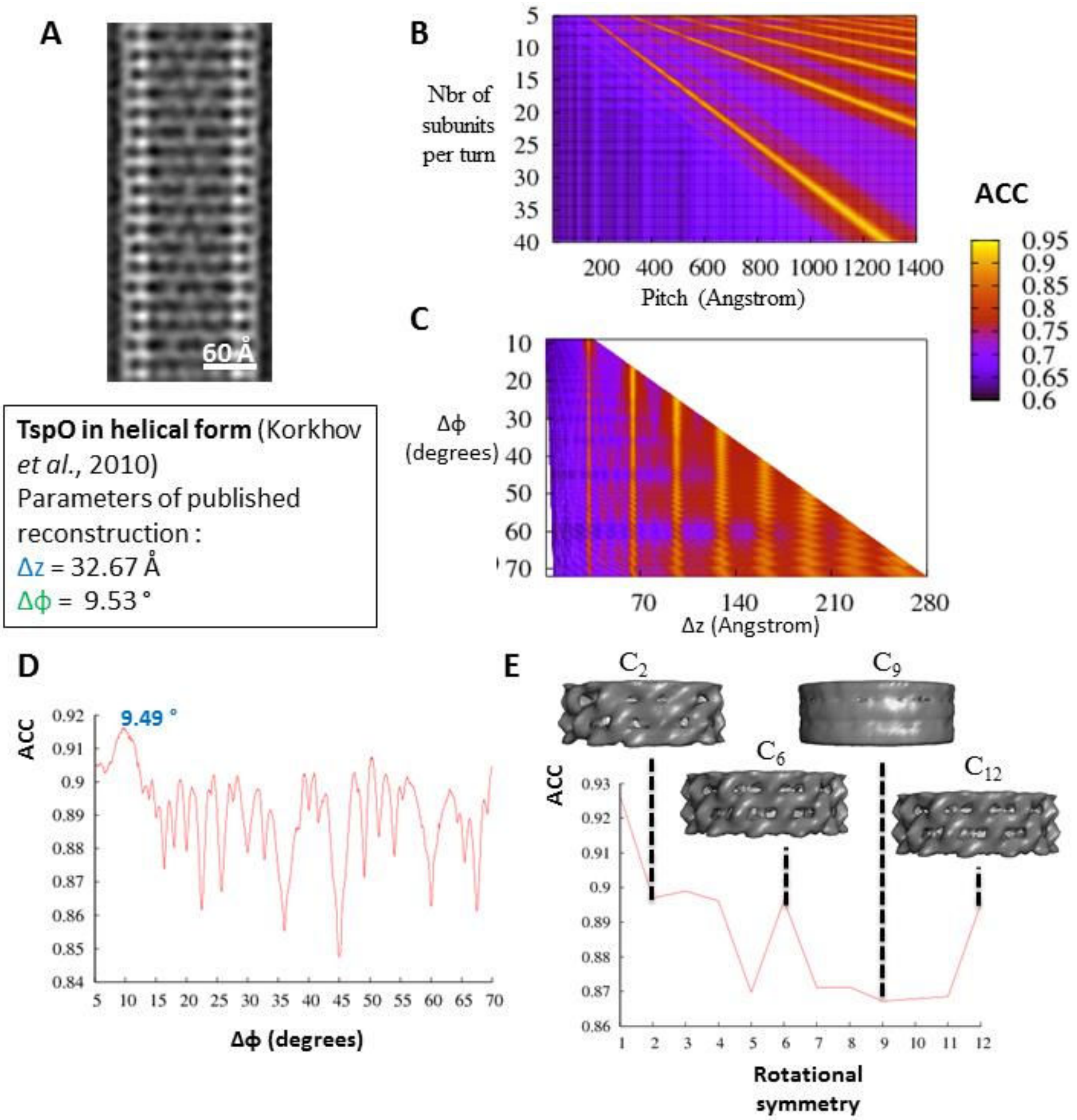


Figure 3.14 : Estimation of symmetry parameters on TspO class-average

A class-average of TspO in helical form (A) obtained by MSA (Korkhov *et al.*, 2010) was used as input to create the ACC plot, expressed a pitch and number of subunits per turn (B) or as the rise per subunit Δz and rotation between subunits Δφ (C). In (D), we can see the ACC plot as a function of rotation between subunits Δφ for a fixed Δz of 32.32 Å, as determined by refining from the ACC plot shown in (C). The panel E shows the effect of imposing various rotational symmetry (C_1 to C_{12}) to the reconstruction with Δz =32.32 Å and Δφ=9.49° , on the ACC plot and on the 3D structure for some chosen rotational symmetries.

As mentioned, the helical parameters are incompatible with only one one-start helix, so we then tried to impose the presence of several one-start helices with the determined parameters. The **Figure 3.14D** shows the ACC plot as a function of the number of starts (rotational symmetry). The highest correlation remains for a 1-start helix, but this could be also due to the lower number of images in this reconstruction. For more starts, two levels of correlation can be observed: a higher one, for 2,3,4,6 and 12 starts, and a lower one for 5,7,8,9,10,11 (and also for the tested values higher than 12 –not shown here-) number of starts. Visually, any imposition of a number of starts other than a divisor of 12 tends to distort the shape of the subunit present in the one-start helix and to smooth the reconstructed volume. On the contrary, imposition of a number of starts that is a divisor of 12 reinforce the subunit density without distortion, with an optimal reinforcement for the 12-start helix. This is a strong indication that the number of start in the whole assembly should be 12, which is indeed the true number in the reconstruction of (V. M. Korkhov et al. 2010).

Together with the fact that among the tests on ideal projection, the only one giving the true solution without ambiguities was the one on a several-start helix (**Figure 3.9**), the present example gives us indications that the method can be particularly successful when the number of starts is higher than one.

In these three tests on experimental cases, we could obtain the helical parameters, either using some additional information (for Measles) or no information at all (TspO). In any case, the time needed to perform the tests is very short. The most time-consuming part is the further analysis of the results, which can be reasonably done within a day. The method is thus, as it is, a valuable alternative for ab initio symmetry determination. Several ideas to improve it and to try to overcome some of its intrinsic limitations, as well as ideas of how to optimally use it in a real case, for a new project, will be discussed in the “conclusion and perpective” part later.

But first we will come back to the ambiguities that were encountered when trying to determine from one projection the helical parameters with our method, and try to answer several questions. How are the ambiguous symmetry parameters distributed? What, in the description of such helices, is ambiguous? How come that volumes apparently that different can indeed share identical projections?

The non- randomness of the ACC pattern

In this part and below, we will focus on the description of the ambiguities for the one-start helices. Firstly because the symmetry determination for the several-start helices was not really ambiguous, and secondly, although it is true that other symmetries than the real one also gave peaks of ACC in a similar fashion as for the one-start helices (**Figure 3.9E**, green vertical lines), the particularities encountered when dealing with the several-start helices, as well as the complications in explaining at the same time the results for the one-start and the several-starts, argue for a separate treatment of these subjects. However, when possible, a link between what we will observe and deduce for the one-start helices and the particularities of the results for several-start helices will be made.

Interestingly, the peaks of ACC corresponding to the ambiguous symmetries are not "randomly" distributed as a function of the symmetry tested. If one takes a closer look at the variation pattern of CC rather than on exact values, it becomes apparent that there is a pseudo-periodicity of the ACC pattern of 2 subunits per turn, and there are pseudo axial symmetries around axes defined by integer numbers of subunit per turn. Through these two operators, the most important peaks are related to the peak corresponding to the true symmetry. Some of these related ACC peaks are indicated by green vertical lines on the ACC plots on **Figures 3.6 to 3.10** (**panels E**) while the true symmetry is indicated by a vertical blue line. For example, considering the test on TMV (**Figure 3.6E**), there is a peak for the true solution at 16.33 subunits per turn, but also at 18.33, 14.33, 12.33... (corresponding to the periodicity of 2 subunits per turn), and there are peaks at 15.67, 17.67, 13.67 subunits per turn (corresponds to the pseudo axial symmetry around integer values). The peaks related to the true one by only few operators show values of ACC almost equal to the one for the true symmetry (variations in ACC of less than 10^{-4}), while for the peaks requiring more operators, the ACC values tend to decrease. This decrease is not always obvious from the figures presented, as the symmetry range which we show here is often too restricted, however we can see the beginning of this decrease on the example of RSV nucleocapsid (**Figure 3.7E**, black arrow). Other important peaks of ACC, not directly related to the true symmetry by the two operators and showing usually lower values of ACC, are also present on the ACC plots. Some of these are indicated by asterisks in **Figures 3.6, 3.7 and 3.8**. A careful empirical analysis of the symmetries corresponding to these peaks reveals however that they are not totally unrelated to the true symmetry. For example, if we inspect the zoom on the ACC plot for the flagella hook (**Figure 3.8F**), we see two peaks (asterisks) for 5.22 and 5.78 subunits per turn, whereas the true symmetry is 5.56. Using the "axial symmetry around integer values" operator, we can go from the true value 5.56 to 6.44, and then from 6.44 to 3.22. These symmetries manifest many common features, the 3.22 symmetry including all the Fourier coefficient of the 6.44 symmetry. Seen in real space, we can imagine the symmetry of 3.22 number of subunits per turn helix being exactly similar to the one of the 6.44 helix, just by considering two adjacent subunits of the 6.44 helix as a single one in the 3.22 helix. Then, from this value of 3.22, we can go to 5.22 by using the same operators as before. Likewise, we can find the 5.78 peak, by following this path: 5.56 => 6.44 => 8.44 => 4.22 => 6.22 => 5.78. This reasoning might be regarded as far-fetched, but all the symmetries being on this path show high values of ACC, and the same kind of relationships is found in all the other examples we could test so far. For RSV (**Figure 3.7F**), the 9.11 peak can thus be obtained by following the path: 9.78 => 10.22 => 5.11 => 7.11 => 9.11 (note: the 8.89 symmetry, directly related to 9.11, also shows a peak of ACC – see **Figure 3.11C**). Similarly as before, the more operations are needed to reach such a symmetry, the lower the corresponding ACC is. On **Figure 3.7E**, for example, we see the peaks indicated by asterisks disappearing rapidly, as the number of subunits increases. The existence of the relationships, involving the N to N/2 number of subunits per turn transition, in addition to the two other operators, also explains why more high ACC peaks are found towards lower number of subunits per turn (this is particularly observable on **Figures 3.7E and 3.8E**). Indeed, each of the symmetry parameters closely related to the true ones will give a related peak at N/2, which will in turn give other closely related peaks by applying the two previously described operators.

Given these empirical observations, we asked ourselves if these ambiguities in symmetry determination and the relationship between the true symmetry and the related ambiguous ones, arise from the method we designed or if they are inherent to the structures (or to their projection) themselves. A search over the literature shows evidence that the second option is true. The most remarkable example we could find comes from (Edward H Egelman 2010), where a test on IHRSR procedure convergence was done using as input images 1000 projections of a TMV structure (Sachse et al. 2007), and starting with different initial helical parameters. The algorithm used by these authors led to several different stable solutions, and the ones shown in the paper (16.33, 15.67, 14.33 and 12.33 subunits per turn) were an **exact**

subset of solutions we found by our method. Many other examples in the literature show the same kind of ambiguities (some are shown in **Figure 3.15**), although surprisingly nobody has, to our knowledge, described these particular relationships between ambiguous symmetries.

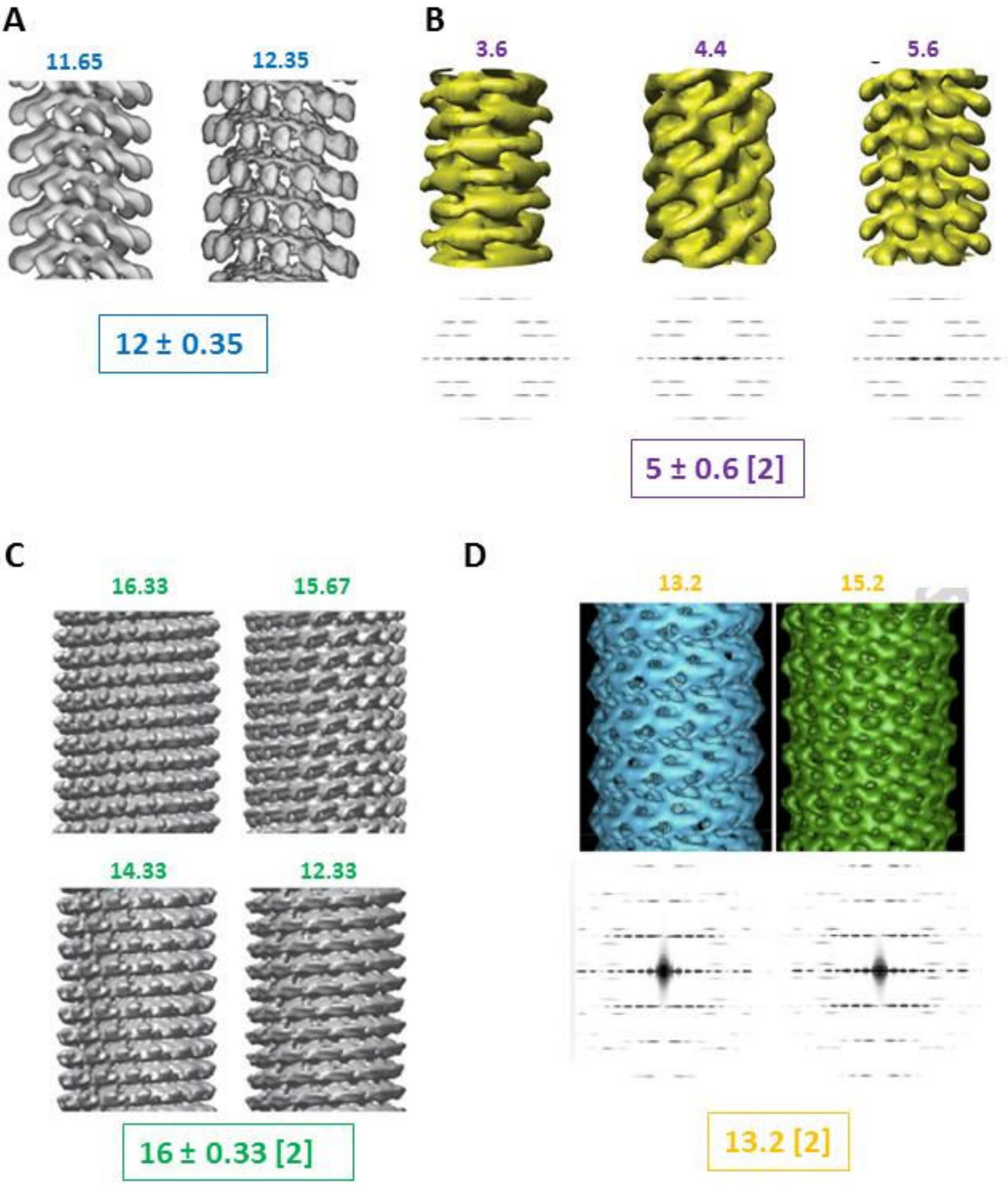


Figure 3.15 : Examples of ambiguities in symmetry parameters determination in the literature

A look over the ambiguities for symmetry determination in the literature show that the empirical rules between ambiguous symmetry parameters that we have determined are indeed verified. Ambiguous reconstructions from Schoehn *et al.*, 2004 (**A**) ; Wang *et al.*, 2006 (**B**) ; Egelman, 2010 (**C**) and Chen *et al.*, 2004 (**D**)(Figure adapted from Egelman, 2006) are shown, as well as corresponding power spectra in (**B**) and (**D**). The number of subunits per turn are indicated above the reconstruction and the relevant correspondence to the rule between ambiguous parameter is shown in boxes

Inspite of the fact that our proposed method does indeed suffer from the same ambiguity problems than for example the IHRSR method, one clear advantage of our method is that there is no need for running many cycles of PM, choosing initial symmetry parameters and symmetry search parameters, etc.. in order to be able to find after hours and hours of calculation and waiting for parameters stabilization that several 3D models can correspond to the images, since with our method we get all the possible ambiguous solutions in one time, from one image.

It would be tempting to find a way to make a distinction between the ACC peak corresponding to the true symmetry parameters and the false-positive ones, directly from our ACC plots, without any additional information. For example, the peaks indicated by asterisks on **Figures 3.6, 3.7 and 3.8** rapidly disappear when looking at their relatives for higher number of subunits per turn, and thus this could be an indication that they are far from the real parameters, but without rigorous mathematical explanations for the behavior of ACC and for the ambiguities, any such empirical choices should be done with great care. Nevertheless, some ideas about how to try to overcome these ambiguity problems will be given in the "conclusions and perspectives" part later below.

First however, some original considerations on both reciprocal and real-space description of such "ambiguous" helices, can still help to understand the sources of the observed ambiguities.

Towards a better understanding of symmetry ambiguities

Some mathematical relationships between ambiguous helices

Let us consider three **one-start** helices A, B and C of the same pitch P composed of one atom per asymmetric unit, and with the number of subunits per turn N(A) , N(B) and N(C) following these rules

$$\begin{cases} N(A) = N - f \\ N(B) = N + f \\ N(C) = N - f + 2 \end{cases} \qquad (1)$$

Where N is an integer and f a non-integer with 0 < f <= 0.5 (f is the fractional part of the number of subunits per turn). The helices A and B represent the cases where we observe what we called an "axial symmetry" of ACC around integer number of subunits per turn, and the helices A and C the observed "periodicity" of ACC of 2 subunits per turn. To simplify, we suppose that these helices have an exact repeat after a distance c in z direction.

The number of turns t to reach this distance for each helix is related to the number of asymmetric units u in the structure by :

$$\begin{cases} t(A) * N(A) = u(A) \\ t(B) * N(B) = u(B) \\ t(C) * N(C) = u(C) \end{cases} \qquad (2)$$

Using (1) :

$$\begin{cases} t(A) * N - t(A) * f = u(A) \\ t(B) * N + t(B) * f = u(B) \\ t(C) * (N + 2) - \ t(C) * f = u(C) \end{cases} \qquad (3)$$

As N and u are integers, these equations imply that $t(A) = t(B) = t(C) = t$ with the product $t * f$ being an integer. As the helices A, B and C have the same pitch P, their repeat $c = t * P$ will thus occur at the same axial distance. The **Figure 3.16** shows the superposition of the helix nets of such helices by taking as example 4.8 (red circles, for helix A, called helix A1), 5.2 (blue circles, for B, called B1) and 6.8 (green circles, for C, called C1) atoms per turn, with a pitch of 20 Å.

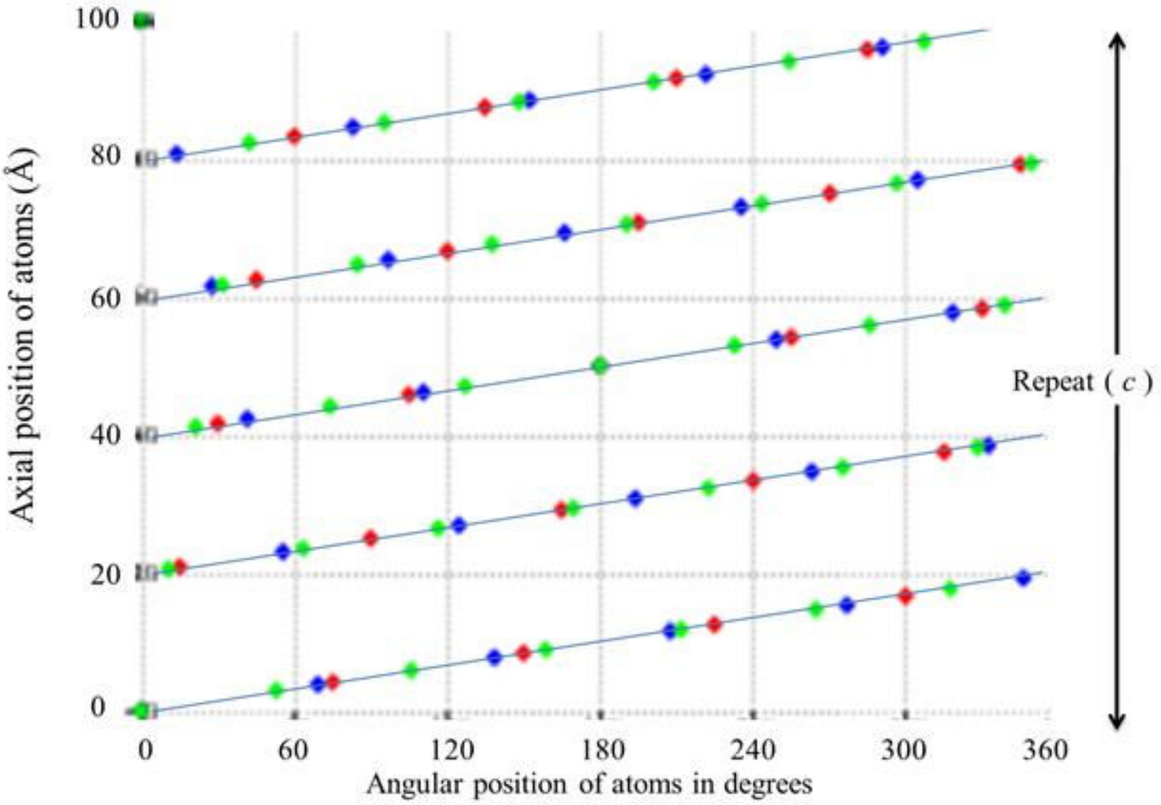


Figure 3.16 : The helix nets of helices A1, B1 and C1
The helix nets of one-start helices containing 4.8 (red dots, helix A1), 5.2 (blue dots, helix B1) or 6.8 (green dots, helix C1) atoms per turn shows that these helices reach a same exact repeat c=100 Å after 5 turns (blue line)

If one uses the same notation as before, the values for N, f, t , c and u for these helices would then be:

$$\begin{cases} N = 5 \\ f = 0.2 \\ t = 5 \\ c = 500\ \text{Å} \\ u(A1) = 24\,; u(B1) = 26\,; u(C1) = 34 \end{cases}$$

According to (Cochran et al., 1952, eq (4)), the transform of a discontinuous helix is finite only in planes at height

$$\zeta = \frac{n}{P} + \frac{m}{p} \qquad (4)$$

with n and m which can assume every integral value, positive or negative, and p being the axial rise per subunits in Angstrom (p is equal to the pitch P over the number of subunits per turn). Solutions for n are the orders of the Bessel functions occurring at this height. As we assumed that there is an exact repeat $= t * P = u * p$, we can multiply equation (4) by $t * P$ and obtain :

$$t * P * \zeta = t * n + u * m = l \qquad (5)$$

where l is an integer that represents the l-th layer line (this formula is the so-called selection rule (Klug, Crick, and Wyckoff 1958)). The transform is thus confined to layers for which

$$\zeta = \frac{n}{P} + \frac{m}{p} = \frac{l}{c} \text{Å}^{-1} \qquad (6)$$

Having the same repeat, the helices A, B and C will have finite transform at the **same heights**, but since they have different axial rise p (and thus also different u), the order of Bessel orders on each layer line must be different. Let us look, as an example, into the solution (n,m) of the equation (6) for values of l=0 , l=1 ,and l=4 for the illustrative helices A1, B1 and C1 (the selection rule for Bessel functions can also be geometrically expressed on the n,l plots, shown in **Figure 3.17**, limited to |n|<=u and |l|<=u).

$$l = 0 \begin{cases} n = \cdots, -48, -24, 0, +24, +48, \ldots \\ m = \cdots, +10\,, +5\,, 0, -5, -10, \ldots \end{cases}$$

$$l = 1 \begin{cases} n = \cdots, -43, -19, 5, +29, +53, \ldots \\ m = \cdots, +9\,, +4\,, -1, -6, -11, \ldots \end{cases}$$

$$l = 4 \begin{cases} n = \cdots, -52, -28, -4, +20, +44, \ldots \\ m = \cdots, +11\,, +6\,, +1, -4, -9, \ldots \end{cases}$$

Helix A1 (4.8 subunits per turn) u = 24

$$l = 0 \begin{cases} n = \cdots, -52, -26, 0, +26, +52, \ldots \\ m = \cdots, +10\,, +5\,, 0, -5, -10, \ldots \end{cases}$$

Helix B1 (5.2 subunits per turn) u = 26

$$l = 1 \begin{cases} n = \cdots - 31, -5, +21, +47, +73, \ldots \\ m = \cdots, +6\,, +1\,, -4, -9, -14, \ldots \end{cases}$$

$$l = 4 \begin{cases} n = \cdots - 46, -20, +6, +32, +58, \ldots \\ m = \cdots, +9\,, +4\,, -1, -6, -11, \ldots \end{cases}$$

$$l = 0 \begin{cases} n = \cdots, -68, -34, 0, +34, +68, \dots \\ m = \cdots, +10\,, +5\,, 0, -5, -10, \dots \end{cases}$$

Helix C1 (6.8 subunits per turn) u = 34

$$l = 1 \begin{cases} n = \cdots - 61, -27, 7, +41, +75, \dots \\ m = \cdots, +9\,, +4\,, -1, -6, -11, \dots \end{cases}$$

$$l = 4 \begin{cases} n = \cdots - 40, -6, +28, +62, +96, \dots \\ m = \cdots, +6\,, +1\,, -4, -9, -14, \dots \end{cases}$$

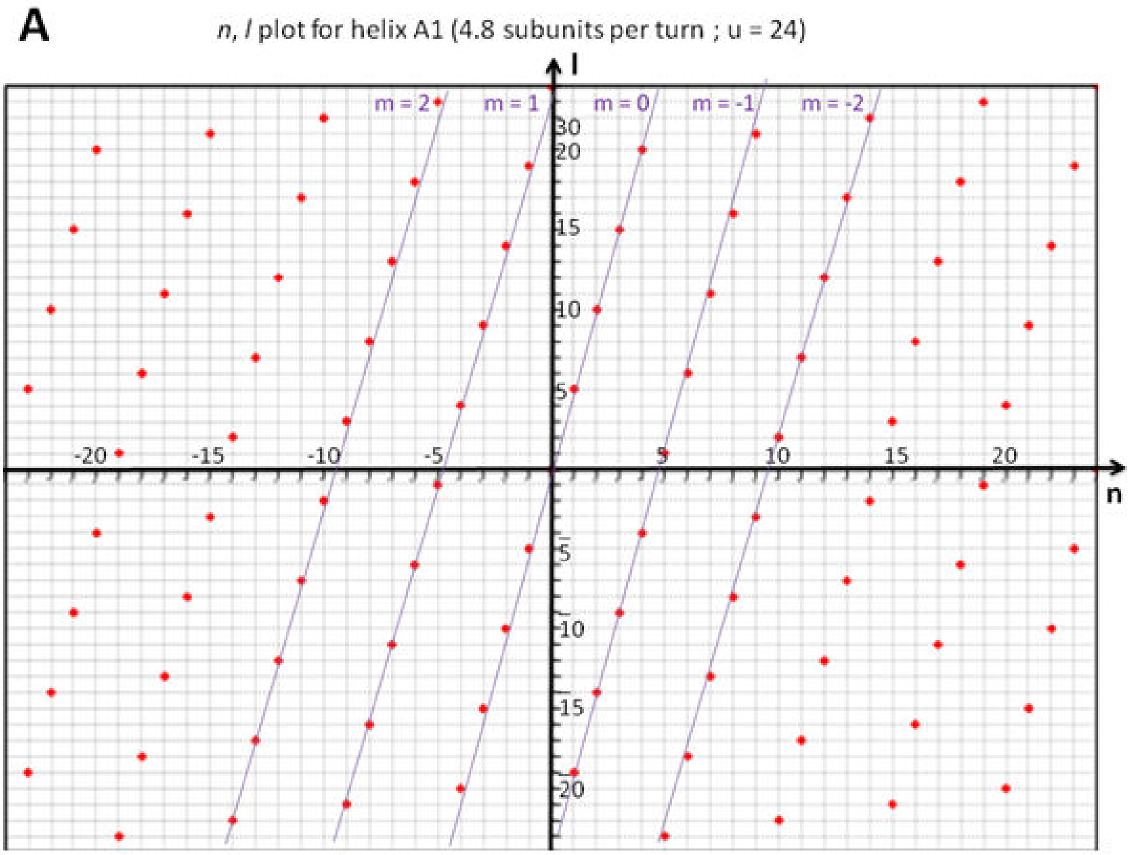


Figure 3.17 (First part): Visual rendering of the selection rule for helices A1, B1 and C1

The n,l plots for helices A1 (**A**), B1 (**B**) and C1 (**C**) are shown for values of n and l limited to |*n*|<=*u* and |*l*|<=*u*. The purple lines and numbers indicate the values of *m*, pointing out the relationships between *m* values for helices A1, B1, and C1. A geometrical interpretation of *m* can be found in Klug *et al.*, 1958.

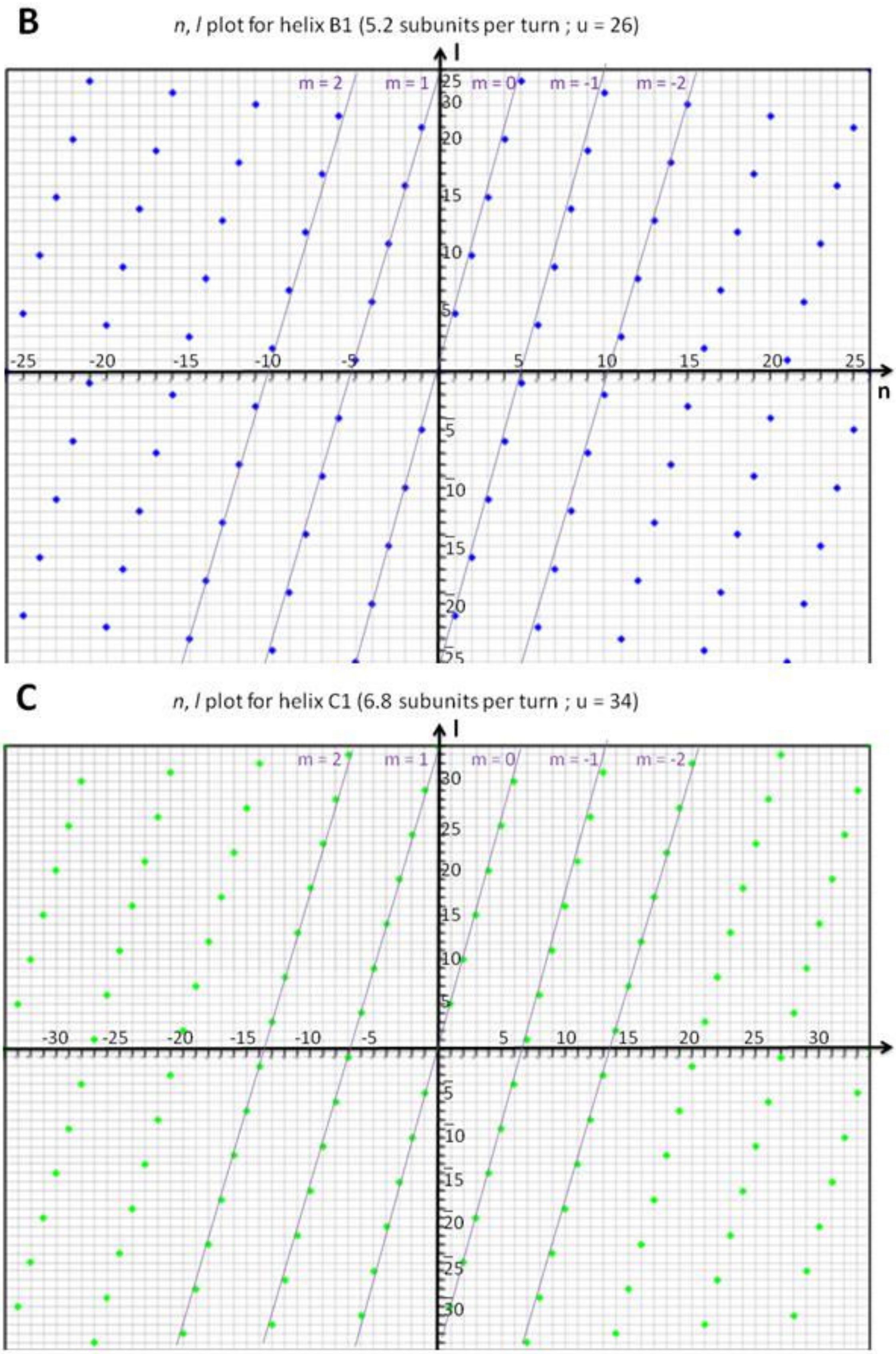


Figure 3.17 (second part)

On a particular layer line, due to the selection rule, the difference between successive values of n is always equal to u, and the difference between successive values of m is always equal to t.

We now can try to deduce relationships for values of n for these different helices. First we can write

$$\begin{cases} \zeta(A) = \frac{n(A)}{P} + \frac{m(A)}{p(A)} \\ \zeta(B) = \frac{n(B)}{P} + \frac{m(B)}{p(B)} \\ \zeta(C) = \frac{n(C)}{P} + \frac{m(C)}{p(C)} \end{cases} \qquad (7)$$

And by replacing the axial rise p by the pitch over the number of subunits per turn as expressed in (X):

$$\begin{cases} \zeta(A) = \frac{n(A)+m(A)N}{P} - \frac{m(A)f}{P} \\ \zeta(B) = \frac{n(B)+m(B)N}{P} + \frac{m(B)f}{P} \\ \zeta(C) = \frac{n(C)+m(C)N}{P} - \frac{m(C)f}{P} + \frac{2m(C)}{P} \end{cases} \qquad (8)$$

If we now look at layer line at same heights, so that

$$\begin{cases} \zeta(B) = \zeta(A) \\ \zeta(C) = \zeta(A) \end{cases} \qquad (9)$$

we can first deduce particular relationships of m values between helices A and B and helices A and C by multiplying left and right side of (9) by P and using the expression of ζ given in (8) :

$$\begin{cases} n(B) + m(B)N + m(B)f = n(A) + m(A)N - m(A)f \\ n(C) + m(C)N - m(C)f + 2m(C) = n(A) + m(A)N - m(A)f \end{cases} \qquad (10)$$

$$\begin{cases} n(B) + m(B)N - n(A) - m(A)N = -f(m(A) + m(B)) \\ n(C) + m(C)N + 2m(C) - n(A) - m(A)N = f\ (m(C) - m(A)) \end{cases} \qquad (11)$$

The left part of equations (11) being an integer, to be always true, we must have :

$$\begin{cases} m(B) = -m(A) \\ m(C) = m(A) \end{cases} \qquad (12)$$

These relationships for values of m can be visually appreciated on **Figure 3.17 (purple lines and numbers)**;so we can now deduce a relationship between the possible orders of Bessel function for helices A, B and C on **each layer line**, using equations (8), (9) and (12):

$$\begin{cases} n(B) = n(A) + 2Nm(A) \\ n(C) = n(A) - 2m(A) \end{cases} \qquad (13)$$

A particularity derived from these relationships is that for each layer line, the Bessel functions occurring for ambiguous helices will have the same parity (because they differ by a multiple of 2). A possible implication of this will be discussed below. If one assigns orders for Bessel function on the PS of helices A1, B1 and C1, we can see that these relationships are indeed verified (example on **Figure 3.18A**). In a more complex case than a helix composed of single atoms, these relationships are also verified, if one considers for example the illustration of the ambiguity in power spectra in ((Egelman 2010), Figure 6.8A-D) (some are shown in **Figure 3.18B**). The indexing of the power spectrum shows for a helix with 15.67 subunits per turn (equivalent to helix A in our example) values of n of 16 and 17 on layer lines 1 and 4, respectively, while for the 16.33 (≈helix B) values are $-16 = 16 - 2 * 16 * (+1)$ and $-15 = 17 - 2 * 16 * (+1)$. The non-verification of relationship for layer line 2 results from an erroneous indexing on the Figure (the -17 order should be a -15) in Egelman's paper.

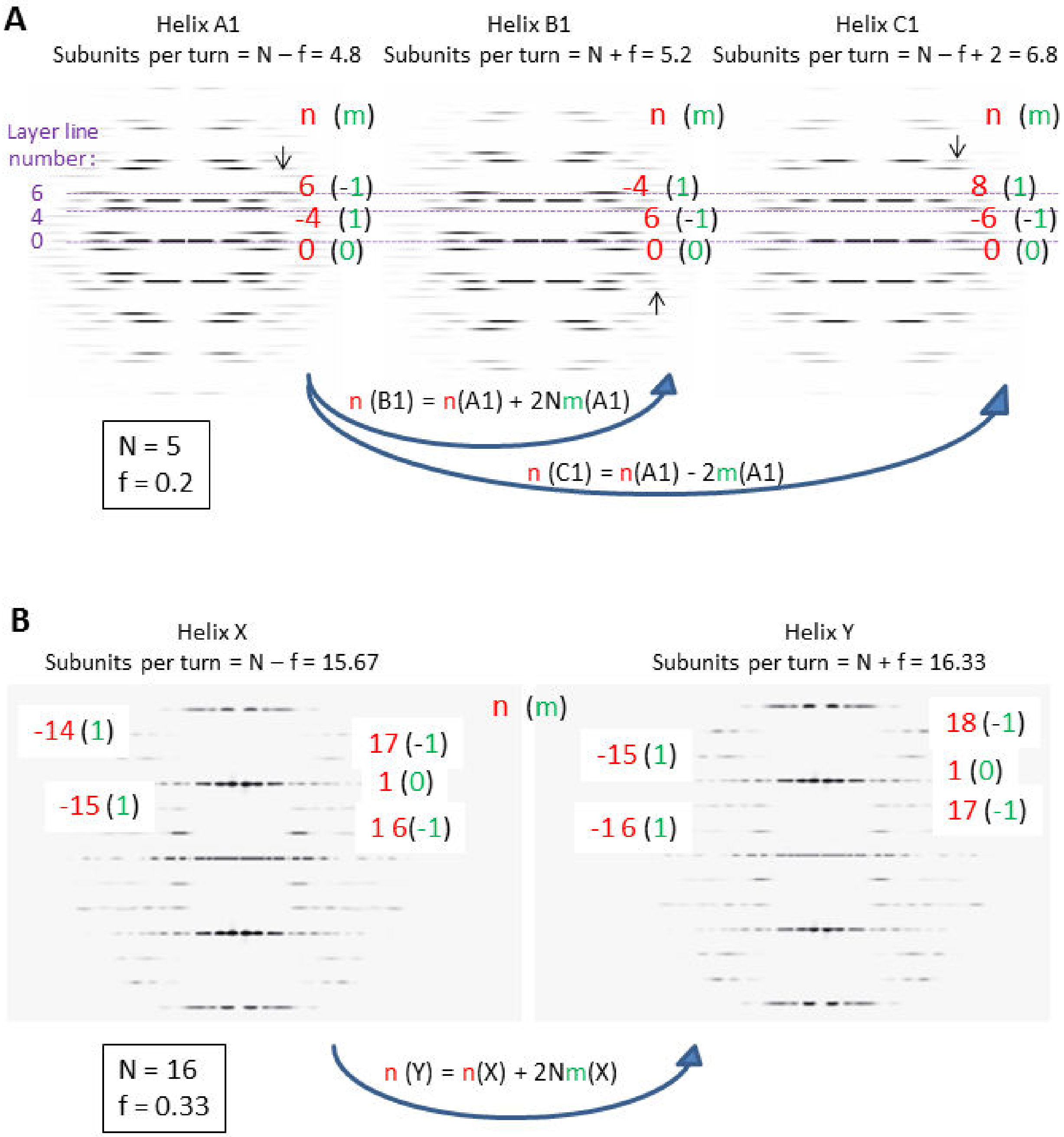


Figure 3.18 : An illustration of the deduced relationships between order of Bessel functions on power spectra of ambiguous helices

The indexing of the power spectra of helices composed of 4.8, 5.2 and 6.8 spheres per turn (**A**) and of the power spectra of helices with 15.67 and 16.33 subunits per turn (**B**) illustrate the relationships of the Bessel orders for helices of type A, B and C as described in the text (values for N and f are indicated in the boxes). The orders *n* of Bessel functions as well as m values are indicated. Black arrows indicate repulse of some Bessel functions.

The need for radial density redistribution

These relationships now pose a problem: how helices that exhibit different orders of Bessel function for each layer line can be ambiguous at some point? If one first consider the modulus of the FT, each layer line is filled with Bessel functions of first kind $Jn(2\pi Rr)$, with **R** being the radius in reciprocal space (distance from meridian) and **r** the radius of the helix in real space. As our ambiguous helices have maximum intensities in FT at the same reciprocal radius **R**, there must be a change in **r**. If we consider our case of helices A, B and C with atoms all placed at same radius, can we make the FT of B similar to FT of A, for example, by a single change in **r** ? We should then have the maxima of $J_n(A)(2\pi Rr(A))$ and $J_n(B)(2\pi Rr(B))$ occurring at the same reciprocal radius for each layer line, while satisfying the relation between n(A) and n(B) from Eq.(13). For example, taking the layer line 4 of helices A1 and B1, and considering their Bessel functions of lowest order, -4 and +6, respectively, if helix A1 has, say, a radius of 50 Angstrom, then, to match the position of the first maximum, helix B1 should have a radius of ~70 Angstrom (**Figure 3.19A**).

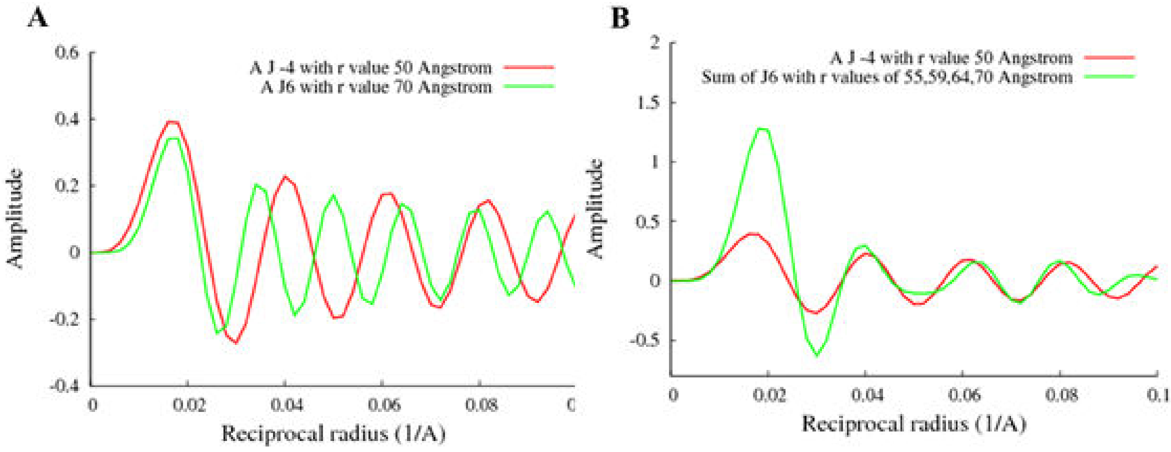


Figure 3.19 : The radial density redistribution can at least partly compensate for variations in Bessel functions orders between ambiguous helices

To simulate the 4th layer line of power spectra of helices of type A1 and B1 (see text), their corresponding Bessel functions of, respectively, orders -4 and +6 , are superposed. In **A**, we simulate the signal in the FT of an helix composed of one atom, where the radius of helix B1 is 70 Å whereas the radius of helix A1 is 50 Å, in order that the first maximum of the Bessel functions matches. In **B**, we show the sum of the contribution of 4 atoms for helix B1 placed at radius *r* in real space 55, 59, 64 and 70 Å . This combination was chosen among all possible combinations of 4 atoms placed at radius between 30 and 80 Å, every 1 Å , to fit at best the Bessel function of order -4 of helix A1.

But as we can see, it is not possible that the next maxima of the function occur at same radial distance. These “repulse” of Bessel functions can however be present in the transform (see **Figure 3.18**, red arrows), even at low resolution for Bessel functions of lower order. So we see that even when considering one simple layer line, there is no ambiguity in the description of helices of type A, B and C. Furthermore, if one wishes to respect the equalities of Bessel functions for each layer line, this would be even more impossible. So why do we still observe these ambiguities? Indeed, the reason lies in the initial settings of the problem itself - we are starting on the 2D projection, without any a priori knowledge about the 3D volume, in particularly concerning the number of atoms in each asymmetric unit, and their radial and angular position.

Taking the example of a helix composed of one atom per asymmetric unit and 4.8 units per turn (**Figure 3.20A**), we thus can see that when an ambiguous symmetry is imposed on its projection (**Figure 3.20B**), for example here 5.2 subunits per turn (**Figure 3.20D**), the reconstructed volume contains a more complex distribution of density than when imposing the original 4.8 subunits per turn (**Figure 3.20C**), with contributions at various radii and at various angular position (**Figure 3.20D**). Due to these contributions, the transform of the helix include the summation of many terms of the form of $J_n(A)(2\pi Rr(A))$. It is not mathematically shown here that such summation could make possible to produce an identical, or very similar, signal in Fourier space, but it is reasonable to admit that the more terms are included, and the more freedom is given for placing the densities, the closer the transform of a such complex helix can be to the one of the original simple helix. The **Figure 3.19B** show how the summation of only 4 Bessel functions of the 6th order with various **r** values (to simulate the 4th layer line of the transform of an a helix with 5.2 subunits per turn, composed of more than one atom in the asymmetric unit) can approximate the position for the maxima of a Bessel function of 4th order (4th layer line of an helix with 4.8 subunits per turn with one atom per unit).

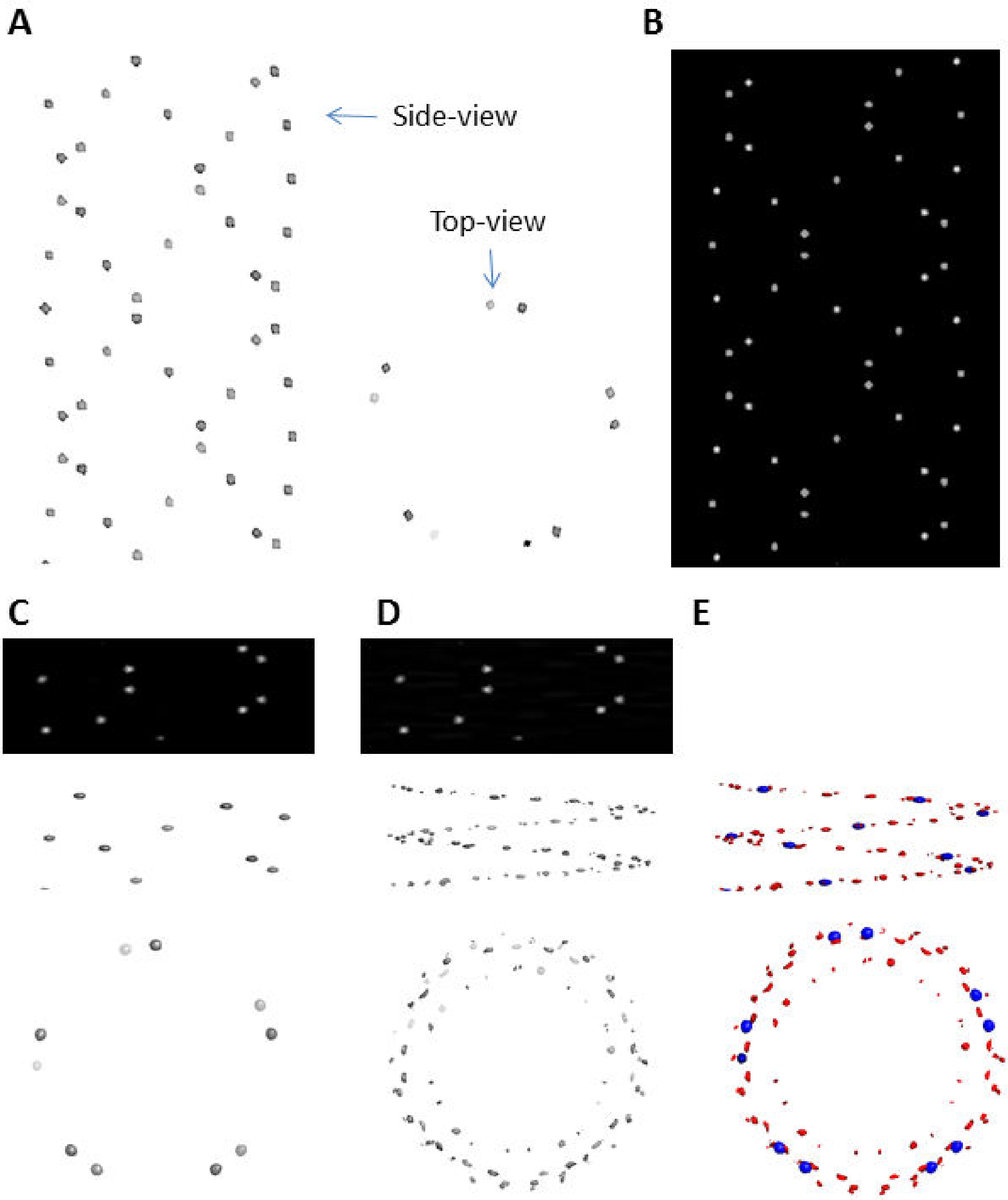


Figure 3.20 : Density redistribution arising from reconstructing a 5.2 subunits per turn helix from a projection of a 4.8 subunits per turn helix

A 3D helix composed of 4.8 small spheres per turn (**A**) was used to calculate a projection (**B**), from which was reconstructed 2 volumes, either imposing a symmetry of 4.8 subunits per turn (**C**) or the ambiguous parameter 5.2 subunits per turn (**D**). In **C** and **D**, one projection of the reconstructed volumes is shown in the upper part : this reprojection is to compare with the bottom part of the projection shown in **B**. The middle and bottom part of **C** and **D** show a side-view and a top-view of the reconstructed volumes, respectively. Panel **E** show a superposition of the reconstructed volumes with 4.8 (blue) and 5.2 (red) subunits per turn.

Of course, our real cases are not as simple as the helices composed of one atom per asymmetric unit, as the true initial volume itself has a complex density distribution, thus with many different contributions to the transform. But the basics are the same : an ambiguous helix to the true one, sharing identical (or very similar) projection, thus identical (or very similar) section in FT, must show a highly different distribution of densities, with in particular the radius of maximum density being adjusted. On **Figure 3.21**, some ambiguous volumes appearing when analyzing the projection of the RSV nucleocapsid structure are shown at a high threshold of visualization, thus revealing the highest density regions that are placed at different radii for the 3 reconstructions. These effects of radius gives thus a possibility for eventually reducing the number of ambiguous volumes for a given projection, using restrictions on the radius of reconstruction, both on inner and outer radius. It is interesting to note that such restrictions can be used in the Egelman's IHRSR method when imposing the symmetry on the reconstruction, thus probably helping to reduce the number of stable solutions. However, our experience of these restrictions using IHRSR show that some ambiguities still persists. A simple test to gain understanding of the radius-restriction related possible decrease of ambiguities would be to compare the ACC plots for different types of radius restrictions.

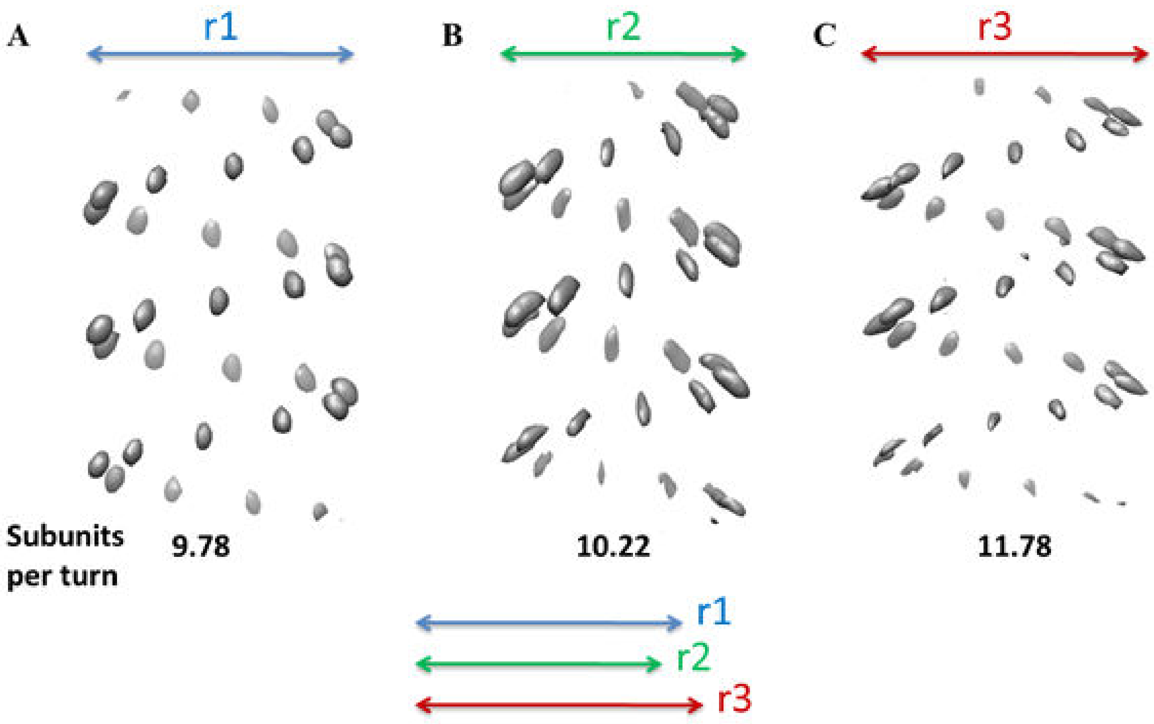


Figure 3.21 : The highest density regions of three ambiguous volumes constructed from RSV nucleocapsid projection show a radial density redistribution
Three volumes corresponding in peaks in ACC plot according to the number of subunits per turn imposed (from RSV nucleocapsid projection, see Figure 5) are displayed at high threshold of visualization. In comparison to the true reconstruction (**A**), the ambiguous reconstructions **B** and **C** show a different radial position of the highest density regions.

Comparison with experimental observations

Now, how these relationships and findings help us to explain our observation, and the ACC profiles obtained through the different tests ? First, as shown, the ambiguous helices, due to their symmetries, give raise to layer lines at same heights. The relationships of Bessel functions for each layer line, that must then be respected, induce an adjustment of the densities (both radially and angularly) in the volumes. Considering the one-start helices, the more "operators" (axial symmetry around integer values and periodicity of 2) are needed to go from the true symmetry to the ambiguous ones, the more potentially not respectable equalities of Bessel functions orders appear, whatever radii r in real space are given as argument of $J_n(A)(2\pi Rr(A))$. Moreover, not every r values are possible, as the width of the box, when segmenting, is one physical limitations for it. This is likely why we observe a decrease of ACC after a number of use of these "operators".

For the helices with additional rotational symmetry (several starts), that were not described in this section, there is an additional restraint on the Bessel functions that occur at each layer line : their order must be an integral multiple of the number of starts (the right side of equation (2) would then by multiplied by the number of starts). Thus, even if different several-start helices can give rise to transform with layer lines at same heights in the same manner as for the one-start helices, with some defined relationships between Bessel orders, it is more difficult for the densities to rearrange in a way that the peaks on all the layer lines are situated at same positions. However, we can still observe other ACC peaks than the true one, related by the same "operators" as described before, even if at lower ACC values (**Figure 3.9E**), identically as for the one-start when a lot of operators were needed to go from the true symmetry to such peaks, so when not all equalities could be respected.

Case of other helices giving rise to diffraction at the same heights

Another question arising from the observation of the ACC plots and subsequent theoretical consideration is why other symmetries that are not related to the true symmetry by the described "operators", but that still give rise to signal in Fourier space at the only same layer lines, are not associated to high values of ACC. The simplest case is when a symmetry is imposed with exactly one more (or one less) subunits per turn than for the true helix. Then it can be shown that between these symmetries, the relationships of Bessel function order are such that the orders parity is not always the same on each particular layer line, depending on the value of m (even m : same parity ; odd m : different parity) in contrast to what we calculated for our ambiguous helices (Eq. 13). As we saw in introduction, in a projection image, the phases along a layer line on opposite sides of the meridian are constrained to differ, theoretically, by either 0° or 180°, depending on the order of the Bessel function on this layer line. If n is even, the difference is 0° whereas if n is odd the phase will differ by 180°. Thus, two 3D volumes having such symmetry relationship (N and N+-1), due to the differences of parity of n, have 2D projections that are fully incompatible each other. Of course, if all m values, for all layer lines (giving signal at the resolution we are considering) on both projection were even, it would be different, but we can in practice ignore such case. Depending on the symmetry of the true volume at the origin of the projection we are studying, it is possible that other symmetries than the (N+1) give raise to transform with layer lines at same height. For examples, if the original symmetry is 9.3 subunits per turn (repeat after ten turns), the 9.1 symmetry also have same repeat, and thus layer lines at same heights. Such cases were not rationally studied, that should be done, for example using deducible relationships for the fractional part of the number of subunits per turn, to see if one can prove –or not- that such pairs of symmetry don't have always same parity of n values for each layer lines, but every cases examined upon there showed parity differences.

Consideration on resolution

Resolution seems to be an important point to mention here: intuitively, one would say that the higher the resolution, the less possible it would be to construct ambiguous models. This hypothesis is presented as an affirmation in (Edward H Egelman 2010), however neither he or ourselves have proofs for this (otherwise he could have shown that the IHRSR procedure always converges when the projections of the high resolution TMV structure are not filtered). Within our tests, the structure of the highest resolution is also the one of TMV (**Figure 3.6**), and the ambiguities are present to the same extent as in the tests on projection of lower resolution structures (for example RSV). We should also keep in mind that the ambiguities might arise not directly from the low resolution of projections, but from the low resolution of the resulting reconstruction, which is constructed from a too limited number of views of the asymmetric unit.

What happens to the views that could not be included in the reconstruction

What is also important to note, is that the demonstrated relationships between the Bessel functions for each layer lines, and the corresponding adjustments of density distribution do not need to be verified on all the Fourier space, only for the central sections corresponding to the assigned on-axis views to the segmented images. However, even in the case of the projection of the very long constructed model of RSV (**Figure 3.3**), that is an optimum case, in term of length (at least much better of what we could have experimentally) to fill the Fourier space when assigning the views –whatever the symmetry is- , we observed the same ambiguities for determining the symmetry (results not shown), showing that many FT central sections can be similar for ambiguous volumes.

This idea that not all Fourier coefficients had to respect the demonstrated relationships, lead us to the question of what happened to the central sections that do not correspond to any view included in the reconstructions, so where the coefficient are only obtained by extrapolations from the adjacent planes that correspond to views that were included in the reconstruction. The **Figure 3.22** shows a comparison between projections of the original helix with 4.8 atoms per turn (**Figure 3.20A**) and projections of reconstructed volumes, imposing either 4.8 subunits per turn or the ambiguous 5.2 symmetry. The projections were made with out-of-plane angles up to 6° every 2° and the on-axis angle views were made all around the helix axis with a 2° step, so that almost all the projections that we are looking at, are along views that were not included when reconstructing these two helices. When one look at the projections with view angles that are close to that of a view that was actually included in the reconstruction (**Figure 3.22A** : same out-of-plane angle, on-axis view ~2° far), both reconstruction with 4.8 and 5.2 subunits per turn show very similar projections, and also, as expected, similar to the original one. Concerning the 4.8 subunits per turn helix (**Figure 3.22, 2nd column**), when one looks at views that are farer from included ones, the quality of reprojections decreases and artifacts become visible. The departure of an on-axis view from an included one cause mostly, when little (**Figure 3.22B**) or no (not shown) out-of plane angle is imposed, a stretching of the projection of the small spheres perpendicularly to the axis of the helix, due to lack of information in this direction. Every increase of out-of-plane angle used for reprojecting the volume causes then a decrease of the quality of the reprojections, and artifacts are also visible along the direction of helix axis (**Figure 3.22C and D**). But, at least, the position of the reprojection of the spheres is respected in regard to the one for the original helix, that is **not** the case in the reprojections of the helix with 5.2 spheres per turn (**Figure 3.22, 3rd column**). Indeed, only a few projection of spheres are visible where there are expected (orange arrows), some are placed between two expected densities (red arrows), and the other are not really visible at all. So, as could be expected, the helix 5.2 does not make any sense at other planes that the one that were included. Of course these effects might depend, in other cases, on the true -and ambiguous- symmetries, that will influence the filling of view space. Also the fact that we look here at an extreme example, where most of the densities in the initial true volume are 0, might have an effect of the strongness of the artifacts. However, these observations go in the same sense than an empirical observation made in (E. H. Egelman 2007) were it is stated that the wrong ambiguous structures give, when used as a model for PM against raw images, an uneven distribution of number of image per reference. In the light of what we observed here, we can suppose that the references that correspond to views for which the projection of the ambiguous structure was **indeed** similar to the one of the true symmetry will be preferred to the one corresponding to views for which significant differences in the projection should occur.

Taken together, all these observations raises interesting perspectives for the method, and a number of them will be reviewed now, after a summary of the obtained results.

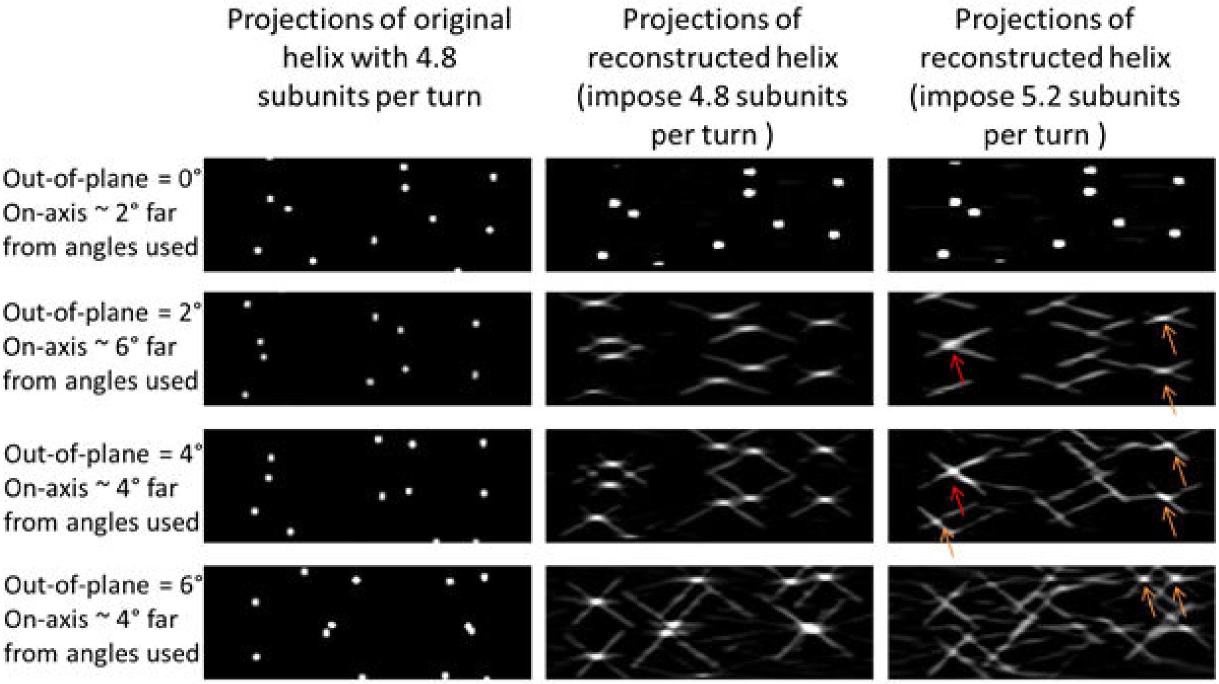

Projections of original helix with 4.8 subunits per turn
Projections of reconstructed helix (impose 4.8 subunits per turn)
Projections of reconstructed helix (impose 5.2 subunits per turn)
Out-of-plane = 0° On-axis ~ 2° far from angles used
Out-of-plane = 2° On-axis ~ 6° far from angles used
Out-of-plane = 4° On-axis ~ 4° far from angles used
Out-of-plane = 6° On-axis ~ 4° far from angles used

Figure 3.22 : The reprojection of reconstructions made from one single projection show artifacts when there is a departure of reprojection angles from angles included in the reconstruction, especially when a false symmetry is imposed

Comparison of projection of the original helix (first column of images) with projection of reconstructed volumes from the projection shown in Figure 20 B, either imposing the true symmetry (second column) or an ambiguous symmetry (third column), for various projection angles indicated on the left. More descriptions are given in the text.

Conclusion on symmetry determination on single 2D projection and further perspectives

Results summary, positive and negative points

We showed through the use of the described method on several ideal cases and experimental cases that the true symmetry is, in almost every case, related to an increase of the measured ACC between reprojections of reconstructed volume and original segmented images. We thus have a way, from a single image, to measure a reliable "probability" that this image corresponds to a particular symmetry. This approach of pseudo-exhaustively imposing the symmetries to test and measuring their plausibility is to our knowledge new, at least for helices projection (one can mention here the ab initio approach to reconstruct models from images of icosahedral objects from (Navaza 2003). Although the underlying theory is totally different, the idea of being fully ab initio and the exhaustive search for the view angles share some similarities with our approach).

For helices presenting additional rotational symmetry, we had cases with unique solutions detectable by ACC measure. An on-going experimental project in the lab, that present a 6-fold rotational symmetry (it is a bacteriophage tail) tends to confirm the success of the method on images (class-average) of such objects. However, this should be further confirmed by more tests and the theoretical description of the ambiguities (or not) for such helices should be done. For helices without rotational symmetry, several solutions gave comparable ACC values and indeed, the projections corresponding to these ambiguous symmetries are undistinguishable (at least in the way that we "look" at them). This confirms many observations and predictions made in the literature. Through the playing with concepts related to the description of ambiguous helices in reciprocal and real space, we were able to gain understanding of the sources of ambiguities. Signal in reciprocal space is confined at the same layer lines, and then the relationships between orders of Bessel functions on these layer lines are such that it is possible that different arrangements of densities in the 3D volume, especially concerning radial positions of densities, give raise to similar central sections of their 3D FT (similar projections). However, the fact that the relationships between Bessel functions's order cannot be always respected by such rearrangement means that we have only a limited number of ambiguous symmetries. Thanks to that, using simple additional informations that we might dispose (top-views, information on subunit, literature), it is possible in experimental cases to deduce from the ACC plot and inspections of the different plausible volumes to decide for the true symmetry. Of course, using more additional information that provides for example information on surface lattice (metal-shadowing, AFM, quick-freeze/deep-etch EM) or mass per unit length measurement (knowing the MW of the subunit), we could even more easily discard many if not all of the ambiguous solutions. However, this would destroy one main advantage of the method, that is its simplicity and rapidity of application.

Beyond the ambiguities problems, as a negative point, we also need to mention the object-dependant effects that we had to face. The success of the use of the method thus depends on the information that we have. Projection length, number of different views that it contains and how these views fill the angular space (these parameters being symmetry-dependant). In that regard, a case like the TMV was ideal, whereas an helix with almost an integer number of subunit per turn is less (although the method still worked for the non-digested nucleocapsid of Measles).

One major positive point of this method development, and particularly of the analysis of the kind of results that could be obtained, is that it lead to a new way to describe ambiguities in helical symmetry determination, that were often observed in the literature –but not really explained-, and to establish relationships between ambiguous symmetries. Thus, even if using another method to determine the helical symmetry, like the Fourier-Bessel approach, one would be able to predict which other symmetries, that might not have been detected in a first place, are likely to be true, and then for example try to reconstruct with these symmetries to see if they make more sense. Furthermore, these descriptions helped to point out some critical points that can help to reduce the problems of symmetry determination, like restrictions on reconstructions radii. We also could deduce particular relationships between Bessel orders for each layer lines of FT of ambiguous helices, and this could help when working with experimental cases. As an example, one could start from the possible solutions given by our method, and by looking at experimental FTs while taking in account what we predict to be present in the FTs, like the relatives orders of Bessel functions for several layer-lines for different ambiguous symmetries, and looking at the intensities in the diffraction peaks (as this is for example done in (Y. A. Wang, Yu, et al. 2006), one could distinguish which solution is more likely to be true. More generally, the understanding of ambiguities that we gained through our analysis gives us the possibility to predict what to expect when analyzing experimental data.

Advantages of the method over existing procedures

With regards to the Fourier-Bessel approach

Until now, the only well-known method that really aims to determine the helical symmetry ab initio is the classical Fourier-Bessel method. It is to note that after the writing of this manuscript part, a paper describing an alternative method for reconstructing helices, and providing also possible ways to determine the helical symmetry was published (Lee, Doerschuk, and Johnson 2011) but cannot be discussed for comparison here due to its complexity and novelty). Despite the development of many programs that makes the task easier, this approach still require time and a good understanding of the underlying theory to be successful. In comparison to it, our approach is very fast and simple, although we saw that analyzing the complex ACC plots resulting from it is not always straightforward. Possibilities for improving this part will be discussed below. The Fourier-Bessel method also requires, being able to do the indexing, quite long and well-diffracting helix portions (this requiring sometimes computational straightening of images). Although in our approach the length as well as the rigidity of the projection that we analyze can also influence the quality of the results, we are not that much limited: as an example, running the method on “low-diffracting” small class-averages of negatively stained measles nucleocapsid images provided us the needed symmetry information. Concerning the ambiguities in symmetry determination in the classical method, there is not that much description of it in literature (maybe the refractory cases were not published?). In the light of what we saw, it is anyway evident that somehow similar ambiguities problems will be encountered in our method and the classical, as the FT of projection of ambiguous volumes are similar. As an example, to assign Bessel orders to particular layer lines in the FT, one has to use a value of the radius of the helix, and the one that can be normally easily measured on the input image is the outer maximal radius, that only help to define a maximum limit for the values of n, thus inducing ambiguities in the indexing.

With regards to the IHRSR approach

Although IHRSR approach is not really originally designed to give an ab initio determination of the symmetry, it can be seen as a way of determining the helical symmetry, as by starting with more or less roughly determined parameters, the method should be able to converge to the true parameters. Thus, starting from many different points, one could in theory also “quasi-exhaustively” sample the parameter space and find solution(s). However, as we saw, not only the starting helical parameters are crucial regarding the final solution found, but also the parameters for symmetry search (increment for search of rise and rotation). In the end this makes many variable that one should test, and for each of them it would implies calculating and refining many reconstructions by PM procedures, so that it would be very time consuming and require to bring together many information to study the convergence of helical parameters. In comparison, we are getting the possible solutions with our method from single images (or class-averages) very quickly, and we do not depend on initial parameters or variable for symmetry search once we have defined a range of reasonable parameters to test and a sufficiently fine search step. Our method is thus complementary to IHRSR, as it would give us starting points for symmetry parameters refinement that are almost already exact (thus speeding up the refinement procedure), and we would be able to predict in which different possible local solutions the IHRSR procedure may lead us, thus gaining significant time in the analysis.

Possible improvements and applications

While globally keeping the method as it is

Improving display of the results

Up to now, the basic output of the script for symmetry determination is a text table containing the symmetry parameters (number of starts, rise and rotation per subunit, and corresponding pitch and number of subunits per turn), the average correlation ACC associated to these parameters, the number of images included in the reconstruction and the standard deviation of the CCs for each symmetry tested. From this table can easily be extracted the symmetry parameters giving the highest correlations, but we saw that looking only at the absolute values of ACC was not so informative. Instead, one can use a plotting program to display the 2D or 3D profiles of ACC as a function of the symmetry parameters, and analyse these profile to extract the potential solutions. Optionally, one can also tell the script to keep the segmented image stacks, the reconstructions, and the reprojections stack for each of the tested parameters. This is usually done once we have detected potential solutions to visually inspect each corresponding reconstructions (and eventually also reprojections). In practice, our current way of analyzing the results present some weakness. One need to plot the result, to manually record which parameters are associated to local maxima of ACC (and it can be quite a lot for a large search), then for each of these parameters one has to re-launch the script by using an option to keep the reconstruction, then open each of the reconstruction with a visualization program (like pymol, chimera), and finally compare them, while trying to keep an eye to which point in the ACC plot there are related. All these steps are very time consuming, usually longer than the generation of the ACC profile itself, and not straightforward (like visualizing in the same time a reconstruction and corresponding position of symmetry parameters in the ACC plot). This will cause us to tend to avoid to look at too much reconstructions for each analyzed image, for example the one that correspond to minor peaks in ACC, that could be a major problem when dealing with very noisy data and/or data with unfavorable symmetry (in regard to the method). Furthermore, it prevents us to do the analysis on a larger scale (many class-averages) in a reasonable time scale. So we propose, for a real improvement of the efficiency of data analysis rather than for any superficial aesthetic reasons, to create a dedicated visualization program, with the required following characteristics:

-For plotting the ACC values as a function of the symmetry parameters, it should make possible to get interactively (mouse) and in real-time those different values when moving through the graph, as well as additional information (number of images included in the reconstruction, standard deviation of CCs, etc).

-It should give the possibility to detect and record the symmetry parameters associated with a given number of local ACC maxima, to display them and save corresponding reconstructions

-It should make possible to display side-by-side the results for several images, using automatically calculated compensatory factors to be able to compare ACC plots that have different range of absolute values (this was usually the case when looking at results for experimental data)

-When the symmetry search is done by varying two parameters in the same time (like pitch and number of subunits per turn), one should be able to interactively pick 2D slices of the resulting 3D plot, and superpose them (for example to look at the ACC profile according to the pitch for various chosen number of subunits per turn imposed, that can help to improve the strategy for parameters search)

-And **most importantly**, for each symmetry tested, 3D surface representation (at a few different visualization thresholds) as well as representative slices (like a top-view and a side view) should be recorded, without writing to the disk all the corresponding reconstruction (it can becomes a huge amount of data, when testing thousands of parameters). Then the user could move through the plot of ACC (2D or 3D) and **directly** look at these representative views of the corresponding reconstruction. Our experience showed that the visualization part was of great importance to discard or retain certain symmetries, and this kind of tool would totally change the potential of this method. Even more, if one have any idea of the shape of the subunits, such a fast exploration of possible 3D volumes that can be obtained from a 2D image, helped by an objective measure such as the ACC, could lead to a more easy determination of the symmetry.

Adopting a more exhaustive scheme of data set analysis

In the present report, we showed results on experimental data set for only one class-average for each of the chosen object. Of course, we have tested the method and analyze the results on several class-averages for each of them, but regarding to the total amount of data that we had in our hands (several hundreds of good class-averages), our tests were done on only a very small fraction of the whole data. We propose that this method should be applied on almost all the class-averages that are obtained ab initio from the raw images (after discarding class-averages of very poor quality, showing for example high degree of bending). This might thus be an automatic and ab initio way for sorting images according to the symmetry, when heterogeneity is present in the data, and that do not depend on any initial model. However, due to the fact that a part of the class-averages will correspond to images that have an out-of-plane angle different than 0, such exhaustive test would first require that the ability of the method to determine the out-of-plane angle as well is proven, that is currently under test.

Another way of analyzing more exhaustively the data set would be to treat individually every individual raw filament, for a sorting purpose. The low signal over noise ratio in the raw images would thus be compensated by the fact that the length of the analyzed projections would then be much greater, that is an advantage for several reasons like for example a better filling of angular space.

Both of these approaches require experimental validation.

Improving the analysis of ACC peaks

As we saw through the analysis of the relationships between helical parameters giving rise to ambiguity in symmetry determination on 2D projection, one can now predict for each symmetry parameters, which are the other parameters that may produce equivalent projection images, and particularly one can predict relationships between Bessel orders on each layer line for a set of ambiguous symmetry. These relationships implies that even if a pair of symmetry parameters are related to the true one by the described "operators", it might be that the underlying relationships between Bessel orders cannot be compensated by redistribution of the densities, especially when Bessel function of very high orders are implicated. In such cases, we predicted a decrease of ACC for corresponding parameters. This hypothesis, although observed experimentally, should be more objectively confirmed, by analyzing the relative heights of the ACC peaks related to the true parameters and correlate this with the possibility or not to verify the relationships of Bessel orders (until a certain resolution). Once such correlation is confirmed, one could use this to determine, among a population of ACC peaks, which is the one more likely to correspond to the true parameters by using a reasoning like : "If these parameters are true, then one should observe a decrease of ACC for those other related parameters, because the relationships between Bessel orders could not be verified, but we don't observe this, so we move to the next ACC peak and repeat the same reasoning..etc.. until the predictions match the observations at best"

Looking for helices other than the elementary one

In the examples that we have shown, we were interested in finding **one** pair of helical parameter : it was the one of the elementary helix, that is the one associated with smallest distance (taking in account rise and rotation) between one subunit to the next, and that is very often also the one that is the most obvious when visually looking at an helix, at least without rotational symmetry (For the helices with rotational symmetry, those parameters correspond to the helix running the most parallel to the helical axis, that is with the smallest rotation between subunits). However, one can construct one helical assembly by using many other ensembles of helices than the elementary one, and the parameters of all those helices depends on the parameters of the elementary helix. Thus, for each potential solution of elementary helix parameters found by the first analysis of the ACC profile, one could predict which other parameters corresponding to the other related non-elementary helices should give high ACC, and verify at which extent it is the case. One has to note that this method would be limited by the number of views that can be inserted in the reconstruction that will be even lower than when imposing the parameters of the elementary helix. If this approach would be able to reduce the ambiguities to a only a few solutions (or at best only one), should be verified experimentally.

Measuring the resolution of reconstructions

This is something that was not tried, mainly because we were dealing with small amount of images, that one don't like to separate into even smaller ensemble, and because there was no clear way of dealing with the fact the number of images was dependant on the symmetry parameters, that could have influence the measure of resolution by FSC. However, one can imagine to take, for this specific measure of FSC, the same number of images for all reconstructions (of course this is better when the parameters search range in not so huge so that some parameters give rise to only very few segmented images), and separate into two datasets. The way of separating images, "one over two" image for each dataset, or images corresponding to the two halves of the original input image, should be appreciated with tests on known cases, as will be the positive effects of adding this measure to the current measures.

Adding a correction factor to the measured ACC

When looking at which factors had an influence on the measured ACC, we understood that not only the departure of the imposed parameters to the true one (and related ambiguous one), but other factors had an influence like the number of images included in the reconstruction and the filling/ sampling of angular space, that are both symmetry dependent. Thus, some trials were attempted to "correct" the measured ACC in order to limit the influence of such factors. However, no really good way of doing it was yet found. Dividing the ACC by the standard deviation of CCs between individual images included in the reconstruction and reprojections (we might expect higher standard deviation for false parameters) gave in some cases interesting results, but not systematically. Other way of correcting the ACC, like correcting the measured values by values obtained from a random noise image without helical symmetry (that may suffer from effects like number of images and filling of angular space as well) should be attempted.

Exploiting the views that could not be included in the reconstruction

We saw earlier that projecting reconstructions corresponding to the true or to ambiguous symmetry parameters along views that were not used for reconstructing the volumes might be a way to differentiate between true or false ambiguous parameters, and may explain empirical observations made in (Egelman 2007).

We should verify this by using reconstructions of the different ambiguous volumes (first for an ideal test case, then on experimental data) constructed from one projection and use them as references for PM against many projection of the known structure, and establish a correlation between how far apart are view parameters to one used for reconstructing the reference from the single projection and views preferences after PM. Of course, we already used models with different ambiguous symmetries of measles for PM, and no difference in the global correlations were found to permit to distinguish one of the symmetry as being the true one. However, no particular care was taken about the angles of projection of these models and further reference distribution. Thus, even if there was certainly for the wrong models, projection views far from one used in reconstruction (thus of bad quality), the raw images might have shift in the helix direction to match with the closest projection that had an angle close to one used when reconstructing the model.

A possible way of using the method would then be to compare each of the possible solutions to the raw images by PM, and carefully analyze the evenness of reference distribution.

While changing important points of the method

Decreasing a deleterious effect of including very few images in reconstructions

As we it was already noted, if one look at the plots of ACC according to the symmetry parameters, the absolute values that are measured are usually all very high, whatever the symmetry is imposed. If one look at the examples on experimental class-averages of measles (**Figures 3.12 and 3.13**), all ACC values are above 0.96, with less as 0.02 difference between highest and lowest value ! We can attribute this in part to the fact that, particularly when a low number of images is included in the reconstruction, each input image contribute itself highly to the CC, because of conservation of information during back-projection and reprojection, thus biasing the measure. Moreover, the way that the other images included in the reconstruction influence this high contribution depends on the imposed symmetry : when views are close each other, one can expect a stronger influence than when views are more a part each other. This symmetry-dependent effects and the fact that anyway the individual CCs are too much influenced by the windowed segments themselves, are some things that we would like to avoid. Thus, one can propose to calculate, for each tested symmetry parameters, a number of reconstruction equal to the number of windowed images, while always avoiding using one different windowed image. For example, if for a pair of symmetry parameter, we segment the input projection into four images numbered 1 to 4, one would calculate one reconstruction with images 1,2,3 ; another with images 1,2,4 ; another with 1,3,4 and finally a last one with images 2,3,4. Then, in that case, to calculate the CC of reprojection of the reconstruction with the image 1, one would use the reconstruction made with images 2,3,4 ; etc…

More generally, one would avoid using a reconstruction including the segmented image X to calculate the CC with segmented image X. This would of course lead to a large increase of the number of reconstruction to calculate, but the benefits might be sufficient to try this.

Using a more appropriate reconstruction algorithm

In the current implementation of the proposed method, the algorithm of reconstruction is a back-projection algorithm (spider command BP 3F) adapted to any single particle of any symmetry. We propose here to use a helical-symmetry oriented reconstruction strategy, which could greatly improve both the speed, and more importantly the capacity of the method to produce the expected results. The most evident of such a strategy is the classical Fourier-Bessel method, which is somehow paradoxal as one of the first aims for developing our method was to avoid using the classical method. However, here, of course, it wouldn't consist in indexing the FT of the input images, but assuming every symmetry parameters that one would like to test, and automatically derive a 3D reconstruction using the Fourier-Bessel algorithms.

This would actually be a fully different approach: in our current way of calculating the reconstruction, the images are « forcing » the reconstruction to "look like" them when reprojected. In this alternative approach, as we are taking in account only the Fourier coefficients that correspond to the symmetry that we want to test (layer line extraction), if the input image don't follow this particular symmetry, no meaningful reconstruction can be calculated in most of the cases. To illustrate this with an extreme case, if the input image is a projection of a perfect helix without noise (that was almost the case in some ideal examples presented above), and if we try to impose many of other symmetry than the true one, one would only pick Fourier coefficient equal to 0, and thus the correlation of the reprojection of such a reconstruction with the original image would be 0 % (in comparison, we had at least 88 % for the projection of TMV at 5 angstrom resolution with our current method…). Of course in a real case, this effect would not be as dramatic, but one could anyway hope for a much better contrast in the measured CCs because most of the non-helical noise would be eliminated, thus facilitating the analysis of the results.

Not only the contrast of correlation would be improved, but many of the above discussed problems could be solved :

- No effects of number of images included in the reconstruction
- No problems of the symmetry dependent uneven sampling of angular space (and sparse filling of Fourier space)
- Reduced effects of interpolations (no need of shifting the original image to window it into smaller segments)
- Increased calculation time (many of the Fourier coefficients are just not taken in account)

One problem should unfortunately still remain : the ambiguous symmetry solutions. As we could show, the ambiguous symmetries have layer lines at same heights. Thus, when imposing ambiguous symmetries using the classical Fourier-Bessel method, one would anyway extract Fourier coefficient containing information, and the reconstructed volumes would probably make sense and have reprojection similar to the initial image. To which extent the number of ambiguous solutions could be reduced by using this new way of calculating reconstruction, is something that need to be verified experimentally. Anyway, if almost any other problems that we encountered when using our method in its current implementation are reduced, this proposed perspective is still one of the most promising.

PART 3 : Towards 3D reconstruction

During this work, methodological developments mostly concerned ab initio determination of helical symmetry parameters and sorting of helical segments by classification. As for the 3D reconstruction procedure as such, we mostly used well described methods like the iterative helical real-space reconstruction method IHRSR (E H Egelman 2000) and a rigorous alignment parameters validation strategy (Sachse et al. 2007). A part of my work consisted in understanding these methods, evaluating their strengths and weaknesses, applying them, and setting up a pipeline for image processing primarily for in-house usage. This resulted in a fairly universal script for helical reconstruction which will be described later (part “Introduction into the developed scripts”). Generally, the methods we used for reconstruction are described in our article about the measles virus nucleocapsid (Desfosses, Goret, Farias Estrozi, Ruigrok, & Gutsche, 2011) and the VSV N-RNA bullets (Desfosses, Ribeiro, Schoehn, Blondel, Guilligay, Jamin, Ruigrok and Gutsche, in preparation) included in the appendix of the present manuscript. Here I will provide more extensive comments on several important aspects of the reconstruction methods and give some perspectives for reconstruction of helical objects.

Measles : reconstruction using IHRSR

Roughly speaking, we used Egelman’s IHRSR method, with some additional steps of image selection and other adjustments (see “Introduction into the developed scripts” part). In our first attempts, we were facing many difficulties to obtain a correct reconstruction, mainly because the refinement of the symmetry parameters leads to multiple solutions, most of which, if not all of them were actually wrong (**Figure 4.1**). We realized the huge importance of the starting point for symmetry search, much more critical than was suspected from the literature. Furthermore, the ‘search step’ parameters for axial rise and angular rotation, as required by *hsearch* program to define the grid range and spacing that will be used to determine the best fitting helical symmetry on the reconstructed volume, had strong effects on final results and were crucial in order find the correct symmetry of the structure.

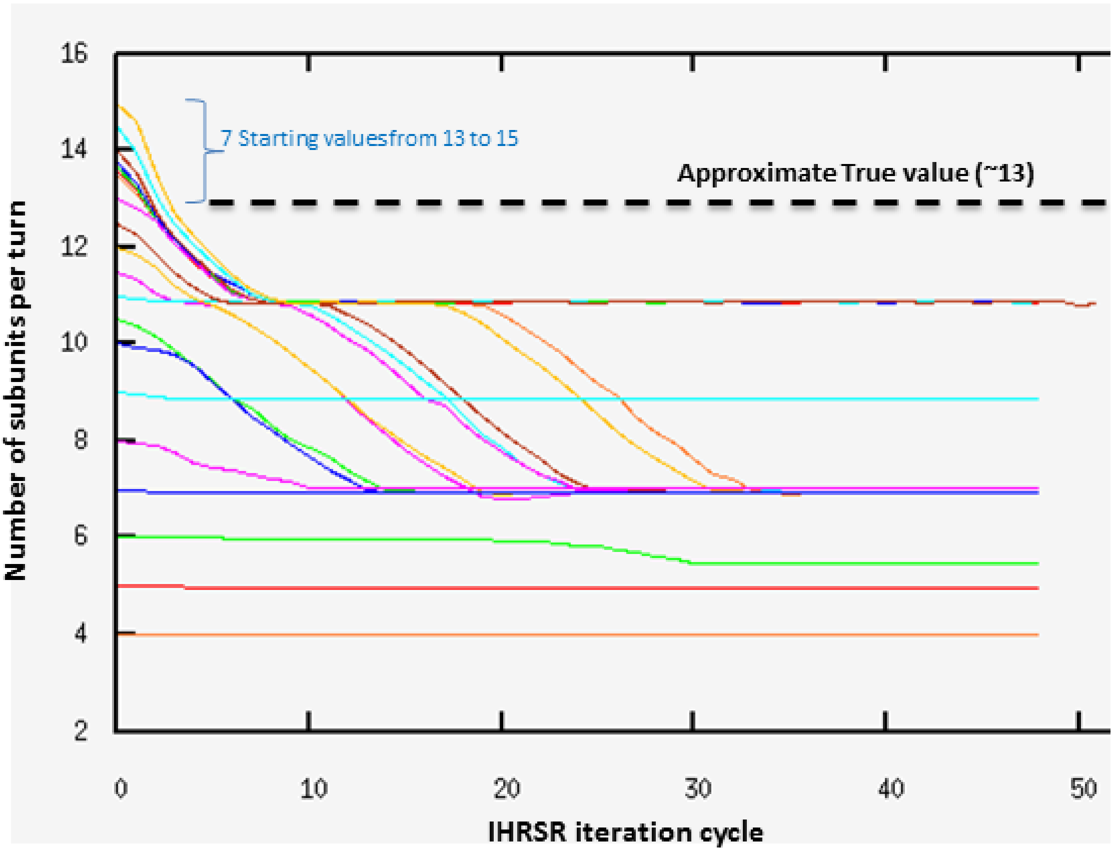


Figure 4.1 : Illustration of lack of convergence and ability of finding true helical parameters during structure refinement using the IHRSR method

An IHRSR refinement was done on a MevND data set after sorting of images according to pitch. The evolution of number of subunits per turn as found by the procedure is shown as a function of IHRSR iteration number. Although we finely sampled the initial guesses of number of symmetry per turn in the range between 11 and 15, none of those solutions converged to the true solution (dashed black line). More surprisingly, the true solution was "crossed" by the trajectory of helical parameters from several refinement trials, without being detected.

After using the classification of 2D images and developping methods of the symmetry estimation based on 2D class-averages which were described earlier in this manuscript, we could finally use the IHRSR approach with success. As an example, for MeVD, we already mentioned how a not yet explained heterogeneity, detected on the PS class-averages solution (See classification, **figure 2.19**), prevented the IHRSR refinement from convergence to a stable and reproducible symmetry, although the variability turned out to be very small (two population of either 12.38 or 12.33 subunits per turn). The problem of convergence in cases of heterogeneity is reported in several IHRSR-based papers, for example in (Y. A. Wang et al. 2006). The use of our method for symmetry parameter estimation, which gave us precise starting points (~12.3 for MeVD and ~13 for MeVND), allowed us to restrict the starting point to a very narrow range and use smaller values for the ‘search step’ parameters, which made possible to obtain stable symmetry parameters.

To conclude on this part, we can say that the Measles case teaches us useful lessons on the use of the IHRSR method, which were sometimes not clear from the literature. The IHRSR method was, in the initial paper describing it, advertised as “a robust algorithm for the reconstruction of helical filaments using single-particle methods” (E H Egelman, 2000), mainly because one could start the reconstruction with a featureless initial model as a smooth cylinder (E. H. Egelman 2007). It is noted in another 2007 paper that the “The reconstructed volume […] will be almost indistinguishable (at 12 Å resolution) for a large range of different initial reference volumes and starting symmetries, which is why the algorithm is called ‘robust’” (E H Egelman 2007). In my hands, the robustness of the method was not so clear. I spent some times to exhaustively test the effects of starting symmetry parameters and the ‘search step’ parameters on a relatively homogeneous data set (MeVD after the classifications step), and I realized how precise and “lucky” one has sometimes to be in order to find the combination of parameters that will allow a correct refinement. Interestingly, a more recent paper on the method (Edward H Egelman 2010) highlights these critical points, which should be taken into account when using IHRSR.

VSV nucleocapsids : reconstruction without symmetry imposition

For determination of the structure of the VSV N-RNA bullet trunks we were facing two main difficulties. First, the data set was highly heterogeneous as could be judged from diameter variability (see classification part). Even after the classification steps, it was never completely clear if we had finally succeeded in obtaining a homogeneous subset of segments. The final relatively low-resolution reconstruction would tend to show that it was not the case. If it were to be done again, I would try to push the sorting of the dataset even more, even if at the end would be left only with a couple of filaments (which would be, due to the very high number of subunits per turn, already enough to get a better resolution than what the one we currently have). The second difficulty arose from the fact that we had no indication of the symmetry(ies). Clearly, the smaller diameter of our reconstituted bullets in comparison to the full virion structure (Ge et al. 2010) indicated that we had less subunits per turn. We tried to estimate this value by establishing a relationship between diameter (=>

circumference) and number of subunits per turn for the known structures (either the viral nucleocapsid or the crystallized N-RNA ring (Green et al. 2006)), and extrapolating this value to our structures. However, the uncertainties related to the measures lead only to a rough estimate between ~31 and ~35 subunits per turn. Using these estimations as starting points for the IHRSR method to converge to the true symmetry parameters was always unsuccessful (results not shown): if the 'search step' values were too small, the initial parameters remained virtually unchanged, so that we ended up with as many final parameters as initial guesses. When these values were set higher, the procedure systematically converged to non-relevant solutions (~ 20-22 subunits per turn). The very small angular rotation and axial rise (due to the high number of subunits per turn), may be one of the reasons for this high sensitivity, in addition to the fact that remaining heterogeneity may be present.

To circumvent those difficulties, I used an approach that does not require initial symmetry guess (only the pitch, which can be easily determined from the images), which is based on reconstruction without symmetry imposition, in a way that is, to my knowledge, not described in the literature for helical samples. In their 2001 paper, (Narita et al. 2001), use an approach without symmetry imposition for reconstructing the quasi helical actin-troponin/tropomyosin complex, but they started from a helically symmetric initial model, whereas we started from a smooth helix with only a defined pitch. The main difficulty that I encountered in my first attempts to reconstruct without symmetry imposition, was that the volumes became so asymmetric upon reconstruction iterations that parts of the helix were very badly defined, or deformed (**Figure 4.2 B,C**). This was due to the fact that the distribution of views per on-axis angle became more and more uneven upon PM iterations (**Figure 4.2A**). A way to solve this problem was therefore to limit the number of images for each on-axis angle bin before including them in a new reconstruction. We therefore included this possibility in our reconstruction pipeline (see part "Introduction into the developed scripts"). In order not to lose too many images during this additional selection step, we also tried to understand what made the distribution of on-axis views so uneven. In our stack of images, we necessarily have a quasi-even distribution of the views, due to the helical symmetry (especially for VSV N-RNA bullet trunks which have a very high number of subunits per turn) and the presence of many different filaments. However, when looking at the y-shifts distribution as determined by projection matching (the shifts along the helical axis), we realized that many images had big y-shifts, despite of the fact that many different on-axis view references were created. This seemed illogical when one considers that if the number of on-axis references is high, each image should be able to find a reference it would match without requiring a big y-shift. Remarkably however, some projections are systematically preferred and cause images to shift in y-direction more than they should in order to match with them. A potential explanation of this phenomenon might lie either in an uneven density distribution in the reference volume, or in interpolations effects (on-axis views at 0, 90, 180 and 270 degrees have for example a general tendency to be preferred, especially when the reference structure is a smooth helix). In the SPIDER release that I used during the thesis (version 17.05), only one value for both x and y shifts search ranges could be given, therefore in order to restrict the search range to a lower value (1 to 3 pixel), I opted for a preliminary rigorous centering, which then allowed to reduce the y-shifts found by PM. This was thus included in our procedure (see “Introduction into the developed scripts”), and made possible to obtain a more even on-axis view distribution while keeping enough images per view. More recent SPIDER releases and some other packages (e.g. EMAN2), already include a possibility of having different search ranges for x and y shifts. However, the centering of the segments will always help to reduce the x-shift search range thus reducing the computation time.

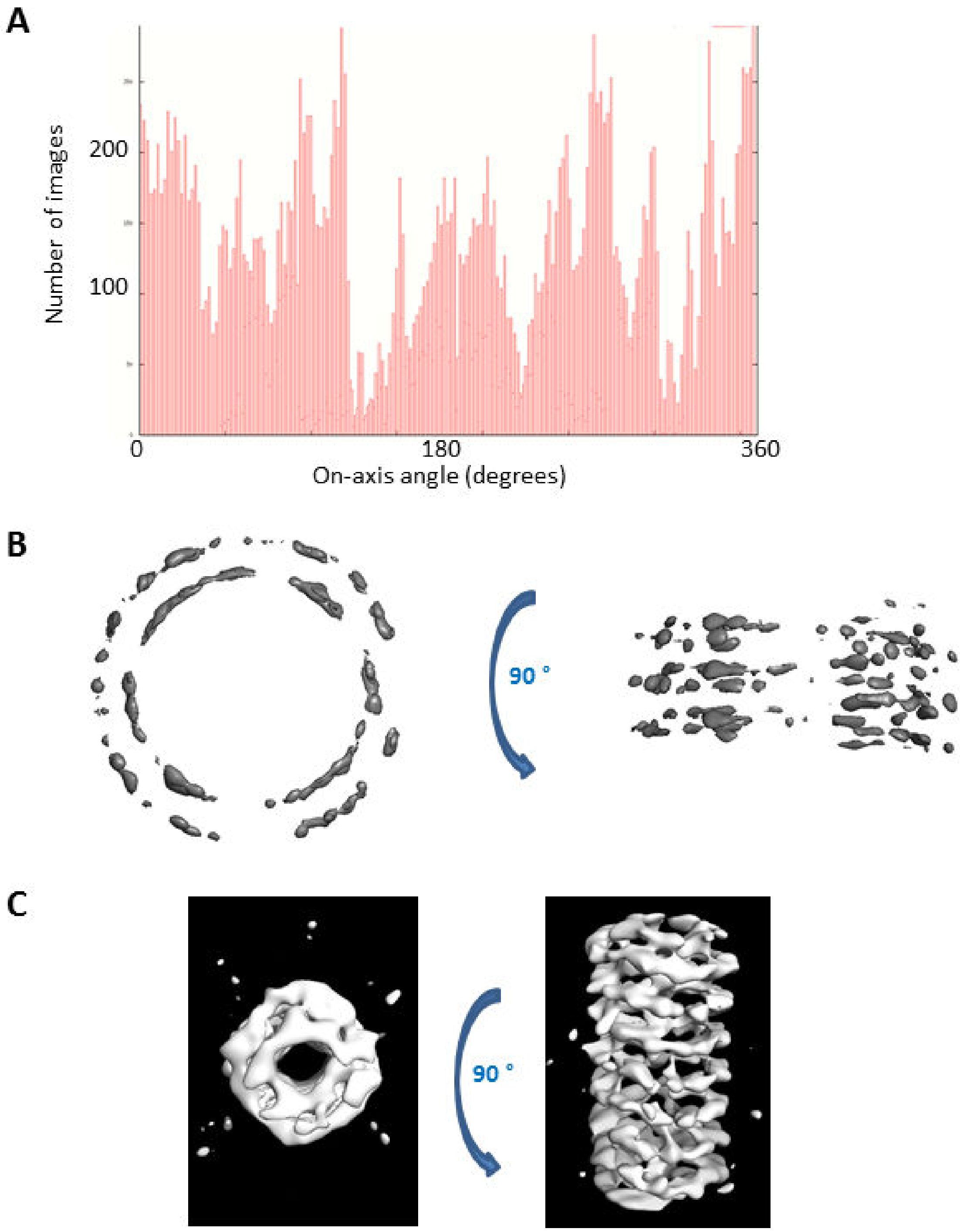


Figure 4.2 : Illustration of difficulties encountered for refining structures without imposing helical symmetry

When one try to iterate projection matching cycles starting from a smooth helix, and not particularly taking care of view distribution, one rapidly ends up with a highly biased on-axis view distribution (**A**), which will become worst and worst through iterations. This lead to reconstruction were some parts in the structure are missing, as seen on a VSV bullets example in **B.** The right part is a view along a poorly represented on-axis views. Also, the circular shape of the helix as seen from the top is can be not respected anymore, as seen on the top view of an asymmetric MeaslesD reconstruction (**C**, left).

By applying the selection of an even number of images per on-axis view in addition to the centering of the segments, we could, starting from a smooth helix, obtain reconstructions for VSV N-RNA bullet trunks, with or without the M protein added. After several PM iterations, the symmetry of the reconstructed volume became apparent (**Figures 4.3 and 4.4**). This symmetry was then imposed on the volume and the structures were "refined" using IHRSR with very small symmetry search range parameters.

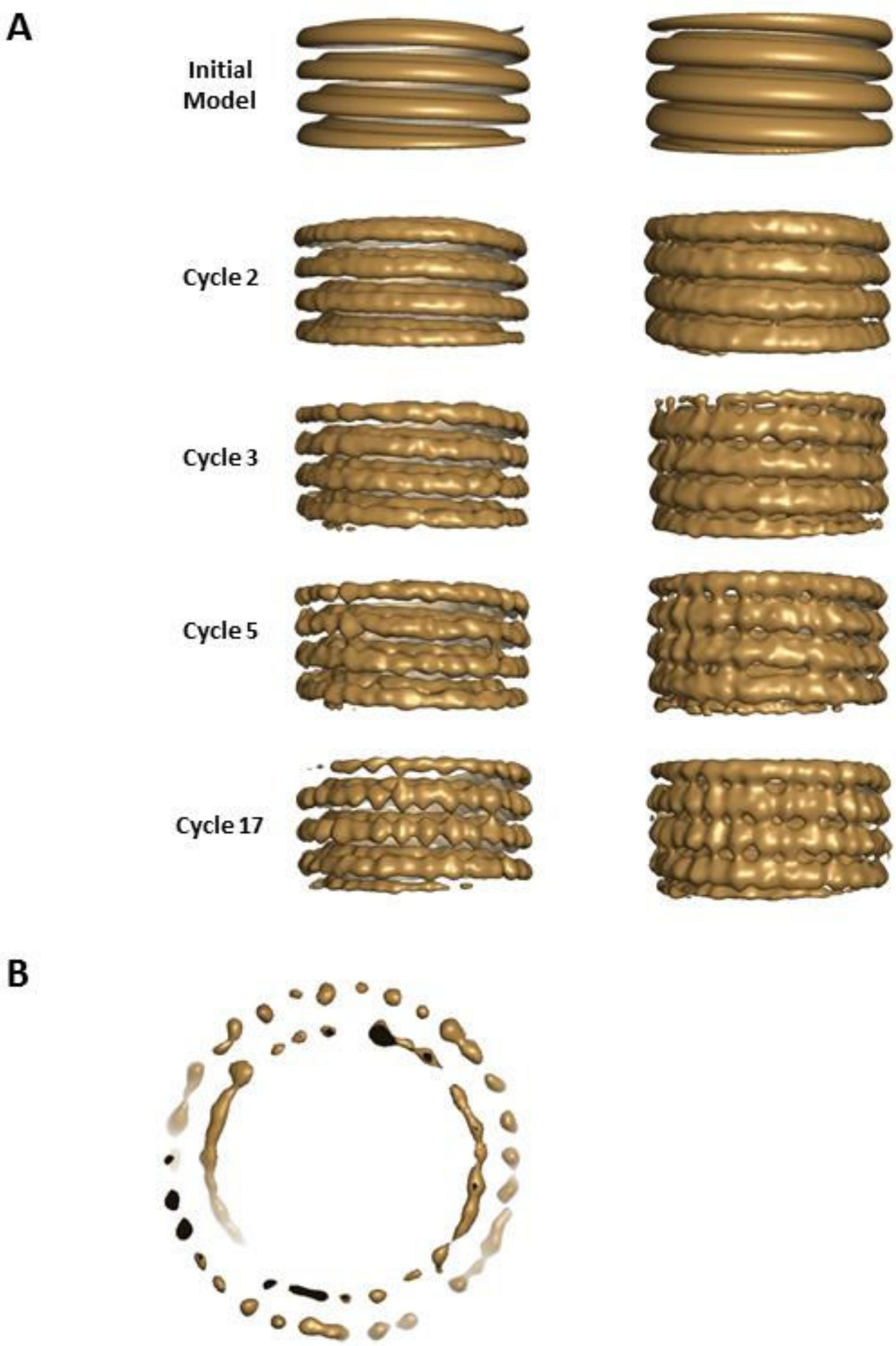


Figure 4.3 : Reconstruction of VSV N-RNA reconstituted bullets using no symmetry imposition step

The panel **A** shows the evolution of the reconstruction (visualized at high threshold on the left and low threshold on the right) from the smooth helix used as initial model to the final reconstruction (cycle 17), on which the helical symmetry appears. The text indicates the PM iteration cycle number. A close-up top-view of this final reconstruction visualized at very high threshold is shown in the panel **B**.

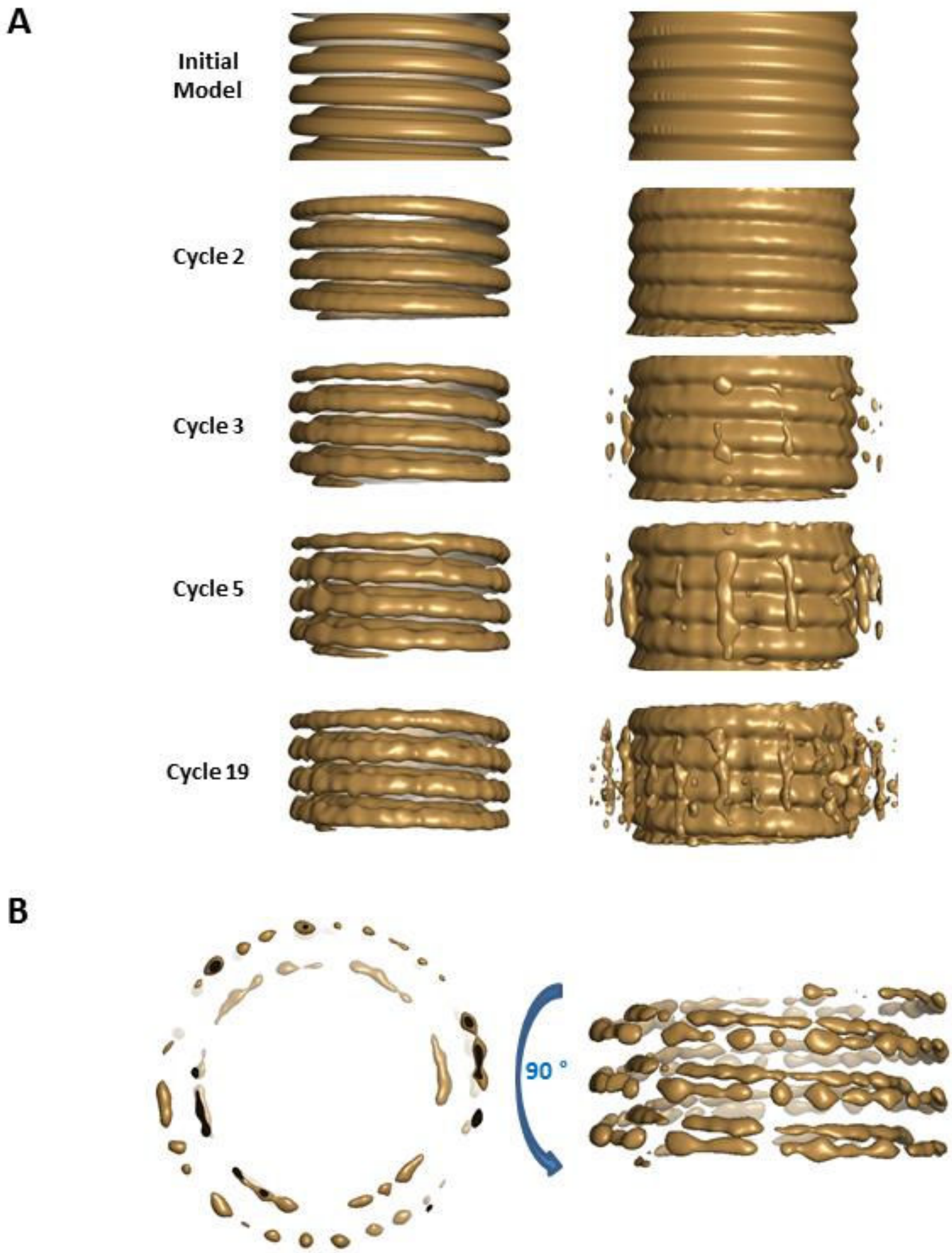


Figure 4.4 : Reconstruction of VSV N-RNA reconstituted bullets after addition of the Matrix protein, by using a procedure with no symmetry imposition step

The panel **A** shows the evolution of the reconstruction (visualized at high threshold on the left and low threshold on the right) from the smooth helix used as initial model to the final reconstruction (cycle 19). The text indicates the PM iteration cycle number. Although the symmetry appears less clearly than on the reconstruction without M, the close-up view of the final reconstruction visualized at very high threshold shown in the panel **B** shows individual densities appearing.

Perspectives

General remarks

Obviously, there is room for improvement of our current reconstruction procedures. For example, the symmetrisation based on multiple inclusion of images according to the helical symmetry (Sachse et al. 2007), which is an appropriate way of taking into account the symmetry, was not completely included in our reconstruction pipeline (some troubleshooting is still required to make it work properly). Additionally, a recent high-resolution work on TMV (Ge and Zhou 2011) introduces several modifications of the original IHRSR procedure, notably the inclusion and the use of a better version of the himpose program and a new method to guide the generation of the reference volume projections taking into account the helical symmetry. A quantitative comparison between this approach and the one of (Sachse et al., 2007) on the exact same data set, especially concerning the two symmetrisation methods, is necessary in order to know what to use in the future. One can also cite another recent methodological paper that proposes a completely new and promising view on the reconstruction of helical objects (Lee, Doerschuk, and Johnson 2011). However, both the complexity of this paper and the fact that no more recent articles applying this method have been published make it impossible to correctly discuss it here. What I would like to briefly discuss now, as a perspective, are some ideas that emerged during the writing of this manuscript and that should be relatively easy to test, and may improve single-particle approaches for helical reconstruction.

Combining the classical helical reconstruction method and single-particle approaches ?

When I reviewed the literature on helical reconstruction, one thing that surprised me was that the separation between the “classical method”, or “Fourier-Bessel reconstruction” (DeRosier & Klug, 1968; DeRosier & Moore, 1970) and the various more recent single-particle approaches (E H Egelman, 2000; Sachse et al., 2007) was so strong. In some papers (e.g. Schoehn et al., 2004), both approaches are used, for example, via determining the symmetry parameters and a low-resolution model by the classical approach and then refining the structure with a purely single-particle approach. However, both methods are never really mixed. The reason why this surprised me, is that, clearly, both methods have strengths and weaknesses (see introduction) and it seems reasonable to combine them to give rise to a stronger more general method. For example, in the single particle approach, the decision if a particle should be included in a reconstruction or not is usually independent of the helical symmetry (e.g. based on cross-correlation coefficient with model's projection), whereas in the "classical method", particularly precise symmetry-adapted criteria exist. I can cite as an example the work of (Wakabayashi et al. 1975) which makes use of several possible selection criteria like selection of images with symmetrical layer-lines or calculation of the difference of phase angles of the amplitude peaks on the opposite sides of the FT in comparison to what is expected from helical symmetry (DeRosier and Moore 1970). I do not see any reason why these checks of preservation of helical symmetry couldn't be done in addition to other selection criteria used in a classical single-particle approach (plus other selection criteria adapted to the rough geometry of filaments, see part "Introduction into the developed scripts").

Another big difference between classical and single-particle approaches lies in the reconstruction process itself: in the former, the reconstruction is done by using only the Fourier coefficients found on the layer lines, whereas the later uses all Fourier coefficients as for an asymmetrical object. For the single-particle approach the question is the following: if, given a certain helical symmetry that is assumed at some point of the procedure, many Fourier coefficients of the images are not relevant (actually they should be 0 in an ideal noise-free projection), then why do we include them in the 3D reconstruction ? Is this inclusion useless, does it only introduce more noise ?

To take this into account, one may, for example, refine particle orientation by using the single-particle approach, and compute reconstruction on each segment using Fourier-Bessel approach. But then one could go further and ask: why refining particle orientation using all Fourier coefficients and not only the relevant ones ? Of course, the precise position of the relevant Fourier coefficients depends on the orientation, which is the parameter we want to improve using only the positions of relevant Fourier coefficients, which depend on the precise orientation… etc… "like a dog chasing its tail" ! However, we are usually in a slightly different case: the orientation parameters are already roughly known with a precision that can be judged from the resolution that we can obtain by combining our images in a reconstruction. Thus, depending on resolution, one can assume an average error in orientation of the segments, in particular in the in-plane and out-of-plane angles which are relevant here, because the on-axis view doesn't influence the layer lines position. One can take this average error into account to keep more Fourier coefficients than one would keep if the orientation was perfectly known, and use only these coefficients to refine particle orientation. To summarize, what I propose is to do a Fourier-space masking of images in order to keep only relevant Fourier coefficients for a given helical symmetry, thereby reducing effects of the noise from the images. These coefficients are positioned on the layer-lines but not all along each layer line because depending on the Bessel function order on each layer-line, the coefficients near the meridian can also be non-relevant (DeRosier and Moore 1970). In order to take into account the uncertainty in particle orientation, one should then "blur" the mask to avoid removing useful information.

The idea is actually similar to what electron microscopists were doing since the beginning of EM by using optical (and later computational) Fourier-filtering (Klug and DeRosier 1966), not only for helical specimen (**Figure 4.5A ;** (DeRosier and Klug 1968)), but also for 2D crystals (**Figure 4.5B ;** (Kiselev, Lerner, and Livanova 1971)) and for projections along symmetry axis of other type of symmetrical objects. (**Figure 4.5C ;** (Baker, Drak, and Bina 1989)).

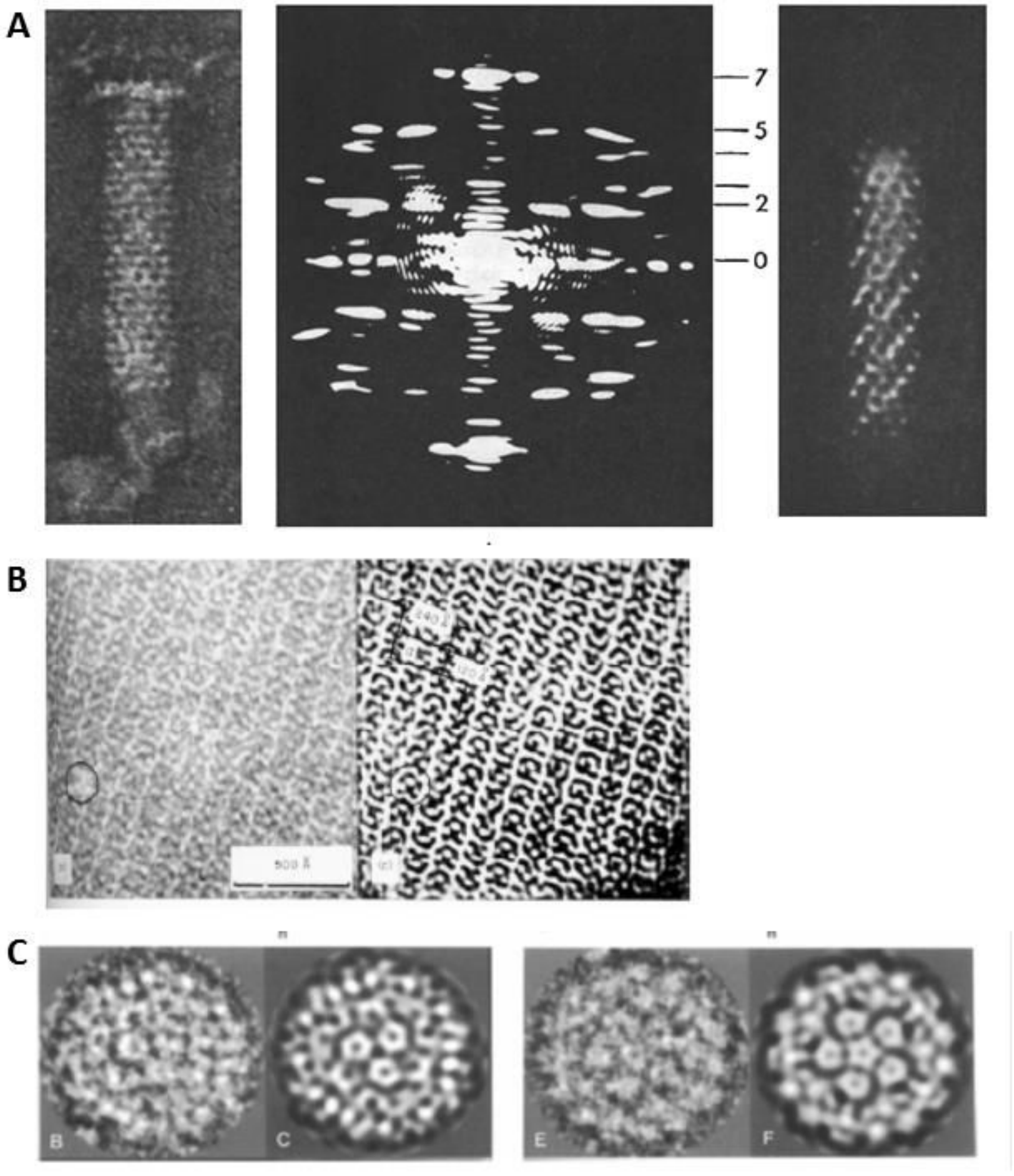


Figure 4.5 : Examples of Fourier Filtering

Example of Fourier Filtering of a helical sample (**A** ; DeRosier & Klug, 1968), a 2D crystal (**B** ; Kiselev, Lerner, & Livanova, 1971), and a icosahedron viewed along symmetry axis (**C** ; Baker, Drak, & Bina, 1989). Precise figure legends can be found in corresponding articles. Briefly, in A : left is the unfiltered image of a phage tail, middle is the Fourier transform of left image on which detected layer lines are numbered and right is the Fourier-filtered image admitting only the diffracted rays from the far side of the particle. In B, left is the unfiltered 2D crystal image of phosphorylase b (protein is black) and right is the same image after filtering. Note the missing particle surrounded by a black circle left appearing in the filtered image right. In C, from left to right : unfiltered image of a SV40 particle viewed along 3-fold symmetry axis ; same after Fourier filtering ; a view along 5-fold symmetry axis ; same after Fourier filtering.

The main difference between those earlier applications of Fourier-filtering and our proposal is that, in particular in case of helices, they were applying this method mostly in order to remove noise for visualization purpose (R A Crowther and Klug 1975) or to make appear separately near and far side of the helical net (Klug and DeRosier 1966), whereas we propose to include the Fourier-masking as a part of the single-particle reconstruction approach for helical structures. One way to include it in the reconstruction process would be:

(0) a 3D reconstruction is first obtained using one's favorite single-particle approach, views are assigned to each image, and symmetry parameters are determined

(1) from the resolution of the reconstruction, one then estimates the average error of views determination (in-plane and out-of-plane)

(2) each image is padded into a larger image, ideally into the largest image as possible to have a finer frequency sampling in Fourier space, and then Fourier-transformed

(3) using the symmetry parameters, the errors on views, the out-of-plane and in-plane angles found for each image, a binary Fourier-mask of the size of the padded image in (2) is created for each image which only contains relevant "blurred" layer-lines and on each layer-line only relevant Fourier coefficients (given order of Bessel function).

(4) The FT of the image (2) is multiplied by the created Fourier-mask

(5) The multiplication product is back-Fourier-transformed and an image of the original size is cropped out of this large Fourier-filtered image

(6) these Fourier-masked images are used for a refinement of view determination by projection matching using the previous structure as reference, a new 3D structure is calculated and the steps (1) to (6) are repeated using newly determined values for the views and resolution. The process stops when no changes in views determination and 3D structure are detected.

Additionally, one could also Fourier-filter the projections of the reconstruction used as reference for PM. The advantage would be the exact knowledge of the views. All details of the method should be explored and tested using simulated and real data.

PART 4 : Introduction into the developed scripts

Preliminaries

One of the aims of this thesis, since the host laboratory had only little experience in helical reconstruction at the time of its beginning, was to set up a dedicated pipeline for the image processing. It consisted thus in setting up the known/used procedures and eventually to improve them and add new processing procedures. Thus, part of the work consisted in writing scripts to use this pipeline in an efficient manner by any user in the lab. I will detail here, for some of the most important scripts, the input arguments to give, as well as the output files. Furthermore, under the light of what was written in the main text of the manuscript, some advices on how to optimally use the scripts, and the critical points to take care of, will be given.

Considering the preprocessing of images, a set of scripts was written to box particles with chosen parameters (size of box, overlap, distance of boxes to extremities of filament), determine and correct for CTF (based on CTFFIND3), Fourier-filter images, eventually verticalize or mask the images… but these script will not be described here as this part of processing was not yet optimally designed, and as it is a part that any user may want to perform in his own manner (for example the way of correcting the CTF). Some simple scripts were also designed for image classification with IMAGIC, and others to extract from the IMAGIC classification outputs, the information needed to create files telling the SPIDER package which images had to be used in the reconstruction (selection file), as a function of their class number. Those scripts are not coupled together to form a defined pipeline because they still require intermediate manual steps, and therefore they will not be presented here. The ones that attained a certain degree of maturity, and that will be detailed in this part, are:

-the master script to make the 3D reconstructions by projection matching (PM) and all related steps

-the script for helical symmetry parameters estimation on 2D projections, that runs the method largely described in the part 2 of this manuscript.

-a script for ab initio helical symmetry parameters estimation on 3D volume

Reconstruction pipeline using the script helix_rec.csh

General organization / purpose

Once the images are prepared in the desired way (CTF-corrected, filtered, masked, verticalized or not), and once an initial model is available (a solid cylinder, a smooth helix, or a model derived from the symmetry determination on 2D projections), one is ready to begin with the reconstruction procedure. In order to adapt it to different projects and strategies, one single script dedicated to the reconstruction and offering a lot of flexibility to the user was created. Once this master script is launched, the user does not need to stop it until a final reconstruction is obtained because it gives the possibility to vary any parameter between each projection matching cycle, and to test different ways to calculate the reconstruction with the possibility to undo some steps if necessary. Therefore, one of the advantages of using this script is that the user does not need to work with many different scripts and manually edit parameters inside these scripts. However, most of the scripts are usable independently of the master script if necessary. The general organization of the processing pipeline and the way the most important scripts are connected to the master script are shown on **Figure 5.1**. Comparing to the original IHRSR procedure (Egelman, 2000), the main advantages are :

-the interactivity (via the subroutine eliminate_images.csh detailed later)

-the possibility to use many different images selection parameters

-the multiple ways of taking symmetry into account or not

-at each step of the processing, a trace of which exact parameters were used is kept

-statistics are kept in an easily readable format for each iteration cycle of PM

-the plotting interface to assess data quality/consistency

-no need to open and edit any script (SPIDER or other) manually

-thus the possibility of easily testing different sets of parameters and their effects

-a more appropriate way not to overestimate resolution by FSC in case of overlapping boxed segments from filaments

-the implemented parallelization scheme for use on CPU clusters.

There are still many possibilities to improve this part of the processing, but I only show here what is ready to use accompanied by comments on what can be modified.

The basic workflow of the master script is shown at **Figure 5.1**: 3D model projections (step 1), alignment parameters search by projection matching (step 2), alignment of images (step 3), calculation of a normalized CC between aligned images and corresponding projection (step 4), selection of images to include in the reconstruction and parameter setting (step 5), and reconstruction taking the symmetry into account (steps 6 and 7). When launching the script for the first time, the user has to enter parameters for steps 1 to 4, but all the values can be changed later if wished during step 5. There is otherwise no user prompting between steps 1 and 5, only optionally after step 6 (reconstruction). As we will see, it is also possible to deactivate the interactive prompting at any time to run the scripts in a more automatic manner.

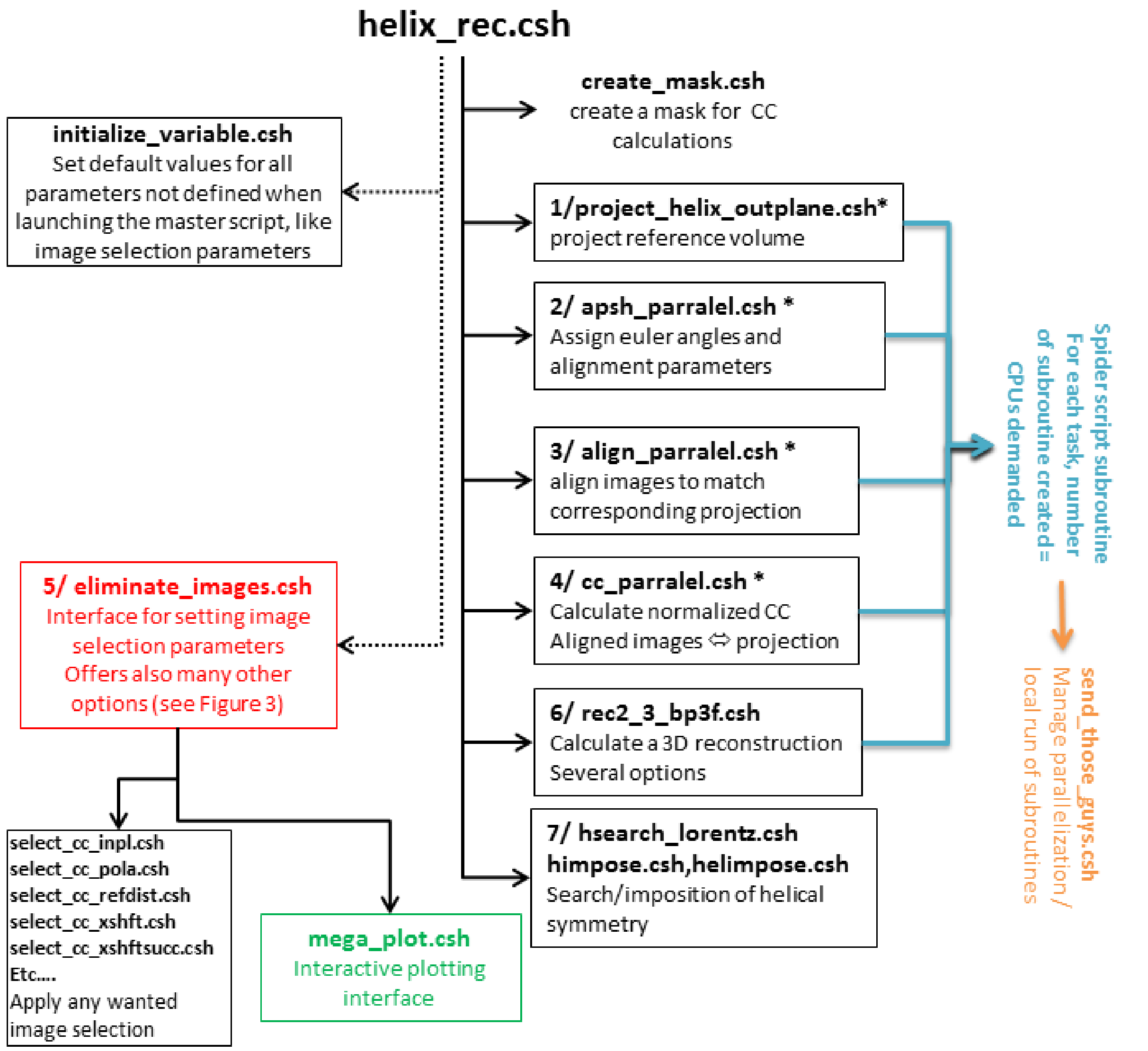

helix_rec.csh
create_mask.csh
create a mask for CC calculations
initialize_variable.csh
Set default values for all parameters not defined when launching the master script, like image selection parameters
1/project_helix_outplane.csh*
project reference volume
2/ apsh_parralel.csh *
Assign euler angles and alignment parameters
3/ align_parralel.csh *
align images to match corresponding projection
4/ cc_parralel.csh *
Calculate normalized CC
Aligned images ⇔ projection
5/ eliminate_images.csh
Interface for setting image selection parameters
Offers also many other options (see Figure 3)
6/ rec2_3_bp3f.csh
Calculate a 3D reconstruction
Several options
7/ hsearch_lorentz.csh
himpose.csh,helimpose.csh
Search/imposition of helical symmetry
select_cc_inpl.csh
select_cc_pola.csh
select_cc_refdist.csh
select_cc_xshft.csh
select_cc_xshftsucc.csh
Etc....
Apply any wanted image selection
mega_plot.csh
Interactive plotting interface
Spider script subroutine
For each task, number of subroutine created= CPUs demanded
send_those_guys.csh
Manage parallelization / local run of subroutines

Figure 5.1 : Relationships between the master script for reconstruction of helical specimen helix_rec.csh and related most important subroutines

The distance of the subroutines to the master script helix_rec.csh indicates their relative launching time. As will be clarified in the text, some steps are facultative. Steps indicated by an asterisk are adapted for parallelization on a cluster.

→ Call subroutine

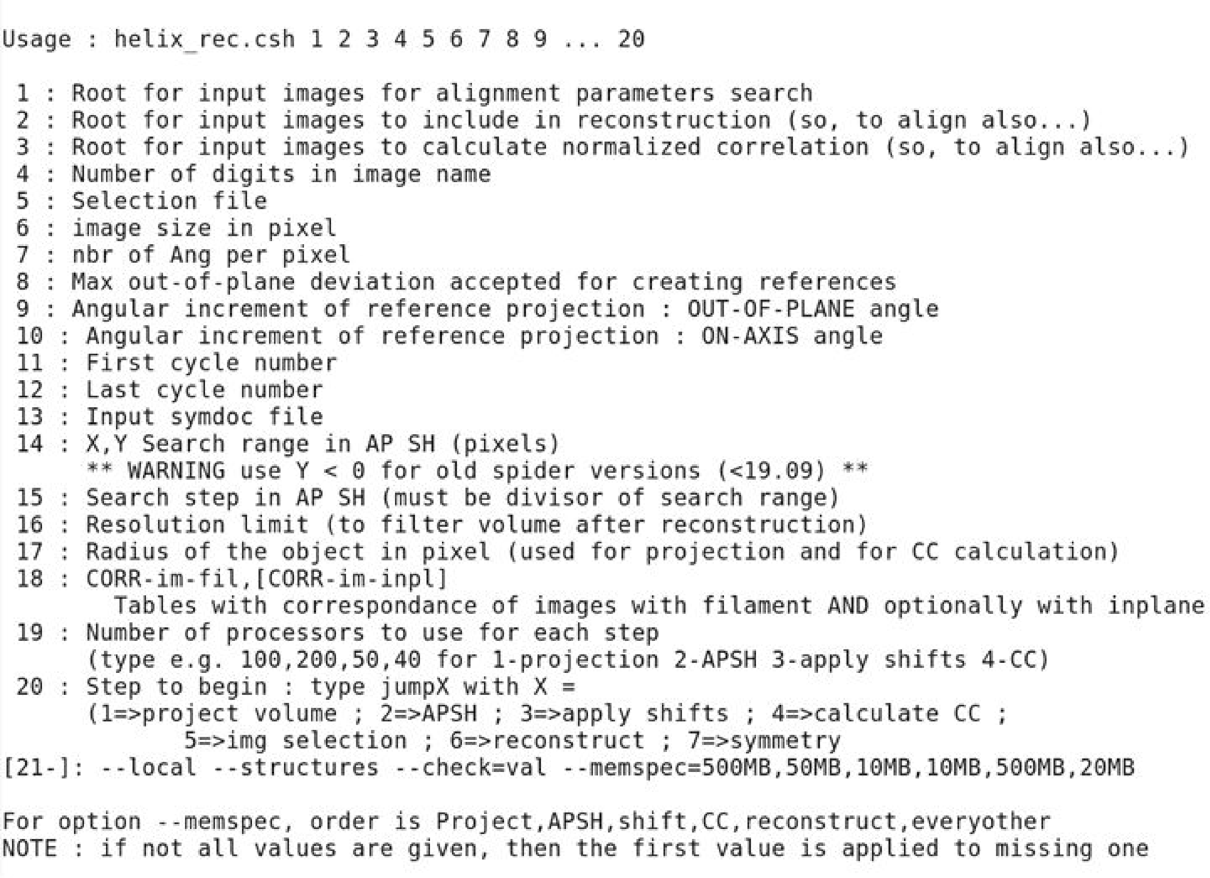
⋯> Source subroutine (to transfer variable content)

⇒ Create subroutine from an existing template (replace patterns by wished arguments)

⇒ Give as input

Launching the master script for the first time

I will now detail the parameters that need to be entered as arguments when launching the script for the first time (**Figure 5.2**). All these parameters can be changed later. All other parameters are set up later (step 5). To keep a trace of the options used, all arguments entered are stored in a file called my_options_x.txt with x being a digit incremented for each new instance of the script launched.



```
Usage : helix_rec.csh 1 2 3 4 5 6 7 8 9 ... 20

 1 : Root for input images for alignment parameters search
 2 : Root for input images to include in reconstruction (so, to align also...)
 3 : Root for input images to calculate normalized correlation (so, to align also...)
 4 : Number of digits in image name
 5 : Selection file
 6 : image size in pixel
 7 : nbr of Ang per pixel
 8 : Max out-of-plane deviation accepted for creating references
 9 : Angular increment of reference projection : OUT-OF-PLANE angle
 10 : Angular increment of reference projection : ON-AXIS angle
 11 : First cycle number
 12 : Last cycle number
 13 : Input symdoc file
 14 : X,Y Search range in AP SH (pixels)
      ** WARNING use Y < 0 for old spider versions (<19.09) **
 15 : Search step in AP SH (must be divisor of search range)
 16 : Resolution limit (to filter volume after reconstruction)
 17 : Radius of the object in pixel (used for projection and for CC calculation)
 18 : CORR-im-fil,[CORR-im-inpl]
        Tables with correspondance of images with filament AND optionally with inplane
 19 : Number of processors to use for each step
      (type e.g. 100,200,50,40 for 1-projection 2-APSH 3-apply shifts 4-CC)
 20 : Step to begin : type jumpX with X =
      (1=>project volume ; 2=>APSH ; 3=>apply shifts ; 4=>calculate CC ;
             5=>img selection ; 6=>reconstruct ; 7=>symmetry
[21-]: --local --structures --check=val --memspec=500MB,50MB,10MB,10MB,500MB,20MB

For option --memspec, order is Project,APSH,shift,CC,reconstruct,everyother
NOTE : if not all values are given, then the first value is applied to missing one
```

Figure 5.2 : Terminal printout when launching the master script helix_rec.csh without enough arguments

Details for each argument are given in the text

1, 2 and 3 : These arguments are the root to the input images for alignment parameters search (1), reconstruction (2) and calculation of the normalized cross-correlations (CC) (3). Very often, one uses the same images for these three steps, but one may want to do differently, like for example by using masked images for alignment parameters search and CC calculations, but including unmasked images in the reconstruction. Alternatively, one may wish to use different Fourier filters for the images for alignment parameters search and for the ones to include in the reconstruction, etc.

The format of input images is the SPIDER format, and they should be individual files (not in a stack), with a numbering containing always the same number of digits (like for e.g img_00001.spi ; img_00100.spi etc.) that is given in argument **4**. The root to the images consists in the absolute or relative path to the images without the extension and without the digits defining the image numbers.

5 : A SPIDER selection document file (Joachim Frank et al. 1996). This file contains the list of images to consider for all steps of the processing. If one wishes to consider all the images given as input, one can use the script mk_seldoc.csh to automatically generate a selection file with all images present in a specified folder. Otherwise, it can contain only some images, for example, the ones selected by a classification step. At each PM cycle, a different selection file can be used (if the user asks for), or created according to a selection criteria from alignment parameters and statistics.

6 and 7 : Image size in pixels and pixel size (in Å per pixel) respectively.

8 : This is the maximum out-of-plane angle (in degrees) that will be used to create the references for projection matching. Entering 0 here means that no out-of-plane will be used, similarly, by entering 12, only references with out-of-plane of maximum plus or minus 12 degrees will be created. Usually, this value cannot be estimated a priori, but one can check the distribution of images according to the out-of-plane angle of references after one cycle of projection matching to verify if the value entered here was big enough. One can also choose to limit this value to gain calculation time. In this case however, images matching the highest out-of-plane angle imposed should be excluded from the reconstruction because they would also probably contain images that have a higher out-of-plane angle.

9 and 10 : This is the angular increment used for the out-of-plane rotation angle , and for the on-axis rotations, respectively, to generate reference projections from the input volume. A way to calculate this value is to decide which resolution one is aiming to and use the geometry of a single-axis tilt series as in (R. A. Crowther, DeRosier, and Klug 1970): the number of equally spaced projections needed for an object of diameter **D** to obtain a resolution of **d** is ***πD/d***. The minimum increment (in degrees) is therefore **(180*d)/(*πD*)**, but one would typically use a slight oversampling. With helices, it is often the case that the dimension along the helix axis is bigger than the direction perpendicular to the helix axis, and so one may want to use a different value for these arguments **9** and **10**. One may also accept to have a lower resolution at the top and bottom edges of the structure (considering that the projection has the helical axis aligned vertically) to reduce the angular increment for the out-of-plane angle in order to reduce the calculation time.

11 and 12 : respectively first and last projection matching cycle number. The script uses a numbering to identify which files were created at a given PM cycle. At the cycle number 'i', the volume called rec_sym_{i-1}.spi is used as input. **Thus, although not entered as an argument, an initial volume called rec_sym_{i-1}.spi must be present in the current directory.** All the output files, except the reconstructions (before/after the optional symmetrization step), are stored in a folder called c'i'. As for each of the other arguments, the last cycle number can be changed later if needed.

13 : Input file containing initial guess for symmetry parameters, in a format readable by Egelman's programs hsearch and himpose. Although this file is not always needed (when one does not wish to impose and/or search for the symmetry of the 3D volume), it is asked here as an obligatory argument. This will be changed in the future.

14 : This is the **search range in x and y** in pixels for alignment parameters search (SPIDER command AP SH). In the SPIDER versions before the 19.09 release (the one used during this work), only one value defining both x and y translation search range is used by the alignment parameters search command APSH. Thus, one should enter here one value as the first part of the argument (x) and a negative value for y (not taken into account). The minimum translation search range is usually not easily predictable. Therefore one should run a first PM cycle with a big search range and use the resulting distribution of x shifts (perpendicular to helix axis) to restrict this value (the y shifts, along helix axis, are not so informative for helices as they do not reflect the centering of helix segments). Using a range limit slightly bigger than the maximum ensures that every image can be properly aligned.

15 : This is the **search step** in pixels for alignment parameters search. It should be a divisor of both the search range in x and y. To set this value, one has to take into account the desired resolution (basically one can multiply the search step in pixels by the pixel size to have an idea of the precision of the search for a given step, although the real precision is higher, as the algorithm in AP SH makes a sub-pixel refinement).

16 : This is the resolution limit in Å to filter the reconstruction during each cycle of PM. The input images are not affected by this filter.

17 : Radius of the object in pixels. This value is used to restrict the projection of the volume to this size and to create a mask for CC calculations. This is useful when one expects some flexibility in the filament of interest and wants the alignment to focus on the central part of the images because the 3D reconstruction might be of poor quality on its extremities.

18 : Here one, or optionally two, tables have to be entered, with a number of lines corresponding to the number of images to use. The first one is a table that assigns each image to a filament number. It is created during the preprocessing, after particle boxing, in order to keep a trace of the correspondence images/filaments (using the script correspondence_img_fil.csh). It can be useful to calculate statistics for individual filaments and it is used during the reconstruction process to separate images into two sets to calculate the FSC while avoiding to include overlapping images of the same filament into the two different volumes (whereby avoiding overestimation of the resolution). It is also used for other image exclusion criteria as will be shown later. The second table is the table that associates each image to an in-plane angle, as could be deduced from the boxing: the angle assigned to each segment of a filament is the one defined by the position of the extremities of the filament. This table is also created during preprocessing (by the script correspondence_img_inpl.csh), and it will be used during the 3D reconstruction procedure to check if the in-plane angle found by projection matching is not too far from the one we already roughly know (Sachse et al., 2007), as well as for polarity checks (Fujii, Kato, and Namba 2009b). This table is optional, as one can use the script on images already verticalized, in which case such table would be useless.

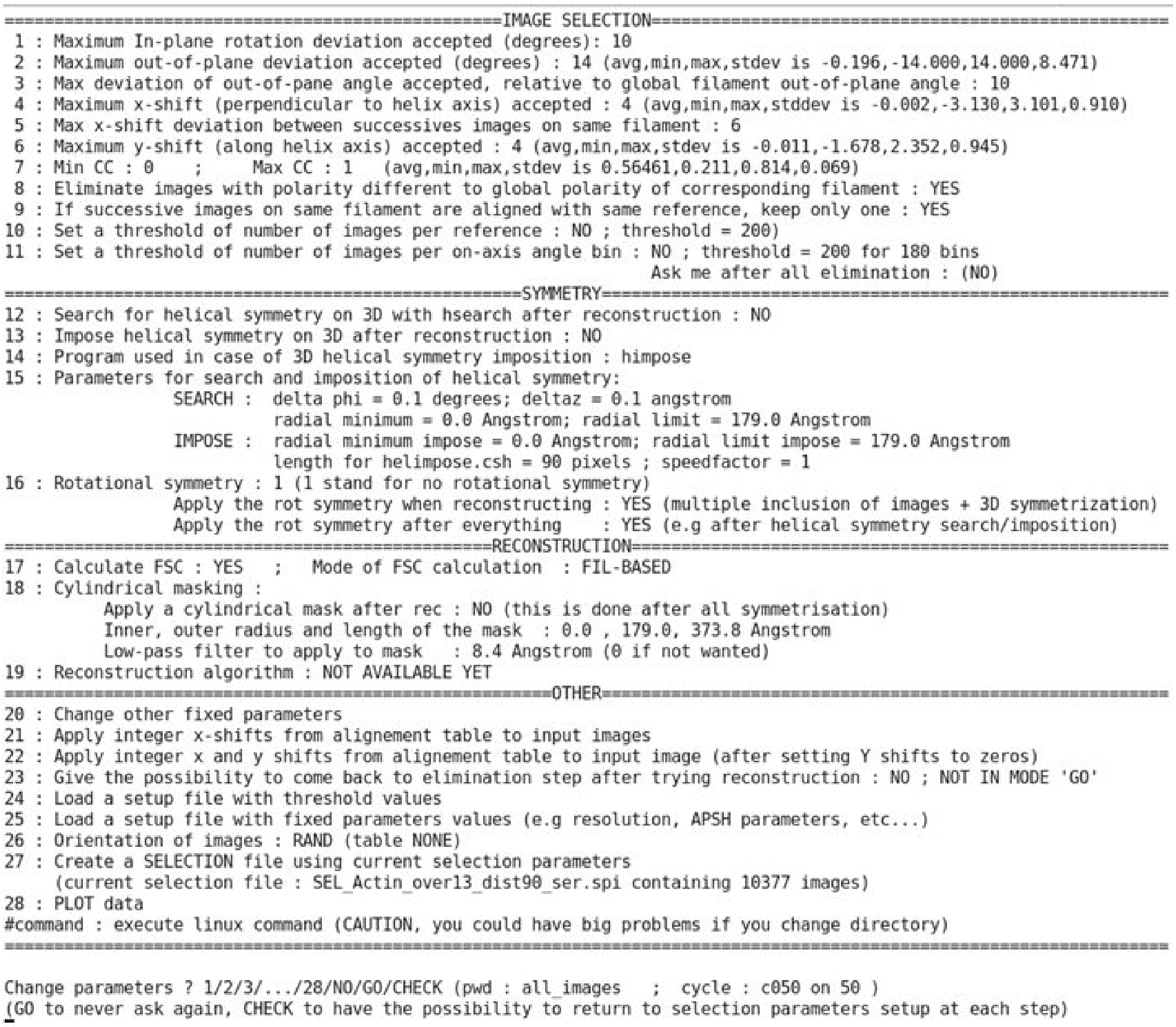
19 : Number of processors to use, for parallelization purpose. Almost every step of the process is parallelized, i.e. projection of the volume, alignment parameters search, alignment of images and normalized CC calculation. One can thus give a different value for each of those steps, separated by commas. If only one value is given, then the same value is used for all steps. Usually the most time consuming step is the alignment parameters search. That would thus require a higher value than the other steps. For the other steps, setting to a too large value can lead to reduced global computation time due to overhead.

20 : This specifies the step of the processing where the script should begin. If the script has, for any reason, being interrupted in the middle of a PM cycle, one can restart it at the step where it has been stopped by typing jumpX where X is the step to begin with, as indicated on **Figure 5.2**. If one starts normally, one should thus type “jump1”.

21 and over : These are optional arguments, that will not be detailed here, and that concern the currently used parallelization systems (like memory requirement specification). The one to note, however, is the option --local that ensures that every step runs locally, eventually on several CPUs if asked accordingly in option 19.

The user-interaction interface and the modes of interactivity

Once the first 4 steps of the first cycle of PM had been done, and later between every PM cycle if the interactive mode is still on (we will see how to set it to off), a histogram of CC between reprojection of current model and each aligned image is displayed fur the current cycle, and the interactive interface shown in **Figure 5.3** appears (it corresponds to the terminal output of the subroutine eliminate_images.csh shown in **Figure 5.1**).



```
=================================================IMAGE SELECTION=================================================
 1 : Maximum In-plane rotation deviation accepted (degrees): 10
 2 : Maximum out-of-plane deviation accepted (degrees) : 14 (avg,min,max,stdev is -0.196,-14.000,14.000,8.471)
 3 : Max deviation of out-of-pane angle accepted, relative to global filament out-of-plane angle : 10
 4 : Maximum x-shift (perpendicular to helix axis) accepted : 4 (avg,min,max,stddev is -0.002,-3.130,3.101,0.910)
 5 : Max x-shift deviation between successives images on same filament : 6
 6 : Maximum y-shift (along helix axis) accepted : 4 (avg,min,max,stdev is -0.011,-1.678,2.352,0.945)
 7 : Min CC : 0    ;     Max CC : 1   (avg,min,max,stdev is 0.56461,0.211,0.814,0.069)
 8 : Eliminate images with polarity different to global polarity of corresponding filament : YES
 9 : If successive images on same filament are aligned with same reference, keep only one : YES
10 : Set a threshold of number of images per reference : NO ; threshold = 200)
11 : Set a threshold of number of images per on-axis angle bin : NO ; threshold = 200 for 180 bins
                                                                  Ask me after all elimination : (NO)
====================================================SYMMETRY=====================================================
12 : Search for helical symmetry on 3D with hsearch after reconstruction : NO
13 : Impose helical symmetry on 3D after reconstruction : NO
14 : Program used in case of 3D helical symmetry imposition : himpose
15 : Parameters for search and imposition of helical symmetry:
                SEARCH :  delta phi = 0.1 degrees; deltaz = 0.1 angstrom
                          radial minimum = 0.0 Angstrom; radial limit = 179.0 Angstrom
                IMPOSE :  radial minimum impose = 0.0 Angstrom; radial limit impose = 179.0 Angstrom
                          length for helimpose.csh = 90 pixels ; speedfactor = 1
16 : Rotational symmetry : 1 (1 stand for no rotational symmetry)
                Apply the rot symmetry when reconstructing : YES (multiple inclusion of images + 3D symmetrization)
                Apply the rot symmetry after everything    : YES (e.g after helical symmetry search/imposition)
=================================================RECONSTRUCTION==================================================
17 : Calculate FSC : YES   ;   Mode of FSC calculation  : FIL-BASED
18 : Cylindrical masking :
         Apply a cylindrical mask after rec : NO (this is done after all symmetrisation)
         Inner, outer radius and length of the mask  : 0.0 , 179.0, 373.8 Angstrom
         Low-pass filter to apply to mask   : 8.4 Angstrom (0 if not wanted)
19 : Reconstruction algorithm : NOT AVAILABLE YET
=====================================================OTHER=======================================================
20 : Change other fixed parameters
21 : Apply integer x-shifts from alignement table to input images
22 : Apply integer x and y shifts from alignement table to input image (after setting Y shifts to zeros)
23 : Give the possibility to come back to elimination step after trying reconstruction : NO ; NOT IN MODE 'GO'
24 : Load a setup file with threshold values
25 : Load a setup file with fixed parameters values (e.g resolution, APSH parameters, etc...)
26 : Orientation of images : RAND (table NONE)
27 : Create a SELECTION file using current selection parameters
    (current selection file : SEL_Actin_over13_dist90_ser.spi containing 10377 images)
28 : PLOT data
#command : execute linux command (CAUTION, you could have big problems if you change directory)
=================================================================================================================

Change parameters ? 1/2/3/.../28/NO/GO/CHECK (pwd : all_images   ;  cycle : c050 on 50 )
(GO to never ask again, CHECK to have the possibility to return to selection parameters setup at each step)
```

Figure 5.3 Terminal printout for the interactive step of image selection and parameters setup, from script helix_rec.csh (subroutine eliminate_images.csh)
See text for details

As shown on the **Figure 5.3**, the interface is composed of four parts. The first part sets up the parameters used for selection of images to be included in the reconstruction and/or eventually to create a new selection file (see option **27**). The second is to set up parameters about the way helical symmetry is taken into account (or not), and/or additional rotational symmetry around Z. The third is related to reconstruction parameters like FSC calculation, and the fourth offers various other options as seen in Fig 3. To choose an option or modify parameters, one needs to enter the corresponding number (as asked at the bottom part of the interface), and questions will then be asked to the user through the terminal. To each option there are one or more related questions. For the sake of space, I cannot detail each of those here. Nevertheless, I tried to make the questions as clear as possible to the user.

When one finishes the setup of parameters (detailed below), one out of three modes can be selected :

-'NO', and no other questions will be asked until the next PM cycle at the same elimination and parameters setting step. An exception is when one has set the option **23** to 'YES', in which case the user has the choice to come back to the elimination and parameters setting step after a reconstruction trial.

-'CHECK' : During the setup of the image selection parameters, one can at any time check how many images will be included in the reconstruction, how many images are excluded by each selection step and how the CC histogram profile (and CC statistics) will be affected by each elimination type. To do this one can type "CHECK" and the mentioned information will be displayed on the terminal and plotted in separated windows.

-'GO' if one wishes to keep all parameters as they are, and just iterate the PM cycles until the last cycle. This will unset the interactive mode and no more prompting will be done unless there is a crash in the procedure (e.g. if zero images are selected for reconstruction by the current selection parameters), or if the user stops the mode "GO" by erasing the file called "CURRENTMODE" located in the folder where helix_rec.csh is running. In this case, in the next image elimination step, the usual user prompting will be done. Alternatively, it is possible to force the script to be in mode "GO", even before reaching the elimination (step 5). To do so, put a tcsh file called "thrsinput" containing all needed variables (a template file for "thrsinput" is available) in the current folder: this file will then be sourced at the elimination step and no prompting will be done anymore. In this case, if the user wants to quit the mode "GO" later in the PM iterations, she/he should remove this file in addition to "CURRENTMODE".

Before setting up the parameters for image selection, one usually needs to plot the information available from the alignment parameters table and check the evolution of PM statistics through the iterations. This is done using the option **28** that will open a new interface with the different plotting options (see **Figure 5.4** for the plotting interface).

Image selection parameters

Especially when dealing with helices, many different image selection criteria can be added to the basic CC criteria. The selection options that are currently included in the script are described below. For each option, default values (as "reasonable" as possible) are shown at the first printout of the "menu" as illustrated by **Figure 5.3**. Thus, none of the options described below are "mandatory fields". If some parameters were interactively changed (by choosing the corresponding option number and by answering the appearing question(s)), the menu is reprinted on the terminal with updated values. The last line of the parameters entries, beginning with '#command', specifies that one can also type any command using the shell syntax starting the line with the character '#' (e.g. '#ls'). This command will be executed and its output printed on the terminal, and the elimination and parameters setting interface will be reprinted.

The list of options :

1 : The **maximum in-plane angle deviation** allowed, in degrees, either from 0 or 180° if the input images are vertical or from the mean in-plane angle of the corresponding filament, if the images were not verticalized. To tell the script if the images are verticalized or not, one needs to set up option **21** and, if the images are not verticalized, one then needs to give a table with the correspondence between images and the mean in-plane angle of the filament (from boxing) if this was not given when launching the script helix_rec.csh. A too large deviation of in-plane angle is an indication of a wrong alignment of images or a high flexibility of the filament.

2 : This is the maximum **out-of-plane** angle deviation allowed **from 0°**. On the right side of the text line corresponding to this option some statistics for this parameter is displayed (between brackets) : the average value (avg), the minimum and maximum value (min and max) and the standard deviation (stddev). This statistics is also displayed for parameters **4**, **5** and **7**. This information is useful for detect weird ion of behavior of the alignment, for example, when the average (for parameters **2**, **4** and **5**, out-of-plane, x and y shifts) is far from 0. To setup a limit for the out-of-plane angle deviation, one needs to plot the distribution of images according to the out-of-plane angle. If this distribution shows no clear fall-off towards higher out-of-plane angles, with nearly no images attributed to the highest angles, one has to stop and analyze. Such a behavior probably means that the chosen maximum out-of-plane angle to create the reference projections was not big enough, so the images with bigger out-of-plane angle had fallen into the highest out-of-plane subset, and/or that some proportion of “bad” particles (wrongly aligned) has fallen in this subset. In any case, one should then set the limit to the last angle that manifests a notable fall-off of the distribution of images per out-of-plane angle. It should also be noted that if a set of images is heterogeneous in pitch, it might affect the distribution of out-of-plane angles (images with a lower pitch than the model will match projections with higher out-of-plane angles). Therefore, this issue should be clarified before the reconstruction step.

<u>**3 :**</u> Maximum deviation of **out-of-plane** angle **from the global** out-of-plane angle of corresponding filament. The global out-of-plane angle is defined as that of the majority of segments for each filament. This criteria is important as a big deviation can indicate a wrong alignment of an image or a curvature of a particular filament out of the plane (a flexible filament). Setting this value to 0 ensures that the selected images will correspond to filaments that are straight enough (especially when this selection is coupled to another selection based on the in-plane angle, see **1**), and that the alignment of these images is consistent within the filament orientation (a sign of good alignment).

<u>**4 :**</u> **Maximum x-shift** (perpendicular to helix axis) allowed, in pixels. Again, one should first plot (option **24**) the x-shift distribution before setting this value. If the distribution shows that many images have a x-shift that is at the limit of what was allowed by the alignment parameter search range, it may indicate that these images are indeed strongly shifted, or that they are badly aligned and thus should be excluded from the reconstruction. Ideally, no images should be shifted by the maximum x-shift allowed for the search, as this maximum should be set up slightly over the expected maximum.

<u>**5 :**</u> **Maximum x-shift** (perpendicular to helix axis) allowed **between successive segments from same filament**, in pixels. Due to the geometry of helices, successive images on the same filament should have very similar shifts perpendicular to the helix axis (x-shifts). If the projection matching has found too different successive x-shifts, it can be a sign of wrong x-shift assignment. However, for curved filaments, or to take in account an imprecise boxing (with the axis defined by the center of successive boxes along the filament deviating from the true axis of the filament), here one can allow a small difference between x-shifts of successive segments. One should also consider that for small distances between the boxed filament segments (higher the overlap between boxes), this threshold should be small.

6 : **Maximum y-shift** allowed (parallel to helix axis), in pixels. Big y-shifts values or y-shifts at the limit of the search range are, contrary the x-shifts, not an indication of a bad alignment. However, allowing too big y-shifts can produce a bias in reference distribution. Indeed, when an image shifts in the y direction to match a projection, it matches with an on-axis view that is further away from the on-axis view of the unshifted image (because y translations are equivalent to on-axis rotations). Given the fact that some on-axis views naturally tend to give higher correlation with images (due to interpolations, for example for views at 0° , 90° , etc ; or to a slightly asymmetric reconstruction), the images will tend to shift in y direction to match with these projections, whereby leading to a bias of the on-axis distribution. However, one can choose to not limit the y-shifts but to take care of the on-axis distribution (see selection option **11**).

7 : **Minimum and maximum CC** accepted between images and projections of the current model. This is an obvious selection criteria, although one has to keep in mind that the CC might also depend on factors like for example the defocus used when acquiring the micrograph. To set up these two threshold values, one should look at the histogram of CC that is automatically displayed when the script arrives at the step 5, and also have a look at the CC as a function of image number (which illustrates the variations among different micrographs, telling us for example in which extent the defocus influences the CC in our set of images). It can seem surprising to give the possibility to set an upper limit to the CC, but this can be used for example when the histogram of CC distribution show two (or more) distinct populations among images, to reconstruct separately the low CC population and the high CC population.

8 : This option is to ensure that every image from the same filament included in the reconstruction shows the **same polarity** (orientation in regard to the filament orientation). As the global polarity of each filament is not known in advance, it is defined as the one of the majority of the segments for each filament during each alignment. This option should naturally not be used in case of apolar filaments. Alternatively, one can see how many images would be discarded when this option is set to YES (using the CHECK mode): in the case of a high number (close to half of the total number of images), one would have an indication of an apolar structure.

9 : Especially when the translation search range is large, successive overlapping images might match with the same projection while being translated by a different y-shift. In that case, I offer here the possibility to keep only one of the successive images with this characteristics. Indeed, they would otherwise just be included twice.

<u>10 :</u> This option is used to set a maximum number of images per matching reference to keep. It was created when no out-of-plane angle was used in the procedure, such as setting a limit of number of images matching with each reference was equivalent to limit the number of images per on-axis angle (option **11** is now used for this purpose). Thus, it is now somehow obsolete, except if one needs for any reason to keep the same amount of images per out-of-plane angle for example.

<u>11 :</u> This option is used to set a maximum number of images to keep per on-axis angle. As we saw in the main text, this selection step is particularly crucial to be able to make reconstructions without symmetry imposition while avoiding that the model becomes more and more asymmetric, with some view angles that are always more populated. In adduition, even when using a symmetry search (and imposition) step, a model that is more regular in terms of on-axis views distribution will better preserve the symmetry, therefore making the symmetry search task easier. To set a threshold value for the number of images to keep per on-axis angle, since the distribution of angles is not continuous, I rather ask for a number of images to keep per on-axis bin : that means that if, for example, 90 bins are chosen, the 360° view space is divided in 90 windows of 4 degrees each. Only the desired number of images is then retained for each window: the selection criterion between images in each window is simply the CC, as no better obvious criterion was found. To level out the number of images per on-axis angle in an optimal way, one should use a number compatible with the angular increment used for the on-axis angle variation for projecting the model, by dividing the 360° on-axis view space by this increment: this optimal number is the default value for this option. It is also asked here if one wishes to set this number of images to keep per on-axis bin "now" or after performing all the eliminations according to the other selection criteria. If one choses "now", a histogram showing the distribution of the number of images is displayed (with the number of bins in the histogram equal to the number of bins one wishes to consider to eliminate images), and one can set a threshold value. However, this distribution will consider all the images, and thus can be different from the one obtained after all the other image selections, in particular the ones concerning the absolute numbers of images in each bin. To overcome this problem, one can ask the script to select the threshold value later: in that case, the distribution histogram is displayed after the other eliminations are done, and the interactive threshold setting is done using this updated distribution.

Symmetry related parameters

In addition to the selection parameters, one can select different options concerning the symmetry (**12** to **16**), if the default values are not suitable. Again, all the values that will be entered here can be changed during the PM cycles if wanted.

12 : Here one can choose to **use** or not use the **hsearch** program (E H Egelman, 2000) to search for a pair of refined helical parameters from initial symmetry guess parameters (defined in the symdoc file given as input to the script, which can be changed at any time by option **15**). Differently to what is proposed in the original IHRSR procedure (E. H. Egelman, 2007), one can use here the program for symmetry search independently from the program that imposes the symmetry. This can be useful for example if one wishes to wait until the symmetry parameters found by the hsearch program are stable through successive PM iterations before applying them, which would help to avoid imposing a symmetry far from reality.

13 : This options specifies if one wishes to **impose** the helical **symmetry** on the reconstructed 3D volume. The symmetry that will be imposed will be read from the symdoc file given as input to helix_rec.csh, and eventually updated by the hsearch program. Therefore, one can enforce a symmetry in every cycle, e.g. if one has determined it using other means, without using the hsearch program at all.

14 : Up to now, two possibilities to impose the symmetry are proposed. The user can opt to run either the himpose program (E H Egelman, 2000) or a script (helimpose.csh) that applies the symmetry through averaging of many volumes generated from the input volume by rotations and translations as defined by the symmetry. The himpose program is much faster, but our experience showed that it can crash in some particular situations (depending of the size of the input/output volumes, distribution of density in the initial volume, etc), so that the helimpose.csh script can be preferred. I personally could not detect major differences in the output volumes generated by both approaches. To impose the symmetry using multiple inclusion of each 2D input image as described in (Sachse et al., 2007), I designed individual scripts, but more tests and troubleshooting are needed before including this option in the presented script.

15 : Here the user can enter parameters for search and imposition of the helical symmetry for programs hsearch, himpose and helimpose.csh. These include :

- The **increment** that is used when searching around the rotation per subunit value (**delta phi**, in degrees) and around the rise per subunit value (**delta z**, in Å) in hsearch . One should note that the search is made over a range of -10 to +10 times the search increment.

- The **inner** and **outer radius** for the helical search and imposition with hsearch and himpose / helimpose.csh. In general, the inner radius will be 0.0, except for hollow tubes, where it will have a value greater than 0. Contrary to the original IHRSR procedure, one can use a different value for the search and for the imposition. This allows to perform the search on a part of the helix that follows the expected symmetry, for example in the case of a partially decorated filament, and to impose the symmetry on the whole structure.

- The **length** of the volume to consider for **helimpose.csh** in pixels. The script helimpose.csh can use only a sub-volume of the input reconstruction to calculate all the translated-rotated versions to average. The length of this sub-volume (i.e. distance along helical axis) is given here. The "**speedfactor**" parameter specifies that the translated-rotated volumes to average will be created using the transformation (speedfactor * axial rise) and (speedfactor * angular rotation). Although quicker, using a value greater than 1 here will affect the quality of the symmetrized map (the helical symmetry will not be perfectly respected).

16 : When the structure shows an additional **rotational symmetry around the helical axis**, it can be entered here. The symmetry can be imposed when reconstructing the volume, using multiple inclusions of each image and additional 3D symmetrisation, and/or after helical symmetry has been imposed on the 3D volume. In the future, the possibility to enter a rotational symmetry perpendicular to the helical axis will be added to the script.

Reconstruction/FSC/3D-masking related parameters

The options 17 to 19 set miscellaneous parameters related to the reconstruction process:

17 : This option is used both to tell the script if one wishes to calculate a Fourier Shell Correlation (FSC) curve at the reconstruction step and how the images have to be split to

generate the two volumes used for the FSC calculation. Although FSC-based criteria (Harauz and van Heel 1986) are widely used to assess resolution of reconstructions (and also very widely discussed (Marin van Heel and Schatz 2005)), we do not necessarily need a resolution estimation at each cycle of reconstruction, especially during earlier projection matching iterations. Furthermore, its calculation significantly slows down the procedure because three reconstructions have to be calculated instead of one, thus systematically setting this option to YES may be inefficient. Splitting the images for calculating "independent" reconstructions is often done by separating the data set into "odd and even numbered" images, or by splitting into two halves (first half of the images in one reconstruction, second half in the other). In the processing of helical objects, we often use overlapping segments as input images: thus, two subsequent images that would be included in the two reconstructions for FSC calculation would contain a large part of overlapping data, which means that the two reconstructions will be not independent. This in turn would result in a serious overestimation of the resolution (in particular due to noise correlation). In our hands, it could lead to almost no crossing of the 0.5 threshold limit, even at the Nyquist frequency. The "two-halves" splitting method also does not seem always right because it may be influenced by the way the micrographs were acquired (e.g. using defocus series). Therefore, I give to the user the choice of using a different way of splitting the images, adapted to the filamentous nature of the sample (option "FIL-BASED"): the images from the same filament are all included in one of the two independent reconstructions to ensure that they do not contain overlapping correlated data. Because the filament length might be very variable, and because we want to include as many images in each reconstruction as possible, we check before adding the images of a new filament to one reconstruction's "image list" which list contains less entries and the images are added to the smallest list. Although this filament-based splitting method seems more appropriate for the case of helical samples, I still leave to the user the possibility to use an "odd an even" splitting scheme or a random image separation.

Finally, the user should keep in mind that the FSC criterion is reliable only when truly independent reconstructions are calculated, and when the images included in the two structures have never "seen" each other. In our current procedure (and in most EM publications), this is not true because the particles from both half data sets are aligned to a single, overall reference 3D reconstruction that is derived from a previous alignment cycle (see (Grigorieff 2000) for discussion on the consequences of this procedure). Although it is not yet done routinely by the EM community, one should from the very first PM iteration start with different initial models derived from different set of images, ideally coming from two batches of proteins preparation, and refine each structure independently, with different sets of images. Only in this case one would be able to speak about independent reconstructions, and a FSC calculation between the two refined structures would be truly meaningful.

18 : I give here the possibility to apply a cylindrical mask to the reconstructed volume. The mask can help to erase densities in the middle of the helix (by specifying an inner radius) and/or densities at the outside (outer radius and length). To avoid applying sharp edges (= strong features) on the reconstruction, I propose to smooth the mask via low-pass filtering (for which a value conform to the pixel size is given by default). The aim of the masking is mainly to get rid of persistent noise in regions where no density is expected, and/or eventually to get rid of densities arising from proteins (or lipids, etc.) in the images which we want to mask out for example because they do not follow the helical symmetry. It is usually not necessary to use this option with himpose program because it already offers similar constraints (but with a sharp-edged mask).

19 : This option provides the possibility of using different reconstruction algorithms available in SPIDER. We now commonly use back-projection using interpolation in Fourier space (command BP 3F) but weighted back-projection or various iterative reconstruction algorithms would be worth testing.

The next options (**20** to **28**) offer various possibilities to the user:

20 : This option is used to change any parameters that were set up in the first launching instance of the script (see **Figure 5.2** and comments above).

21 : Reducing the y-translation search range for alignment is crucial for reconstruction without symmetry imposition. This prevents images from aligning to references requiring a too big translation parallel to helical axis, which in turn can cause clustering of image distribution to certain references thus leading to more and more asymmetric 3D reconstructions. Before SPIDER release 19 (I used version 17 during the thesis work), only one value was used for both x and y search range. Because x-shifts must always be searched over a relatively large range (due to imprecise centering during boxing), I added this option of centering filaments to progressively reduce the x-shifts search over PM iterations. This is done by giving here the possibility to center the images by **applying** the **x-shifts** found in the current PM iteration and re-boxing (extracting) the images. To avoid useless and harmful interpolations, the closest integer of each x-shift is actually used. This option works only for already verticalized images. To ensure a good centering, one can center the images using the alignment parameters obtained from a first model (even a featureless model like a smooth helix), and repeat the centering one or more times after the model has been refined. In any case, taking care of the distribution of x-shifts is a way to verify that the centering was well done. Using this centering a few times, we can reduce the search range for alignment parameters to as low as one or two pixels. When this option is used, it is asked for a minimum of CC that each image should present before applying the x-shifts to it. Indeed, one could consider that images with a very low CC have not found the correct x-shift. It will also be asked after shifting if one wishes to change the input images of the script to the shifted ones.

<u>22 :</u> This is the equivalent of the previous option (**16**), but adapted to **non-verticalized images**, so that the applied integer shifts to correct the centering are both in x and y directions. However there should be no shifts applied in the direction of the helical axis, as this would change the initially chosen partial overlap between images and possibly lead to 100% overlap for some images. Note that this option is not yet perfectly working, and some images are not centered properly.

<u>23 :</u> As we saw, there are many parameters for selection of images and some other parameters for symmetry search and imposition, and thus the number of possible combinations is high, so the user might wish to **test** the effect of different parameters on the **reconstructions** before using one of those for the next round of PM. This is possible by entering YES to this option. For each parameter trial, the reconstruction, as well as the FSC curve and the statistics of elimination, are kept. When the interactive mode is unset (when ‘GO’ was entered at the selection step) this option is not available and set automatically to “NO”.

<u>24 :</u> At each cycle, once all parameters for **image selection** and **symmetry and reconstruction options** are set, a file containing the values for all these variables is created and placed in the directory of the corresponding PM cycle (named cX/ where X is the cycle number). This file is useful to keep a trace of values that were used, but it can also be **loaded** by the script using this option to load the parameters in the current instance of the script. This can be convenient for example if one runs a reconstruction from two different set of images in parallel but wishes to use strictly the same selection and symmetry parameters: then, for one of the runs, one can simply load this file created by the other run.

25 : This option allows to **load** a similar file as described above but containing all the "**fixed parameters**", which means the ones that were set up when running the script for the first time (and which can be eventually changed later using option **20**).

26 : This option is used to tell the script if one works on verticalized or non-verticalized images. In the case of working with non-verticalized images, one then needs to enter the name of the file that contains the correspondence between images and in-plane angles as recorded during the boxing. This information is used to check filament polarity (option **8**) and to limit the deviation of the in-plane angle (option **1**) in respect to the filament.

27 : Once one is confident enough about the reconstruction, one can decide to reduce the number of images taken into account for the next steps of the procedure by creating a new "selection file" containing only the best images. This is done by applying a discard scheme on the current selection file using all the image selection criteria as set up in options from **1** to **11** (using current PM alignment). This is a way for example to discard images that are never properly aligned vertically (high in-plane angle deviation) or centered.

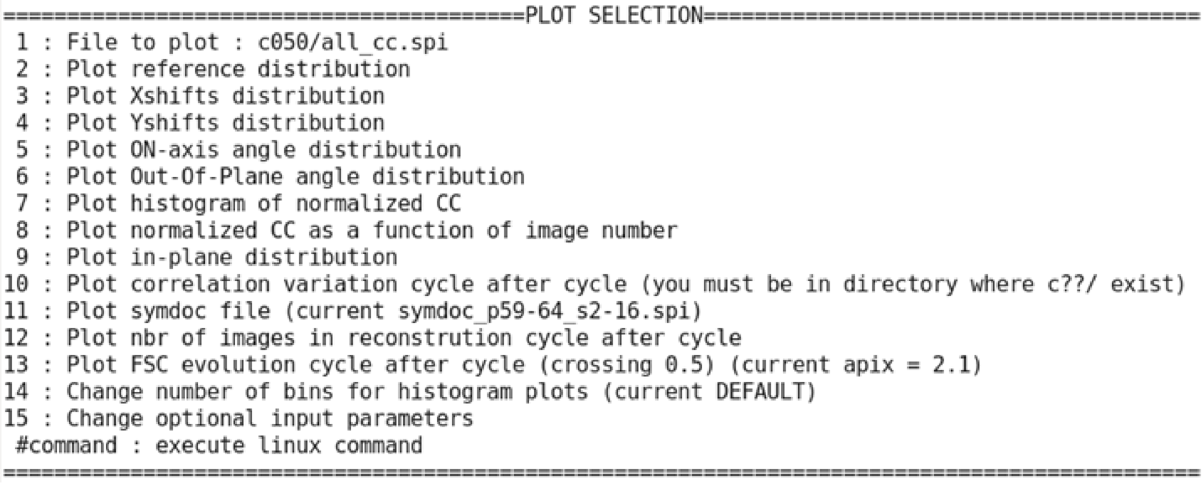
28 : This option opens the plotting interface (**Figure 5.4**). This plotting interface is a part of the process that should be easily improved in the future by adding new plotting options. As a non-exhaustive list, we have:

-plot characteristics of selected individual filaments

-plot in-plane angles as a function of the distance along the filaments (Sachse et al., 2007) for individual filaments or groups of filaments to assess the curvature and eventually to associate this to a new selection criterion.

-plot distribution of x-shift differences from one segment to the next

-for individual filaments or groups of filaments, plot on-axis angle as a function of the translation along the filament axis in comparison with the expected one when symmetry is imposed. This would highlight problems of deviation from the expected symmetry, or wrongly aligned segments.



```
=========================================PLOT SELECTION=========================================
 1 : File to plot : c050/all_cc.spi
 2 : Plot reference distribution
 3 : Plot Xshifts distribution
 4 : Plot Yshifts distribution
 5 : Plot ON-axis angle distribution
 6 : Plot Out-Of-Plane angle distribution
 7 : Plot histogram of normalized CC
 8 : Plot normalized CC as a function of image number
 9 : Plot in-plane distribution
10 : Plot correlation variation cycle after cycle (you must be in directory where c??/ exist)
11 : Plot symdoc file (current symdoc_p59-64_s2-16.spi)
12 : Plot nbr of images in reconstrution cycle after cycle
13 : Plot FSC evolution cycle after cycle (crossing 0.5) (current apix = 2.1)
14 : Change number of bins for histogram plots (current DEFAULT)
15 : Change optional input parameters
 #command : execute linux command
================================================================================================
```

Figure 5.4 : The plotting interface mega_plot.csh

The plotting interface can be used within the helix_rec.csh processing pipeline (subroutine eliminate_images.csh, option 28), in which case the required arguments are automatically entered. It can be either completely independently used, in which case one should enter at least an alignment parameter table (option 1). The plotting options 10, 12 and 13 require the plotting script to be launch in the same directory where helix_rec.csh was run (so where folders called c_X with X the PM iteration number(s) exist), and a pixel size value has to be given for option 13. The different options should be mostly self-explanatory. When one uses the option 11 (plot symdoc file), the evolution through PM iterations of axial rise and angular rotation, but as well expressed as pitch and number of subunits per turn, will be plotted.

Once the user has finished with plotting, the terminal prompting returns to the elimination and parameters setting interface.

Outputs of helix_rec.csh

Terminal output :

During the run of helix_rec.csh, the terminal displays the current step of the process, the associated relevant parameters used, and in the case of parallelization over multiple CPUS the number of terminated jobs over the total number of jobs. Any error message related to wrong parameters entries or crash of some part of the procedure will also be displayed.

Outputs files :

As already mentioned, the script helix_rec.csh uses a numbering system (three-digit number that we will call X below) to refer to the projection matching iteration cycle. The basic output files of the script are the reconstructions, called rec_nosym_X.spi (before symmetrization) and rec_sym_X.spi (after symmetrization). In the case where no symmetrization was performed, those volumes are the same. A copy of rec_sym_X.spi in CCP4 format and with a correct pixel size is also created. Then, a repertory called c_X is created at each PM cycle and it contains:

-the angles used to calculate projections

-the original alignment parameters file (Euler angles, shifts, matching projection) as created by SPIDER

-the same alignment parameters but containing the normalized CC coefficients between the input images and the projections used for alignment parameters search.

-the same file, but filtered according to any selection parameters used during step 5. These files are useful to see the effects of the selections on the statistics of the remaining images (e.g. in-plane distribution, out-of-plane, etc..). These files can be given as input to mega_plot.csh for visualization.

-if asked so, a file containing the FSC data as given by SPIDER, as well as a curve in postscript format with additional labels like the FSC 0.5 crossing resolution estimate. The two reconstructions used for FSC calculation are also stored.

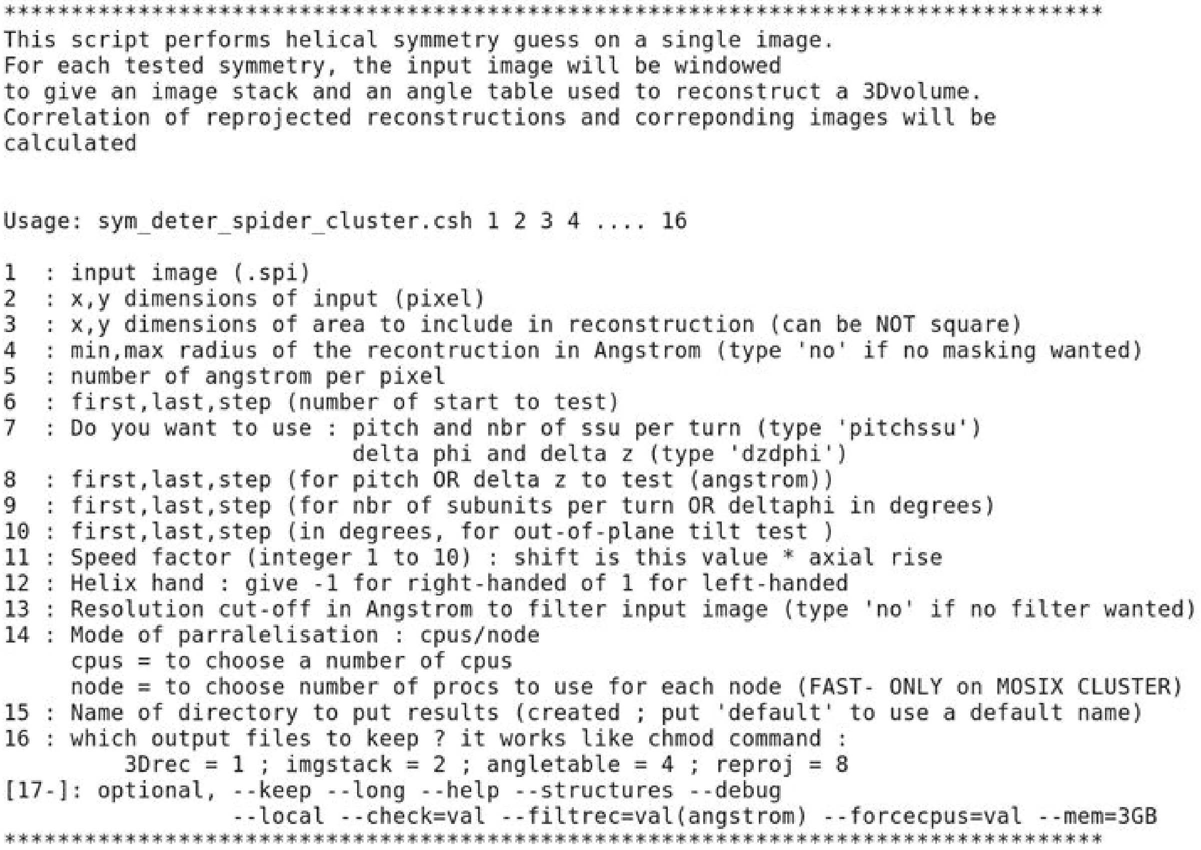
-the files containing all the parameters and selection thresholds used for current cycle. These files can be loaded by another instance of the script using the options 24 or 25 at step 5. Additionally, the file containing the threshold values (and symmetry + reconstruction options) can be copied in any directory where helix_rec.csh is running and named ‘thrsinput’ to automatically switch this particular instance of helix_rec.csh to the non-interactive mode ‘GO’ by using the given values.

-a text file containing some statistics extracted from the alignment parameters files before and after selection of images.

Determination of symmetry on 2D projection

Input arguments and general advices

The new method for automatic determination of helical symmetry parameters based on 2D projection was largely described in the dedicated chapter. Here I will just show the script that does this in practice and in particular the arguments that are needed. The **Figure 5.5** shows the initial print-out of the script which arguments are described below.



```
********************************************************************************
This script performs helical symmetry guess on a single image.
For each tested symmetry, the input image will be windowed
to give an image stack and an angle table used to reconstruct a 3Dvolume.
Correlation of reprojected reconstructions and correponding images will be
calculated


Usage: sym_deter_spider_cluster.csh 1 2 3 4 .... 16

1  : input image (.spi)
2  : x,y dimensions of input (pixel)
3  : x,y dimensions of area to include in reconstruction (can be NOT square)
4  : min,max radius of the recontruction in Angstrom (type 'no' if no masking wanted)
5  : number of angstrom per pixel
6  : first,last,step (number of start to test)
7  : Do you want to use : pitch and nbr of ssu per turn (type 'pitchssu')
                          delta phi and delta z (type 'dzdphi')
8  : first,last,step (for pitch OR delta z to test (angstrom))
9  : first,last,step (for nbr of subunits per turn OR deltaphi in degrees)
10 : first,last,step (in degrees, for out-of-plane tilt test )
11 : Speed factor (integer 1 to 10) : shift is this value * axial rise
12 : Helix hand : give -1 for right-handed of 1 for left-handed
13 : Resolution cut-off in Angstrom to filter input image (type 'no' if no filter wanted)
14 : Mode of parralelisation : cpus/node
     cpus = to choose a number of cpus
     node = to choose number of procs to use for each node (FAST- ONLY on MOSIX CLUSTER)
15 : Name of directory to put results (created ; put 'default' to use a default name)
16 : which output files to keep ? it works like chmod command :
         3Drec = 1 ; imgstack = 2 ; angletable = 4 ; reproj = 8
[17-]: optional, --keep --long --help --structures --debug
                 --local --check=val --filtrec=val(angstrom) --forcecpus=val --mem=3GB
********************************************************************************
```

Figure 5.5 : Terminal printout of the script for helical symmetry determination on 2D projection when not enough arguments are given as input

Details for each argument are given in the text

__1 :__ The **input** projection image in SPIDER format (single file – not a stack). The helix axis must be aligned with the Y axis of the image and centered. Note that only an input image with an odd number of pixels in x direction can ensure a true centering.

__2 :__ Self-evident.

__3 :__ When cutting the input image into successive segments, a window with the given **dimensions** is applied on the image. The length of this window (Y dimension) is particularly important as it will define the number of turns included in the reconstruction, as well as the possible number of segments that can be cut out from the input image (the bigger this dimension, the less segments can be extracted without that the upper and lower edge of the window getting out of the input image).

4 : Once a reconstruction is calculated, a minimum **inner radius** and maximum **outer radius** can be imposed using this option. Alternatively, if 'no' is given here as argument, no masking will be done. Masking is done by multiplying the reconstruction by two binary cylinders with the correct dimensions. These values have a great importance as a tight masking produces less ambiguous peaks on the average cross-correlation (ACC) (see the corresponding part of the manuscript).

5 : Pixel size in Å. This value should be reasonable in respect to the expected quality of the reconstruction as well as to the expected precision of symmetry parameters determination. It must be kept in mind that every segment extracted from the input image will suffer interpolation due to rotations and to shifts of non-integer pixels and thus a value of at least 3 times less than the expected resolution of reconstruction should be used. However, a lower value can be used to achieve a higher precision of symmetry parameters determination (see comments on precision for argument **9**).

6 : Range and step for the **number of starts** to try. By number of starts I mean the number of sub-helices composing the whole assembly which are related by rotational symmetry around the helix axis. This additional symmetry is imposed by the multiple use of the extracted segments and assignment of proper on-axis angles. Unless one knows in advance that there is a particular rotational symmetry in his object, one should first try to determine the helical parameters of the one-start helix that is repeated N times in the assembly. Indeed, imposing an additional rotational symmetry can increase the ambiguity of parameters determination, in particular for the angular rotation between subunits.

7 : The script accepts as input either range and steps for pitch and number of subunits per turn or axial rise (delta z) and angular rotation between subunits (delta phi). The way to express the symmetry parameters must be given here by giving '**pitchssu'** or '**dzdphi'** as argument. In practice, the SPIDER subroutine that runs all the needed steps for each symmetry, uses the "axial rise and angular rotation" expression of helical parameters, and thus a conversion will be done when using here the 'pitchssu' mode. The precision of this conversion is not a limiting factor (10^{-5} range) because the precision of the shifting operation (command RT SQ) made by SPIDER for segmenting the input image is much lower (10^{-2} range).

8 : Range and step, in Angstroms for **pitch** OR **axial rise** to test (depending on the argument given in 7). Due to the precision of the SPIDER shifting operation, using a step for axial rise search (in ‘dzdphi’ mode) lower than 0.01 has no sense, as the test would then just be redundant.

__9 :__ Range and step for **number of subunits per turn** OR **angular rotation** (in degrees) to test (depending on the argument given in 7). Again, the precision of the angles used by SPIDER for reconstructing is in the order of 10^{-3}, therefore one should adapt the search step consequently. Evidently the minimal step would thus be 10^{-3} for angular rotation. For the number of subunits per turn, one could theoretically use a smaller step until reaching the precision limit of the angles assignment in SPIDER. For example a step of 10^{-3} can lead to a difference of about 0.002° (360/12.333 – 360/12.334 = 0.0024).

One may argue here that there is no sense of choosing a too fine angular and translational search grid, as both the input image (raw image or a class-average), and the obtainable reconstruction do not usually contain enough information (or resolution) to consider such fine details. However, especially when the input image contains a long portion of helix, a small difference of assigned rotation angles between subunits or a small difference in translation can make a notable difference after many turns. If one considers for example the image of TMV shown in **Figure 3.6**, that contains ~10 turns of helix (~163 subunits), an inaccuracy of only 0.05° for the angular assignment and 0.05 Å for translations (which is what can be obtained when using a step for search of 0.1 for these values) would lead to errors of more than 8 degrees and 8 Å for the last segments included in the reconstruction. Comparing to the actual parameters of the helix , ~22° between subunits and ~1.4 Å rise per subunit, the errors would represent, for these last few segments, respectively 36 % and 570 % ! Moreover, our experience showed that the variations of ACC as a function of the helical parameters can be quite rapid. The zoom on the ACC plot for the Flagellar hook example (**Figure 5.6**) shows for instance that if one would have chosen a step of 0.1 for the subunit per turn search (that might seem reasonable at a first glance), depending on the starting point of course, one may not detect the peak of ACC corresponding to the right parameters.

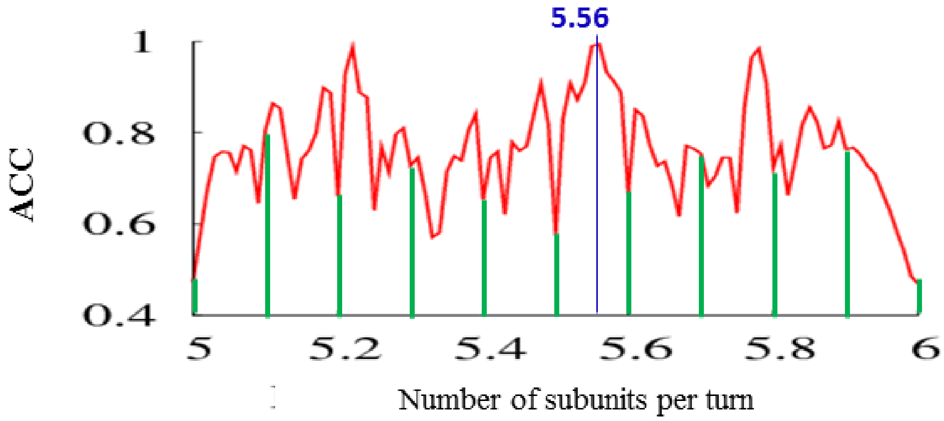


Figure 5.6 : Sampling of helical parameter search can be critical to detect the true symmetry
A zoom on the ACC plot as a function of number of subunits per turn obtained from the Flagellar Hook projection (see Figure 3.8) shows that if the sampling of the number of subunits per turn search would have been 0.1 (green vertical lines) between 5 and 6, the obtained profile would not have make possible to detect the true symmetry (blue vertical line)

10 : Range and step for **out-of-plane** to test. This option is not yet implemented (one should enter 0,0,1 here).

The ranges and steps chosen as arguments 6, 8, 9 and 10 will determine the total number of parameters to test. For each of these arguments the number of different parameters N is N(arg)=(last-first)/step, and the total number of combinations is therefore N(6)*N(8)*N(9)*N(10). This can easily give rise to huge numbers, so the user should not refine all parameters at the same time. A good way to proceed, at least for helices without rotational symmetry, is to check first that the rough ACC profile as a function of the pitch shows the same maximum, whatever the number of subunits per turn imposed (that is usually the case), and then refine the pitch by imposing an arbitrary number of subunits per turn. Then one can use the determined pitch to precisely sample the search for the number of subunits per turn.

11 : The **spacing factor** argument defines the spacing along the helix axis for segmentation of the input image. When its value is set to 1, a segment is windowed every ‘axial rise’ pixels along the helix axis, thus allowing to extract as much segments as possible from the input image. Larger values will lead to a segmentation every ‘spacing factor’*‘axial rise’ pixels, that will reduce the number of total segment and thus decrease the computing time for segmentation, reconstruction, re-projection and CC calculation. However this would result in a loss of information and as we are already limited by the length of the projection, a value other than 1 should be only used for preliminary tests .

12 : The **helix handedness** cannot be determined from a projections and the cross correlation between projections of two volumes of opposite handedness and the input images are strictly equal. Nevertheless the script offers the possibility to create right or left-handed volumes (for visualization purposes or usage in other processing steps). In practice, this is done by either positively or negatively incrementing the on-axis view of extracted segments.

13 : **Resolution cut-off** in Angstroms. A Fourier Gaussian low-pass filter is applied to the input image before segmenting (SPIDER command FF). Thus even a non-filtered raw image extracted from a micrograph can be analyzed. Filtering attenuates some negative effects of interpolation when shifting the input image for segmenting and it makes the available resolution more reasonable in respect to what can be expected for the reconstructed volume.

14 : The **mode of parallelization** refers to the way the processes are distributed among the processors. Using the option 'no' will cause all the processes to run on the local node (no parallelization) using all its available processors. To take advantage of a computer cluster, the script is currently made for using the mosix parallelization system, and the user can either define a total number of cpus to use (option 'cpus'), or choose how many cpus will be used on each available node (option 'node'). In practice the option 'node', for a same number of cpus used, offers much faster calculation times. When parallelization is used, then the total number of parameters to test is distributed among the requested number of cpus. Because different parameters will result in a very different the number of extracted segments and therefore greatly influence the calculation time, I designed a way to distribute the parameters which takes this disparity into account, such as the global mean time of calculation for each CPU will be similar.

15 : **Output directory**

16 : This **option** serves to **keep** extra output files that are not kept by default when 0 is put as argument. For each symmetry tested, it can be : stack of segmented images, 3D reconstructions, re-projections of reconstructions, and the angles table used for each reconstruction. The files that are kept are defined by the number that will be entered here : a

binary value corresponds to each optional output, as indicated. This argument contains the sum of these values corresponding to all output files one wishes to keep. For example, putting here 15 (= 1 + 2 + 4 + 8 = 2^0 + 2^1 + 2^2 + 2^3) would mean to keep all optional outputs .

17 and over : These are optional arguments that will not be exhaustively detailed here. The option --local ensures that every step runs locally. In this case, the reconstruction is done using the MPI implementation in SPIDER. Thus, once one has chosen a small subset (or one pair) of helical symmetry parameters to analyze its effect on the corresponding reconstruction, it will be quicker to run the command locally. The option --filtrec=val will cause the script to create an additional version of the reconstruction(s) that are low-pass filtered using the 'val' cut-off in Angstroms.

Outputs of the script

The main output of the script is the text table containing:

-the symmetry parameters: number of starts, rise and rotation per subunit, and corresponding pitch and number of subunits per turn

- the average cross correlation (ACC) associated to these parameters

-the number of images included in the reconstructions

-the standard deviation of the CCs for each symmetry tested.

This table gives a second one, sorted by ACC, and a third one which format is compatible with the use of the 'pm3d' plotting mode in gnuplot (3D plots with colored scale).

A file containing the arguments used to launch the script is also kept.

The other optional outputs are listed in the description of option **16** of the script.

Determination of helical symmetry parameters on 3D volume

Method description

Classically, in the IHRSR method, one starts the reconstruction process by assuming a starting helical symmetry, and the symmetry parameters are refined over the projection matching iterations by applying a least-square fit algorithm (E H Egelman, 2000) on the reconstructed non-perfectly symmetric volume using the program hsearch. In most of the cases, I could successfully use this program, but for two reasons we were brought to develop our own way for helical symmetry determination on a 3D volume that will be briefly described here. The first is that for some of our volumes, depending on their dimensions and their density distribution, the hsearch program crashed. The second reason is methodological: the hsearch program uses initial helical parameters, and also a search step, for its refinement. Thus, one has to know relatively precisely the symmetry in advance, that is not necessarily the case (e.g. for the VSV N-RNA bullet trunks). In addition, the search step of hsearch can also play an important role: sometimes, a solution is just not "seen" by the algorithm.

As a solution to these issues, we proposed to use an "exhaustive" approach for symmetry determination on 3D volumes (**Figure 5.7**). Given the fact that due to the helical symmetry a pair of translation/ rotation along/around the helical axis will bring a volume to an equivalent position, one can take the input volume and calculate the CC of this volume with itself after translation and rotation (**Figure 5.7A**) and expect a higher CC when the correct transformation is imposed. By testing all rotations between 0 and 360 degrees and a reasonable translation range, one can then obtain a map of CC coefficients as a function of these two parameters (**Figure 5.7B**). On this map, one can already have a picture of the helical net, follow the helical path, and eventually count the number of subunits per turn. To have a more quantitative estimation of the helical parameters, one can then calculate a power spectrum of the 3D translational and rotational autocorrelation plot (**Figure 5.7C**) and measure the position of the peaks: their reciprocal distance (along horizontal or vertical axis) to the origin will be directly related to the main repetitive distances that are the rotation between subunits and the pitch (I restricted the method to one-start helices). Of course, for noisy and far from perfectly symmetric reconstructions (exemplified for VSV reconstruction with no symmetry imposition on **Figure 5.7D, E, F**), the very first two pairs of peaks in the power spectrum might not correspond to the right helical parameters, for example due to artifacts like high intensity region in the power spectrum, and we have to consider more pairs then the first intense peak. The method gave rise to a simple script (**Figure 5.8**) which will now be described in more details.

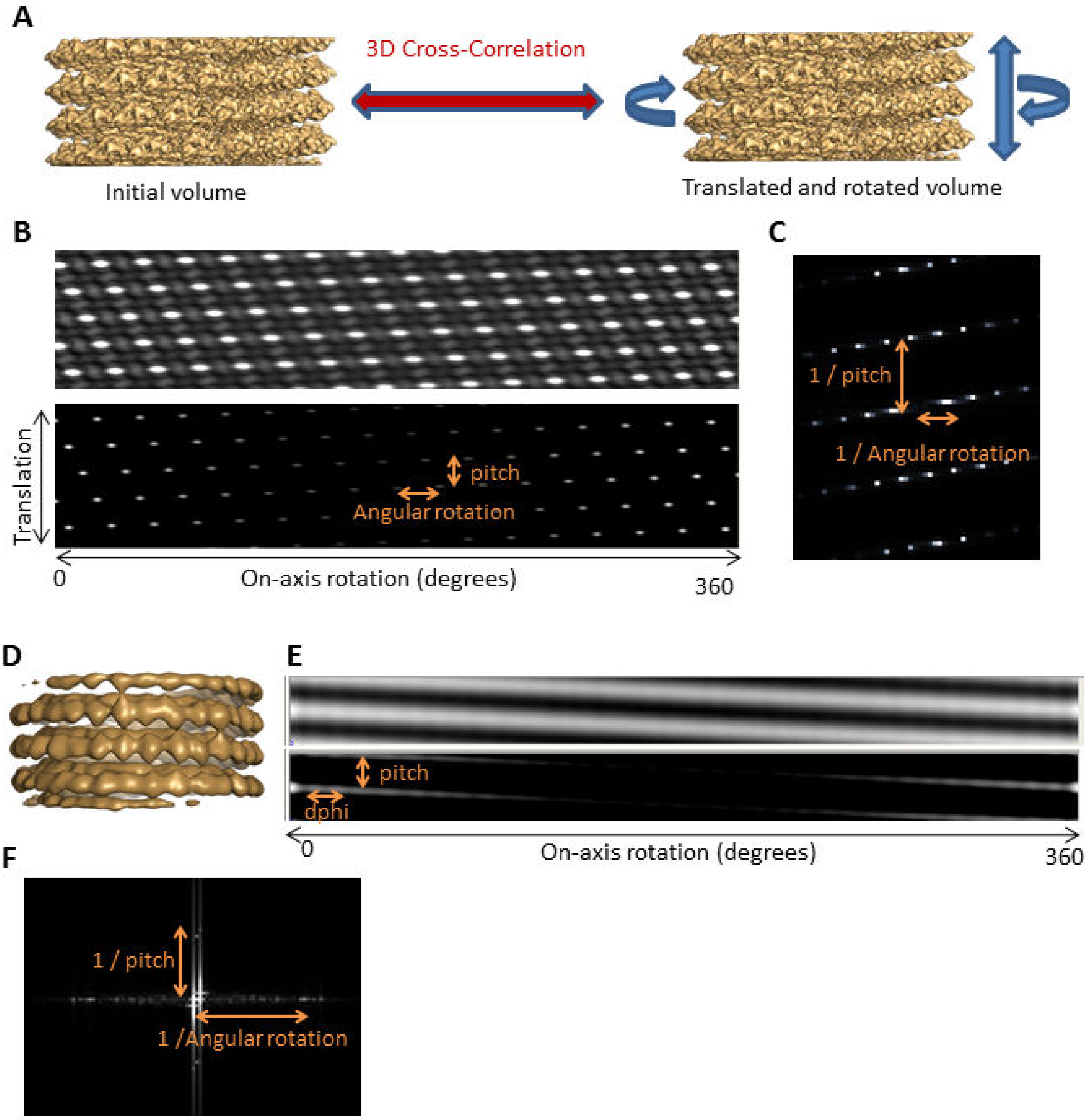
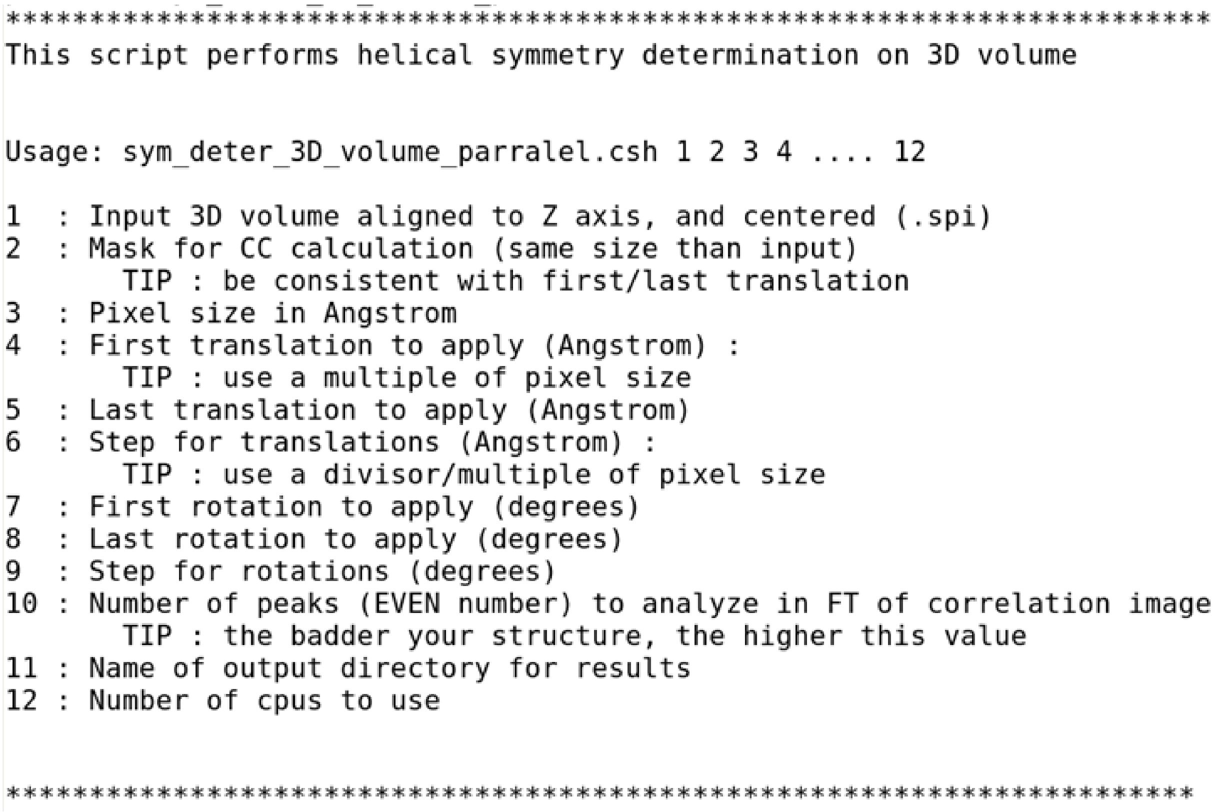


Figure 5.7 : Helical parameters determination on 3D volume based on translational and rotational autocorrelation analysis

The input volume is compared with himself by CC after applying a large range of translation along the helical axis and all rotations around the helix axis (**A**). The panel **B** shows the resulting autocorrelation image which represent the helical net (using the example of TMV structure depicted on panel **A**), at two different thresholds (top and bottom). The panel **C** shows the central part of the power spectra of autocorrelation image from panel **B**. To illustrate the method on a real case, we applied it to a reconstruction of VSV bullets-shaped N-RNA (panel **D**). Although more noisy, the autocorrelation image (panel **E**, two different thresholds) shows clearly the pitch pattern and a fainter signal for rotation between subunits. Panel **F** is the central part of the power spectra of autocorrelation image of panel **E**.



```
************************************************************************
This script performs helical symmetry determination on 3D volume


Usage: sym_deter_3D_volume_parralel.csh 1 2 3 4 .... 12

1  : Input 3D volume aligned to Z axis, and centered (.spi)
2  : Mask for CC calculation (same size than input)
       TIP : be consistent with first/last translation
3  : Pixel size in Angstrom
4  : First translation to apply (Angstrom) :
       TIP : use a multiple of pixel size
5  : Last translation to apply (Angstrom)
6  : Step for translations (Angstrom) :
       TIP : use a divisor/multiple of pixel size
7  : First rotation to apply (degrees)
8  : Last rotation to apply (degrees)
9  : Step for rotations (degrees)
10 : Number of peaks (EVEN number) to analyze in FT of correlation image
       TIP : the badder your structure, the higher this value
11 : Name of output directory for results
12 : Number of cpus to use


***********************************************************************
```

Figure 5.8 : Terminal printout of the script for helical symmetry determination on 3D volume, when not enough arguments are given as input

Details for each argument are given in the text

Input arguments

1 : The **input** volume in SPIDER format. The helix axis must be aligned with the Z axis and the volume should be centered. Again, only an input volume with an odd number of pixels in x and y directions can ensure a true centering.

2 : Mask for CC calculation (SPIDER format, same size as the input volume, it should contain voxels of values 0 and 1 only). The cross-correlations between input volume and the translated and rotated input volume will be calculated only within the volume defined by the mask. Therefore, the size of the mask should take into account the first and the last translations to apply to the initial 3D volume for CC calculations and the size of the initial volume. For example, if the input volume is 400 Å in length and one wishes to apply translations between -100 Å and + 100 Å, then the mask should be 200 Å in length (this corresponds to the overlapping part between all translated volumes and the initial volume).

3 : Self-evident

4 : First translation to apply to the volume, in Å. I recommend using a multiple of the pixel size, to avoid interpolations when translating the volume, combined to the use of a step for translations that is a divisor or a multiple of the pixel size. The first and last translations should also take into account the volume length and the mask size (see comments for argument **2**).

5 : Last translation to apply to the volume, in Å.

The optimum distance between the first and the last translations to apply should be a multiple of the pitch, to have an exact number of repetitions in this direction in the CC image, and thus to have less artifacts in the FT such as spreading of peaks in vertical direction (E H Egelman and DeRosier 1992). This length (last translation – first rotation), can thus be optimized in two steps, first to get an idea of the pitch and second to be set as a multiple of the pitch (and eventually more steps to refine this value if needed).

6 : Step for translations to be applied to the input volume for CC calculations. Again, using divisor/multiple of pixel size is an advantage.

7 : First rotation to apply to the volume, in degrees. For the determination of parameters to work properly, 0° should be given there.

8 : Last rotation to apply to the volume, in degrees. For the determination of parameter to work properly, 360° should be given there.

9 : Step for rotations to apply to the volume, in degrees.

10 : Once every translation and rotation is applied to the volume and all cross-correlations between the transformed volumes and the initial volume are calculated, the script generates a 2D image representing the CCs as grey values with the rotations in X and the translations in Y. This image is then Fourier transformed and the position of peaks in the FT are calculated using SPIDER command BBBB. This command only searches for a given **number of peaks** that should be given here as argument. As the FT is symmetric, the number to put here should be even. The position of peaks in FT is related to the repetitive pattern in the 2D CC image, i.e. to the symmetry parameter (repetitions in translations and rotations). For highly symmetric 3D volumes, the first peaks are sufficient to determine the symmetry parameters, and thus a low value can be used here. In the case of volumes with a less clear symmetry, other artifacts leading to high values in FT before the peaks that are related to the symmetry can be observed in practice. Therefore the value to enter here should be higher.

11: Name of **output** directory.

12: Number of cpus to use (for parallelization)

Output files :

The script will first calculate all the CCs between translated-rotated volumes and the initial volume, transform this information into a 2D image, calculate a power spectrum (PS) and extract the position of the maximum of the PS. Then, based on the result, it will calculate possible pairs of helical parameters (pitch, number of subunits per turn), print them on the terminal and store them in a text file. The 2D image representing all CCs as well as its PS is also stored. Sometimes this real 2D image provides even more information about the the symmetry parameters than the proposed values. For each pair of parameters, the handedness of the corresponding one-start helix is also given.

Conclusion and perspectives

In this work, we used single-particle approaches for helical reconstruction and obtained low resolution three-dimensional reconstructions of two forms of MeV nucleocapsids and of reconstituted bullets of VSV N-RNA. In this section we will summarize the results and the main steps that lead to them, give perspectives for their improvements, and discuss possible directions of research for the biological questions addressed in the present work. Much more detailed perspectives can also be found in the corresponding specific parts of the manuscript.

Image processing of helical specimen

In the image processing part, we have tried to grasp from the literature the most relevant methods. We also added our own steps in order to build an effective processing pipeline that can be now used in the laboratory by others. This includes the use of particle classification based on their real-space and reciprocal-space representation, a novel approach for estimation of helical parameters from a 2D projection, and the implementation of a user-friendly script for helical 3D reconstruction.

2D-Classification

The nucleocapsids of negative strand RNA viruses are generally rather flexible helices. In particular, we had to face helical axis bending and pitch variability for MeV non-digested nucleocapsids, high heterogeneity of diameters of reconstituted VSV bullets, and at least two different symmetries for digested MeV nucleocapsids. Two-dimensional classification methods are common processing tools to sort out heterogeneity/flexibility issues in EM images, and constitute a necessary step before 3D reconstruction. However, for helical particles these steps are only very sparsely documented in the literature. Thus, in this manuscript, we analyzed and explained in detail how one can use 2D-classification for flexible helical specimens, while giving examples of results based on our data. In particular, we described possible interpretation of eigenimages from the real-space data to identify the main source of variability within the data sets in order to sort the images in more homogeneous populations.

In addition to classification of the real-space representation of the images, we introduce an original method of classification based on the power-spectra (PS) of the images. We show that, for helical specimen, this method enables detection and separation of different helical conformations, while getting rid of some of the drawbacks of classification of real-space representation of images (like the influence of translational and on-axis rotation variability). As we show for images of digested MeV nucleocapsids, this method can eventually allow to detect small variability of helical parameters that cannot be identified based on the real-space class-averages. Furthermore, this classification method produces reciprocal space representation of the data with a higher signal to noise ratio in the PS-class-averages than obtained by just summing up the power-spectra of the entire data set, and detects departure from helical symmetry (e.g. one-side stained filaments). As a perspective, we can say that the method of PS classification requires deeper insights and further improvements. In particular, new strategies must be developed in order to minimize the influence of the CTF on the classification outcomes, to improve the in-plane alignment, and to give a relatively bigger weight to the higher resolution terms of the data.

Altogether, the different classification steps allowed us the obtain results for the 3D reconstruction that we couldn't obtain before. Among others, for the two types of Measles Virus nucleocapsids, we could obtain convergence of symmetry parameters to relevant solutions using IHRSR refinement, which was not the case with the entire unsorted data set. For reconstituted VSV bullets, which were completely refractory to IHRSR refinement, the sorting upon variable diameter made it finally possible to obtain a reconstruction without any symmetry imposition which allowed us to roughly determine the number of subunits per turn.

As we further specified in the corresponding part of the manuscript, a general remark and perspective for the 2D classification step is that it must be better integrated into the global reconstruction process, in order to test more systematically the effects of various classification strategies on the final reconstruction(s). One promising way to deal with such complexity may be the use of databases, where all possible information on each segmented image, each corresponding filament, each class of any classification trial, each reconstruction test, would be stored.

Ab initio helical symmetry parameter determination

Virtually all currently used methods of helical reconstruction critically rely on the precise knowledge of the helical symmetry parameters. These can be obtained by analysis of the Fourier transform of images, but this step can be difficult or even impossible, and requires a lot of manual intervention. A major part of the thesis was dedicated to description, validation, and use of a new method for ab initio determination of helical parameters. This method is based on a 2D projection image of a helix: a series of symmetry related views is extracted from this image, corresponding orientations are assigned in order to reconstruct a 3D model, and the "quality" of this model is assessed by calculating cross-correlation (CC) between it reprojections and the original image. We have shown using several artificial examples and real cases (with known answer) that, in most of the cases, we indeed obtain a maximum of CC for the right helical parameters. However, when the parameter search is extended over a larger range, we systematically observed (except in cases with several start helices) other maxima of CC of a similar magnitude. A deeper look at the ambiguous parameters, as well as theoretical considerations, enabled us to give an original point of view on the ambiguities in helical parameters determination from a 2D projection, and to propose general rules to predict ambiguous solutions, that happened to coincide very well with data available from the literature, when failure of helical reconstruction to find a unique solution was described.

The determined rules appear to be very useful in practice: if one obtains helical parameters for an unknown object (e.g. by FT analysis or IHRSR refinement), one would directly know which other parameters are also likely to be true, and eventually test them as alternatives for structure refinement. To give a concrete example, the reconstruction of RSV nucleocapsid made in (Tawar et al. 2009) showed 9.8 subunits per turn, but neither fitted very well with the parameters extrapolated from the ring crystal structure (10.35) nor achieved the expected resolution. Using our rules, one would know that it is worth trying refining the reconstruction around the values of, at least, 10.2 and 11.8 subunits per turn. Another concrete example of the utility of knowing those rules, is the interpretation of the results from (Schoehn et al. 2004) on digested Measles nucleocapsid reconstruction from cryo-EM images. The authors obtained two reconstructions: a first reconstruction with 12.35 subunits per turn at 12 Å resolution and a second reconstruction with 11.64 subunits per turn at 25 Å resolution, the latter being difficult to interpret. As this second solution is predicted by our "rules", we now know that it most probably is a result of the intrinsic symmetry parameters ambiguity, and do not consider it further for interpretations.

We give in the corresponding part of the manuscript many specific ideas of how to improve this method and its application, including a better display of the results in order to enrich their interpretation, a more systematic application to different class-averages for real data sets, developing ways of reducing the ambiguities (e.g. exploiting the sections in the 3D Fourier-Transform of the reconstruction which only arises from interpolation from adjacent planes), or using other reconstruction algorithm. Another important perspective would be to deeper analyze cases where the helical parameter determination failed to find a maximum of CC for the true parameters, which are not extensively documented in the present manuscript. Understanding why and in which conditions this happens will be an advantage for using this method.

Reconstruction

Improving the 3D reconstruction step was not really an objective of this work. The main contribution was to build an easy-to-use processing pipeline (see next paragraph) in an effort to include most of the known methods, i.e. IHRSR refinement (E H Egelman 2000) and extensive validation of alignment parameters (Sachse et al. 2007).

The only original approach that we used for reconstruction was for reconstituted VSV bullets, were we show that it is possible to obtain a reconstruction from which the helical parameters can be at least roughly determined, without imposing any symmetry on the starting model. This was made possible mostly by carefully restraining to a constant number the amount of images per on-axis view included in the reconstruction, in order to avoid progressive assymetrisation of the reconstruction through projection matching refinement.

The perspectives for 3D reconstruction are numerous. First we should incorporate in the processing pipeline a symmetrisation based on multiple inclusion of each 2D image (Sachse et al. 2007), which we tried, but found out that it required some trouble-shooting before including it in the stable version of the reconstruction script. We also proposed that “single-particle” approaches should be more tightly coupled to classical FT-based approaches. This can be done through the inclusion in the reconstruction pipeline of well-known

(Wakabayashi et al. 1975) image/filament selections based on the property of their FT (e.g. phase checks), or through the use of Fourier-filtered images for alignment parameters refinement, for which we proposed a protocol. In addition to these perspectives, we would like to cite recent publications (Bharat, Davey, et al. 2012; Bharat et al. 2011) that introduce new methods combining cryo-Electron Tomography and single-particle approach to solve the structure of highly heterogeneous helical specimen. Briefly, for each micrograph acquired for single-particle reconstruction, a tilt series of the same area is also recorded and used for calculating a tomographic reconstruction. The tomogram is used both to assess the quality of the filaments (e.g. flattening and bending), and to determine the symmetry parameters of each individual filament. This allows to strictly select the images to process for the single-particle reconstruction and provide the necessary helical parameters. Thus, this approach addresses major problems that we have described in this manuscript: heterogeneity/flexibility, filaments distortions, and facing unknown multiple helical parameters.

Building a user-friendly interface for 3D reconstruction and other steps of the processing

One of the aims of this thesis, since the host laboratory had only little experience in helical reconstruction at the time of its beginning, was to set up a dedicated pipeline for the image processing. Thus, part of the work consisted in writing scripts to use this pipeline in an efficient manner by any user in the lab. The idea is that a user should not have to open or edit any script (especially the not-so-friendly scripts using SPIDER syntax) to go through all the processing steps, but instead uses a series of commands which are documented so that each input argument is clear. Ultimately, I wanted to group most of the processing in a single user-friendly “master” script which would guide the user from the raw micrographs to the final reconstruction.

This aim was not fully accomplished, but an important part of the processing, the iterative projection-matching for 3D reconstruction and accompanying image selection, gave rise to a relatively advanced interactive script, which we have described in detail, providing a thorough explanation on each parameter and giving recommendation for its use. This includes suggesting many options for model projection, alignment parameters search, reconstruction, resolution estimation, data analysis through a plotting interface, and most importantly offering different ways of taking (or not) the symmetry into account and different possibilities for image selection based on various criteria. This script is also meant to be used in a flexible fashion by making it possible to vary parameters between projection matching iterations and to test the effect of different combinations of image selection criteria on the quality of the 3D reconstruction.

In addition, the method for ab initio symmetry determination based on 2D projection has also been implemented in such a way that it can be easily used, as well as another script for symmetry determination on 3D volume based on exhaustive search. We described both scripts, providing explanations on the input arguments, and gave advise for their use.

An extensive use of the designed scripts was done by a post-doc in Dr. Guy Schoehn group (Gregory Effantin), both for symmetry determination and for reconstruction steps. Structures of ESCRT-III proteins polymers, resulting from these applications were recently published (Effantin et al. 2012), and a new article, on a helical bacteriophage tail, is in preparation (Gregory Effantin, personal communication).

Measles and VSV nucleocapsids structure

Orientation of Measles Virus nucleoprotein in the nucleocapsid

A remarkable difference between the Rhabdovirus and RSV N-RNA rings is that the RNA binds at the inside of the rings for Rhabdoviruses, and the outside of the ring for RSV. The internal position of the VSV RNA suggested by the crystal structure was confirmed in the virus particle (Ge et al. 2010), where the helical turns with the smallest diameter (at the tip of the bullet) have a very similar structure to those of the recombinant N-RNA rings. Here, we corroborate the external position of the RNA at the outside of the RSV N-RNA by docking the RSV N-RNA crystal structure in our helical reconstruction of measles N-RNA, in its two forms (digested and non-digested).

This reconstruction also showed that the C-terminal domain of recombinant N points towards the inside of the helical coil, which was the second major question that we wanted to address. The N_{TAIL}, the extreme C-terminal part of the protein (residues 400-525), presumably natively unfolded (Longhi et al. 2003), plays a central role in viral replication and transcription by providing the site for binding of the polymerase-phosphoprotein complex. We could not directly resolve the precise location of this domain by comparing our two maps of MeV nucleocapsids (+/- N_{TAIL}), but the global orientation of the subunits already had important implications: as the inside of this helix is much too narrow to accommodate the 13 N-tails per helical turn of the nucleocapsid, this explains why the intact nucleocapsid forms a loose coil (and removal of N_{TAIL} result in a tight coil), and predicts that the N_{TAIL} must escape the interior of the helix between helical turns (prediction strengthened by unfolding of the helix by antibodies directed against a His-tagged $N_{TAIL).}$

Subsequently to the MeV publication (Desfosses et al. 2011), an article from a group of collaborators (Jensen et al. 2011) reported an in situ structural characterization of N_{TAIL} in the context of the entire nucleocapsid based on NMR and SAXS data. They demonstrate that N_{TAIL} is highly flexible in intact nucleocapsids and that the phosphoprotein binding site (MoRE; residues 485–502) is in transient interaction with N_{CORE}. Together with our docking results, they were able to build a model explaining both how the C-terminal part of N_{CORE} can be oriented toward the helix interior while maintaining the binding site for the polymerase cofactor accessible. In this model, the first 50 disordered amino acids of N_{TAIL} form an articulated spacer that allows the MoRE to escape from the interior of the capsid via the confined interstitial space between successive turns of the helix (see publication in the annex). In this model, the N_{TAIL} is placed in the close vicinity of the RNA, providing a mechanistic rationalization of the entire disordered domain of the nucleocapsid. The remaining residues (~500-525) are again more mobile. When we added anti-His antibody to the nucleocapsid containing a histidine-tagged nucleoprotein at its extreme C-terminus, we observed that most of the nucleocapsid could not fold into a usual helical assembly. This suggests that the last residues of N probably folds back toward the interior of the nucleocapsid, and are not completely facing the solvent.

A higher resolution EM structure of both Measles digested and non-digested nucleocapsids would eventually make possible to locate the MoRE bound to N_{CORE}, if it has a specific binding site. More perspectives on resolution improvement are given in the last section of this part.

Finally, to bridge the gap between isolated nucleocapsids and the transcription/ replication complex, structures of nucleocapsids bound to the N-binding domain of the phosphoprotein P (C-terminal domain) would be very useful (the full phosphoprotein being too flexible). During this work, a step toward this goal was done by cloning and expressing this domain of P and by performing preliminary binding assays and observations by negative staining. We found that the domain effectively binds, and observed an effect on nucleocapsid morphology. To go further, first the binding conditions should be optimized (toward a stoichiometry of ideally 1 to 1), and then, similarly to what was done for the Measles intact nucleocapsid, a screen of observation conditions by negative staining should be done to obtain a low-resolution reconstruction. Optimally, cryo-EM on the complex between the nucleocapsid and the C-terminal of P should then be attempted.

VSV reconstituted bullet-shape N-RNA

Our discovery that information necessary for packing of viral genetic material into helical bullets is contained in the nucleoprotein alone opens up new perspectives for studying of nucleocapsids as excitingly versatile nanomachines controlled by pH and ionic strength. We now can attempt a thorough step-by-step analysis of the virion assembly mechanism. The issue of tip nucleation is still unresolved. Presently we tend to consider the tip-to-trunk transition in the light of the quasi equivalence concept originally conceived for icosahedral assemblies. This attractive direction needs to be explored both experimentally and theoretically and higher-resolution structures are clearly necessary in order to transform the current speculations into a reliable model.

At the present stage, we were unable to induce a notable change in symmetry parameters of the helical trunk by adding the matrix protein M. The reason for this behavior might lie in the acidic pH necessary for the in vitro bullet folding. Alternatively, it might also reside in the way of sample preparation and the resulting sub stoechiometric decoration of the bullets by the M protein. We have several ideas of different preparation strategies to improve the binding stoechiometry. A co-expression of M and N in insect cells can also be envisioned.

Finally, one should keep in mind that the virion also contains a non-negligible amount of the phosphoprotein P. The exact position of the P-L complex in the viral particle is yet unknown and can eventually be addressed by electron tomography. Based on the crystal structure of the decameric N-RNA ring decorated by the C-terminal N-binding domain of P, the phosphoprotein binds between loops of the C-terminal domains of two neighbouring N protomers (Green and Luo 2009). Thus, a docking of the crystal structure of N into the bullet trunk reconstruction of (Ge et al. 2010) would suggest that P is located inside the bullet. One can therefore imagine that at the outside the bullet is rigidified by a helical scaffold of M, whereas at the inside it is additionally maintained by a network of P. We already started to analyze the morphology of VSV N-RNA in the presence of P, and this study needs to be pursued.

Finally, even if the reconstituted VSV bullets are a very useful tool for in vitro analysis and reconstitution based on purified components, it would be interesting and important to study different in vivo intermediates of virion assembly. This can be addressed with the help of mutant VSV viruses devoid of fusogen envelope glycoprotein G (Avinoam et al. 2011).

Toward higher resolution

One major drawback of our structural studies of the nucleocapsids of negative strand RNA viruses is the poor resolution of the obtained 3D reconstructions, which limited the biological interpretation. For the Measles project, it is not so surprising due to the use of the negative staining technique, but there is no such easy justification for VSV. For VSV, one main reason might be the strong heterogeneity of the data set which could have probably be even more extensively addressed, even although our classifications procedures finally made it possible to obtain a reconstruction without symmetry imposition which indicated the rough symmetry of the majority of the particles. If the data set had to be reanalyzed, a finer sorting of the segments according to their symmetry would be worth trying, even if this would significantly lower the number of images included into the final reconstruction. Tests that were made a posteriori showed that our method of helical parameter determination could be applied to portions of raw images, eventually giving a tool for a finer sorting of images. Other recently described methods for selection only of the best preserved tubes and of sorting according to symmetry by combining tomography and single-particle approaches should also be considered (Bharat, Davey, et al. 2012).

For Measles, the rigid digested nucleocapsids are certainly a good target for acquiring new cryo-EM data sets and trying to apply all possible new standards for image processing to improve the reconstruction of (Schoehn et al. 2004). The relatively precise knowledge of the symmetry will also largely help to achieve this goal. In a higher resolution structure, the RNA molecule may become visible and thus definitely confirm the docking of the crystal structure into the map. If the non-digested nucleocapsids are clearly too flexible to be directly studied with cryo-EM (at pH 7.5, 150 mM NaCl), the observation that we made from condition screening by negative staining clearly showed that pH and ionic force have an influence on the morphology of the nucleocapsid, eventually making them more regular. This shows that a screening of condition for cryo-EM image acquisition have a chance to give an improvement in the rigidity of nucleocapsids, potentially to a state where a 3D reconstruction can be attempted. Together with a higher resolution of the digested nucleocapsid, this may provide the information needed for localizing parts of the N_{TAIL} domain like the MoRE and precise the model provided by NMR data (Jensen et al. 2011).

Finally, more effort should be spent on trying to isolate and crystalize Measles N-RNA rings, which would then provide crucial missing information to combine with higher resolution structures of the helical assemblies.

Annex

Publications directly relevant to the thesis manuscript

Published (1st Author) : Nucleoprotein-RNA Orientation in the Measles Virus Nucleocapsid by Three-Dimensional Electron Microscopy

The following document contains in a row the main text and the supplementary information.

JOURNAL OF VIROLOGY, Feb. 2011, p. 1391–1395 Vol. 85, No. 3
0022-538X/11/$12.00 doi:10.1128/JVI.01459-10

Nucleoprotein-RNA Orientation in the Measles Virus Nucleocapsid by Three-Dimensional Electron Microscopy▽†

Ambroise Desfosses,[1] Gaël Goret,[2] Leandro Farias Estrozi,[1] Rob W. H. Ruigrok,[1] and Irina Gutsche[1]*

UMI 3265 UJF-EMBL-CNRS, Unit for Virus Host Cell Interactions, Grenoble, France,[1] and UMR 5075 CEA-CNRS-UJF, Institut de Biologie Structurale Jean-Pierre Ebel, Grenoble, France[2]

Received 13 July 2010/Accepted 15 November 2010

Recombinant measles virus nucleoprotein-RNA (N-RNA) helices were analyzed by negative-stain electron microscopy. Three-dimensional reconstructions of trypsin-digested and intact nucleocapsids coupled to the docking of the atomic structure of the respiratory syncytial virus (RSV) N-RNA subunit into the electron microscopy density map support a model that places the RNA at the exterior of the helix and the disordered C-terminal domain toward the helix interior, and they suggest the position of the six nucleotides with respect to the measles N protomer.

The RNA genome of nonsegmented negative-strand RNA viruses is tightly and regularly encapsidated by the viral nucleoprotein N, providing flexible helical templates for viral transcription and replication. Upon heterologous expression, nucleoproteins associate not only with long cellular RNAs to form helical nucleocapsids undistinguishable from the viral ones but also with short cellular RNAs that noncovalently close up into N-RNA rings. In the rings, N-RNA is sterically constrained in a biologically inactive form, but the rings have an advantage of being rigid enough for X-ray crystallography. Conversely, the helical assemblies are challenging for electron microscopy (EM) analysis because of their flexibility but are the biologically relevant ones.

The atomic structures of N-RNA rings of rabies virus and vesicular stomatitis virus (both rhabdoviruses) (1, 10) reveal the shielding of RNA between two domains of N in a positively charged cleft situated inside the rings. Extended N- and C-terminal domains reach out to neighboring N protomers in order to stabilize and rigidify the structure. Recently, the structure of N-RNA rings of respiratory syncytial virus (RSV; a paramyxovirus) was determined (24). The global architecture of the nucleoprotein is very similar to that of the rhabdoviruses, although there are 7 ribonucleotides (nt) instead of 9 bound to each N protomer. However, the lateral contacts between adjacent N subunits of the ring confer to it an opposite curvature, which results in an outward RNA groove location. RSV N has an N-terminal exchange domain similar to that of rhabdovirus N, but the C-terminal domain is slightly different, as it is not clearly involved in contacts between subsequent N protomers. Is this inversion of the subunit orientation due simply to steric constraints in the ring, or does it also take place in a helical nucleocapsid? Tawar and coworkers modeled an RSV N-RNA helix but could not directly dock the atomic structure of RSV N into their helical EM reconstruction (24).

A sequence alignment between RSV N and measles virus N (MeV N), both paramyxovirus nucleoproteins, is difficult to interpret because of the lack of amino acid identity. However, a comparison of the secondary structure elements observed in the RSV N structure, with a secondary structure prediction for MeV N (6) (Fig. 1), shows even more similarity than that between rhabdovirus and RSV N. This comparison also shows that the β-hairpin projecting from the distal end of the RSV N protomer (24) is conserved between these two paramyxovirus nucleoproteins. One important difference lies in the length of the highly disordered C-terminal domain, the N tail, that is 31 residues long (360 to 391) for RSV N (24) but 126 residues long (400 to 525) for MeV N (16). A short sequence in the MeV N tail (residues 489 to 506) folds into a dynamic helical structure that is stabilized by binding of the viral phosphoprotein that carries the viral RNA-dependent RNA polymerase

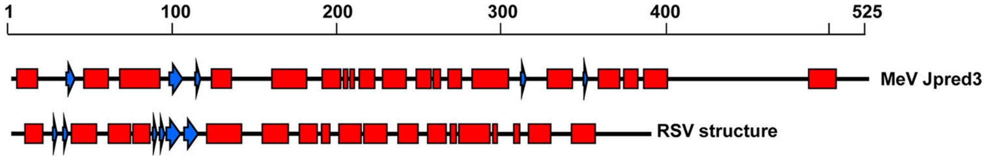


FIG. 1. Predicted secondary structure of MeV N compared to secondary structure elements in the atomic structure of RSV N. α-Helices are represented as red boxes, β-strands as blue arrows.

* Corresponding author. Mailing address: UVHCI, UMI 3265 UJF-EMBL-CNRS, BP 181, 38042 Grenoble, Cedex 9, France. Phone: 33476209463. Fax: 33476209400. E-mail: gutsche@embl.fr.

† Supplemental material for this article may be found at http://jvi.asm.org/.

▽ Published ahead of print on 24 November 2010.

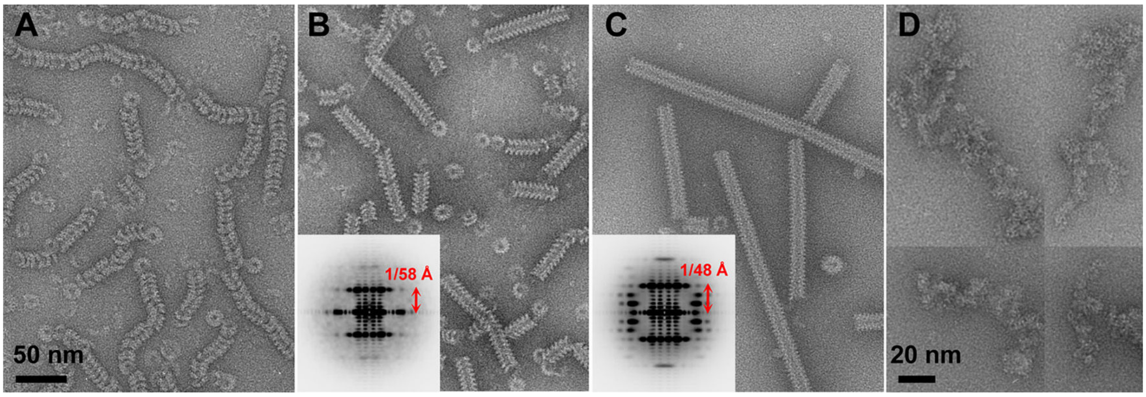


FIG. 2. Fields of view of negatively stained MeV nucleocapsids. (A) Intact nucleocapsids with 2% uranyl acetate and a single carbon layer. (B and C) Intact (B) or digested (C) nucleocapsids with NanoW in a double carbon layer and a representative class average of power spectra. (D) Recombinant C-terminally His-tagged nucleocapsids bound to anti-His-tagged antibody.

(12, 13, 16). The N tail is also involved in binding host proteins, such as hsp70 (5, 26) and interferon regulatory factor 3 (14, 15). So far, the location of the N tail on the helix is not known, although it is usually shown on the outside in cartoons that illustrate transcription and replication of paramyxoviruses (see Fig. 9 in reference 3). The helical model derived from the recombinant N-RNA ring structure of RSV, however, would place the N tail toward the helix interior, which would have consequences for the contacts between subsequent helical turns.

The helical structure of the intact measles virus N-RNA under cryoelectron microscopy (cryo-EM) conditions is highly flexible and difficult to determine by Fourier-Bessel image analysis or even by single-particle-based approaches (2, 21). However, once the N tail is removed by proteolysis, the structure becomes more regular and rigid and thus amenable to helical reconstruction by cryo-EM (21). Here, we show that the nondigested nucleocapsid structure can be addressed in negative-stain electron microscopy by trapping the sample between two layers of carbon film and by using NanoW stain (from Nanoprobes) instead of the more traditional uranyl acetate (Fig. 2). This preparation technique enables to image intact measles virus nucleocapsids as well as their trypsin-digested counterparts and has the advantage

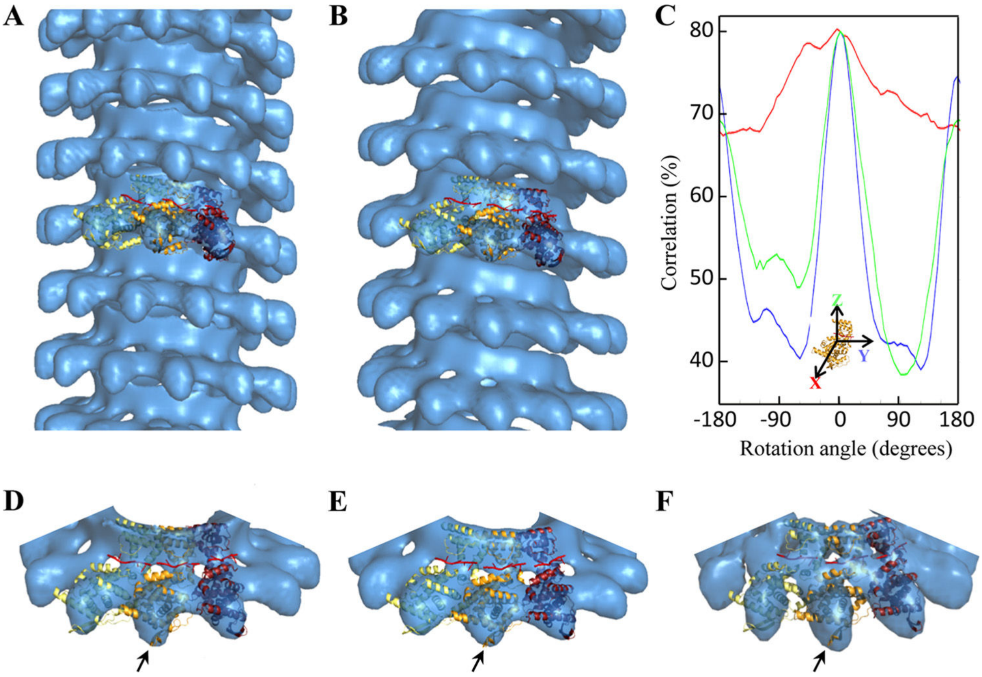


FIG. 3. Three-dimensional reconstructions of MeV nucleocapsids. (A and D) Digested nucleocapsid, (B and E) intact nucleocapsid, (F) cryo-EM reconstruction of digested MeV N (21). The fit of the RSV N-RNA atomic structure is shown as an overlay. The N-terminal β-hairpin fits nicely into the density (arrow). The RNA is shown as a red ribbon. (C) Docking precision for panel A. Shown is the correlation upon rotation of the monomer (see the supplemental material).

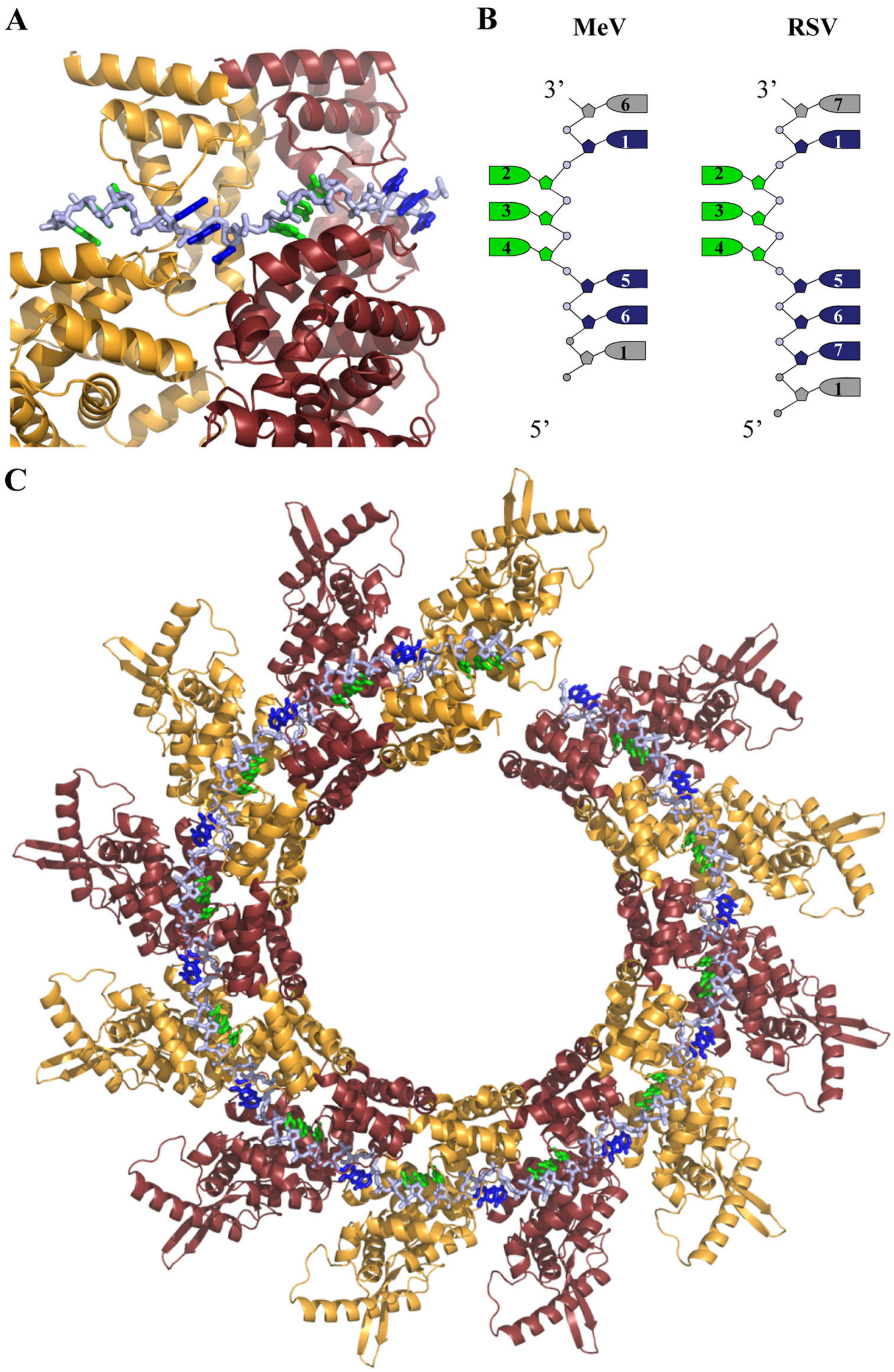


FIG. 4. RNA binding to MeV N based on RSV N-protomer fitting and energy minimization for solvent-exposed bases. Protein-oriented bases are in green, solvent-oriented bases are in blue, and the backbone is in light blue. (A) Enlarged view of RNA binding. (B) Schematic diagram for a comparison of RNA interaction with MeV N and RSV N. The numbering is as in reference 24. The gray nucleotides are on the neighboring N protomers. (C) Top view of the helical fit.

of maintaining the helix in a more rigid state. For this analysis, recombinant MeV N was produced, and a fraction of it was trypsinated as described previously (21) and imaged with a transmission electron microscope. Overlapping segments of the visually most rigid helices were selected with Boxer (17), contrast transfer function (CTF) corrected with CTFFIND3 (18) and Bsoft (11), and aligned and classified with Imagic (25). An additional classification of power spectra of individual image frames and a sorting based on artificial smooth helical volumes improved the homogeneity of different subsets separated according to diameter and helical parameters. The major subsets were used for angular assignment and three-dimensional (3D) reconstruction in an iterative projection-matching procedure similar to that for IHRSR (7, 8) with the SPIDER package (9, 22), starting from a smooth helix of a chosen pitch as the initial model (for details, see the supplemental material).

Final three-dimensional reconstructions of trypsin-digested and intact measles nucleocapsids at a resolution of 25 Å are shown in Fig. 3. Removal of the N tail leads to a compaction of the helix, with the pitch shortening from 57.2 Å to 48.7 Å and a diameter constriction from 200 Å to 190 Å, in line with the previous cryo-negative-stain EM work (2). The number of subunits per turn in the digested nucleocapsid was found to be 12.33, the same as that previously obtained for such species under cryo-EM conditions by Fourier-Bessel analysis of the most regular helix coupled to IHRSR (21). Thus, in this case, the double-carbon layer negative-stain microscopy and the NanoW stain seem to maintain the helical structure without modifying the helical parameters. The intact nucleocapsid helix accommodates a nearly integer number of 12.92 subunits per turn, which agrees with the 5% increase in diameter. The overall shape of the nucleoprotein subunit is nevertheless very similar in both reconstructions, corroborating previous arguments in favor of the intrinsic N-tail disorder (2, 16).

Given the predicted structural similarity between RSV and MeV N, the atomic model of the RSV nucleoprotein monomer (Protein Data Bank [PDB]accession number 2WJ8) was used for fitting into the obtained 3D volumes with VEDA (http://mem.ibs.fr/VEDA), a new graphical version of URO (19) (Fig. 3). For fitting, MeV nucleoprotein helices were considered to be left-handed based on previously published metal shadowing results (21), and a modified PDB file of the RSV N protomer with only 5 nt corresponding to nt 2 to 6 was used, given that MeV N-RNA contains 6 nt per N protomer (4, 23) and not 7. Interestingly, without any constraints imposed during fitting, the fit ensures the continuity of RNA bound to measles virus N. Atomic coordinates of two RNA segments bound to consecutive subunits were extracted from the thus-obtained MeV nucleocapsid model, an additional ribonucleotide (corresponding to number 7 in Fig. 1D in reference 24) was inserted, and energy minimization was performed with VEGA software (20) to obtain a continuous, physically realistic RNA molecule. Since bases 2, 3, and 4 bind in a cavity on the RSV-N protein, their coordinates were kept fixed, while those of the solvent-facing ribonucleotides, 5, 6, and 1, were optimized. Figure 4 illustrates the possibility of easily constructing a 6-nt RNA with three bases facing the protein, as in the RSV N-RNA rings, and three bases stacked and pointing away from the protein into the solvent.

This fit of the atomic structure of RSV N into the negative-stain EM reconstructions is also consistent with the previously published cryo-EM structure of the digested MeV nucleocapsid (21) (Fig. 3F) and the RNA position predicted therein by *cis*-platinum RNA labeling. It suggests that the RNA is indeed localized at the exterior face of the helix, as in the RSV N-RNA rings, and not as in rhabdoviral N-RNA rings. Although the disordered C-terminal domain could not be resolved in the reconstruction of the intact nucleocapsid, the fit suggests that this crucial domain would point toward the helix interior. In addition, binding of anti-His-tagged antibody to C-terminally His-tagged nucleocapsids prevents correct helix formation (Fig. 2D) (see the supplemental material), indicating that the N tail domain may come out at a site where it interferes with contacts between two subsequent turns of the N-RNA helix, contributing to its flexibility.

We thank Francine Gérard and Euripedes Ribeiro de Almeida from the group of Marc Jamin (UVHCI) for help with initial production of MeV N and Guy Schoehn from UVHCI-IBS for discussions and advice.

A.D. was financed by a Ph.D. fellowship from the Rhône-Alpes region. This work was financed in part by the FINOVI Foundation.

REFERENCES

1. **Albertini, A. A., et al.** 2006. Crystal structure of the rabies virus nucleoprotein-RNA complex. Science **313:**360–363.
2. **Bhella, D., A. Ralph, and R. P. Yeo.** 2004. Conformational flexibility in recombinant measles virus nucleocapsids visualised by cryo-negative stain electron microscopy and real-space helical reconstruction. J. Mol. Biol. **340:**319–331.
3. **Bourhis, J. M., B. Canard, and S. Longhi.** 2006. Structural disorder within the replicative complex of measles virus: functional implications. Virology **344:**94–110.
4. **Calain, P., and L. Roux.** 1993. The rule of six, a basic feature for efficient replication of Sendai virus defective interfering RNA. J. Virol. **67:**4822–4830.
5. **Couturier, M., et al.** 2010. High affinity binding between Hsp70 and the C-terminal domain of the measles virus nucleoprotein requires an Hsp40 co-chaperone. J. Mol. Recognit. **23:**301–315.
6. **Cuff, J. A., and G. J. Barton.** 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins **40:**502–511.
7. **Egelman, E. H.** 2007. The iterative helical real space reconstruction method: surmounting the problems posed by real polymers. J. Struct. Biol. **157:**83–94.
8. **Egelman, E. H.** 2000. A robust algorithm for the reconstruction of helical filaments using single-particle methods. Ultramicroscopy **85:**225–234.
9. **Frank, J., et al.** 1996. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J. Struct. Biol. **116:**190–199.
10. **Green, T. J., X. Zhang, G. W. Wertz, and M. Luo.** 2006. Structure of the vesicular stomatitis virus nucleoprotein-RNA complex. Science **313:**357–360.
11. **Heymann, J. B., G. Cardone, D. C. Winkler, and A. C. Steven.** 2008. Computational resources for cryo-electron tomography in Bsoft. J. Struct. Biol. **161:**232–242.
12. **Jensen, M. R., et al.** 2008. Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein. J. Am. Chem. Soc. **130:**8055–8061.
13. **Kingston, R. L., W. A. Baase, and L. S. Gay.** 2004. Characterization of nucleocapsid binding by the measles virus and mumps virus phosphoproteins. J. Virol. **78:**8630–8640.
14. **Laine, D., et al.** 2005. Measles virus nucleoprotein induces cell-proliferation arrest and apoptosis through NTAIL-NR and NCORE-FcgammaRIIB1 interactions, respectively. J. Gen. Virol. **86:**1771–1784.
15. **Laine, D., et al.** 2003. Measles virus (MV) nucleoprotein binds to a novel cell surface receptor distinct from FcγRII via its C-terminal domain: role in MV-induced immunosuppression. J. Virol. **77:**11332–11346.
16. **Longhi, S., et al.** 2003. The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. J. Biol. Chem. **278:**18638–18648.
17. **Ludtke, S. J., P. R. Baldwin, and W. Chiu.** 1999. EMAN: semiautomated software for high-resolution single-particle reconstructions. J. Struct. Biol. **128:**82–97.
18. **Mindell, J. A., and N. Grigorieff.** 2003. Accurate determination of local defocus and specimen tilt in electron microscopy. J. Struct. Biol. **142:**334–347.
19. **Navaza, J., J. Lepault, F. A. Rey, C. Alvarez-Rua, and J. Borge.** 2002. On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation. Acta Crystallogr. D Biol. Crystallogr. **58:**1820–1825.
20. **Pedretti, A., L. Villa, and G. Vistoli.** 2002. VEGA: a versatile program to convert, handle and visualize molecular structure on Windows-based PCs. J. Mol. Graph. Model. **21:**47–49.

21. **Schoehn, G., et al.** 2004. The 12 Å structure of trypsin-treated measles virus N-RNA. J. Mol. Biol. **339:**301–312.
22. **Shaikh, T. R., et al.** 2008. SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. Nat. Protoc. **3:**1941–1974.
23. **Sidhu, M. S., et al.** 1995. Rescue of synthetic measles virus minireplicons: measles genomic termini direct efficient expression and propagation of a reporter gene. Virology **208:**800–807.
24. **Tawar, R. G., et al.** 2009. Crystal structure of a nucleocapsid-like nucleoprotein-RNA complex of respiratory syncytial virus. Science **326:**1279–1283.
25. **van Heel, M., G. Harauz, E. V. Orlova, R. Schmidt, and M. Schatz.** 1996. A new generation of the IMAGIC image processing system. J. Struct. Biol. **116:**17–24.
26. **Zhang, X., et al.** 2005. Hsp72 recognizes a P binding motif in the measles virus N protein C-terminus. Virology **337:**162–174.

Nucleoprotein-RNA orientation in the measles virus nucleocapsid by three-dimensional electron microscopy

Ambroise Desfosses[1], Gaël Goret[2], Leandro Estrozi[1], Rob W.H. Ruigrok[1] and Irina Gutsche[1]*

1; UMI 3265 UJF-EMBL-CNRS, Unit for Virus Host Cell Interactions, Grenoble, France.
2; UMR 5075 CEA-CNRS-UJF, Institut de Biologie Structurale "Jean-Pierre Ebel," Grenoble, France.

* Corresponding author. Mailing address: UVHCI, UMI 3265 UJF-EMBL-CNRS, BP 181, 38042 Grenoble, Cedex 9, France. Phone: + 33476209463. Fax: +33476209400. E-mail: gutsche@embl.fr

Supplementary online information

Supplementary electron microscopy and image analysis methods

Negative stain electron microscopy

For preparation of negatively stained grids, the sample was applied to the clean side of a thin carbon film on the carbon-mica interface and stained either with 2 % (w/v) uranyl acetate or with 2% NanoW stain (Nanoprobes). For the preparation of double layer carbon grids, a carbon film with the absorbed sample was floated on a drop of NanoW. A 400-mesh copper grid was put on top of the floating carbon film and the whole was turned upside down and used to catch a second layer of carbon film floating on another drop of NanoW. Thus the sample was entirely and uniformly stained and trapped between two thin layers of carbon. The grids were observed under low-dose conditions with a JEOL 1200 EX II transmission electron microscope with a tungsten filament at 100 kV. Images were recorded on Kodak SO-163 films at a nominal magnification of 40,000 times. Selected negatives were then digitised on a Zeiss scanner (Photoscan TD) at a step size of 7 micrometer giving a pixel size of 1.75 Å at the specimen level.

Image processing software

Image processing was carried out in an integrated approach, combining different software packages for different steps in the analysis procedure. In particular, the EMAN software package (6) was used for particle selection; CTFFIND (7) for contrast transfer function determination, BSOFT (5) for the CTF correction; Imagic (15) for multivariate statistical analysis, classification and multireference alignment steps; Spider (4, 13) for projection matching and 3D reconstruction; the hsearch and himpose programs from the IHRSR package (2, 3) for symmetry search and imposition; URO (Navaza et al, 2002) and VEDA (http://mem.ibs.fr/VEDA) for crystal structure fitting; VEGA for RNA modelling (9); Pymol (http://www.pymol.org/) and Chimera (10) for visualisation.

Image preprocessing

Micrographs were selected based on concentration, length and apparent rigidity of measles virus nucleocapsids present and on the CTF quality and parameters determined for each micrograph by CTFFIND3. Extremities of filaments were selected with the helix option of the EMAN Boxer tool, while paying attention to avoid picking too flexible and/or discontinuous segments. Originally, 400*400 pixel overlapping segments were extracted every 6 pixels along the filament axis. On the whole, 25244 segments of 1798 filaments of the intact nucleocapsids and 73794 images of 1461 filaments of the digested nucleocapsids were selected. These individual images were corrected for CTF by phase-flipping with the bctf program from BSOFT, and then each helical segment was verticalised using the in-plane rotation angle calculated from the coordinates of filament extremities, clipped into 200*200 pixel images to remove empty areas caused by verticalisation, normalised and band-pass filtered (low frequency cutoff = 350 Å ; high frequency cutoff = 16 Å). A version of binned images (final size 35nm*35nm) was generated for the reconstruction steps.

Classification of helical segments

The 200*200 pixel images were iteratively aligned and classified in the IMAGIC software package as typically done for single particles and as described for example in (11). This permitted removal of slightly curved and discontinuous segments and to initially separate images into subsets of different diameter, pitch, out of plane tilt and azimuthal angle.

Classification of power spectra

The 200*200 pixel images were padded into 800*800 pixel images for the calculation of the raw power spectra (PS), from which a 200*200 pixel clipped version (corresponding to frequencies up to 1/14 $Å^{-1}$) was created for multivariate statistical analysis (MSA) and classification in the IMAGIC software package. Class-averages of power spectra showed increased signal of helical diffraction. These class-averages were rotationally averaged, and each raw PS was divided by rotationally-averaged version of the corresponding class-average to amplify the weak features at higher resolution. These modified raw PS were then rotationally aligned against the class-averages and the procedure of MSA and classification was reiterated. The mask used for MSA was first a simple filled circle, and after one cycle of classification, a more complex mask was created around areas containing diffraction peaks, in order to force the classification to reflect the precise position of the peaks and not the position of the Thon rings of the CTF. This classification step enabled separation of the data set into more homogeneous subsets of segments of different helical symmetry. Only classes clearly showing the second layer line were taken for further analysis.

Sorting according to pitch

A set of smooth helices with a fixed diameter corresponding to the one measured on class-averages of individual helical segments and with different pitches was created for both digested nucleocapsids (explored pitch range between 44 and 54 Å to conform with the total sum of power spectra) and for native ones (explored pitch range between 54 and 72 Å). Reference projections were made from each model and used in a projection matching procedure with Spider to assign each raw image to a pitch class. This enabled to further refine the subsets provided by classifications of helical segments and their power spectra. The major subsets (pitches of 47 and 48 Å for the digested and of 56 to 62 Å for the native nucleocapsid) were used for angular assignment and 3D reconstruction. Each pitch subset (two subsets for digested and three subsets for the intact nucleocapsid) was treated independently up to the final reconstruction.

3D Reconstruction

Either the entire image set or subsets of images derived from the various classification procedures described above, where used in an iterative projection matching procedure similar to IHRSR, starting from a smooth helix of chosen pitch as initial model. At each projection matching cycle, the aligned images were selected according to correlation, in-plane rotation and shift. Moreover, only helical segments of the same polarity as the one of the original filament were included in the reconstruction. The reconstructed volumes were filtered to 16 Å before every new projection matching cycle.

The initial guesses for the number of subunits per turn were either used as variables to study the convergence of IHRSR procedure, or assessed by ab initio symmetry estimation based on corresponding class averages. In the second case, less iterations were needed to achieve convergence. Briefly, a set of 3D volumes was created for each representative class average by applying different helical parameters (i.e. shifts and rotations) followed by back projection. The helical parameters used to create the volume which reprojection correlated best with the original class average, were chosen as an initial estimate for IHRSR. For the digested nucleocapsid, helical parameters of 12.33 subunits per turn and a pitch of 48.7 Å could be unambiguously determined. For the native nucleocapsid, two solutions of 12.92 subunits per turn and 57.2 Å axial rise or 10.95 subunits per turn and 56.6 Å axial rise appeared possible and projections of the two obtained reconstructions were indistinguishable at our level of resolution. However, the shape of the subunit in the 10.95 subunits per turn reconstruction and the inter-subunit contacts were very different from those obtained for the digested nucleocapsid and from all previously described analysis of measles virus N (1, 12) and of RSV N (14). Thus, these helical parameters were rejected as an erroneous solution. Each pitch class gave similar number of subunits per turn. The reconstructions shown correspond to the most populated class. On the whole, 6305 segments were included in the final reconstruction of the native nucleocapsid, and 11309 segments were included into the reconstruction of the digested state. The resolution of the reconstructions was estimated by splitting the data into two independent sets (all segments of the same filament were included in the same reconstruction to be sure that the two data sets were indeed independent and that no bias was introduced by overlapping data) and calculating two reconstructions with which the Fourier Shell Correlation was calculated as a function of resolution. The conservative FSC=0.5 criterion was used to estimate the resolution, which was found to be around 25 Å for both reconstructions.

Docking of the atomic model of RSV nucleoprotein N into the electron density maps

The atomic model of the RSV nucleoprotein monomer was extracted from the published Xray crystal structure of the RSV-N decameric ring (pdb id 2wj8). The residues 2 to 35 and 361 to

375, as well as ribonucleotides 1 and 7, were removed prior to docking into the EM densities. We used VEDA (http://mem.ibs.fr/VEDA), the graphical version of URO (8), to fit the atomic model while taking symmetry into account. The resolution used during the fit was limited to 25 Å. The atomic model of RSV N was placed at eight different initial positions (with the RNA facing the inside or the outside of the helix, the C-terminal domain of N pointing inwards or outwards, the N subunit tilted to the left or to the right), and final positions with optimized correlations between the EM density map and the atomic model were calculated. To get a better insight into the docking precision, the variation of the correlation upon rotation of the best fit (RNA at the exterior, C-terminal domain of N pointing inwards and the outer tip of the N-protomer tilted to the left) around its principal axes of inertia is plotted in **Figure 3C**. The best fit with a correlation of 80.2 % at 25 Å resolution corresponds to the RNA localisation towards the exterior of the helix and the disordered C-terminal tail oriented towards the helix interior. As a control, the previously published 12 A cryoEM structure of digested MeV nucleocapsids (12) was subjected to the same fitting procedure **(Figure 3F)**. Based on the continuity of RNA and on the interprotomer contacts, this fit proved itself to be the best model at 12 Å resolution as well. Thus, the same fit of the RSV N-RNA atomic structure is valid both for negative stain and cryoEM reconstructions.

Supplementary methods of expression and purification of recombinant measles virus N–RNA and N_{H6}-RNA

Full-length measles virus N, Halle strain was expressed in Sf21 insect cells and purified as described (12). In addition, for the purpose of C-terminal localisation, the same N was cloned into a pet22b vector (from Novagen) with or without an added hexahistidine tag fused at the C-terminus. These constructs are referred to as pet22b/N_{H6} and pet22b/N. The sequence of the coding region was checked by sequencing (MWG). BL21(DE3)RIL E. coli strain was transformed with pet22b/N_{H6} or pet22b/N and grown overnight to saturation in LB medium containing 100 µg/ml ampicillin and 34 µg/ml chloramphenicol. An aliquot of the overnight culture was diluted 1/100 in LB medium and grown at 37°C. At OD_{600} of 0.6, isopropyl β-D-thiogalactopyranoside was added to a final concentration of 0.1mM, and the cells were grown at 37°C for 3h. The induced cells were harvested, and collected by centrifugation. The resulting pellets were resuspended in 150 mM NaCl, 20 mM Tris–HCl (pH 7.5) in the presence of the protease inhibitor cocktail completee-EDTA free (from Roche) and sonicated. N-RNA was further purified as the one from insect cells. Trypsin treatment of recombinant measles virus N–RNA was done as described (12).

Supplementary results of interaction of $anti_{H6}$ antibody with N_{H6}-RNA

Based on the position of the C-terminus of RSV N in the fit, the C-terminal domain of MeV nucleoprotein would point towards the interior of the helix. However, a possibility remains that the flexible linker could extend the C-terminal domain to the exterior of the helix. A complementary argument for the internal location is provided by engineering a full-length measles virus N with a hexa-histidine tag fused at the extreme C-terminus. Expressed in E. coli, this construct was incubated with an anti polyhistine-tag antibody and centrifuged through a glycerol cushion to eliminate unbound antibody. More precisely, an excess of anti polyhistine-tag antibody (from Sigma) was added to purified N_{H6}-RNA (and to non tagged N-RNA purified from E. Coli for control). This mixture was deposited on top of a 450 µl 30% (v/v) glycerol cushion and centrifuged for one hour at 45,000 rpm and 4 °C in an SW55 rotor (Beckman; 192,000g). The pellet was resuspended in 200 ml of buffer and the remaining glycerol dialysed away in 150 mM NaCl, 20 mM Tris–HCl (pH 7.5). The presence of antibody and nucleoprotein in the supernatants and pellets was checked by SDS-PAGE. The dialysed re-suspended pellets were used for negative staining.

Whereas typical helical nucleocapsids could be observed in the tagged preparation prior to antibody binding as well as in the non tagged control, the antibody bound C-terminally tagged nucleocapsids gave rise to clearly distorted aggregates **(Figure 2D)**, reinforcing the proposal that the N-tail is indeed located towards the helix interior. Its interaction with the antibody prevents correct helix formation.

Supplementary references

1. **Bhella, D., A. Ralph, and R. P. Yeo.** 2004. Conformational flexibility in recombinant measles virus nucleocapsids visualised by cryo-negative stain electron microscopy and real-space helical reconstruction. J Mol Biol **340:**319-31.
2. **Egelman, E. H.** 2007. The iterative helical real space reconstruction method: surmounting the problems posed by real polymers. J Struct Biol **157:**83-94.
3. **Egelman, E. H.** 2000. A robust algorithm for the reconstruction of helical filaments using single-particle methods. Ultramicroscopy **85:**225-34.
4. **Frank, J., M. Radermacher, P. Penczek, J. Zhu, Y. Li, M. Ladjadj, and A. Leith.** 1996. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J Struct Biol **116:**190-9.
5. **Heymann, J. B., G. Cardone, D. C. Winkler, and A. C. Steven.** 2008. Computational resources for cryo-electron tomography in Bsoft. J Struct Biol **161:**232-42.
6. **Ludtke, S. J., P. R. Baldwin, and W. Chiu.** 1999. EMAN: semiautomated software for high-resolution single-particle reconstructions. J Struct Biol **128:**82-97.
7. **Mindell, J. A., and N. Grigorieff.** 2003. Accurate determination of local defocus and specimen tilt in electron microscopy. J Struct Biol **142:**334-47.
8. **Navaza, J., J. Lepault, F. A. Rey, C. Alvarez-Rua, and J. Borge.** 2002. On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation. Acta Crystallogr D Biol Crystallogr **58:**1820-5.
9. **Pedretti, A., L. Villa, and G. Vistoli.** 2002. VEGA: a versatile program to convert, handle and visualize molecular structure on Windows-based PCs. J Mol Graph Model **21:**47-9.
10. **Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin.** 2004. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem **25:**1605-12.
11. **Ramey, V. H., H. W. Wang, and E. Nogales.** 2009. Ab initio reconstruction of helical samples with heterogeneity, disorder and coexisting symmetries. J Struct Biol **167:**97-105.
12. **Schoehn, G., M. Mavrakis, A. Albertini, R. Wade, A. Hoenger, and R. W. Ruigrok.** 2004. The 12 A structure of trypsin-treated measles virus N-RNA. J Mol Biol **339:**301-12.
13. **Shaikh, T. R., H. Gao, W. T. Baxter, F. J. Asturias, N. Boisset, A. Leith, and J. Frank.** 2008. SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. Nat Protoc **3:**1941-74.
14. **Tawar, R. G., S. Duquerroy, C. Vonrhein, P. F. Varela, L. Damier-Piolle, N. Castagne, K. MacLellan, H. Bedouelle, G. Bricogne, D. Bhella, J. F. Eleouet, and F. A. Rey.** 2009. Crystal structure of a nucleocapsid-like nucleoprotein-RNA complex of respiratory syncytial virus. Science **326:**1279-83.
15. **van Heel, M., G. Harauz, E. V. Orlova, R. Schmidt, and M. Schatz.** 1996. A new generation of the IMAGIC image processing system. J Struct Biol **116:**17-24.

Submitted (co-1st Author) : Self-organization of the Vesicular Stomatitis Virus Nucleocapsid into a Bullet Shape

The following document contains in a row the main text and the supplementary information.

SELF-ORGANISATION OF THE VESICULAR STOMATITIS VIRUS NUCLEOCAPSID INTO A BULLET SHAPE.

Ambroise Desfosses[1,2,#], Euripedes A. Ribeiro Jr[1,#], Guy Schoehn[1,3], Danielle Blondel[4], Delphine Guilligay[1], Marc Jamin[1], Rob W. H. Ruigrok[1] and Irina Gutsche[1]*.

[1] UJF-EMBL-CNRS UMI 3265, Unit of Virus Host Cell Interactions, Grenoble, France,
[2] Structural and Computational Biology Unit, EMBL Heidelberg, Germany,
[3] CEA-CNRS-UJF UMR 5075, Institut de Biologie Structurale-Jean-Pierre Ebel, Grenoble, France,
[4] CNRS UPR 3296, Virologie Moléculaire et Structurale, Gif-sur-Yvette, France
# contributed equally
* correspondence to: gutsche@embl.fr

Online abstract:

The typical bullet shape of Rhabdoviruses is thought to rely on the matrix protein stabilising the nucleocapsid coil. We reconstitute the bullet shaped nucleocapsids of Vesicular Stomatitis Virus in vitro, analyse their nucleation and growth, and provide cryoEM reconstructions of the bullet tip and the helical trunk. These findings bridge the gap between the isolated N-RNA in form of an undulating ribbon, and the tight bullet shaped virion skeleton.

Vesicular Stomatitis Virus (VSV), a Mononegavirales and the prototype Rhabdovirus, encloses a bullet-shaped skeleton made up of a helical trunk topped by a conical tip. The skeleton contains a nucleocapsid template for viral replication and transcription formed by the negative-strand viral RNA coated with nucleoprotein N. Cryo-electron microscopy (cryoEM) analysis of the entire virion recently culminated in a 10 Å 3D resolution reconstruction of the skeleton trunk where the viral matrix protein M bridges the consecutive turns of the N-RNA helix[1]. Here we analyze the polymorphism of purified viral and recombinant N-RNA (Fig. 1, Supplementary Figs. 1 and 2), show that it can fold into flexible bullet-shaped structures in the absence of other viral components and provide cryoEM 3D reconstructions of both the tip and the trunk (Fig. 1, Supplementary Methods and Supplementary Figs. 3, 4 and 5).

At neutral pH, purified N-RNA forms an undulating ribbon at 150 mM NaCl (Fig. 1a) but auto-assembles into a unidirectional necklace of conical tips at low ionic strength (Fig. 1b, c). The tip reconstruction suggests that tip nucleation may start with a ten subunit-turn compatible with the crystallized recombinant N-RNA decameric rings[2]. The tip features ~5 turns (Fig. 1e) with the diameter of the cone's base reaching ~390 Å. These measures of the in vitro reconstituted N-RNA tips agree with the 450 Å outer diameter of the virion N-RNA trunk proposed to be achieved after ~7 turns based on the 2D class averages[1].

Protonation of N at pH 5 and at low ionic strength allows the conical tips to progress into full bullets morphologically similar to the viral skeletons (Fig. 1g, h). The distribution of trunk diameters illustrates their flexibility and ranges from 370 to 415 Å while centered at ~390 Å consistent with the five-turn tips (Supplementary Fig. 4). A ~25 Å resolution 3D reconstruction of this diameter set (Fig. 1f) contains about 33 N subunits per turn in agreement with estimations based on observed top views (Fig. 1d) and on the dihedral angle between N subunits in the virion bullet[1] (Supplementary Methods).

The polymorphism of the VSV nucleocapsid, and in particular the ribbon-to-tip and the tip-to-trunk transitions can be considered in the light of the quasi-equivalence concept conceived for capsid proteins of icosahedral viruses[3] and expanded to helical arrangements[4]. Quasi-equivalent subunit assembly is thought be based on molecular switches that include environment sensitive elements and are often comprised of disordered segments at subunit interfaces[5]. Here we highlight the role of electrostatic interactions in both tip nucleation at neutral pH and in tip-to-trunk transition, probably triggered by neutralization of carboxyl clusters at low pH.

The proper assembly of icosahedral capsids with large T numbers and of the Mononegavirales nucleocapsids involves auxiliary proteins. The N-RNA of Ebola virus from the filovirus family of Mononegavirales requires the matrix protein for condensation into a flexible helix, further stabilized by additional viral proteins[6]. As for the bullet trunk of the VSV virion, its is actually composed of two nested helices: an inner N-RNA helix and an outer M-protein helix supposed to confer the bullet shape architecture to the nucleocapsid core[1,7,8]. Here we show that isolated nucleocapsids can adopt a bullet-shaped structure solely under the effect of pH and ionic strength, and rule out the requirement of other viral components. However, the conformational variability and/or flexibility of the reconstituted nucleocapsids, in particular in terms of their diameter and exact helical symmetry (Supplementary Fig. 4), indicates that the role of an outer scaffold of M in the virion skeleton might be to rigidify the nucleocapsid fixing it precisely at 37.5 subunits per turn as observed in the cryoEM reconstruction[1]. Partial decoration of nucleocapsids by M at pH 5 tightened the N-RNA diameter distribution without modifying the global subunit arrangement (Supplementary Methods and Supplementary Figs. 4 and 5). The 14-GKKSKK-19 residues of M[9] may play a cementing role at neutral but not at low pH where the negative charges of N are already neutralized and N-RNA forms bullets on its own. Our study demonstrates that the information necessary for packaging of the VSV genetic material into bullets is contained in the nucleoprotein alone thus providing a tool for step-by-step analysis of the virion assembly.

References:

1. Ge, P., Tsao, J., Schein, S., Green, T. J., Luo, M. & Zhou, Z. H. Science **327**, 689-693 (2010).
2. Green, T. J., Zhang, X., Wertz, G.W. & Luo, M. Science **313**, 357-360 (2006).
3. Caspar, D.L. & Klug, A. Cold Spring Harb Symp Quant Biol. **27**, 1-24 (1962).
4. Caspar, D.L. Biophys J. **32**, 103-135 (1980).
5. Johnson, J.E. & Speir, J.A. J. Mol. Biol. **269**, 665-675 (1997).
6. Bharat, T.A., Noda, T., Riches, J.D., Kraehling, V., Kolesnikova, L. et al., Proc. Natl. Acad. Sci. U. S. A. **109**, 4275-4280 (2012).
7. Newcomb, W. W. & Brown, J. C. J. Virol. **39**, 295-299 (1981).
8. Newcomb, W. W., Tobin, G. J., McGowan, J. J. & Brown, J. C. J. Virol. **41**, 1055-1062 (1982).
9. Dancho, B., McKenzie, M.O., Connor, J.H. & D.S. Lyles J. Biol. Chem. **284**, 4500-4509 (2009).

Acknowledgements:

We thank Leandro Estrozi and John Briggs for discussions. This work was supported by a CNRS PEPS 2009 grant, the French ANR-07-001-01, the FINOVI foundation, Lyonbiopôle and the Rhône-Alpes region.

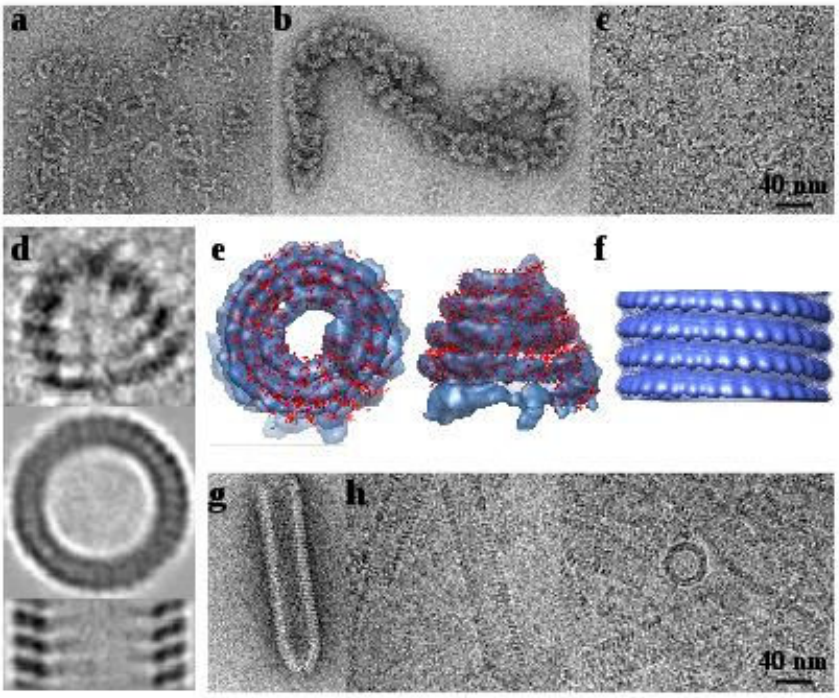


Fig. 1. VSV N-RNA polymorphism. a. Loosely coiled N-RNA ribbons (negative stain). b. Strings of tips (negative stain). c. Strings of tips (cryoEM). d. Representative class averages of tips (from c) and helical trunks (from h, top view of a ~33 subunit/turn bullet and side view). e. 3D CryoEM reconstruction of conical tips (blue) with N subunits (red) placed based on the crystal structure of the VSV N-RNA ring deformed to account for the change in tip radius and subunit inclination. f. 3D CryoEM reconstruction of the ~33 subunit/turn helical trunk. g, h. N-RNA bullets (cryoEM).

Supplementary Materials:
Supplementary Methods
Supplementary Figures 1, 2, 3, 4, 5
References (10-27)

Supplementary Methods

Biochemical Sample Preparation

N-RNA purification from virus infected cells

VSV nucleocapsids were isolated from virus infected cells. Cells were harvested 3 days post-infection and collected in 2 ml of hypotonic buffer (50 mM NaCl, 10 mM Tris-HCl pH 7.4, 1 mM EDTA). The cells were lysed and the supernatant was loaded onto a 20-40% CsCl gradient in 150 mM NaCl, 20 mM Tris-HCl pH 7.5 (buffer A) and centrifuged for 16 h at 30000 r.p.m. at 4 °C in an SW41 rotor. Nucleocapsids were recovered by puncturing the tube at the level of the visible band and were dialyzed against the same buffer without the CsCl. These purified nucleocapsids contained less than 0.001 % of M compared to the M:N ratio in purified virus as determined by Western blot analysis.

Recombinant N-RNA production in insect cells

cDNA of VSV-N protein of the Indiana laboratory strain (Orsay) originally cloned in a pBluescript II vector[10], was amplified by PCR and introduced into the pFastBac HTB plasmid, using RsrII and XhoI restriction sites. The Bac-to-Bac baculovirus (AcMNPV) expression system (Invitrogen) was used to generate recombinant virus. For protein production, Spodoptera Frugiperda Sf 21cells were grown in suspension in SF-900 serum free medium (Gibco BRL) to 0.5 x 10^6 cells/mL and then infected with AcMNPV encoding VSV N protein with a ratio of 1% (volume of virus / volume of cell culture). Protein expression was monitored using the fluorescent marker eYFP, co-integrated in the virus with the gene of the protein[11]. When the specific signal of eYFP reached a plateau 4-5 days after infection, cells were harvested by pelleting at 800g for 10 min and then suspended in buffer A (10 mL/L of cell culture) containing complete^{TM} protease inhibitor cocktail tablets (Roche) and DNAse I (Sigma). Cells were disrupted by three cycles of freezing in liquid nitrogen and thawing at 37°C. Debris was removed by centrifugation for 20 min at 16000 g, 4 °C and the supernatant was layered onto a continuous CsCl gradient as described above. The N-RNA was then dialyzed in buffer A, layered onto a 15% glycerol cushion (v/v in buffer A) and centrifuged as described for the CsCl gradient above. The capsid in the pellet was resuspended in buffer A and stored at 4°C. Protein concentrations were measured by absorbance spectroscopy using the Bio-Rad Bradford assay.

Formation of N-RNA bullets

Immediately before EM analysis, N-RNA samples were extensively dialyzed against MilliQ water at room temperature. The sample was centrifuged at 16,000 g for 1 min at room temperature and the quality of the preparation in the supernatant was checked by SDS-PAGE. The pH of the preparation was then adjusted to 5 or to 7.5 by adding NaAc or Tris-HCl buffer respectively up to 5 mM final buffer concentration. The ribbon-bullet rearrangement relies neither on minor viral contaminants nor on the viral RNA but on the nucleoprotein alone, because it takes place no matter if the N-RNA is purified from virus or if the nucleoprotein is expressed in insect cells (Supplementary Fig. 1) where it binds cellular RNA[12].

Interaction of N-RNA with M and M-N-RNA bullet preparation for EM

VSV M was purified from virus by solubilisation with CHAPS as described[13]. Before the interaction studies with N-RNA, M protein was subjected to a dialysis against MilliQ water, then centrifuged at 80,000 g for 20 min in an Airfuge ultracentrifuge (Beckman Coulter) equipped with a A100/18 rotor in order to remove nucleation sites for self polymerization. We first tested if M could bind to N-RNA at both pH 5 and 7 at low salt. For this, N-RNA in 5 mM NaAc pH 5 or 5 mM Tris-HCl pH 7 was incubated with M at 20°C during 10 min. The final N-RNA concentration was 5 μM. M was added to N-RNA in a 1:3 M:N-RNA molar ratio. The mixtures (20 μL) were loaded on top of a 15% (v/v) glycerol cushion of 400 μL in the buffer of the binding conditions and centrifuged at 25,000 rpm for 2h at 20°C (SW55Ti rotor using Ultra Clear tubes of 0.8 mL). After centrifugation, 20 μL of sample from the top of the glycerol cushion were recovered and the pellet was suspended in 5 μL SDS-PAGE sample buffer. 10 μL of each sample was loaded on a 12% SDS-PAGE and proteins were detected by silver staining. Free N-RNA pelleted through the cushion whereas free M remained at the top. At both pH values M was found in the pellet when mixed with N-RNA (Supplementary Fig. 2). For the EM analysis of the interaction of N-RNA bullets with M, water-dialysed N-RNA and water-dialysed M were mixed at a 1:1 ratio and the pH adjusted to 5 with NaAc buffer (5 mM final buffer concentration). Addition of more M resulted in too much background noise in the EM images.

Sample preparation and Electron microscopy

Negative stain EM

For preparation of negatively stained grids, the sample was applied to the clean side of a thin carbon film on the carbon-mica interface and stained with 2 % (w/v) uranyl acetate.

CryoEM

N-RNA was vitrified as described[14] on carbon-coated quantifoil 3.5/1 grids (Quantifoil Micro Tools GmbH, Germany). The grids were observed with a Phillips CM200 transmission electron microscope with a LaB_6 filament at 200 kV. Images were recorded under low electron dose conditions at 27,500x magnification on Kodak SO-163 films and negatives were digitized with a Zeiss scanner (Photoscan TD) to a pixel size of 2.55 Å at the specimen level. The defocus of the images used for further analysis was approximately 2 to 5 μm as determined from the power spectra. The 88 best micrographs were selected for further analysis.

CryoEM image analysis

Image processing software

Image processing was carried out on a 40 processor Linux cluster in an integrated approach, combining different software packages for different steps of analysis. In particular, the EMAN software package[15] was used for particle selection; CTFFIND[16] for contrast transfer function determination, BSOFT[17] for the CTF correction; Imagic[18] for multivariate statistical analysis, classification and multireference alignment steps; Spider[19,20] for projection matching and 3D reconstruction; the hsearch and himpose programs from the IHRSR package[21,22] for symmetry search and imposition; URO[23] and its graphical version VEDA[24], were used for crystal structure fitting; Pymol[25] and Chimera[26] for visualisation. Supplementary Figure 3 summarizes the flowchart of the image analysis of the helical trunk of the bullet.

2D image processing (steps 1 and 2 in the flowchart Supplementary Figure 3)

Coordinates of the extremities of the helical trunks were recorded with the helix option of the EMAN Boxer tool[15], while paying attention to avoid picking flexible and/or discontinuous fragments. Originally, 612*612 Å overlapping segments were extracted every 10.2 Å along the trunk axis. Approximately 1000 trunks of N-RNA were selected, which resulted in ~65000 segments. The original in-plane angle and filament assignment of the images were recorded for further alignment validation[27]. Individual images were corrected for CTF by phase-flipping, low-pass filtered to 15 Å, normalised (mean=0; sigma=1), and masked by an accordingly rotated smooth-edged rectangular mask of 200 Å length and 560 Å width. A vertical version of these images was created for the classification steps.

The vertical masked images were iteratively aligned perpendicularly to the helical axis and classified in IMAGIC. The eigenvector describing the variability in diameter was used for separation of the initial dataset into subsets of different diameters of 390±40 Å (Supplementary Fig. 4). This is significantly lower than the 450 Å outer diameter of the entire virion nucleocapsid[1] but is consistent with the fewer curls of the in vitro observed tips (see main text). If the dihedral angle between adjacent N subunits in the reconstituted bullets corresponds to the one in the virion, then one can estimate that the in vitro N-RNA bullet trunks would contain between ~31 and ~35 subunits per turn as opposed to 37.5 subunits per turn determined for the virion skeleton[1]. For further analysis, a subset of ~20000 images was obtained by merging classes with diameters of 390±5 Å.

3D Reconstruction (steps 3 and 4 in the flowchart Supplementary Figure 3)

The first trials to perform reconstructions using procedures with helical symmetry search and enforcement (IHRSR) did not converge and were very sensitive to initial symmetry parameters as well as to symmetry search parameters. We therefore decided to start with a reconstruction without any symmetry constraints. A smooth and continuous helix of the pitch determined from the class-averages was used as a starting model for projection matching (PM). The first cycle improved the filament boxing in the direction perpendicular to the helical axis by shifting the images by the integer number of pixels closest to the translation value determined by PM. During the subsequent PM cycles, this boxing was repeated when translations obtained by PM significantly deviated from zero. The images that could not be unambiguously centred were discarded. This rigorous centring procedure was crucial for minimization of the translational search range during PM, which prevents images from alignment with references requiring too large a translation parallel to helical axis (which in turn can cause clustering of image distribution to certain references thus leading to more and more asymmetric 3D reconstructions). For the same purpose, for reconstructions after each PM cycle, the same number of images per on-axis view (perpendicular to the helical axis) was selected based on correlation with the reference. Cross-correlation based selection and other standard single particle-based selection procedures that impose alignment restrains were applied[27].

After approximately 20 cycles, individual subunits were clearly distinguishable (Supplementary Fig. 5b) and the helical symmetry could be visually assessed as ~33 subunits per turn consistent with the rotational symmetry determination of the top view class averages (Fig. 1d) and with estimations based on the entire virion nucleocapsid diameter and symmetry. The stacking of the N subunits in subsequent turns appeared quasi vertical, which is notably different from the 37.5 subunits/turn symmetry of the helical trunk of the entire virus[1]. ~12000 segments of N-RNA were included in the final reconstruction. The PM-reconstruction cycling was then continued with the same image selection criteria as above, supplemented with a step of refining (with the hsearch program of the IHRSR package) and imposing (with the himpose program of the IHRSR package) the symmetry after each iteration. The symmetry parameters converged to 32.8 subunits per turn. The final reconstruction including ~6000 images of N-RNA was filtered to 22 Å (Supplementary Fig. 5a).

Analysis of M-N-RNA images and 3D Reconstruction without symmetry application.

The methodology applied was the same as for the N-RNA in the absence of the matrix protein M. From ~1000 trunks and ~80000 initially selected segments of M-N-RNA, ~20000 segments were included in the final reconstruction obtained without any symmetry constraints which converged to ~32 subunits per turn for the M-N-RNA complex (Supplementary Fig. 5c). Contrary to the N-RNA reconstruction, in the case of M-N-RNA additional stripes of density were noticed all around the map and always at the same distance to the nucleoprotein helix. A comparison with the map of the M-N-RNA helix in the intact virion[1] indicated that these densities might correspond to the M protein. The relatively low density of M in comparison to the nucleoprotein part can be explained by a non-uniform and non-stoechiometric binding of the matrix protein to the nucleoprotein. This would agree with the raw images, where the M density was not clearly visible, and with biochemical data indicating that only partial decoration of the N-RNA bullets could be achieved in vitro because of the previously described self-polymerization of M[13]. The reason for the gap between N and M lies in the strong Fresnel fringes surrounding the nucleoprotein helix. Since the decoration of nucleocapsids by the M protein was clearly sub-stoechiometric, a further symmetry refinement and application (step 4 in the flowchart Supplementary Figure 3) seemes inappropriate.

Reconstruction of separate bullet tips

Negatives were recorded and scanned as for bullets and binned to 5.1 Å at the specimen level. A generous semi-automatic particle selection with the EMAN boxer routine lead to an extraction of a total of 6928 subframes of 128*128 pixels containing individual tips which were CTF-corrected with CTFFIND and BSOFT and low-pass-filtered at 15 Å with Imagic. The data set was translationally but not rotationally aligned relative to the rotationally averaged total sum of the individual images. This translationally centred data set was subjected to multivariate statistical analysis and classification. Characteristic class averages were then used as a set of references for multi reference alignment of each sub frame with Spider[19,20] and the new translational parameters were used to update the boxer coordinates and extract better centred particles. This procedure was repeated several times until the classes became stable and the individual frames well centred. Representative class averages were examined and circular top views as well as typical side views of the tip containing five prominent striations (as the one presented in Figure 1d in the main text) were identified. Five class averages which looked most reminiscent of a side view of the tip were each assigned 180 different angles while keeping the views perpendicular to the tube axis. Thus, five crude 3D volumes of the tip were created by back projection and then averaged together to produce a start model for iterative projection matching with Spider. After ~20 cycles the 3D reconstruction was stable and showed a notable helicity, even if individual N subunits could not be visualised. 4400 particles were included in the final reconstruction which resolution was estimated via Fourier shell correlation to be around 40 Å according to the 0.5 criterium. The X-ray crystal structure of the N10 VSV-N-RNA ring[2] (2GIC.pdb) was placed at the top of the tip for visual comparison with Pymol (Fig. 1e).

Supplementary Fig. 1. Negative stain EM images of bullets formed from recombinant VSV N expressed in insect cells and bound to cellular RNA.

Supplementary Fig. 2. Binding of the Matrix protein (M) to N-RNA at pH 5 (a) and 7 (b). N-RNA at a final concentration of 5 µM was incubated in the presence of M at a molar ratio of 1:3 M:N-RNA for 10 min at room temperature. After centrifugation of the mixture through a 15% (v/v) glycerol cushion at pH 5 (A) or 7 (B), samples of supernatant (S) and pellet (P) were resolved on a 12% SDS-PAGE and detected by silver staining. Lane 1, protein standards with molecular mass indicated on the left; lanes 2 and 3, N-RNA alone; lanes 4 and 5, N-RNA:M in a 3:1 molar ratio; lanes 6 and 7, M alone.

Supplementary Fig. 3. Flowchart of the image analysis procedure for the helical bullet trunk.

Supplementary Fig. 4. Histograms of diameter distribution for the trunk segments. The upper panel shows the diameter distribution of the N-RNA helical trunks in the absence of M (red), the lower in the presence of M (turquoise).

Supplementary Fig. 5. 3D cryoEM reconstruction of the N-RNA helical trunk reconstituted in presence and absence of M. a. The front half of the 3D volume of the N-RNA bullet trunk (left) and a 20 Å thick central slice through this volume (right). b. the same as in a but shown for the intermediate volume calculated before helical symmetry refinement, i.e. after step 3 in the flowchart Supplementary Figure 3) (shown for a comparison with c). c. The front half of the non symmetrised 3D volume of the M-N-RNA bullet trunk (left) and a 20 Å thick central slice through this volume (after step 3 in the flowchart Supplementary Figure 3)

References:

1. Ge, P., Tsao, J., Schein, S., Green, T. J., Luo, M. & Zhou, Z. H. Science **327**, 689-693 (2010).
2. Green, T. J., Zhang, X., Wertz, G.W. & Luo, M. Science **313**, 357-360 (2006).
3. Caspar, D.L. & Klug, A. Cold Spring Harb Symp Quant Biol. **27**, 1-24 (1962).
4. Caspar, D.L. Biophys J. **32**, 103-135 (1980).
5. Johnson, J.E. & Speir, J.A. J. Mol. Biol. **269**, 665-675 (1997).
6. Bharat, T.A., Noda, T., Riches, J.D., Kraehling, V., Kolesnikova, L. et al., Proc. Natl. Acad. Sci. U. S. A. **109**, 4275-4280 (2012).
7. Newcomb, W. W. & Brown, J. C. J. Virol. **39**, 295-299 (1981).
8. Newcomb, W. W., Tobin, G. J., McGowan, J. J. & Brown, J. C. J. Virol. **41**, 1055-1062 (1982).
9. Dancho, B., McKenzie, M.O., Connor, J.H. & D.S. Lyles J. Biol. Chem. **284**, 4500-4509 (2009).
10. Iseni, F., Baudin, F., Blondel, D. & Ruigrok, R.W. RNA **6**, 270-281 (2000).
11. Bieniossek, C., Imasaki, T., Takagi, Y. & Berger, I. Trends in Biochemical Sciences **37**, 49-57 (2012).
12. Iseni, F., Barge, A., Baudin, F., Blondel, D. & Ruigrok, R.W. J. Gen. Virol. **79**, 2909-2919 (1998).
13. Gaudin, Y., Barge, A., Ebel, C. & Ruigrok, R.W. Virology **206**, 28-37 (1995).

14. Dubochet, J., Adrian, M., Chang, J.J., Homo, J.C., Lepault , J. , McDowall, A.W. & Schultz, P. Q. Rev. Biophys. **21**, 129-228 (1988).
15. Ludtke, S.J., Baldwin, P.R. & Chiu, W. J. Struct. Biol. **128**, 82-97 (1999).
16. Mindell, J.A. & Grigorieff, N. J. Struct. Biol. **142**, 334-347 (2003).
17. Heymann, J.B., Cardone, G., Winkler, D.C. & Steven, A.C. J. Struct. Biol. **161**, 232-242 (2008).
18. van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R. & Schatz, M. J. Struct. Biol. 116, 17-24 (1996).
19. Frank, J., Rademacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M. & Leith, A. J. Struct. Biol. 116,190-199 (1996).
20. Shaikh, T.R., Gao, H., Baxter, W.T., Asturias, F.J., Boisset, N., Leith, A. & Frank, J. Nat. Protoc. **3**, 1941-1974 (2008).
21. Egelman, E.H. Ultramicroscopy **85**, 225-234 (2000)
22. Egelman, E.H. J. Struct. Biol. **157**, 83-94 (2007)
23. Navaza, J., Lepault, J., Rey, F.A., Alvarez-Rúa, C. & Borge, J. Acta Crystallogr. D Biol. Crystallogr. **58**, 1820-1825 (2002)
24. http://mem.ibs.fr/VEDA
25. http://www.pymol.org/, DeLano 2002
26. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. & Ferrin, T.E. J. Comput. Chem. **25**, 1605-1612 (2004).
27. Sachse, C., Chen, J.Z., Coureux, P.D., Stroupe, M.E., Fändrich, M. & Grigorieff, N. J. Mol. Biol. **371**, 812-835 (2007).

Fig. S1

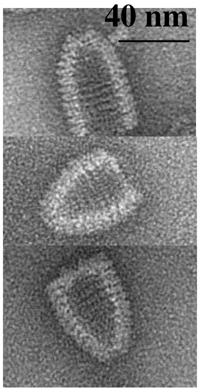

40 nm

Fig. S2

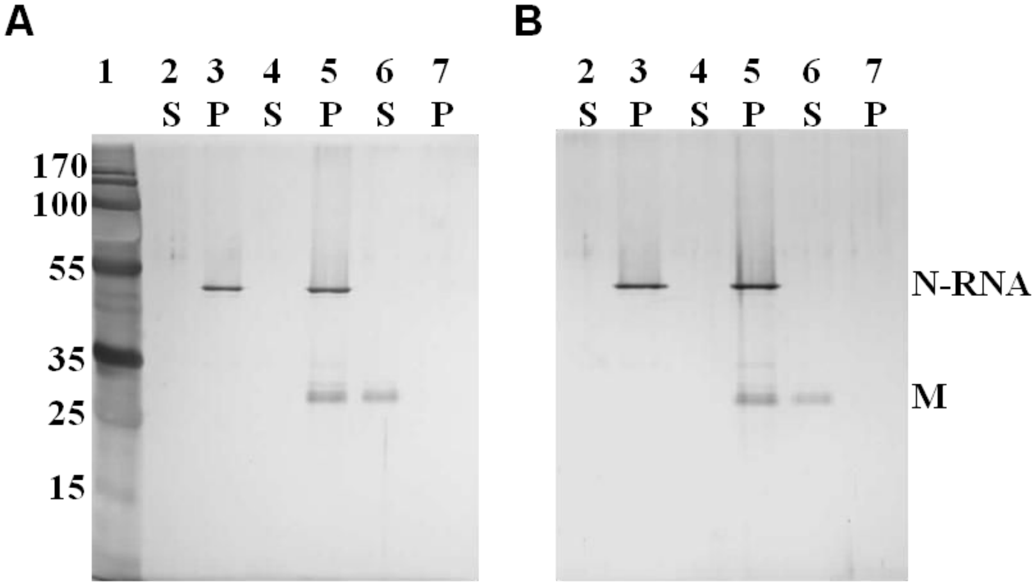

A
1
2
S
3
P
4
S
5
P
6
S
7
P
170
100
55
35
25
15
B
2
S
3
P
4
S
5
P
6
S
7
P
N-RNA
M

Fig. S3

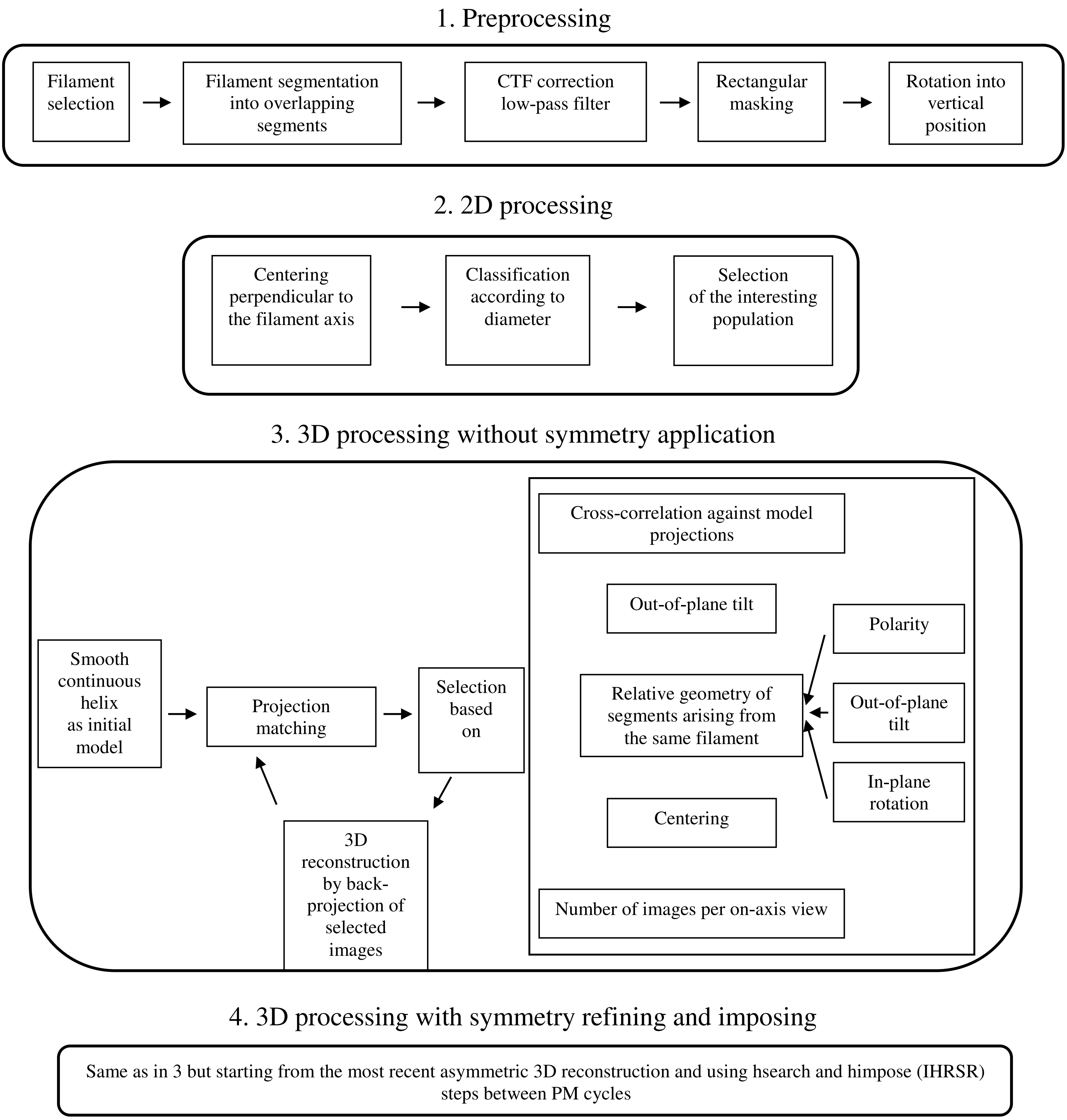
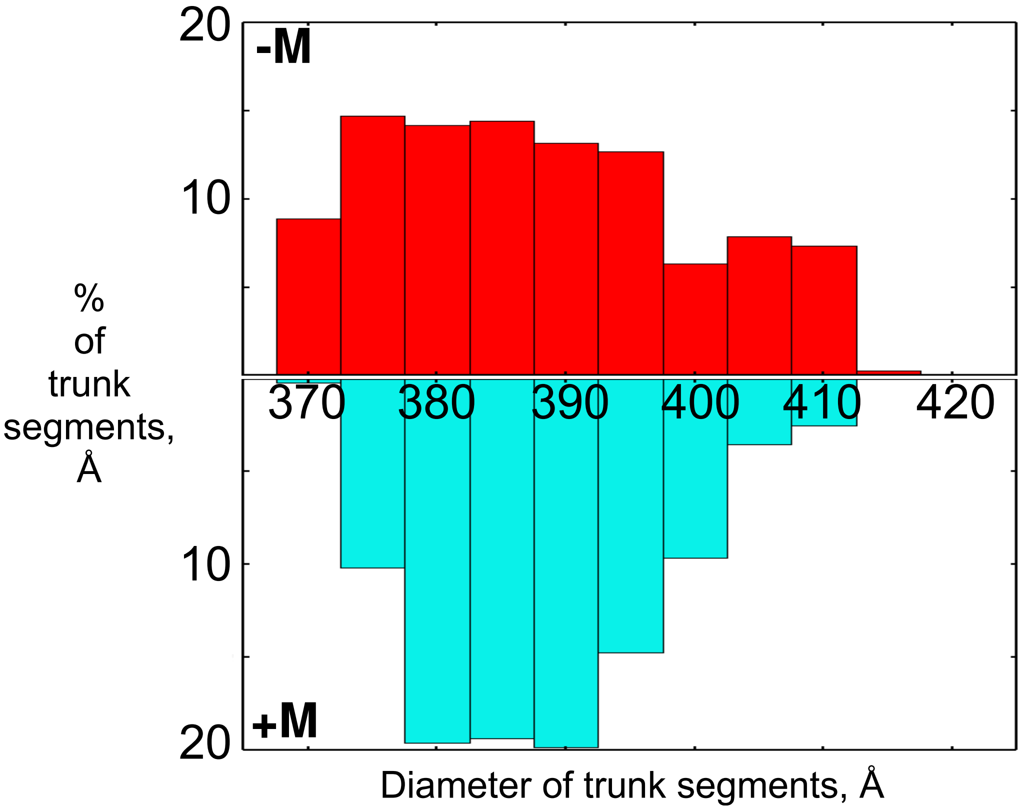
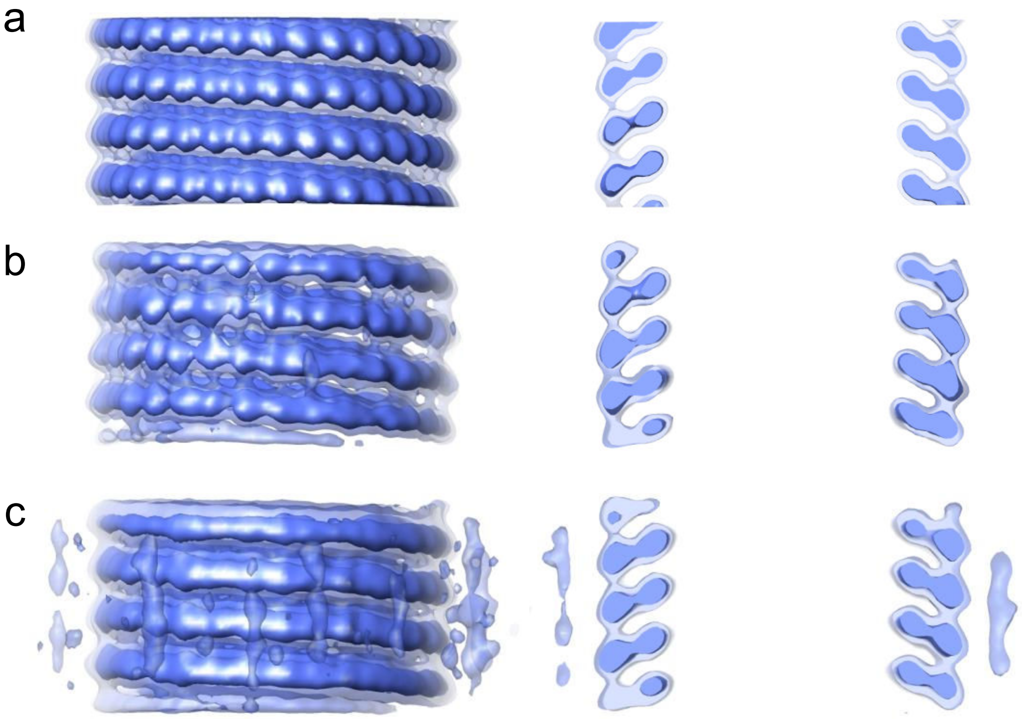

1. Preprocessing
Filament selection
Filament segmentation into overlapping segments
CTF correction low-pass filter
Rectangular masking
Rotation into vertical position
2. 2D processing
Centering perpendicular to the filament axis
Classification according to diameter
Selection of the interesting population
3. 3D processing without symmetry application
Smooth continuous helix as initial model
Projection matching
Selection based on
3D reconstruction by back-projection of selected images
Cross-correlation against model projections
Out-of-plane tilt
Relative geometry of segments arising from the same filament
Polarity
Out-of-plane tilt
In-plane rotation
Centering
Number of images per on-axis view
4. 3D processing with symmetry refining and imposing
Same as in 3 but starting from the most recent asymmetric 3D reconstruction and using hsearch and himpose (IHRSR) steps between PM cycles

Fig. S4



20
-M
10
%
of
trunk
segments,
Å
370
380
390
400
410
420
10
+M
20
Diameter of trunk segments, Å

Fig. S5



a
b
c

Published (5th Author) : Intrinsic disorder in measles virus nucleocapsids

The following document contains the main text and the supplementary information.

Intrinsic disorder in measles virus nucleocapsids

Malene Ringkjøbing Jensen[a], Guillaume Communie[a,b], Euripedes Almeida Ribeiro, Jr.[a,b], Nicolas Martinez[b], Ambroise Desfosses[b], Loïc Salmon[a], Luca Mollica[a], Frank Gabel[a], Marc Jamin[b], Sonia Longhi[c], Rob W. H. Ruigrok[b], and Martin Blackledge[a,1]

[a]Institut de Biologie Structurale Jean-Pierre Ebel, Commissariat à l'Energie Atomique, Centre National de la Recherche Scientifique, Université Joseph Fourier, 41, Rue Jules Horowitz, 38027 Grenoble, France; [b]Unit for Virus Host Cell Interactions, Centre National de la Recherche Scientifique, Université Joseph Fourier, European Molecular Biology Laboratory, 6, Rue Jules Horowitz, 38042 Grenoble, France; and [c]Architecture et Fonction des Macromolécules Biologiques, Centre National de la Recherche Scientifique, Universités d'Aix-Marseille I and II, 163 Avenue de Luminy, 13288 Marseille Cedex 09, France

Edited by Angela M. Gronenborn, University of Pittsburgh School of Medicine, Pittsburgh, PA, and approved April 25, 2011 (received for review February 28, 2011)

The genome of measles virus is encapsidated by multiple copies of the nucleoprotein (N), forming helical nucleocapsids of molecular mass approaching 150 Megadalton. The intrinsically disordered C-terminal domain of N (N_{TAIL}) is essential for transcription and replication of the virus via interaction with the phosphoprotein P of the viral polymerase complex. The molecular recognition element (MoRE) of N_{TAIL} that binds P is situated 90 amino acids from the folded RNA-binding domain (N_{CORE}) of N, raising questions about the functional role of this disordered chain. Here we report the first in situ structural characterization of N_{TAIL} in the context of the entire N-RNA capsid. Using nuclear magnetic resonance spectroscopy, small angle scattering, and electron microscopy, we demonstrate that N_{TAIL} is highly flexible in intact nucleocapsids and that the MoRE is in transient interaction with N_{CORE}. We present a model in which the first 50 disordered amino acids of N_{TAIL} are conformationally restricted as the chain escapes to the outside of the nucleocapsid via the interstitial space between successive N_{CORE} helical turns. The model provides a structural framework for understanding the role of N_{TAIL} in the initiation of viral transcription and replication, placing the flexible MoRE close to the viral RNA and, thus, positioning the polymerase complex in its functional environment.

NMR | SAXS | ensemble description | dynamics | unfolded protein

Measles virus (MeV) is a member of the *Paramyxoviridae* family of the *Mononegavirales* order of negative sense, single stranded RNA viruses. The viral genome is encapsidated by multiple copies of the nucleoprotein (N) forming a helical nucleocapsid. Transcription and replication of the viral RNA are initiated by an interaction between N and the polymerase complex, composed of the phosphoprotein (P) and the RNA-dependent RNA polymerase (1). N consists of two domains: N_{CORE} (residues 1–400), responsible for the interaction with the viral RNA and for maintaining the nucleocapsid structure, and a long intrinsically disordered domain, N_{TAIL} (residues 401–525) serving as the anchor point for the polymerase complex (2, 3). The molecular recognition element (MoRE) (residues 485–502) of the disordered N_{TAIL} interacts with the C-terminal three-helix bundle domain, XD, of P (residues 459–507) (4) and thereby recruits the polymerase complex onto the nucleocapsid template (5, 6).

The realization that intrinsically disordered proteins (IDPs) are functional despite a lack of structure (7–9) has revealed entirely new paradigms that appear to redefine our understanding of the role of conformational flexibility in molecular interactions (10–12). Until now most IDPs have been studied in isolation, or in the presence of a single interaction partner, although it is evident that a real physiological environment could influence the nature and relevance of apparent intrinsic disorder. In this context resolving the question of whether the protein is actually disordered in situ is of paramount importance. In this case the mechanistic role of the extensive disorder present in N_{TAIL} is particularly intriguing, because the MoRE is located at a distance of 90 apparently unfolded amino acids away from the folded N_{CORE} domain that binds the RNA (13). In order to resolve the mechanism by which the remote interaction between N_{TAIL} and the polymerase complex initiates transcription and replication, it is necessary to develop an atomic resolution understanding of molecular disorder in the context of the intact nucleocapsid. Here we use Nuclear Magnetic Resonance (NMR) spectroscopy, small angle scattering (SAS), and electron microscopy (EM) to describe the conformational behavior and mechanistic role of N_{TAIL} in situ.

Results

N_{TAIL} Populates a Dynamic Equilibrium Comprising Preencoded Helical Conformers at the Phosphoprotein Recognition Site. In this study we have developed an atomic resolution ensemble description of isolated N_{TAIL} from MeV using recently developed tools designed to provide quantitative descriptions of conformational equilibria in IDPs on the basis of experimental NMR data (14–16). Chemical shifts (17, 18) and residual dipolar couplings (RDCs) (19, 20), measured in a weakly ordering alignment medium were combined to directly probe the level and nature of residual structure in N_{TAIL}, revealing that while the majority of N_{TAIL} behaves like an intrinsically disordered chain, the MoRE exists in a rapidly interconverting conformational equilibrium between an unfolded form and conformers containing one of four discrete α-helical elements situated around the interaction site (Fig. 1, Fig. S1, Tables S1 and S2). All of these α-helices are found to be stabilized by N-capping interactions mediated by side chains of four different aspartic acids or serines that precede the observed helices (21, 22). N-capping stabilization of helices or turns represents an important mechanism by which the primary sequence encodes prerecognition states in disordered proteins, and has been observed in the proteins Tau (23), Sendai virus N_{TAIL} (19), the N-terminal transactivation domain of p53 (24), and the ribosomal protein L9 (25).

A crystal structure of the chimeric complex between a short construct of N_{TAIL} and XD shows that N_{TAIL} docks as a helix between residues Q486 and A502 (26). This helix is similar to the longest of the four helical elements present in isolated N_{TAIL}. Changes in chemical shifts and RDCs (Fig. 2) confirm that upon binding to XD, the MoRE of N_{TAIL} folds into a helix. However the decreasing values of secondary structure propensity (SSP) (17) and the RDCs towards the ends of the helix indicate

Author contributions: M.R.J., F.G., R.W.H.R., and M.B. designed research; M.R.J., G.C., E.A.R.J., N.M., A.D., L.S., L.M., F.G., M.J., S.L., and M.B. performed research; M.R.J., G.C., F.G., and M.B. analyzed data; and M.R.J., F.G., R.W.H.R., and M.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: martin.blackledge@ibs.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1103270108/-/DCSupplemental.

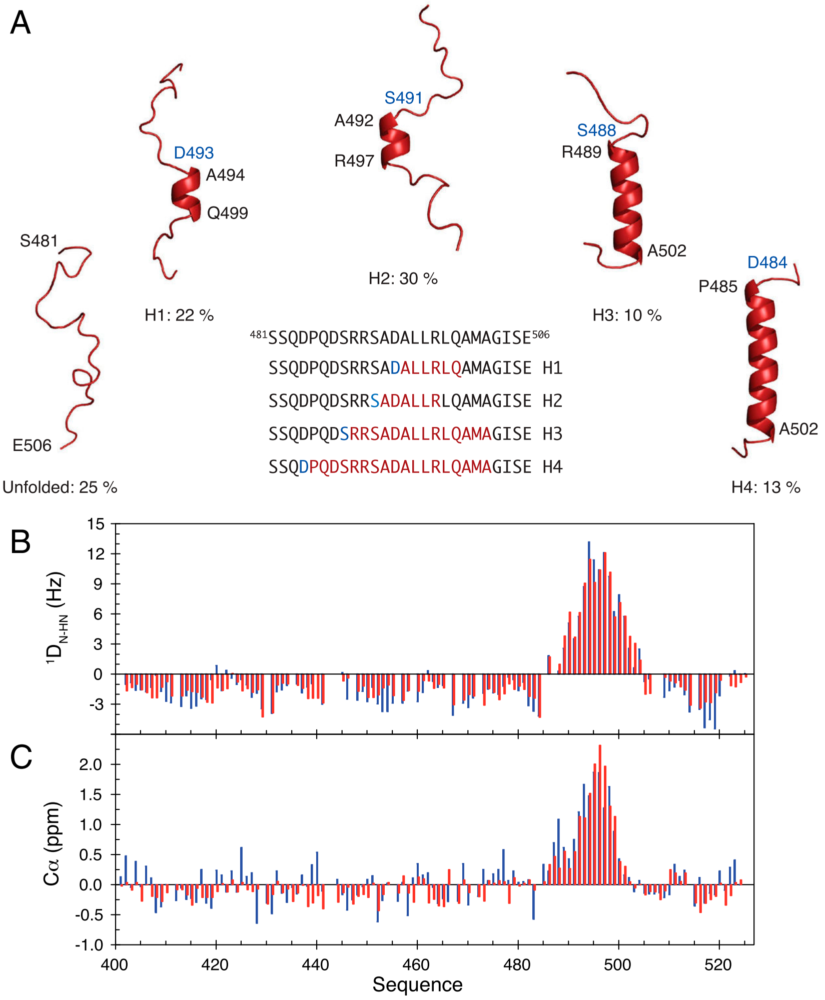


Fig. 1. Ensemble description of the MoRE of N_{TAIL}. (*A*) N_{TAIL} preferentially adopts a dynamic equilibrium between a completely unfolded state and different partially helical conformations each represented by a single cartoon structure for clarity. All helices are stabilized by N-capping interactions through aspartic acids or serines (blue residues). The location of the helices within the MoRE is shown in the primary sequence. (*B*) Comparison of experimental (blue) and back-calculated (red) $D_{N\text{-}HN}$ RDCs from the model of N_{TAIL} shown in (*A*). (*C*) Comparison of experimental (blue) and back-calculated (red) Cα secondary chemical shifts from the model of N_{TAIL} shown in (*A*).

some residual degree of dynamics in the complex. In addition, exchange line broadening persists for residues surrounding the two smallest helices (H1 and H2) present in the conformational equilibrium, even for a large excess of XD compared to N_{TAIL}. There is therefore evidence that both conformational selection from the equilibrium free-form ensemble, and coupled folding and binding, drive the interaction between N_{TAIL} and XD, testifying to the complexity of this highly dynamic interaction.

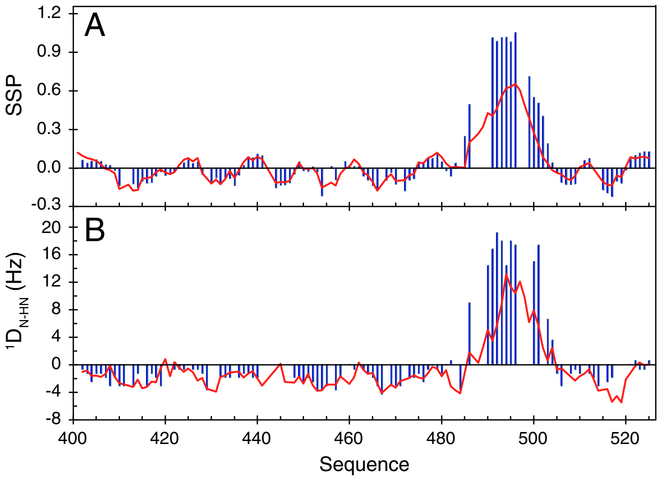


Fig. 2. The MoRE of N_{TAIL} folds upon binding to the XD domain of P protein. (*A*) SSP (17) of N_{TAIL} obtained from experimental Cα and Cβ chemical shifts in free (red) and P (XD) bound (blue) form. (*B*) N-HN RDCs in free (red) and bound (blue) form of N_{TAIL}.

N_{TAIL} Remains Flexible in Intact Nuclecapsids and Binds Transiently to the Capsid Surface. Although the MoRE folds upon binding, the remainder of the 90 amino acid long N-terminal chain between the interaction site and N_{CORE} remains flexible (Fig. 2), again raising the intriguing question of the functional role of this long strand. To extend the investigation of N_{TAIL} to a physiologically relevant environment, we have therefore used solution state NMR to characterize the conformational behavior and flexibility of ${}^{15}N$, ${}^{13}C$ labeled nucleocapsids. From EM (Fig. 3) we estimate the molecular mass distribution of the objects in the NMR sample to fall in a range between 2 to 50 Megadalton that would normally preclude detection of solution state NMR signals of a folded globular protein (27). The heteronuclear single quantum coherence (HSQC) spectrum of the intact capsids however reveals that N_{TAIL} remains flexible when attached to the nucleocapsid. Comparisons of ${}^1H$-${}^{15}N$ (Fig. 3), and ${}^{13}C$-${}^{13}C$ (Fig. S2) correlation spectra of the isolated N_{TAIL} domain and intact nucleocapsids show that the NMR resonances superimpose, demonstrating that the local conformational behavior of residues 450–525 of N_{TAIL} is

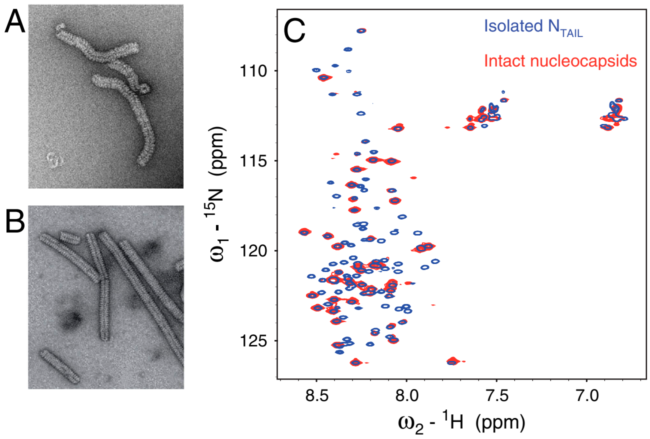


Fig. 3. Electron microscopy and NMR studies of Measles virus nucleocapsids. (*A*) Electron micrograph (negative staining) of the ^{13}C, ^{15}N labeled nucleocapsid sample used for solution NMR studies. (*B*) Electron micrograph of trypsin-digested ^{13}C, ^{15}N labeled nucleocapsids. The solution NMR spectrum of this sample was empty. (*C*) Superposition of the ^{1}H-^{15}N HSQC spectrum of isolated N_{TAIL} (blue) and intact nucleocapsids (red).

retained in situ. However, signals for the first 50 amino acids (residues 401–450) are absent, while large variations of peak intensities indicate differential flexibility along the remainder of the chain, with the MoRE having particularly low intensities (Fig. 4*A*).

To further probe the conformational dynamics of N_{TAIL}, we have measured ^{15}N R_2 spin relaxation rates in isolated N_{TAIL} and intact nucleocapsids (Fig. 4*B*). Isolated N_{TAIL} shows uniform R_2 relaxation rates throughout the sequence, except in the MoRE where the presence of residual helical structure results in elevated rates. ^{15}N R_2 values of N_{TAIL} in the capsid exhibit a very different profile. In the center of the MoRE (around residue 495) R_2 values are similar to the rates in the isolated N_{TAIL} domain, indicating that the MoRE is in slow exchange, while the larger relaxation rates observed at the edges of the MoRE indicate that the same exchange rate appears faster (smaller chemical shift differences) for these sites. These results suggest that the MoRE of N_{TAIL} slowly exchanges on and off the surface of the nucleocapsids. Analysis of the intensity of the peaks shows that more than 95% of the MoRE population is bound. The R_2 values increase dramatically around residue 460, which, combined with the absence of signals of the first 50 residues of N_{TAIL}, indicates that the first stretch of 50 amino acids of the unfolded domain is conformationally restricted. We note that the C terminus of the protein also interacts, either directly with the capsid, or folds back onto the MoRE as it interacts with the capsid.

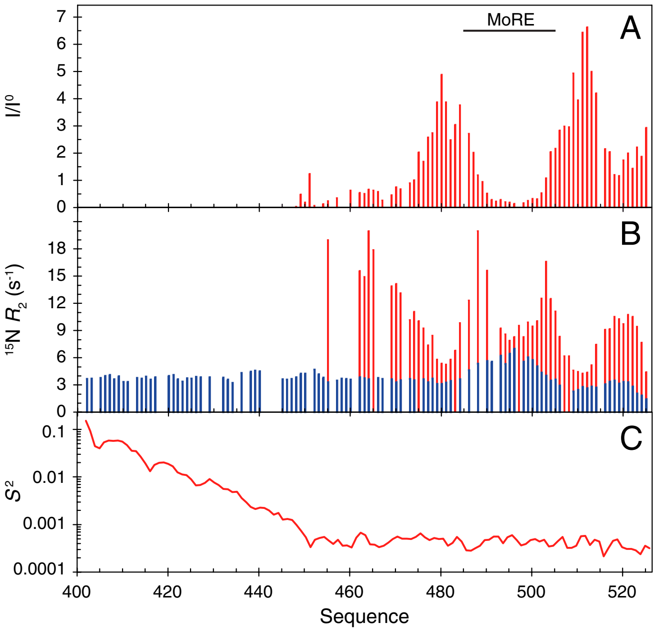


Fig. 4. Dynamics of N_{TAIL} in intact capsids. (*A*) Intensity profile of the ^{1}H-^{15}N HSQC spectrum of intact nucleocapsids. The intensity profile was calculated as the ratio of the intensities (*I*) in the capsid spectrum and the intensities in the spectrum of the free N_{TAIL} domain (I^0). (*B*) Comparison of ^{15}N R_2 relaxation rates measured on a 1 GHz spectrometer in the free form of N_{TAIL} (blue) and in intact nucleocapsids (red). (*C*) N-H angular order parameter S^2 averaged over an ensemble of 5,000 conformers of N_{TAIL} that were calculated as shown in Fig. 5 and described in the *Methods* section.

N_{TAIL} Exfiltrates from Inside to Outside of the Capsid Helix Through the Interstitial Space Between Successive N_{CORE} Helical Turns. MeV nucleocapsids have previously been visualized by EM, exhibiting a characteristic herring-bone appearance (5, 28–31). Nothing is known about the location and conformational state of N_{TAIL} in intact nucleocapsids because N_{TAIL} does not appear to contribute coherently to the reconstructed density from EM, however it is apparent that both the structure and dynamics of the nucleocapsids are significantly modulated by N_{TAIL}. Whereas full-length capsids adopt flexible structures, the capsids become significantly more compact and rigid upon cleavage of the disordered tail (Fig. 3 *A* and *B*) (5, 32, 33). EM also reveals that the diameter of the capsid decreases from 200 to 190 Å and that the pitch decreases from 57.2 to 48.7 Å upon removal of N_{TAIL} (34).

The atomic resolution structure of N_{CORE} of MeV is unknown, however, the structure of the N-RNA complex of Respiratory Syncytial Virus (RSV), another member of the *Paramyxoviridae* family, was recently solved using X-ray crystallography (35), and docked into the EM density map of MeV N-RNA on the basis of secondary structural homology (34). Notably, this coarse docking places the C terminus of N_{CORE}, and therefore the N terminus of N_{TAIL}, at the interior of the helix capsid, raising intriguing questions about the position of N_{TAIL} within the capsid. Due to steric hindrance, the 13 copies of N_{TAIL} per turn of the capsid helix cannot reside in the interior of the capsid and remain flexible enough to give rise to NMR signals. We have therefore investigated whether the disordered N_{TAIL} can escape from the interior of the MeV nucleocapsid helix, as reconstructed using EM, by building explicit models that obey random coil statistics for the conformational sampling of the primary sequence, while avoiding the N_{CORE} domains in the capsid. This model (Fig. 5) demonstrates that N_{TAIL} can indeed exfiltrate from the interior of the capsid via the interstitial space between the N_{CORE} moieties. Importantly, reorientational sampling of the chain calculated over the entire ensemble (Fig. 4*C*), demonstrates that maximal angular freedom is only achieved after approximately 50 amino acids, providing a reasonable explanation for the lack of solution NMR signals up to residue 450. In this case the first 50 amino acids of N_{TAIL} retain conformational disorder, which would also explain why they could not be resolved in the EM reconstruction of the capsids (34).

Small Angle Scattering Confirms Transient Binding of N_{TAIL} MoRE to Capsid Surface. Small angle X-ray and neutron scattering (SAS) provides important information concerning the dimensions of N_{TAIL} in intact nucleocapsids. Despite significant polydispersity in terms of length, the cross-sectional radii of gyration, R_C, of the capsids can be accurately determined from these data. SAS analysis of the scattering length density distribution around the nucleocapsid symmetry axis gives R_C values of (78.0 ± 0.6) Å and (69.5 ± 2.4) Å for the intact and cleaved forms respectively (Fig. 6, Fig. S3). The expected value of R_C calculated from the atomic coordinates of RSV N-RNA docked into the reconstructed electron density of the cleaved MeV capsid gives very

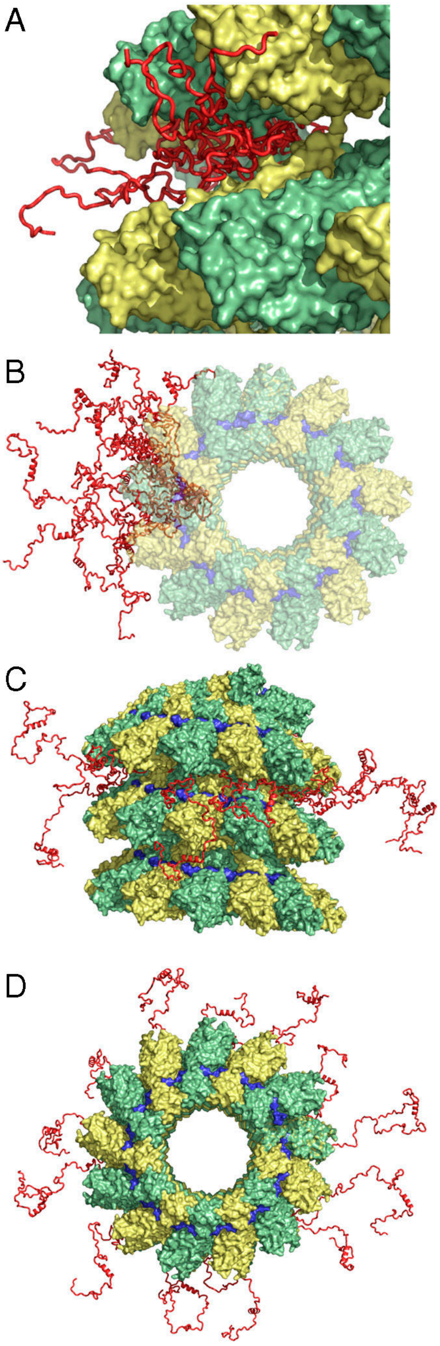


Fig. 5. Proposed model of the location of N_{TAIL} in intact nucleocapsids. The three-dimensional coordinates of the RSV N-RNA subunit docked into the EM density map of MeV N-RNA were used (34). The conformational sampling algorithm *flexible-meccano* was used to build chains from the C terminus of the folded domain of N_{CORE} (successive N_{CORE} monomers are coloured green and yellow). Amino-acid specific conformational sampling allows the chain to escape from the interstitial space of the capsid helix. (*A*) Representation of the conformational sampling of N_{TAIL} from a single N protein in the capsid. Different copies of N_{TAIL} (red) are shown to indicate the available volume sampling of the chain. The first 50 amino acids of N_{TAIL} are shown. (*B*) Representation of the conformational sampling of N_{TAIL} from a single N protein in the capsid, shown along the axis of the nucleocapsid. (*C*, *D*) Representation of the 13 N_{TAIL} conformers from a single turn of the nucleocapsid. In the interests of clarity, (*B*–*D*) deliberately show more conformers outside the capsid, and fewer bound to the surface, than are probable at any one time (see text). The position of the RNA is shown in blue.

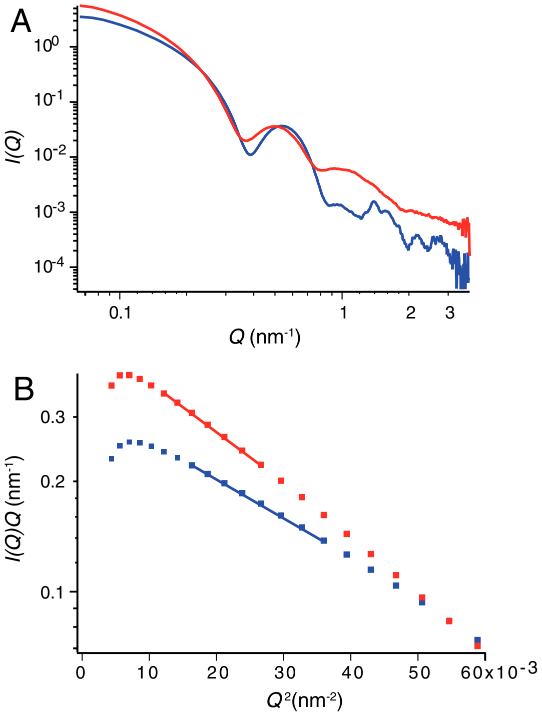


Fig. 6. Small angle X-ray scattering of nucleocapsids. (*A*) Data from intact full-length nucleocapsids (red) and N_{TAIL}-cleaved nucleocapsids (blue). (*B*) Linear fits of $\ln[I(Q)Q] = \ln[I(0)Q] - \frac{1}{2}R_C^2Q^2$ used to extract values of R_C from SAXS and SANS data (Fig. S3) from the cleaved (blue) and noncleaved (red) helical capsids.

good agreement with experiment (68.0 Å), while the calculated model of the capsid with the full-length chain gives a value of 83.8 Å when the MoRE is entirely free, and 78.4 Å when the center of the MoRE is positioned less than 8 Å from any of the folded domains of the capsid. The NMR-based model of a transient interaction between the MoRE and the capsid is therefore strongly supported by the SAS data. These results also provide a steric explanation for the observed decrease in pitch between intact and cleaved capsids (34), as parts of the disordered N_{TAIL} reside in the interstitial space between the N_{CORE} lobes.

Discussion

Measurements of NMR, SAS, and EM on nucleocapsids therefore provide the basis for the development of an in situ ensemble model describing the conformational behavior of N_{TAIL} in intact nucleocapsids. On the basis of this model we are able to provide a structural framework for understanding the dual role of the 125 amino acid intrinsically disordered N_{TAIL} domain. The first 50 disordered amino acids form an articulated spacer that allows the MoRE to escape from the interior of the capsid via the confined interstitial space between successive turns of the helix. The remainder of the chain, on the other hand, is more mobile, and retains the conformational sampling that exists in the isolated form of the protein. This sampling includes the conformational equilibrium of rapidly interconverting helical elements in the MoRE that is predefined by the primary sequence. At the same time the MoRE exchanges on and off the surface of the nucleocapsids, with the majority of conformers in contact with the capsid. The NMR and SAS data indicate that at any given time approximately one of the 13 copies of the nucleoprotein per helical turn is completely free in solution, while the remainder are bound to the capsid surface. While we currently have no information about the position of the binding site, or whether this binding is specific, such a mode of action would provide an efficient mechanism by which N_{TAIL} could "catch" the viral polymerase complex when in free solution, and colocalize the complex on the nucleocapsid surface, thereby initiating transcription and replication of the viral RNA. Interestingly the RNA is sequestered on the outer surface of the RSV and MeV capsids (34, 35), which would place the RNA in the immediate vicinity of N$_{\mathrm{TAIL}}$ as it emerges from the interstitial space, providing a mechanistic rationalization of the entire disordered domain of the nucleocapsid. Further structural and dynamic information will be necessary in order to determine the subsequent sequence of events that follow this initial recognition step, and ultimately lead to transcription and replication.

Methods

Cloning, expression, and purification of the isotopically labeled isolated MeV N$_{\mathrm{TAIL}}$ domain and the C-terminal domain of P (XD) were described previously (13). Cloning, expression, and purification procedures for MeV nucleoproteins are described in *SI Text* (34). Random cellular RNA forms the basis of the reconstituted nucleocapsids which are therefore of variable length.

NMR Experiments. All NMR experiments were carried out at 25 °C. For the measurement of RDCs ^{13}C, ^{15}N labeled N$_{\mathrm{TAIL}}$ was aligned in a liquid crystal composed of poly-ethylene glycol (PEG) and 1-hexanol (36) giving rise to a residual deuterium splitting of 21 Hz. ^{1}D$_{\mathrm{N\text{-}HN}}$, ^{1}D$_{C\alpha\text{-}C'}$ and ^{1}D$_{C\alpha\text{-}H\alpha}$ RDCs were obtained using 3D BEST-type HNCO and HN(CO)CA experiments modified to allow for coupling evolution in the ^{13}C dimension (37). Spectra were acquired with a sweep width of 7.5 kHz and 512 complex points in the ^{1}H dimension and a sweep width of 1.32 kHz and 36 complex points in the ^{15}N dimension. For the ^{13}C dimension, the spectra were acquired with a sweep width of 1.2 kHz and 60 complex points (HNCO-type spectra) and 3 kHz and 60 complex points [HN(CO)CA-type spectrum]. Estimates of experimental errors on the RDCs were obtained through repeated measurements: 1.0 Hz (^{1}D$_{\mathrm{N\text{-}HN}}$), 2.0 Hz (^{1}D$_{C\alpha\text{-}H\alpha}$) and 0.5 Hz (^{1}D$_{C\alpha\text{-}C'}$). Spectra were processed in NMRPipe (38) and analyzed using Sparky (39) and CCPN (40).

The complex between N$_{\mathrm{TAIL}}$ and XD was obtained by preparing a sample containing 0.14 mM ^{15}N, ^{13}C N$_{\mathrm{TAIL}}$ and 1.4 mM unlabeled XD. The complex was aligned in a liquid crystal composed of PEG and 1-hexanol giving rise to a residual deuterium splitting of 26 Hz. ^{1}D$_{\mathrm{N\text{-}HN}}$ were obtained for N$_{\mathrm{TAIL}}$ in the complex using a 2D IPAP SOFAST-HMQC (41) experiment containing 1,024 complex points in the ^{1}H dimension and 150 complex points in the ^{15}N dimension. All RDCs (free and bound form of N$_{\mathrm{TAIL}}$) were measured at a ^{1}H resonance frequency of 600 MHz. ^{15}N R_2 relaxation rates of N$_{\mathrm{TAIL}}$ in its free form and in the context of intact nucleocapsids were measured at a ^{1}H frequency of 1,000 MHz. Standard pulse sequences were used and the spectra were recorded with a sweep width of 14 kHz and 1,024 complex points in the ^{1}H dimension and a sweep width of 3 kHz and 100 complex points in the ^{15}N dimension (42).

Asteroids Description of the Molecular Recognition Element of N$_{\mathrm{TAIL}}$ from NMR Data. Experimental RDCs and Cα chemical shifts were used in a combined approach to obtain an ensemble description of the MoRE of N$_{\mathrm{TAIL}}$ using the minimal ensemble approach (19). A representative ensemble description of the N$_{\mathrm{TAIL}}$ MoRE (defined between residues 485–502) was obtained by generating ensembles of N$_{\mathrm{TAIL}}$ each consisting of 10,000 conformers using *flexible-meccano* (14) with varying helix lengths and positions within the MoRE. One hundred and twenty different ensembles were created to cover the entire MoRE with helices with a minimum length of four residues and a maximum length of 18 residues. Furthermore, an ensemble without helices comprising 50,000 conformers was generated. The alignment tensor of each conformer in the ensembles was calculated using PALES (43, 44) and ensemble-averaged RDCs were obtained for each of the 121 ensembles. Ensemble-averaged chemical shifts were calculated using SPARTA (45) using 1,000 conformers, except for the completely unfolded ensemble where 5,000 conformers were used.

The number of helices, N, necessary to describe the experimental data, and the position and length of the helices, were determined by incrementing N. For each step, the genetic algorithm ASTEROIDS (16) was used to select N helical ensembles and their associated populations such that the predicted population weighted RDCs (Fig. 1*B*, Fig. S1) and chemical shifts (Fig. 1*C*) were in agreement with the experimental values using:

$$O_{\mathrm{CALC}} = \Sigma_{k=1}^{N} p_k O_k + \left(1 - \Sigma_{k=1}^{N} p_k\right) O_U. \tag{1}$$

O_k and O_U are the simulated ensemble-averaged observables for the kth helical and unfolded ensemble, respectively, and p_k is the population associated with the kth ensemble. A χ^2 function is calculated over all residues of the MoRE. Model selection is achieved by optimization of the population and a scaling factor for the RDCs corresponding to the degree of alignment (Table S2). Experimental Cα chemical shift uncertainty used in the combined target function was estimated as 0.3 ppm. The ASTEROIDS selection used 2,000 successive generations and was repeated 10 times for each run to ensure a well defined solution for each value of N (16). Standard F-statistics were used to test the significance of one model over the other (Table S1).

Modelling of N$_{\mathrm{TAIL}}$ in the Context of the Capsid. N$_{\mathrm{TAIL}}$ was built onto the atomic resolution model of N$_{\mathrm{CORE}}$ derived from docking of the RSV N$_{\mathrm{CORE}}$ structure into the EM density of MeV N-RNA capsids. Disordered N$_{\mathrm{TAIL}}$ conformers were built using the *flexible-meccano* algorithm that sequentially constructs peptide chains by randomly sampling amino acid specific dihedral angle distributions (14). Steric clashes are avoided with the folded domains of all copies of N$_{\mathrm{CORE}}$ in the capsid. Angular order parameters relative to the capsid frame were calculated over 5,000 conformers in a single ensemble of independent copies of N$_{\mathrm{TAIL}}$ from the same N protein as described (46).

Small Angle X-Ray Scattering. SAXS experiments were carried out on intact and trypsin-digested nucleocapsids at concentrations of 0.35 mM (intact capsids) and 0.25 mM (digested capsids). All sample volumes were adjusted to 50 µL, and were measured on the high brilliance beamline ID02 at the European Synchrotron Radiation Facility (ESRF) Grenoble, France, using a quartz capillary with 2 mm optical path-length. Scattering data were recorded at a sample-detector distance of 1.5 m at a photon wavelength $\lambda = 0.996$ Å ($E = 12.46$ keV). Both samples and the buffer were exposed for five times 0.1 s. No radiation damage was observed in any case. The corrected one-dimensional intensities $I(Q)$ ($Q = (4\pi/\lambda)\sin\theta$, where 2θ is the scattering angle) from the buffers were subtracted from the respective sample intensities using the SAXS Utilities program (47).

Small Angle Neutron Scattering. Small Angle Neutron Scattering (SANS) experiments were carried out on intact and trypsin-digested nucleocapsids at concentrations of 0.35 mM (intact capsids) and 0.05 mM (digested capsids). All sample volumes were adjusted to 200 µL and were measured on the instrument D22 at the Institute Laue-Langevin (ILL) (Grenoble, France) in Hellma® quartz cuvettes 100QS with 1 mm optical path length. Scattering data were recorded at instrumental configurations (collimator/detector) 2 m/2 m, 8 m/8 m, and 17.6 m/17.6 m at a neutron wavelength $\lambda = 6$ Å. At each configuration, the samples, the buffers, the empty beam, an empty quartz cuvette, as well as a boron sample (electronic background) were measured. Exposure times varied from 30 min to 3 h according to sample and collimator/detector setup. Transmissions were measured during 2 min for each sample. Raw data were reduced using a standard ILL software package (48), normalized to an absolute scale after the various detector corrections and azimuthally averaged to obtain the one-dimensional scattering curve.

Calculation of Cross-Sectional Radius of Gyration from Scattering Data. Assuming capsid structures (hollow cylinders with an overall length much larger than the diameter), scattering curves were analyzed in terms of rod-like shaped particles. Cross-sectional radii of gyration, R_C, were extracted from linear fits of SAXS and SANS data according to (49):

$$\ln[I(Q)Q] = \ln[I(0)Q] - \frac{1}{2} R_C^2 Q^2. \tag{2}$$

$I(0)$ is the cross-sectional part of the scattering. The range of validity of the approximation was reasonably fulfilled in both cases (intact form: $0.92 \le R_C Q \le 1.43$; cleaved form: $0.79 \le R_C Q \le 1.22$).

The experimentally determined cross-sectional radii of gyration (Eq. **2**) were compared to the ones calculated from the atomic resolution structure of RSV docked into the EM density map of MeV N-RNA using the radial coordinates r_i of the N atoms in a unit sectorial element around the cylindrical axis of symmetry:

$$R_C^2 = \frac{1}{N} \Sigma_i r_i^2. \tag{3}$$

ACKNOWLEDGMENTS. We thank Dr. Phil Callow (ILL) for providing local support on the instrument D22, Dr. Theyencheri Narayanan for support using the ID02 beamline at the ESRF, and Stéphanie Costanzo for preparation of the isolated N$_{\mathrm{TAIL}}$ domain and the C-terminal domain of P. We would like to thank the Partnership for Structural Biology for access to biochemical and biophysical platforms, and Irina Gutsche for discussions. This work was supported by the Agence Nationale de la Recherche through

ANR PCV 2007 ProteinMotion (M.B.), ANR "Physico-Chimie du Vivant" ANR-08-PCVI-0020-01 (S.L.), ANR "Jeunes Chercheuses Jeunes Chercheurs" 2010 ProteinDisorder (M.R.J), and the Finovi foundation of Lyon (M.B. and R.W.H.R). M.R.J. benefited from a long-term EMBO fellowship and support from Lundbeckfonden, Denmark. The authors acknowledge use of the 1 GHz spectrometer installed at the French high field NMR center in Lyon through the Très Grandes Infrastructures de Recherche, and acknowledge access to CERM research infrastructure (EC contracts 026145 and 261863).

1. Curran J, Kolakofsky D (1999) Replication of paramyxoviruses. *Adv Virus Res* 54:403–422.
2. Kingston RL, Baase WA, Gay LS (2004) Characterization of nucleocapsid binding by the measles virus and mumps virus phosphoproteins. *J Virol* 78:8630–8640.
3. Curran J, et al. (1993) The hypervariable C-terminal tail of the Sendai paramyxovirus nucleocapsid protein is required for template function but not for RNA encapsidation. *J Virol* 67:4358–4364.
4. Johansson K, et al. (2003) Crystal structure of the measles virus phosphoprotein domain responsible for the induced folding of the C-terminal domain of the nucleoprotein. *J Biol Chem* 278:44567–44573.
5. Longhi S, et al. (2003) The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J Biol Chem* 278:18638–18648.
6. Bourhis J, et al. (2004) The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. *Virus Res* 99:157–167.
7. Dunker AK, et al. (2002) Intrinsic disorder and protein function. *Biochemistry* 41:6573–6582.
8. Tompa P (2002) Intrinsically unstructured proteins. *Trends Biochem Sci* 27:527–533.
9. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6:197–208.
10. Tompa P, Fuxreiter M (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci* 33:2–8.
11. Sugase K, Dyson HJ, Wright PE (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* 447:1021–1025.
12. Bracken C, Iakoucheva LM, Romero PR, Dunker AK (2004) Combining prediction, computation and experiment for the characterization of protein disorder. *Curr Opin Struct Biol* 14:570–576.
13. Gely S, et al. (2010) Solution structure of the C-terminal X domain of the measles virus phosphoprotein and interaction with the intrinsically disordered C-terminal domain of the nucleoprotein. *J Mol Recognit* 23:435–447.
14. Bernadó P, et al. (2005) A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering. *Proc Natl Acad Sci USA* 102:17002–17007.
15. Jensen MR, et al. (2009) Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. *Structure* 17:1169–1185.
16. Nodet G, et al. (2009) Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. *J Am Chem Soc* 131:17908–17918.
17. Marsh JA, Singh VK, Jia Z, Forman-Kay JD (2006) Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. *Protein Sci* 15:2795–2804.
18. Jensen MR, Salmon L, Nodet G, Blackledge M (2010) Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *J Am Chem Soc* 132:1270–1272.
19. Jensen MR, et al. (2008) Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein. *J Am Chem Soc* 130:8055–8061.
20. Jensen MR, Blackledge M (2008) On the origin of NMR dipolar waves in transient helical elements of partially folded proteins. *J Am Chem Soc* 130:11266–11267.
21. Serrano L, Fersht AR (1989) Capping and alpha-helix stability. *Nature* 342:296–299.
22. Serrano L, Sancho J, Hirshberg M, Fersht AR (1992) Alpha-helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. *J Mol Biol* 227:544–559.
23. Mukrasch MD, et al. (2007) Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. *J Am Chem Soc* 129:5235–5243.
24. Wells M, et al. (2008) Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc Natl Acad Sci USA* 105:5762–5767.
25. Luisi DL, Wu WJ, Raleigh DP (1999) Conformational analysis of a set of peptides corresponding to the entire primary sequence of the N-terminal domain of the ribosomal protein L9: evidence for stable native-like secondary structure in the unfolded state. *J Mol Biol* 287:395–407.
26. Kingston RL, Hamel DJ, Gay LS, Dahlquist FW, Matthews BW (2004) Structural basis for the attachment of a paramyxoviral polymerase to its template. *Proc Natl Acad Sci USA* 101:8301–8306.
27. Szymczyna B, Gan L, Johnson J, Williamson J (2007) Solution NMR studies of the maturation intermediates of a 13 MDa viral capsid. *J Am Chem Soc* 129:7867–7876.
28. Finch JT, Gibbs AJ (1970) Observations on the structure of the nucleocapsids of some paramyxoviruses. *J Gen Virol* 6:141–150.
29. Lund GA, Tyrrell DL, Bradley RD, Scraba DG (1984) The molecular length of measles virus RNA and the structural organization of measles nucleocapsids. *J Gen Virol* 65:1535–1542.
30. Fooks AR, et al. (1993) Measles virus nucleocapsid protein expressed in insect cells assembles into nucleocapsid-like structures. *J Gen Virol* 74:1439–1444.
31. Bhella D, Ralph A, Murphy LB, Yeo RP (2002) Significant differences in nucleocapsid morphology within the Paramyxoviridae. *J Gen Virol* 83:1831–1839.
32. Schoehn G, et al. (2004) The 12 A structure of trypsin-treated measles virus N-RNA. *J Mol Biol* 339:301–312.
33. Bhella D, Ralph A, Yeo RP (2004) Conformational flexibility in recombinant measles virus nucleocapsids visualised by cryo-negative stain electron microscopy and real-space helical reconstruction. *J Mol Biol* 340:319–331.
34. Desfosses A, Goret G, Estrozi LF, Ruigrok RWH, Gutsche I (2011) Nucleoprotein-RNA orientation in the measles virus nucleocapsid by three-dimensional electron microscopy. *J Virol* 85:1391–1395.
35. Tawar RG, et al. (2009) Crystal structure of a nucleocapsid-like nucleoprotein-RNA complex of respiratory syncytial virus. *Science* 326:1279–1283.
36. Rückert M, Otting G (2000) Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments. *J Am Chem Soc* 122:7793–7797.
37. Lescop E, Schanda P, Brutscher B (2007) A set of BEST triple-resonance experiments for time-optimized protein resonance assignment. *J Magn Reson* 187:163–169.
38. Delaglio F, et al. (1995) NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6:277–293.
39. Goddard T, Kneller D *SPARKY 3* (University of California, San Francisco).
40. Vranken WF, et al. (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* 59:687–696.
41. Kern T, Schanda P, Brutscher B (2008) Sensitivity-enhanced IPAP-SOFAST-HMQC for fast-pulsing 2D NMR with reduced radiofrequency load. *J Magn Reson* 190:333–338.
42. Farrow N, et al. (1994) Backbone dynamics of a free and a phosphopeptide-complexed SRC homology-2 domain studied by N-15 NMR relaxation. *Biochemistry* 33:5984–6003.
43. Zweckstetter M (2008) NMR: prediction of molecular alignment from structure using the PALES software. *Nat Protoc* 3:679–690.
44. Zweckstetter M, Bax A (2000) Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. *J Am Chem Soc* 122:3791–3792.
45. Shen Y, Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J Biomol NMR* 38:289–302.
46. Markwick PRL, et al. (2009) Toward a unified representation of protein structural dynamics in solution. *J Am Chem Soc* 131:16968–16975.
47. Sztucki M, Narayanan T (2007) Development of an ultra-small-angle X-ray scattering instrument for probing the microstructure and the dynamics of soft matter. *J App Crystallogr* 40:S459–S462.
48. Gosh R, Egelhaaf S, Rennie A (2006) A computing guide for small-angle scattering experiments. Institute Laue Langevin internal report.
49. Porod G (1982) *Small Angle X-ray Scattering*, eds O Glatter and O Kratky (Academic Press, New York).

Supporting Information

Jensen et al. 10.1073/pnas.1103270108

SI Methods

Recombinant Protein Production and Purification. Cloning, expression, and purification of the isotopically labeled isolated measles virus N_{TAIL} domain and the C-terminal domain of P (sometimes known as XD) were described previously (1, 2). Experiments were carried out in 50 mM phosphate buffer at pH 7 with 50 mM NaCl.

Cloning, Expression, and Purification of Intact Measles Virus Nucleocapsids. The cloning procedure of the measles virus nucleoprotein gene (strain Edmonston B) into the expression vector pET22b (+) was described previously (3). The vector was transformed into *Escherichia coli* Rosetta™ (λDE3)/pRARE strain (Novagen) for expression of the recombinant protein. Unlabeled protein was obtained in Luria-Bertani medium, while the uniformly isotopically labeled ^{15}N and $^{15}N/^{13}C$ protein samples were produced in M9 minimal medium supplemented with 1.0 g/L of $^{15}NH_4Cl$, 2.0 g/L of ^{13}C glucose and Minimum Essential Medium (MEM) vitamins (Gibco). The cells were grown at 37 °C until the optical density (OD) at 600 nm reached 0.6 and the protein expression was then induced with 0.5 mM isopropyl-1-thio-β-D-galactopyranoside (IPTG) for 14–16 h at 30 °C. Cells were harvested by centrifugation and then suspended in lysis buffer (10 mL/L of bacteria culture) containing 20 mM Tris-HCl, 150 mM NaCl at pH 7.5 (buffer A), supplemented with 1 mM $MgSO_4$, complete™ protease inhibitor cocktail tablets (Roche), DNAse I (Sigma), and lysozyme (Fluka) and incubated for 30 min on ice. Cells were completely disrupted by sonication on ice and the debris was removed by centrifugation for 20 min at 16,000 × *g*, 4 °C. Typically, 5–8 mL of the supernatant was layered onto a continuous gradient of 23–26 mL of CsCl (20–40% w/w in buffer A). The gradient was centrifuged at 25,000 rpm for 15 h at 12 °C (SW28 Beckman rotor using UltraClear™ tubes of 38.5 mL), and the visible nucleocapsid band was collected by puncturing the tube. The sample was dialyzed into buffer A and layered onto a glycerol cushion 15% (v/v in buffer A) and then centrifuged as described for the CsCl gradient. The capsid on the bottom was resuspended in 50 mM sodium phosphate buffer pH 7.0 with 50 mM NaCl and dialyzed in the same buffer overnight. Sample was centrifuged at 16,000 × *g*, 1 min at 4 °C and the quality of the capsid preparation in the supernatant was checked by SDS-PAGE and electron microscopy (negative staining) as previously described (3). Protein concentrations were measured by absorbance spectroscopy using BioRad Bradford's method based-protein assay. The yield of ^{15}N- and $^{15}N/^{13}C$-labeled measles virus nucleoprotein was about 78 mg/L. The protein solution was frozen in liquid nitrogen and stored at −80 °C at final concentration ranges of 0.2 to 0.4 mM. Trypsin-digested nucleocapsids were obtained as described previously and comprised residues 14–405 (4).

Capsid EM Negative Staining (Sample Quality Control). Noncleaved and cleaved capsids were resuspended and dialyzed in the same buffer used for the NMR studies (e.g., 50 mM sodium phosphate buffer pH 7.0 50 mM NaCl). Samples were centrifuged at 16,000 × *g*, 1 min, 4 °C and the quality of the capsid preparation in the supernatant was checked by SDS-PAGE and electron microscopy. Briefly, the capsids were diluted to a concentration of about 0.1 mg/mL and were adsorbed onto the clean face of a carbon film on mica, negatively stained with 2% (w/v) uranyl acetate and observed under low-dose conditions with a JEOL 1200 EX II microscope at 100 kV and a nominal magnification of 40,000X.

1. Longhi S, et al. (2003) The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J Biol Chem* 278:18638–18648.
2. Gely S, et al. (2010) Solution structure of the C-terminal X domain of the measles virus phosphoprotein and interaction with the intrinsically disordered C-terminal domain of the nucleoprotein. *J Mol Recognit* 23:435–447.
3. Desfosses A, Goret G, Farias Estrozi L, Ruigrok RWH, Gutsche I (2011) Nucleoprotein-RNA orientation in the measles virus nucleocapsid by three-dimensional electron microscopy. *J Virol* 85:1391–1395.
4. Schoehn G, et al. (2004) The 12 A structure of trypsin-treated measles virus N-RNA. *J Mol Biol* 339:301–312.

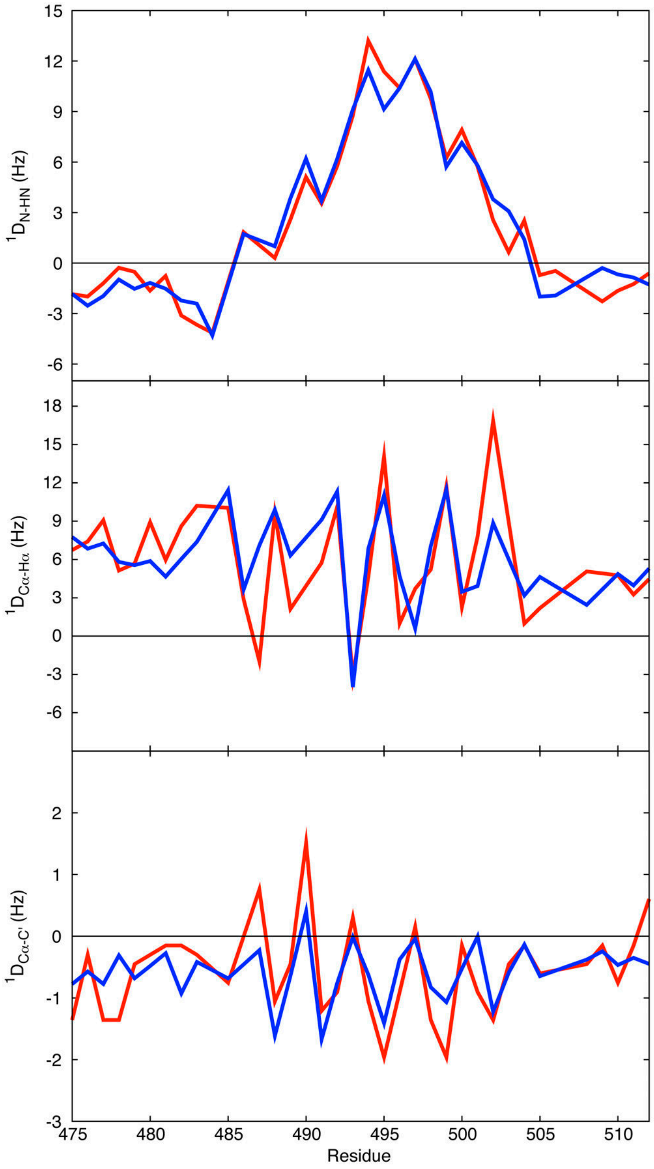


Fig. S1. Comparison of different types of experimental (red) and back-calculated (blue) RDCs in the molecular recognition element of N_{TAIL}. The back-calculated RDCs were obtained as a population-weighted average corresponding to the conformational equilibrium depicted in Fig. 1 (main text).

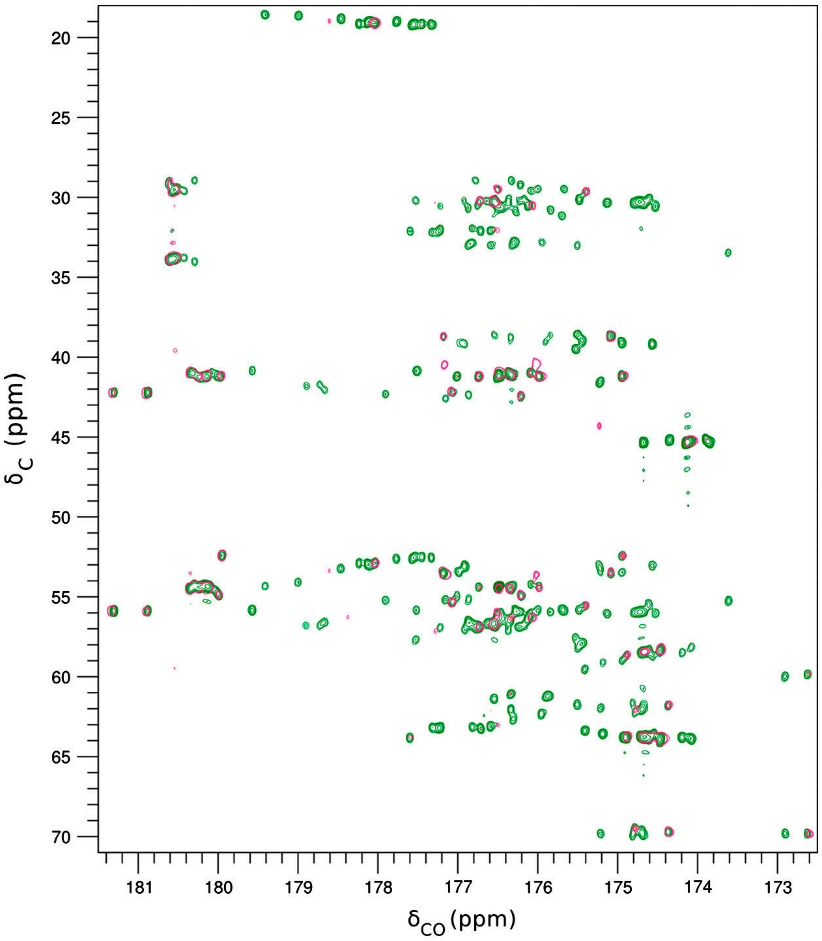


Fig. S2. ^{13}C detected correlation spectra from free N_{TAIL} (green) and full-length capsid (red) recorded at a ^{1}H frequency of 700 MHz and 25 °C. The spectrum of the free N_{TAIL} was acquired using the CBCACO pulse sequence (1, 2) with 1,024 and 192 complex points and sweep widths of 10.5 and 12.7 kHz in the direct and indirect dimensions, respectively. The spectrum of the intact capsid was acquired using the HCBCACO pulse sequence (3) with 1,024 and 192 complex points and sweep widths of 10.5 and 12.7 kHz in the direct and indirect dimensions, respectively.

1 Duma L, Hediger S, Lesage A, Emsley L (2003) Spin-state selection in solid-state NMR. *J Magn Reson* 164:187–195.
2 Bermel W, et al. (2006) Protonless NMR experiments for sequence-specific assignment of backbone nuclei in unfolded proteins. *J Am Chem Soc* 128:3918–3919.
3 Bermel W, et al. (2009) H-start for exclusively heteronuclear NMR spectroscopy: the case of intrinsically disordered proteins. *J Magn Reson* 198:275–281.

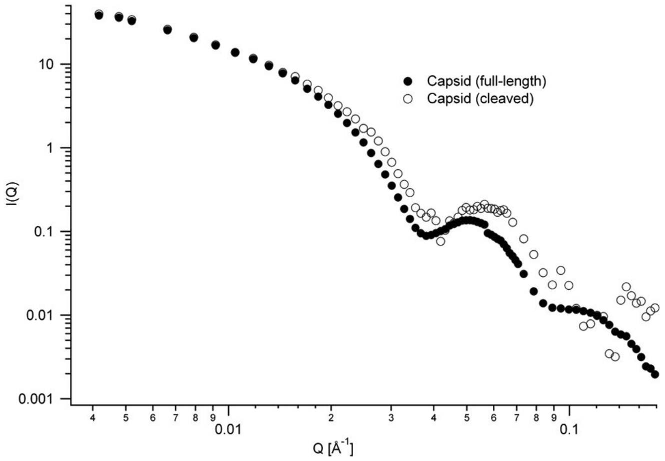


Fig. S3. Small angle neutron scattering data $I(Q)$ of the intact capsid and the cleaved, trypsin-digested form (no N_{TAIL}) in a double-logarithmic representation. The data of the cleaved form are noisier as a consequence of the lower concentration.

Table S1. Data reproduction from ensembles with different combinations of helical conformers

Number of helical conformers	χ^2*	Number of optimized parameters[†]	Helical conformers[‡]	Population (%)[§]	Significance[¶]
1	433	4	485–502	34	
2	231	7	486–498	22	$P < 0.0001$
			492–502	37	
3	186	10	485–502	19	$P < 0.0001$
			492–497	32	
			494–499	23	
4	163	13	485–502	13	$P = 0.0041$
			489–502	10	
			492–497	30	
			494–499	22	
5	154	16	485–502	13	$P = 0.1043$
			489–496	12	
			492–497	19	
			492–502	12	
			494–499	21	

*The target function for the χ^2 included all 114 experimental data points (three types of RDCs and Cα chemical shifts).

[†] One helix implies the optimization of three parameters: starting amino acid, final amino acid, and the population. In addition, a scaling factor is optimized to take into account the absolute level of alignment for the RDCs.

[‡] Range of the invoked helices.

[§] The population of the invoked helices. The remaining conformers are completely unfolded.

[¶] Significance of the improvement of this model as compared to the simpler model calculated using a standard *F*-test.

Table S2. The six best ASTEROIDS solutions assuming that N_{TAIL} samples four specific, helical conformers in conformational equilibrium with a completely unfolded form

Solution	χ^2	Helical conformers	Population (%)
1	163	485–502	13
		489–502	10
		492–497	30
		494–499	22
2	167	485–502	16
		489–499	8
		492–497	30
		494–499	20
3	168	485–502	12
		489–497	17
		492–501	19
		494–499	19
4	169	485–502	11
		489–497	18
		492–502	19
		494–499	21
5	170	485–502	14
		491–495	23
		492–501	17
		494–499	24
6	173	485–502	17
		492–497	23
		492–499	13
		494–499	19

Other publications

Published (2nd Author) : Extracellular complexes of the hematopoietic human and mouse CSF-1 receptor are driven by common assembly principles

Extracellular Complexes of the Hematopoietic Human and Mouse CSF-1 Receptor Are Driven by Common Assembly Principles

Jonathan Elegheert,[1] Ambroise Desfosses,[3] Alexander V. Shkumatov,[4] Xiongwu Wu,[5] Nathalie Bracke,[1] Kenneth Verstraete,[1] Kathleen Van Craenenbroeck,[2] Bernard R. Brooks,[5] Dmitri I. Svergun,[4] Bjorn Vergauwen,[1] Irina Gutsche,[3,*] and Savvas N. Savvides[1,*]

[1]Unit for Structural Biology, Laboratory for Protein Biochemistry and Biomolecular Engineering (L-ProBE)
[2]Laboratory of Eukaryotic Gene Expression and Signal Transduction (LEGEST)
Ghent University, K.L. Ledeganckstraat 35, 9000 Ghent, Belgium
[3]Unit for Virus Host-Cell Interactions, UMI 3265 UJF-EMBL-CNRS, 6 Rue Jules Horowitz, BP 181 38042, Grenoble Cedex 9, France
[4]Biological Small Angle Scattering Group, EMBL, Notkestraße 85, 22603 Hamburg, Germany
[5]Laboratory of Computational Biology, National Heart Lung and Blood Institute, National Institutes of Health (NIH), Bethesda, MD 20892, USA
*Correspondence: gutsche@embl.fr (I.G.), savvas.savvides@ugent.be (S.N.S.)

DOI 10.1016/j.str.2011.10.012

SUMMARY

The hematopoietic colony stimulating factor-1 receptor (CSF-1R or FMS) is essential for the cellular repertoire of the mammalian immune system. Here, we report a structural and mechanistic consensus for the assembly of human and mouse CSF-1:CSF-1R complexes. The EM structure of the complete extracellular assembly of the human CSF-1:CSF-1R complex reveals how receptor dimerization by CSF-1 invokes a ternary complex featuring extensive homotypic receptor contacts and striking structural plasticity at the extremities of the complex. Studies by small-angle X-ray scattering of unliganded hCSF-1R point to large domain rearrangements upon CSF-1 binding, and provide structural evidence for the relevance of receptor predimerization at the cell surface. Comparative structural and binding studies aiming to dissect the assembly principles of human and mouse CSF-1R complexes, including a quantification of the CSF-1/CSF-1R species cross-reactivity, show that bivalent cytokine binding to receptor coupled to ensuing receptor-receptor interactions are common denominators in extracellular complex formation.

INTRODUCTION

Receptor tyrosine kinases (RTKs) are a large family of metazoan-specific cell surface receptors that play essential roles in diverse cellular processes (Lemmon and Schlessinger, 2010). The hallmark of signaling via RTKs lies in cytokine-induced activation of the receptor extracellular segments, which initiates a cascade of intracellular signaling following activation of the intrinsic tyrosine kinase activity of RTKs. Class III RTK (RTKIII) groups four pleiotropic hematopoietic receptors: the prototypic platelet-derived growth factor receptor (PDGFR), colony stimulating factor-1 receptor (CSF-1R), KIT, and fms-like tyrosine kinase III receptor (Flt3). Collectively, intracellular signaling via RTKIII has a major impact in the development and homeostasis of the cellular repertoire throughout the hematopoietic system. RTKIIIs are characterized by a modular structure featuring five extracellular Ig-like domains followed by a single transmembrane helix (TM) and intracellular split kinase domains (Lemmon and Schlessinger, 2010). A remarkable aspect of RTKIII activation is that the cognate protein ligands are all dimeric with similar dimensions despite their grouping into two fundamentally different folds (four helix bundles versus all-β cystine-knot scaffolds) (Jiang et al., 2000; Oefner et al., 1992; Pandit et al., 1992; Savvides et al., 2000; Wiesmann et al., 1997; Zhang et al., 2000). Recently, interleukin-34 (IL-34) was identified as a second ligand to CSF-1R (Lin et al., 2008), thus adding a perplexing dimension to RTKIII signaling because IL-34 bears no sequence similarity to the currently known cytokine ligands for RTKIII/V or other proteins.

Activation of the extracellular segment of human CSF-1R (hCSF-1R) by its two cytokine ligands, hCSF-1 and IL-34, is the cornerstone of signaling cascades central to immunity because CSF-1R:cytokine-signaling complexes are essential for the proliferation, differentiation, and functionality of cells derived from the mononuclear phagocytic lineage, such as monocytes, tissue macrophages, microglia, osteoclasts, and antigen-presenting dendritic cells (Chihara et al., 2010; Chitu and Stanley, 2006; Lin et al., 2008; Wei et al., 2010). Furthermore, signaling via wild-type hCSF-1R and mutants thereof has been implicated in a wide range of pathologies in humans, such as arthritis, atherosclerosis, tumor growth, and metastasis (Chitu and Stanley, 2006).

CSF-1R is arguably the most intriguing member of the RTKIII family for two main reasons: (i) CSF-1R is the only known RTK that is activated by two unrelated protein ligands, and (ii) CSF-1R activation demonstrates restrictive species specificity. For instance mouse CSF-1 (mCSF-1) does not signal through hCSF-1R and other primate CSF-1Rs, yet, hCSF-1 can activate CSF-1R from all primate and nonprimate species tested thus far (Garceau et al., 2010). IL-34, the recently identified second ligand for CSF-1R, appears to follow suit, in that human IL-34 does not activate mCSF-1R, whereas murine IL-34 does signal through hCSF-1R (Wei et al., 2010).

Despite the prominence of hCSF-1R and hCSF-1 in the biomedical literature over the last 3 decades, structural characterization of the extracellular complex has remained elusive, whereas structures of the intracellular kinase domain have only recently become available (Schubert et al., 2007; Walter et al., 2007). Such insights are the missing link to the structural and functional diversity of RTKIII/V extracellular complexes, and would help provide a nearly complete picture of the entire CSF-1 ligand-receptor signaling complex given the available structure of the CSF-1R intracellular kinase domains. A recent flurry of studies of RTKIII/V extracellular complexes led to a structural paradigm for RTKIII/V activation, whereby the receptors bind via their N-terminal Ig-like domains to the activating dimeric cytokine and concomitantly make homotypic contacts between their membrane-proximal domains (Chen et al., 2008; Leppänen et al., 2010; Liu et al., 2007; Ruch et al., 2007; Shim et al., 2010; Verstraete et al., 2011b; Yang et al., 2008, 2010; Yuzawa et al., 2007).

A recent structural study of mCSF-1 in complex with the first three extracellular domains of mCSF-1R (mCSF-1R$_{D1–D3}$) revealed unexpected monovalent binding of mCSF-1 to one mCSF-1R$_{D1–D3}$ molecule leading to a binary complex (Chen et al., 2008), in contrast to predictions based on earlier studies of the homologous murine and human c-kit receptors in complex with stem cell factor (SCF). Although this first structural snapshot of a partial mCSF-1R complex is informative in its own right, it cannot be readily extrapolated to represent CSF-1R activation in general, given the complexity of species cross-reactivity in CSF-1R signaling. Furthermore, the reported binary mCSF-1R$_{D1–D3}$:mCSF-1 complex does not offer realistic insights into possible homotypic receptor interactions, a likely critical element of receptor activation.

Here, we dissect the structural modularity and thermodynamic-binding fingerprints of the extracellular human and mCSF-1:CSF-1R assemblies. Together, our comparative studies provide a comprehensive set of structural and mechanistic insights that now helps to establish a consensus for the assembly of hematopoietic CSF-1 ligand-receptor complexes.

RESULTS AND DISCUSSION

Biochemical and Thermodynamic Characterization of Full-Length CSF-1R Ectodomain Complexes (CSF-1:CSF-1R$_{D1–D5}$)

To enable structural and biophysical studies of human and mCSF-1:CSF-1R$_{D1–D5}$ complexes, we produced recombinant glycosylated human and mouse CSF-1R$_{D1–D5}$ in transiently transfected HEK293T cells in the presence of kifunensine, which limits N-linked glycosylation to Man$_{5–9}$GlcNAc$_2$ glycan structures (Chang et al., 2007). Recombinant human and mouse CSF-1 was produced by in vitro refolding of inclusion bodies after protein expression in *E. coli*. Preparations of purified recombinant hCSF-1 and glycosylated hCSF-1R$_{D1–D5}$ were analytically fractionated by field-flow fractionation (FFF), followed by quantification of their molecular weight (MW) via online multi-angle laser light scattering (MALLS). This led to MW determinations of 35 and 76 kDa, for hCSF-1 and hCSF-1R$_{D1–D5}$, respectively. These values are in excellent agreement with the electrophoretic mobility of dimeric hCSF-1 and monomeric glycosylated hCSF-1R$_{D1–D5}$ on SDS-PAGE (Figure 1A). Titration of hCSF-1R$_{D1–D5}$ with excess molar amounts of cognate CSF-1 resulted in a monodisperse molecular species that exhibited a marked shift in elution profile to a much larger particle (145 kDa as determined by MALLS) when compared to the unbound CSF-1R ectodomain (Figure 1A). Considering the experimental accuracy of MW determination by MALLS, we could infer that the apparent CSF-1:CSF-1R$_{D1–D5}$ complex could be rationalized in terms of one hCSF-1 dimer and two copies of hCSF-1R$_{D1–D5}$.

We employed isothermal titration calorimetry (ITC) to establish the affinity, thermodynamic profile, and stoichiometry of the CSF-1:CSF-1R$_{D1–D5}$ complex. Our results show that the complex is characterized by bivalent binding of hCSF-1 to the receptor ectodomain (one hCSF-1 dimer to two molecules of hCSF-1R$_{D1–D5}$) and that the ensuing high-affinity complex (equilibrium dissociation constant [K_D] = 13.6 nM) is the result of a markedly exothermic binding event coupled to an entropic penalty (Figure 1B; see Table S2 available online). The nanomolar (nM) affinity value we report here for the soluble full-length extracellular complex differs significantly from previously reported K_D values of ~50–100 pM for native hCSF-1R based on cell assays (Roussel et al., 1988). Similar differences have already been observed for a number of systems, including the homologous KIT and Flt3 (Graddis et al., 1998; Lemmon et al., 1997; Lev et al., 1992; Streeter et al., 2001; Verstraete et al., 2011b), and can be attributed to the absence of the TM region and the two-dimensional spatial confinement of the membrane. Upon extending our analysis to the mCSF-1:CSF-1R ectodomain complex, we found that mCSF-1 also binds its cognate mCSF-1R$_{D1–D5}$ in a bivalent fashion to form a high-affinity ternary complex (K_D = 21.7 nM) (Figure 1B) with a similar thermodynamic profile, indicating that the assembly of human and mouse ectodomain complexes is likely based on common principles.

Characterization of the CSF-1 Ligand-Receptor Species Cross-Reactivity

We took advantage of the availability of human and mouse extracellular CSF-1Rs and ligands to quantify their cross-reactivity and to lend further cross-validation to the binding stoichiometries determined for the human and mouse complexes. To our knowledge, this has never been reported while the biomedical literature is heavily populated by studies of hCSF-1 activity in a murine cellular background and vice versa. Such information could have important implications in the design and interpretation of cellular assays testing cytokine:receptor activity from a particular species in a heterologous background. Our experiments revealed bivalent binding of CSF-1 ligands to receptors in both cross-reactivity experiments, consistent with the binding behavior of human and mCSF-1R to their cognate ligands (Figure 2). We calculated a K_D of 66.2 nM for the hCSF-1:mCSF-1R$_{D1–D5}$ interaction, which agrees well with the ability of hCSF-1 to activate all nonprimate CSF-1R tested so far. On the other hand, mCSF-1 binds nearly 500-fold less tightly to hCSF-1R$_{D1–D5}$ (K_D = 2.8 μM) than to its cognate receptor, thus corroborating the observation that mCSF-1 is not able to activate primate CSF-1R in a cellular setting (Figure 2). Together, our binding studies on the assembly of cognate and noncognate

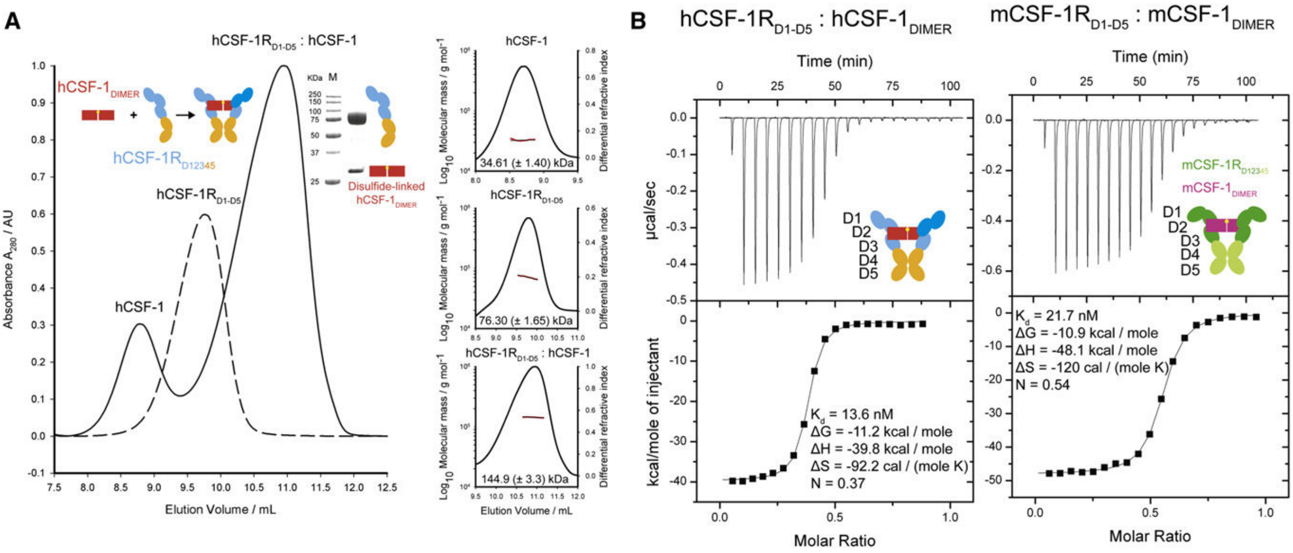


Figure 1. hCSF-1R$_{D1–D5}$ Forms a Ternary Assembly with hCSF-1

(A) Isolation of hCSF-1R$_{D1–D5}$:hCSF-1 by FFF. Formation of the complex leads to a marked shift in elution profile away from the individual protein components after titration with a molar excess of hCSF-1. The different protein components employed are annotated. The inset shows an SDS-PAGE strip of the isolated complex. The disulfide-linked dimeric nature of hCSF-1 is confirmed because the samples are lacking BME. Slight smearing of the hCSF-1R$_{D1–D5}$ band is due to a certain level of heterogeneous glycosylation (Aricescu et al., 2006). The insets show molecular mass determination by MALLS. The measurements confirm the dimeric nature of hCSF-1 and suggest a hCSF-1R$_{D1–D5}$ monomer and a hCSF-1:hCSF-1R$_{D1–D5}$ 1:2 stoichiometry of binding. Derived molecular masses and fits to the experimental LS data are shown.

(B) Titration of hCSF-1 into hCSF-1R$_{D1–D5}$ (left panel) and mCSF-1 into mCSF-1R$_{D1–D5}$ (right panel). Both CSF-1 ligands form a high-affinity ternary complex (n = 1:2) with their cognate receptors.

CSF-1 ligand-receptor complexes show that bivalent cytokine binding to receptor is a conserved mechanistic aspect of the extracellular ligand-receptor interaction.

Electron Microscopy Structure of the Complete Extracellular Assembly of the hCSF-1:CSF-1R Complex

We approached structural characterization of the complete extracellular-signaling complex of hCSF-1R with hCSF-1, based on images of negatively stained hCSF-1R$_{D1–D5}$:hCSF-1 complex obtained by electron microscopy (EM). The recombinant hCSF-1R$_{D1–D5}$:hCSF-1 complex used in the EM analysis was obtained by preparative size-exclusion chromatography (SEC) as a highly monodisperse molecular species. Multivariate statistical analysis (MSA) and classification of circa 18,500 particles indicated the presence of a 2-fold symmetry axis. Thus, an ab initio 3D reconstruction was produced by angular reconstitution with imposed C2 symmetry and further improved by iterative projection matching to generate a 3D reconstruction of the hCSF-1R$_{D1–D5}$:hCSF-1 extracellular complex to ~23 Å resolution (Figures 3A and 3B).

The reconstructed 3D molecular envelope of the hCSF-1R$_{D1–D5}$:hCSF-1 complex reveals a central triangular toroidal structure featuring a pair of appendages extending away from each other at the top of the ring in a plane perpendicular to the toroid, and two in-plane legs of electron density emanating from the bottom of the ring (Figure 3B). Clear features in the electron density strongly suggested that dimeric hCSF-1 binds bivalently to two hCSF-1R$_{D1–D5}$ receptor molecules at the head of the particle, and that the two receptor molecules engage in homotypic interactions away from the ligand-binding epitope. Manual placement of homology models of hCSF-1R$_{D1–D5}$ derived from the structure of the extracellular segment of human KIT (Yuzawa et al., 2007), and of the crystal structure of hCSF-1 (Pandit et al., 1992), into the EM map confirmed this initial interpretation, and showed that the volume of the EM map could readily account for all components of the hCSF-1R extracellular complex. To improve our preliminary model against the experimental EM envelope, we employed a computational approach based on molecular dynamics protocols, which produced 20 different models that were subsequently averaged to yield the final model (Figure 3B; Figure S1A).

The hCSF-1R$_{D1–D5}$:hCSF-1 complex now joins the human KIT$_{D1–D5}$-SCF (Yuzawa et al., 2007) and the human Flt3 ligand-receptor (Verstraete et al., 2011b) complexes as the third complete extracellular RTKIII complex structurally characterized to date, and offers important architectural and functional insights. First, it reveals that the cytokine-binding epitope on hCSF-1R is defined by domains 2 and 3 (Figure 3B). With the exception of the Flt3 ligand-receptor interaction, this feature of receptor-ligand engagement has emerged as a consensus blueprint of RTKIII activation in all other structurally characterized RTKIII complexes thus far (binary mCSF-1R$_{D1–D3}$:mCSF-1 complex, Chen et al., 2008; KIT$_{D1–D3(5)}$:SCF, Liu et al., 2007; Yuzawa et al., 2007; and PDGFR$_{D1–D3}$:PDGF-B, Shim et al., 2010). Second, it shows that receptor homotypic interactions can be attributed to a broad interaction interface between the tandem D4 domains of hCSF-1R, whereas the membrane-proximal D5 domains diverge away to a separation of ~65 Å (Figure 3B). Homotypic receptor interactions have long been considered as the driving force for the cooperative character of

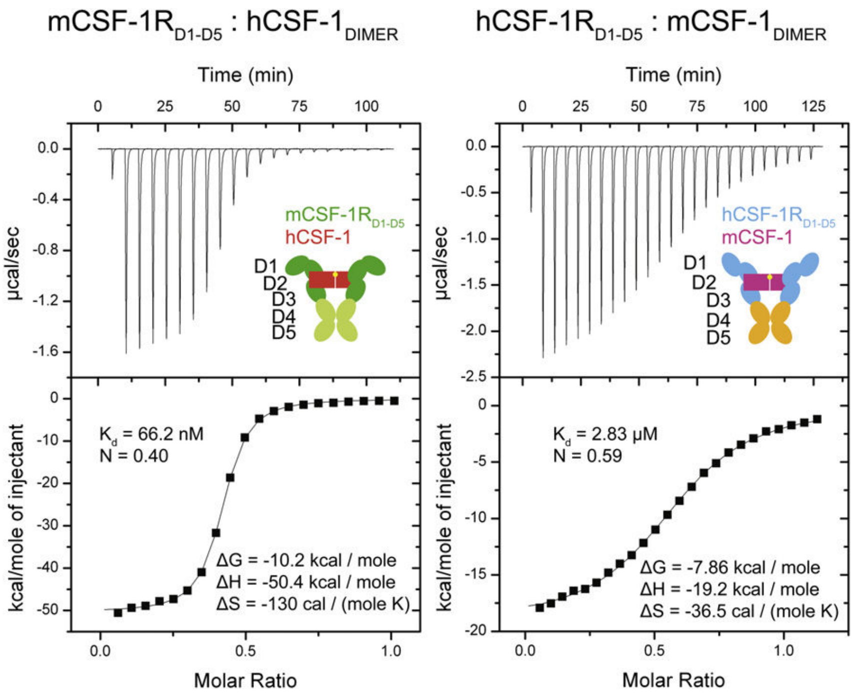


Figure 2. Thermodynamic Characterization of Noncognate Extracellular Human and Mouse CSF-1 Receptor-Ligand Complexes

Thermodynamic measurements of the human and mouse CSF-1R$_{D1-D5}$:CSF-1 species cross-reactivity. In each case CSF-1 was titrated into noncognate CSF-1R. hCSF-1 is able to form a high-affinity complex with mCSF-1R $_{D1-D5}$ (left panel), whereas the mCSF-1:hCSF-1R$_{D1-D5}$ interaction is of much lower strength (right panel). Both complexes display a 1:2 CSF-1:CSF-1R stoichiometry of binding.

extracellular complex formation and activation in RTK. Recent studies on RTKIII receptors KIT and PDGFR showed that receptor contacts mediated by a conserved dimerization sequence fingerprint mapped to the *EF* loop of D4 are important for receptor activation (Yang et al., 2008; Yuzawa et al., 2007) (Figure 3C). Consistent with the proposed key role of the consensus dimerization motif, structural studies on Flt3, the only RTKIII/V receptor lacking this sequence fingerprint, showed that the Flt3 ligand-receptor assembly is devoid of homotypic receptor interactions (Verstraete et al., 2011b).

Whereas our structural studies show that hCSF-1R$_{D4}$ plays a direct role in the CSF-1 extracellular ternary complex, the possible contribution of D5 still remains unclear. The membrane-proximal D5 in KIT$_{D1-D5}$-SCF does not make interactions with its tandem D5 and the corresponding C termini come to 15 Å from each other (Yuzawa et al., 2007). Furthermore, the crystal structure of the complete extracellular Flt3 ligand-receptor complex has recently shown that the two Flt3$_{D5}$ approach each other to about 25 Å (Verstraete et al., 2011b). In hCSF-1R this separation is much larger, thus highlighting the possible conformational diversity of the membrane-proximal domains. Reconciling such interdomain distances in terms of growing evidence on the importance of TM domains in RTK activation (Finger et al., 2009; Li and Hristova, 2006) is not obvious. Yet, it would appear that the linker regions between D5 and the TM domains of RTKIII (typically 10–15 amino acids) would offer the necessary spatial freedom to allow such intramembrane interactions to take place, whereas the D4–D5 interface could help orient such associations. Finally, our studies show that the N-terminal D1 extends well away from the core of the complex without making any interactions with the ligand or other receptor domains. Our computational models show considerable rigid-body flexibility around the D1–D2 linker (Figure S1A). Indeed, the corresponding negative-stain electron density for D1 only became clear in later rounds of image classification. Interestingly, Flt3$_{D1}$ in the Flt3 ligand-receptor complex also emanates away from the core of the complex (Verstraete et al., 2011b). It is currently not clear what the possible role of such flexible D1 modules might be, but it has been suggested that D1 might participate in intermolecular interactions at the cell surface (Verstraete et al., 2011b). However, the apparent conformational independence of D1 in human Flt3 and CSF-1R is not a conserved structural feature within the RTKIII family because structures of the binary mCSF-1R$_{D1-D3}$:mCSF-1 complex, as well as the ternary KIT$_{D1-D3(5)}$:SCF and PDGFR$_{D1-D3}$:PDGF-B complexes, shows that D1 bends downward to interact with D2. We carried out additional measurements on hCSF-1R$_{D1-D5}$:hCSF-1 by small-angle X-ray scattering (SAXS), which consistently corroborate our EM findings, in that the scattering data indicate a P2-symmetric ternary complex with flexible D1 and large divergence of the membrane-proximal D5 (Figure S1B; Table S1).

Structural Plasticity of hCSF-1R$_{D1-D5}$ Revealed by SAXS Analysis of the Unbound Receptor

We carried out measurements on hCSF-1R$_{D1-D5}$ by SAXS to generate structural insights into unbound hCSF-1R and any possible domain rearrangements that might occur upon ligand binding. The X-ray scattering by hCSF-1R$_{D1-D5}$ within a broad concentration range was only consistent with a dimeric species (Figure 4; Table S1). Interestingly, the MW for hCSF-1R$_{D1-D5}$ as determined based on our SAXS data is exactly twice the MW determined via analytical FFF-MALLS measurements conducted at lower concentrations (Figure 1A). This suggests that monomeric and dimeric species for hCSF-1R$_{D1-D5}$ can exist in equilibrium, albeit with a rather poor K$_D$. Molecular envelopes derived from ab initio reconstructions and rigid-body modeling agree remarkably well with each other and point to a well-defined dimeric assembly that lacks internal symmetry (Figure 4). Despite the dramatic deviation from the 2-fold symmetry observed in the receptor:ligand complex (Figure 3B), we note that the extended conformation of the unliganded receptor resembles the bound conformation observed in the EM structure, hinting that preferential structural sampling might facilitate

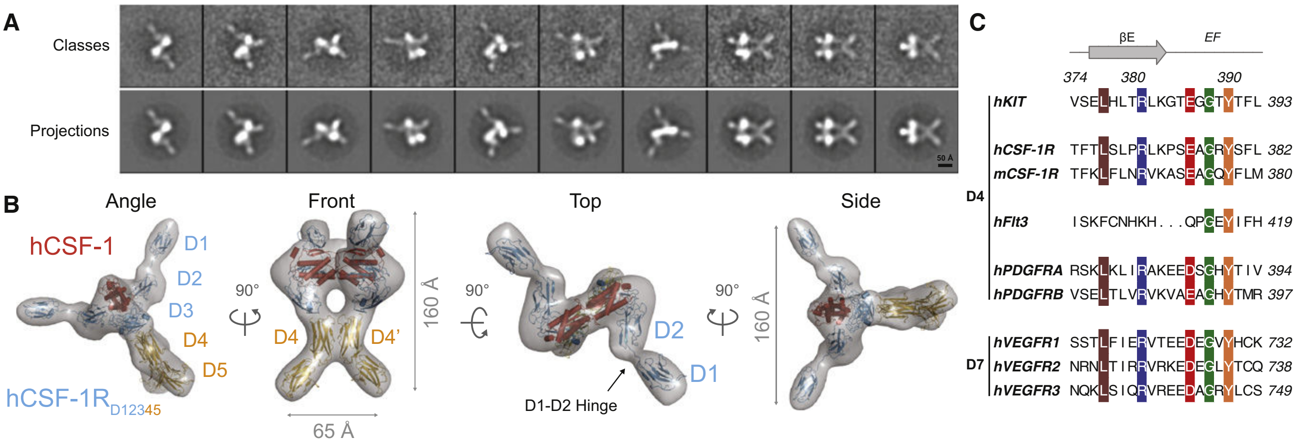

A
Classes
Projections
50 Å
B
Angle
Front
Top
Side
hCSF-1
D1
D2
D3
D4
D5
hCSF-1R$_{D12345}$
90°
D4
D4'
160 Å
65 Å
90°
D2
D1
D1-D2 Hinge
90°
160 Å
C
βE
EF
374
380
390
hKIT VSELHLTRLKGTEGGTYTFL 393
hCSF-1R TFTLSLPRLKPSEAGRYSFL 382
mCSF-1R TFKLFLNRVKASEAGQYFLM 380
D4
hFlt3 ISKFCNHKH. . .QPGEYIFH 419
hPDGFRA RSKLKLIRAKEEDSGHYTIV 394
hPDGFRB VSELTLVRVKVAEAGHYTMR 397
hVEGFR1 SSTLFIERVTEEDEGVYHCK 732
D7
hVEGFR2 NRNLTIRRVRKEDEGLYTCQ 738
hVEGFR3 NQKLSIQRVREEDAGRYLCS 749

Figure 3. Architecture of Liganded hCSF-1R$_{D1–D5}$

(A) Three-dimensional reconstruction of the hCSF-1R$_{D1–D5}$:hCSF-1 complex from EM data. A gallery of representative class averages (above) and reprojections of the final 3D reconstruction (below) under similar orientations is shown.

(B) Angle, front, top, and side orientational views of the reconstructed particle superimposed with computational models of the complex.

(C) Conservation of the D4-D4′ dimerization motif across members of the RTKIII and RTKV families. Residues 374–393 present on the D4 βE strand and *EF* loop of hKIT are aligned with corresponding sequences of h/mCSF-1R, hFlt3, hPDGFR, and hVEGFR. Conserved residues are highlighted. hFlt3 lacks the complete motif and has been shown to be devoid of homotypic receptor contacts (Verstraete et al., 2011).

See also Figure S1.

productive ligand binding. The observed hCSF-1R dimerization in vitro is consistent with previously reported cellular studies that showed the propensity of CSF-1R to form dimers at the cell surface of CSF-1-dependent BAC1.2F5 cells (Li and Stanley, 1991). Thus, the structural view of unbound hCSF-1R analysis of the SAXS data may represent dimeric forms of hCSF-1R at high levels of receptor expression or when the receptors are constitutively activated in disease scenarios. In this respect extracellular receptor predimerization could also play an important role in generating the ultrahigh affinities observed in a physiological setting. Interestingly, a number of other RTKs, such as the IGF1 (Lawrence et al., 2007), EGFR (Chung et al., 2010; Mi et al., 2011), and Eph (Himanen et al., 2007) receptors, do form oligomers in the absence of cytokine ligand. Nonetheless, hCSF-1R$_{D1–D5}$ would have to undergo dramatic domain rearrangements to bind hCSF-1. Such conformational switching has already been observed in the related human KIT (Yuzawa et al., 2007) and human VEGFR (Ruch et al., 2007). Together, our data reinforce the notion that extracellular complex formation is cooperative and relies on an intricate interplay of receptor-ligand interactions, and intramolecular and homotypic receptor contacts.

Human and Mouse CSF-1R$_{D1–D3}$ Can Form Stable Ternary Complexes with Cognate CSF-1 Ligands

A previous structural study of mCSF-1 in complex with the first three extracellular domains of mCSF-1R (mCSF-1R$_{D1–D3}$) revealed an unexpected binary complex, whereby a mCSF-1 dimer binds monovalently to a single mCSF-1R$_{D1–D3}$ molecule (Chen et al., 2008). This is in striking contrast to full-length ectodomain that forms a ternary complex with cognate or noncognate ligand (Figures 1–3). To address this apparent discrepancy in behavior between full-length and truncated receptors and to explore the contribution of the D4–D5 module to the mechanism of ternary complex formation, we produced recombinant glycosylated human and mCSF-1R$_{D1–D3}$ to enable structural and biophysical studies.

Although the full-length ectodomains could readily reach their endpoint assembly even with substoichiometric molar amounts of hCSF-1 using either SEC or FFF methods (Figure 1A), the CSF-1R$_{D1–D3}$ constructs behaved differently (Figure 5). Titrating hCSF-1 with a molar excess of hCSF-1R$_{D1–D3}$ only leads to minor shift on SEC as a shoulder peak of the unbound CSF-1R$_{D1–D3}$ peak (Figure 5A). This behavior is consistent with previous findings (Chen et al., 2008). However, upon titrating hCSF-1R$_{D1–D3}$ with a stoichiometric excess of cytokine ligand, a clear shift can be obtained in the elution profile of hCSF-1R$_{D1–D3}$ on SEC corresponding to a well-defined and markedly larger molecular species (Figure 5A). We sought to obtain more direct evidence into the molecular composition of the two species observed in SEC by attempting to determine their MW via analytical FFF-MALLS. Preparation of the hCSF-1:hCSF-1R$_{D1–D3}$ complex by either a molar excess of hCSF-1 or hCSF-1R$_{D1–D3}$ consistently revealed an ∼65 kDa assembly, consistent with binary complex formation (Figure 5B). This clearly contradicted the chromatographic observation of two different kinds of complexes via SEC (Figure 5A). In an effort to resolve this apparent discrepancy, we applied the peak fraction obtained via SEC by titrating a molar excess of hCSF-1 to hCSF-1R$_{D1–D3}$ to FFF followed by MALLS measurements. This fraction falls apart into two peaks, and the largest molecular species represented a 65 kDa particle as determined by MALLS (Figure 6A). Therefore, we wondered whether the kinetics of molecular diffusion underlying the FFF method combined with a possible instability of the hCSF-1:hCSF-1R$_{D1–D3}$ at such low concentrations might affect the integrity of the complex. To address this, we first subjected the distinct peak of the hCSF-1:hCSF-1R$_{D1–D3}$ complex isolated by SEC (Figure 5A) to crosslinking with formaldehyde followed

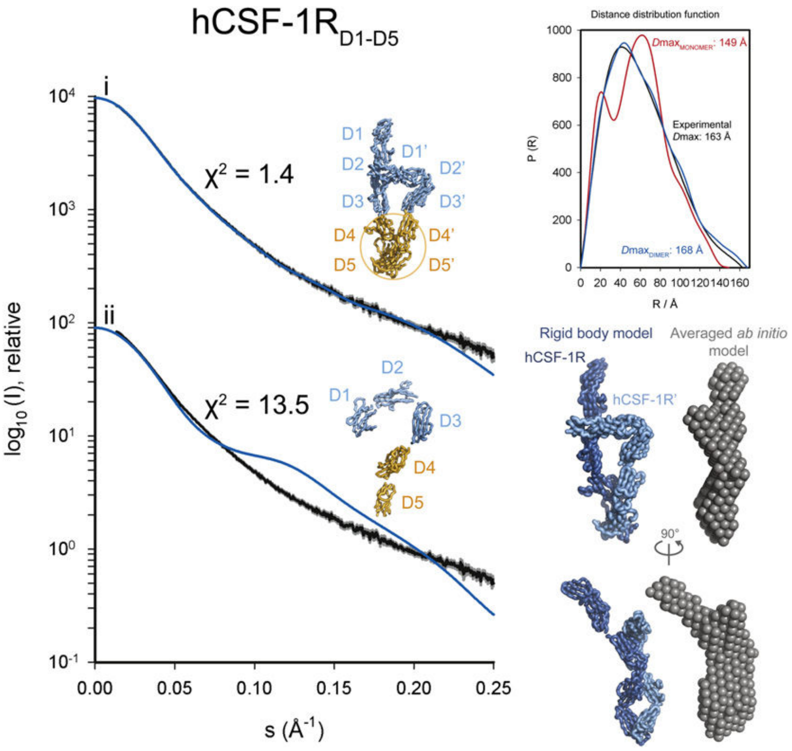


Figure 4. Plasticity of Unliganded hCSF-1R$_{D1–D5}$

Structural analysis of unliganded hCSF-1R$_{D1–D5}$ by SAXS. Experimental scattering curves are shown in black to a maximal momentum transfer of s = 0.25 Å^{-1} (nominal resolution 25 Å), and the individual data:fit pairs are put on an arbitrary y axis to allow for better visualization. Curve "i" shows rigid-body optimized fit of dimeric hCSF-1R$_{D1–D5}$. Modeling was constrained by specifying ambiguous contact distances for the D4–D5 and D4′–D5′ modules (circled). Curve "ii" shows rigid-body optimized fit of receptor domains for monomeric hCSF-1R$_{D1–D5}$. The upper inset shows the calculated distance distribution function for modeled dimeric and monomeric receptors, and their fits with the experimental function. The rigid-body SASREF model and ab initio GASBOR bead model are displayed side to side to highlight agreement in overall shape reconstruction. See also Table S1.

by fractionation via FFF and MALLS measurements. Indeed, this approach led to a dramatically different elution profile on FFF characterized by a single peak corresponding to a molecular species of 109 kDa (Figure 6A). This indicates that both binary and ternary hCSF-1:hCSF-1R$_{D1–D3}$ complexes are possible depending on experimental conditions, and that an apparent prerequisite for the formation and stability of the ternary complex is the presence of a stoichiometric excess of ligand.

We employed ITC to further characterize the interaction between hCSF-1R$_{D1–D3}$ with cognate hCSF-1 and to obtain insights into the contribution of the membrane-proximal module D4–D5 to the extracellular assembly (Figure 6B). First, the binding isotherm could be accurately fitted using a "one set of binding sites" model, and there was no evidence for two sequential or independent binding sites with different affinities. Importantly, the complex displayed a 1:2 stoichiometry of binding revealing bivalent binding of hCSF-1 to hCSF-1R$_{D1–D3}$, in complete agreement with the association mode of the full-length ectodomain complex (Figure 1B). Nonetheless, the strength of the interaction and the corresponding thermodynamic profile differs drastically from that of the hCSF-1R$_{D1–D5}$:hCSF-1 interaction (Figure 6B; Table S2). Notably, hCSF-1 binds 15-fold less tightly to hCSF-1R$_{D1–D3}$ than to full-length extracellular hCSF-1R (K_D = 213 nM versus K_D = 13.6 nM). Thus, the absence of the membrane-proximal module D4–D5 provides a significant enthalpic loss of ~15 kcal mol^{-1} coupled to an entropic gain.

The observation of the bivalent hCSF-1R$_{D1–D3}$:hCSF-1 complex via ITC (n = 0.5) is in stark contrast to the monovalent binding mode reported for the mCSF-1R$_{D1–D3}$:CSF-1 interaction (Chen et al., 2008), thus creating a puzzling paradox with respect to mechanistic aspects of receptor binding and activation. It would indeed seem unlikely that complex formation would bear such fundamental differences in the two homologous systems given the preponderance of conserved sequences on human and mCSF-1 and CSF-1R involved at the interaction epitope (Figure S2). To resolve the apparent disagreement between the two sets of findings, we characterized the assembly of the mCSF-1:CSF-1R$_{D1–D3}$ complex by ITC. Our results based on several experimental replicas show unequivocally that the stoichiometry, corresponding affinities, and thermodynamic profile for mCSF-1R$_{D1–D3}$:mCSF-1 are equivalent to those of the human counterpart (Figure 6B). Furthermore, we conclude that the relative contribution of the membrane-proximal domains to complex formation is similar in the two systems indicating a conserved role for the D4–D5 in the assembly of the extracellular complex. Thus, both the human and mCSF-1 ligand-receptor assemblies appear to share a common interaction mode, based on the inherent capacity of CSF-1 to bind bivalently to its cognate receptor. It is currently unclear why the ITC measurements by Chen et al. (2008) on the mCSF-1R$_{D1–D3}$:CSF-1 interaction deviate so fundamentally from the data we present here. Nonetheless, our combined SEC/FFF/MALLS analysis of the CSF-1:CSF-1R$_{D1–D3}$ complex provides a rationale for the crystallographic observation of the intriguing mCSF-1:CSF-1R$_{D1–D3}$ binary complex (Chen et al., 2008), in the sense that we have shown that both ternary and binary assemblies can be formed for the CSF-1:CSF-1R$_{D1–D3}$ complex depending on experimental conditions.

To provide further structural insights into extracellular complex formation and to investigate further the bivalent mode of CSF-1 binding to CSF-1R revealed by our ITC analysis (Figure 6B), we measured SAXS data for the hCSF-1R$_{D1–D3}$:hCSF-1 and mCSF-1R$_{D1–D3}$:mCSF-1 complexes (Figure 6C; Table S1). Both complexes were prepared via SEC by saturating CSF-1R$_{D1–D3}$ with a molar excess of cognate CSF-1, and were conservatively pooled (Figure 5A). Our data analysis reveals that the crystal structure of the binary mCSF-1:mCSF-1R$_{D1–D3}$ complex (Chen et al., 2008) is grossly incompatible with the SAXS data (Figure 6C, curve i), thereby directly challenging the claim that

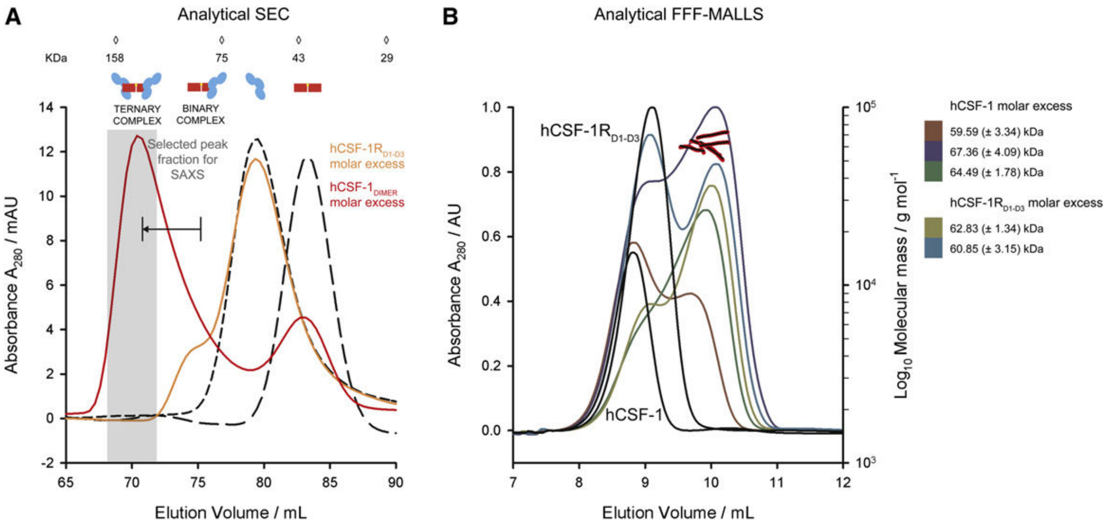


Figure 5. hCSF-1 Can Make Both a Monovalent and Bivalent Complex with hCSF-1R$_{D1-D3}$

(A) Isolation of the hCSF-1R$_{D1-D3}$:hCSF-1 complex by SEC. Titration with either a molar excess of hCSF-1R$_{D1-D3}$ or hCSF-1 leads to different complexes. A marked shift in elution profile away from the individual protein components can only be observed after titration with a molar excess of hCSF-1. The resulting peak fraction has as such been analyzed by SAXS (Figure 6C). The different protein components employed are annotated.

(B) Only the binary complex can be observed by FFF, regardless of stoichiometric excess of any component.

mCSF-1 cannot dimerize mCSF-1R in the absence of the membrane-proximal module D4–D5 (Chen et al., 2008). Both the molecular parameters obtained directly via SAXS and structural modeling of the data showed unambiguously that hCSF-1R$_{D1-D3}$:hCSF-1 and mCSF-1R$_{D1-D3}$:mCSF-1 can form stable ternary complexes with P2 symmetry in solution (Figure 6C, curves ii–iii), thus providing a structural basis for the observed binding stoichiometries via ITC (Figure 6B). Furthermore, we note that the overall features of hCSF-1R$_{D1-D3}$:hCSF-1 in solution are consistent with the corresponding segment in the hCSF-1R$_{D1-D5}$:hCSF-1 EM model, in that D1 points upward, albeit at a slightly different angle (Figure 6C, curve iii).

A Common Assembly Mechanism for Human and Mouse CSF-1 Ligand-Receptor Complexes

The integration of our findings on both the human and mouse CSF-1 ligand-receptor complexes puts our study in position to help resolve a puzzling mechanistic paradox for the assembly of extracellular CSF-1 ligand-receptor complexes that arose from a recent study on the mouse system (Chen et al., 2008). The premise of this study was that mCSF-1 is unable to dimerize its cognate receptor in the absence of the membrane-proximal domains D4 and D5. The authors proposed that formation of a binary complex lowers the affinity of the second binding site on the dimeric cytokine, calling upon a ‘‘negative cooperativity’’ scenario, and extrapolated their reasoning to a distinct mechanistic proposal for CSF-1R activation entailing two steps. In a first step, the ligand and receptor form an initial binary complex with low affinity that can only proceed to the ternary complex upon the simultaneous formation of cytokine-receptor interactions at the opposite binding epitope coupled to homotypic receptor interactions.

The diverse biochemical and structural evidence we reported here illustrates that the assembly of human and mouse extracellular CSF-1 complexes is driven by two common overriding principles. In the first instance, the cytokine ligands have the inherent capacity to offer two receptor binding sites leading to ternary complex formation. Thus, bivalent binding of CSF-1 can take place to the pool of monomeric and dimeric CSF-1R at the cell surface. Second, assembly of the high-affinity complex is dramatically enhanced as a result of well-defined homotypic interactions between extracellular domain 4 modules. This is an example of positive cooperativity, and in the case of CSF-1, this is reflected in a pronounced enthalpy gain upon formation of the ternary complex. This also implies that binding of cytokine ligand to already predimerized CSF-1R would invoke a reorientation of the ectodomains to prime their role in the signaling complex. Together, these two sequential steps constitute a clear consensus for the binding events that lead to the assembly of high-affinity human and mCSF-1 ligand-receptor complexes. It remains to be seen whether IL-34, the newly discovered cytokine ligand for CSF-1R, will follow suit.

EXPERIMENTAL PROCEDURES

Production of Recombinant CSF-1 and CSF-1R Ectodomain Variants and Complexes Thereof

Recombinant human and mCSF-1 were produced as inclusion bodies in a prokaryotic expression system based on a previously described approach (Verstraete et al., 2009) and were purified to homogeneity following in vitro refolding. The fragment encoding residues 1–149 corresponding to the α splice variant of human and mCSF-1 was cloned into the pET-15b vector (Novagen). After expression in the BL21(DE3) CodonPlus-RP (Novagen) *E. coli* strain, h/mCSF-1 accumulated as inclusion bodies. The inclusion bodies were washed three times and then solubilized in 6.5 M GnHCl, 100 mM $NaPO_4$ (pH 8.0), 10 mM Tris, and 10 mM β-mercaptoethanol (BME). Next, denatured h/mCSF-1 was refolded by rapid dilution in refolding buffer (100 mM Tris [pH 8.5], 1 M L-arginine, 3 mM GSH, 1.5 mM GSSG, and 0.2 mM PMSF) at 277 K. The clarified refolding mixture was loaded onto a HisTrap FF 5 ml affinity column, eluted, and subsequently purified by SEC using a Prep-Grade HiLoad

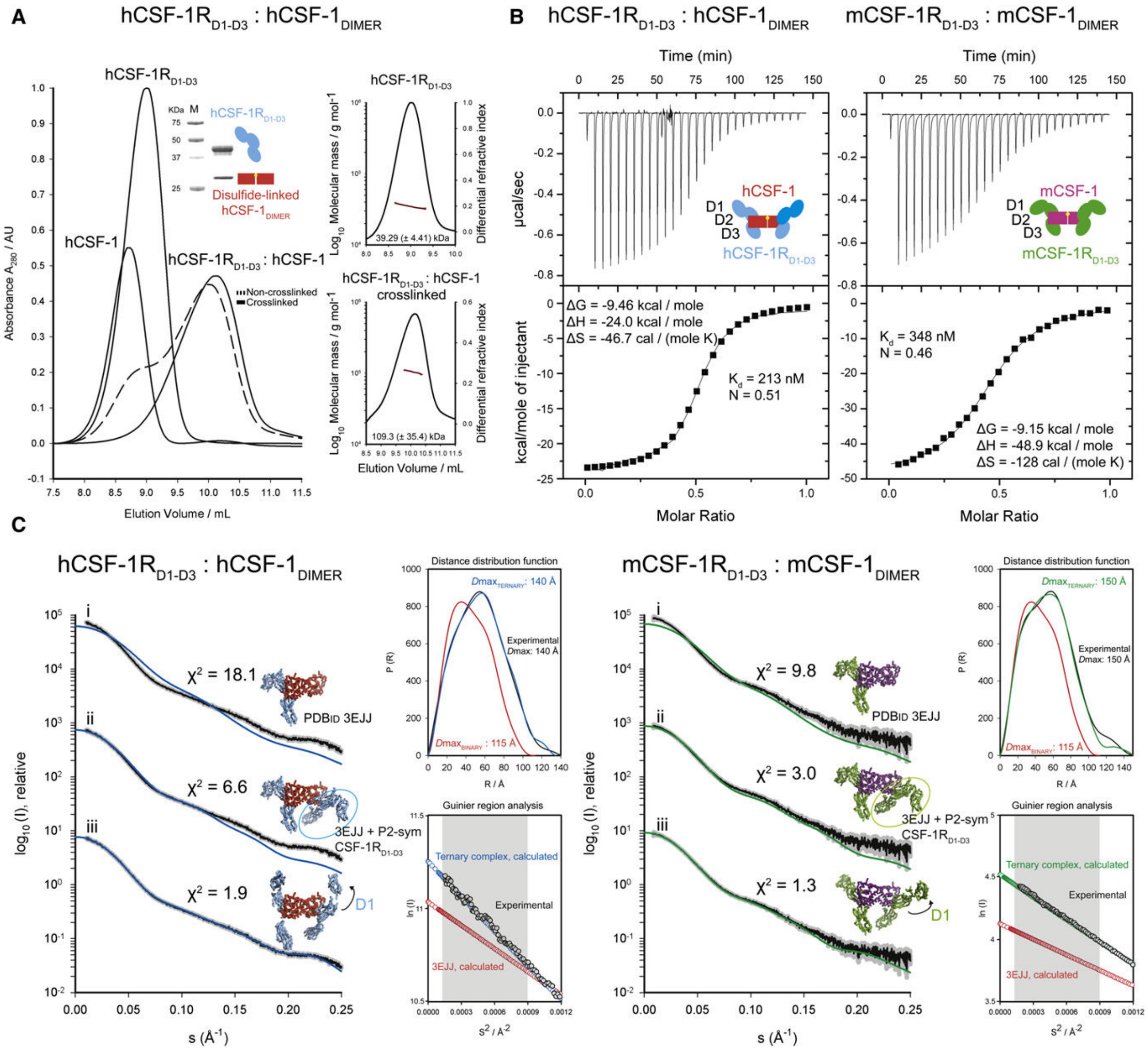


Figure 6. Human and Mouse CSF-1R$_{D1–D3}$ Can Form Ternary Complexes with Cognate CSF-1 Ligands

(A) The ternary hCSF-1R$_{D1–D3}$:hCSF-1 complex is transient on FFF. Injection of the isolated SEC peak fraction (Figure 5A) on FFF reveals a disassembly of the complex. A 110 kDa species indicative of a ternary complex can only be observed after incubation with a crosslinking agent, suggesting that a ternary complex is inherently less stable. The insets show an SDS-PAGE strip of the isolated noncrosslinked complex and molecular mass determination by MALLS.

(B) Thermodynamic profile of hCSF-1R$_{D1–D3}$:hCSF-1 and mCSF-1R$_{D1–D3}$:mCSF-1. Both thermograms can be accurately fitted by a “one set of binding sites” model and display a 1:2 CSF-1:CSF-1R$_{D1–D3}$ stoichiometry of binding.

(C) Structural analysis of hCSF-1R$_{D1–D3}$:hCSF-1 (left panel) and mCSF-1R$_{D1–D3}$:mCSF-1 (right panel) by SAXS after isolation by SEC (Figure 5A). Experimental scattering curves are shown in black to a maximal momentum transfer of s = 0.25 Å^{-1} (nominal resolution 25 Å), and the individual data:fit pairs are put on an arbitrary y axis to allow for better visualization. Curve “i” shows a comparison of the experimental scattering with calculated scattering from the monovalent mCSF-1R$_{D1–D3}$:mCSF-1 structure (PDB code 3EJJ). This binary model lacks significant scattering mass as judged by the gross incompatibility with the lowest angle experimental data. Curve “ii” illustrates a comparison of the experimental scattering with calculated scattering from a bivalent model derived from the mCSF-1R$_{D1–D3}$:mCSF-1 structure (PDB code 3EJJ) in which an additional CSF-1R$_{D1–D3}$ arm was generated by applying a pure 2-fold symmetry operation about the ligand dimer interface (circled). Curve “iii” shows rigid-body optimized fit of the bivalent CSF-1R$_{D1–D3}$:CSF-1 complex with specified CSF-1:CSF-1R$_{D2}$ core contacts and moving domains D1 and D3. The upper insets show the calculated distance distribution function for the modeled ternary complexes (blue or green) and for PDBid 3EJJ (red), and their fits with the experimental function (black). The lower insets display the experimental Guinier region (black) and the calculated Guinier region of the rigid-body refined ternary models (blue or green) and the binary PDBid 3EJJ model (red). The shaded area indicates the range of fitting for R_G analysis ($R_G \cdot S \leq 1.3$).

See also Figure S2 and Table S2.

16/60 Superdex 75 column (GE Healthcare). To remove the N-terminal His tag, h/mCSF-1 was subsequently incubated overnight at room temperature with 1 U of biotinylated thrombin (Novagen) per milligram of h/mCSF-1. Proteolytic cleavage was monitored by SDS-PAGE. Biotinylated thrombin was removed using a streptavidin-agarose column (Novagen). Thrombin-treated h/mCSF-1 was purified using a Source 30Q anion-exchange resin, followed by gel filtration chromatography on a Superdex-75 column (GE Healthcare). The fractions corresponding to h/mCSF-1 were pooled and used for further experiments.

Recombinant glycosylated human and murine CSF-1R ectodomain variants were produced in transiently transfected HEK293T cells in the presence of kifunensine based on established protocols (Aricescu et al., 2006; Chang et al., 2007; Verstraete et al., 2011a). The recombinant CSF-1R ectodomains carried a C-terminal 6xHis tag. h/mCSF-1R was purified by affinity chromatography from the supernatant using a Talon FF column (Clontech). The eluted fractions containing the purified protein were subsequently injected onto a Prep-Grade HiLoad 16/60 Superdex 200 column (GE Healthcare). The fractions corresponding to h/mCSF-1R were pooled and used for further experiments.

Human and murine CSF-1R$_{D1–D5}$:CSF-1 and CSF-1R$_{D1–D3}$:CSF-1 complexes were isolated by gel filtration chromatography on Superdex-200 column (GE Healthcare) after incubation of CSF-1R ectodomains with excess molar amounts of purified cognate CSF-1.

MALLS

The molecular masses of CSF-1, CSF-1R, and the CSF-1R:CSF-1 complexes were determined by MALLS. Protein sample was injected into a HPLC-driven (Shimadzu) FFF module (Wyatt Technology) equilibrated with a 20 mM HEPES (pH 7.5), 150 mM NaCl running buffer. The FFF module was coupled to an online UV detector (Shimadzu), an 18-angle light scattering detector (DAWN HELEOS), and a refractive index detector (Optilab T-rEX) (Wyatt Technology). Typical concentrations used were 1–10 μM of protein species. A RI increment value (*dn/dc* value) of 0.185 ml/g was used for the protein concentration and molecular mass determination. FFF cross-flows were varied to optimize the resolution of separation. Data analysis was carried out using the ASTRA V software.

EM

For preparation of negatively stained hCSF-1R$_{D1–D5}$:hCSF-1 complex, purified sample at ~0.2 mg/ml in PBS was applied to the clear side of carbon on a carbon-mica interface and stained with 2% (w/v) uranyl acetate. Images were recorded under low-dose conditions with a JEOL 1200 EX II microscope at 100 kV and at a nominal 40.000× magnification. Selected negatives were digitized on a Zeiss scanner (Photoscan TD) to a pixel size of 3.5 Å at the specimen level. Image processing was carried out using the boxer routine from the EMAN software package (Ludtke et al., 1999) for particle selection, CTFFIND3 (Mindell and Grigorieff, 2003) for contrast transfer function determination, bctf from the bsoft package (Heymann et al., 2008) for CTF correction, Imagic (van Heel et al., 1996) for MSA, classification, and angular reconstitution, and Spider (Shaikh et al., 2008) for projection matching. UROX (Siebert and Navaza, 2009) was used for structure fitting.

A generous semiautomatic particle selection with the EMAN boxer routine led to the extraction of a total of 18,432 individual particle subframes of 80 × 80 pixels that were corrected with respect to the contrast transfer function, and low-path filtered at 15 Å resolution. The data set was translationally aligned relative to the rotationally averaged total sum of the individual images. The aligned data set was subjected to MSA, which suggested the presence of a 2-fold symmetry axis. Characteristic class averages were used as a set of references for multi-reference alignment (MRA) followed by MSA and classification. After several iterations, representative class averages were selected to generate a crude initial model of the hCSF-1R$_{D1–D5}$:hCSF-1 complex by angular reconstitution in C2 symmetry. Iterative projection matching of the model led to a 3D reconstruction with a well-defined global core corresponding to the ligand and hCSF-1R$_{D2-D5}$, and a protruding weak density cloud, which we interpreted as D1 linked via a flexible linker to D2 in the complex core. To isolate a population of hCSF-1R1R$_{D2-D5}$:hCSF-1 particles with a better-defined orientation for D1, a set of models with the same core fitting the EM envelope, but different orientations of D1 protruding into the weak density cloud, was created based on the mCSF-1R1R$_{D1–D3}$-mCSF-1 crystal structure (Chen et al., 2008). These models were converted into EM density and averaged together, which reinforced the density of the core in comparison to D1, thus supporting the notion that D1 is flexible. The average model was used for more rounds of projection matching, which allowed a better definition for the position of D1. A total of 9,421 particles were included in the final reconstruction, which approached 23 Å resolution as estimated via Fourier shell correlation (FSC) according to the 0.5 criterion.

Modeling of the hCSF-1R$_{D1–D5}$:hCSF-1 Complex into the EM Envelope

A homology model for hCSF-1R$_{D1–D5}$ based on PDB entry 2E9W (Yuzawa et al., 2007) was fit into the 3D envelopes from EM with the EMAP module (Wu et al., 2003) of the CHARMM (Brooks et al., 2009) package to generate initial positions of the complex. A self-guided Langevin dynamics (Wu and Brooks, 2003) simulation of 1,000 ps was performed, including an implicit solvation model, to search the conformational space to reach the conformations satisfying the EM map constraints. The final conformation was minimized with constraints to maintain the C2 symmetry.

SAXS

Data were collected at beamlines X33 of the EMBL at DESY (Hamburg) and ID14-3 at ESRF (Grenoble) using a robotic sample changer (Roessle et al., 2008). The measurements were carried out at 298 K, within a momentum transfer range of 0.01 $Å^{-1} < s < 0.6\ Å^{-1}$, where $s = 4\pi\sin(\theta)/\lambda$, and 2θ is the scattering angle. All samples were measured at solute concentrations ranging from 0.5 to 10.0 mg/ml in 50 mM $NaPO_4$ (pH 7.40), 100 mM NaCl, with intermittent buffer solution (50 mM $NaPO_4$ [pH 7.40], 100 mM NaCl), and the radiation damage was monitored using standard procedures. The data were processed and extrapolated to infinite dilution, and the Guinier region was inspected using the program PRIMUS (Konarev et al., 2003). The radius of gyration (R_g), the forward scattering ($I(0)$), the maximum particle dimension (D_{max}), and the distance distribution function ($p(r)$) were evaluated using GNOM (Svergun, 1992). The molecular masses of the different constructs were calculated by comparison with the reference bovine serum albumin (BSA) samples. DAMMIN (Svergun, 1999) and AUTOPOROD were used to obtain the excluded volume and Porod volume of the particles, respectively. GASBOR (Svergun et al., 2001) was used to obtain the higher resolution ab initio bead models for the unliganded hCSF-1R$_{D1–D5}$; 15 independent runs with an average NSD value of 2.3 were structurally aligned and averaged with DAMAVER (Volkov and Svergun, 2003). X-ray scattering patterns from structural models were calculated using the program CRYSOL (Svergun et al., 1995). Constrained rigid-body refinement of the h/mCSF-1R$_{D1–D3}$:h/mCSF-1 complexes was carried out in SASREF7 (Petoukhov and Svergun, 2005) with imposed P2 symmetry and specified CSF-1:CSF-1R$_{D2}$ contacts. Constrained rigid-body refinement of the hCSF-1R$_{D1–D5}$:hCSF-1 complex was carried out in SASREF7 with imposed P2 symmetry, specified CSF-1:CSF-1R$_{D2}$ contacts, and ambiguous contact distances for the D4-D4′ interface. Constrained rigid-body refinement of the unliganded receptor was carried out in P1 symmetry, and refinement convergence was optimal upon definition of ambiguous distance contacts at the D4-D4′ interface.

ITC

Calorimetric measurements were carried out using purified h/mCSF-1 and h/mCSF-1R samples dialyzed exhaustively against 20 mM HEPES (pH 7.5), 150 mM NaCl. Experiments were carried out using a VP-ITC MicroCalorimeter at 310 K, and data were analyzed using the Origin ITC analysis software package. Titrations were always preceded by an initial injection of 3 μl and were carried out using 10 μl injections applied 300 s apart, with continuous stirring. The data were fit to the "one binding site model," and apparent molar reaction enthalpy (ΔH°), apparent entropy (ΔS°), association constant (K_A), and stoichiometry of binding (N) were determined. Several titrations were performed to evaluate reproducibility.

ACCESSION NUMBERS

The EM map for the 3D reconstruction of the hCSF-1R$_{D1–D5}$:hCSF-1 complex has been deposited in the EMDB under accession code EMD-1977.

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures and two tables and can be found with this article online at doi:10.1016/j.str.2011.10.012.

ACKNOWLEDGMENTS

J.E., K.V., B.V., and K.V.C. are research fellows of the Research Foundation Flanders, Belgium (FWO). A.D. was funded by a PhD fellowship of the Rhone-Alpes Region, France. A.V.S. and D.I.S. were supported by HFSP Research Grant Ref. RGP 55/2006. S.N.S. is supported by the FWO (Projects 3G064307 and G059710) and Ghent University (BOF instrument). We thank the EMBL-Hamburg for synchrotron beamtime allocation at DESY (Hamburg, Germany) and Clement Blanchet for technical support at beamline X33. Access to these synchrotron facilities is supported by the European Commission under the 7th Framework Programme: Research Infrastructures, Grant Agreement Number 226716. We further thank the SWITCH Laboratory at VIB-Belgium for access to FFF-MALLS facilities and Rodrigo Gallardo for help during data collection and analysis. J.E., A.D., A.V.S., X.W., N.B., B.R.B., D.I.S., B.V., and I.G. contributed to experimental design, performed research, analyzed data, and contributed to the writing of the manuscript. K.V. and K.V.C. provided experimental tools and technical assistance with tissue culture and mammalian protein expression. S.N.S. conceived and directed the study, contributed to experimental design, analyzed data, and contributed to the writing of the manuscript.

Received: May 12, 2011
Revised: September 13, 2011
Accepted: October 6, 2011
Published: December 6, 2011

REFERENCES

Aricescu, A.R., Lu, W., and Jones, E.Y. (2006). A time- and cost-efficient system for high-level protein production in mammalian cells. Acta Crystallogr. D Biol. Crystallogr. *62*, 1243–1250.

Brooks, B.R., Brooks, C.L., 3rd, Mackerell, A.D., Jr., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., et al. (2009). CHARMM: the biomolecular simulation program. J. Comput. Chem. *30*, 1545–1614.

Chang, V.T., Crispin, M., Aricescu, A.R., Harvey, D.J., Nettleship, J.E., Fennelly, J.A., Yu, C., Boles, K.S., Evans, E.J., Stuart, D.I., et al. (2007). Glycoprotein structural genomics: solving the glycosylation problem. Structure *15*, 267–273.

Chen, X., Liu, H., Focia, P.J., Shim, A.H., and He, X. (2008). Structure of macrophage colony stimulating factor bound to FMS: diverse signaling assemblies of class III receptor tyrosine kinases. Proc. Natl. Acad. Sci. USA *105*, 18267–18272.

Chihara, T., Suzu, S., Hassan, R., Chutiwitoonchai, N., Hiyoshi, M., Motoyoshi, K., Kimura, F., and Okada, S. (2010). IL-34 and M-CSF share the receptor Fms but are not identical in biological activity and signal activation. Cell Death Differ. *17*, 1917–1927.

Chitu, V., and Stanley, E.R. (2006). Colony-stimulating factor-1 in immunity and inflammation. Curr. Opin. Immunol. *18*, 39–48.

Chung, I., Akita, R., Vandlen, R., Toomre, D., Schlessinger, J., and Mellman, I. (2010). Spatial control of EGF receptor activation by reversible dimerization on living cells. Nature *464*, 783–787.

Finger, C., Escher, C., and Schneider, D. (2009). The single transmembrane domains of human receptor tyrosine kinases encode self-interactions. Sci. Signal. *2*, ra56.

Garceau, V., Smith, J., Paton, I.R., Davey, M., Fares, M.A., Sester, D.P., Burt, D.W., and Hume, D.A. (2010). Pivotal advance: avian colony-stimulating factor 1 (CSF-1), interleukin-34 (IL-34), and CSF-1 receptor genes and gene products. J. Leukoc. Biol. *87*, 753–764.

Graddis, T.J., Brasel, K., Friend, D., Srinivasan, S., Wee, S., Lyman, S.D., March, C.J., and McGrew, J.T. (1998). Structure-function analysis of FLT3 ligand-FLT3 receptor interactions using a rapid functional screen. J. Biol. Chem. *273*, 17626–17633.

Heymann, J.B., Cardone, G., Winkler, D.C., and Steven, A.C. (2008). Computational resources for cryo-electron tomography in Bsoft. J. Struct. Biol. *161*, 232–242.

Himanen, J.P., Saha, N., and Nikolov, D.B. (2007). Cell-cell signaling via Eph receptors and ephrins. Curr. Opin. Cell Biol. *19*, 534–542.

Jiang, X., Gurel, O., Mendiaz, E.A., Stearns, G.W., Clogston, C.L., Lu, H.S., Osslund, T.D., Syed, R.S., Langley, K.E., and Hendrickson, W.A. (2000). Structure of the active core of human stem cell factor and analysis of binding to its receptor kit. EMBO J. *19*, 3192–3203.

Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch, M.H.J., and Svergun, D.I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. J. Appl. Crystallogr. *36*, 1277–1282.

Lawrence, M.C., McKern, N.M., and Ward, C.W. (2007). Insulin receptor structure and its implications for the IGF-1 receptor. Curr. Opin. Struct. Biol. *17*, 699–705.

Lemmon, M.A., and Schlessinger, J. (2010). Cell signaling by receptor tyrosine kinases. Cell *141*, 1117–1134.

Lemmon, M.A., Pinchasi, D., Zhou, M., Lax, I., and Schlessinger, J. (1997). Kit receptor dimerization is driven by bivalent binding of stem cell factor. J. Biol. Chem. *272*, 6311–6317.

Leppänen, V.M., Prota, A.E., Jeltsch, M., Anisimov, A., Kalkkinen, N., Strandin, T., Lankinen, H., Goldman, A., Ballmer-Hofer, K., and Alitalo, K. (2010). Structural determinants of growth factor binding and specificity by VEGF receptor 2. Proc. Natl. Acad. Sci. USA *107*, 2425–2430.

Lev, S., Yarden, Y., and Givol, D. (1992). A recombinant ectodomain of the receptor for the stem cell factor (SCF) retains ligand-induced receptor dimerization and antagonizes SCF-stimulated cellular responses. J. Biol. Chem. *267*, 10866–10873.

Li, E., and Hristova, K. (2006). Role of receptor tyrosine kinase transmembrane domains in cell signaling and human pathologies. Biochemistry *45*, 6241–6251.

Li, W., and Stanley, E.R. (1991). Role of dimerization and modification of the CSF-1 receptor in its activation and internalization during the CSF-1 response. EMBO J. *10*, 277–288.

Lin, H., Lee, E., Hestir, K., Leo, C., Huang, M., Bosch, E., Halenbeck, R., Wu, G., Zhou, A., Behrens, D., et al. (2008). Discovery of a cytokine and its receptor by functional screening of the extracellular proteome. Science *320*, 807–811.

Liu, H., Chen, X., Focia, P.J., and He, X. (2007). Structural basis for stem cell factor-KIT signaling and activation of class III receptor tyrosine kinases. EMBO J. *26*, 891–901.

Ludtke, S.J., Baldwin, P.R., and Chiu, W. (1999). EMAN: semiautomated software for high-resolution single-particle reconstructions. J. Struct. Biol. *128*, 82–97.

Mi, L.Z., Lu, C., Li, Z., Nishida, N., Walz, T., and Springer, T.A. (2011). Simultaneous visualization of the extracellular and cytoplasmic domains of the epidermal growth factor receptor. Nat. Struct. Mol. Biol. *18*, 984–989.

Mindell, J.A., and Grigorieff, N. (2003). Accurate determination of local defocus and specimen tilt in electron microscopy. J. Struct. Biol. *142*, 334–347.

Oefner, C., D'Arcy, A., Winkler, F.K., Eggimann, B., and Hosang, M. (1992). Crystal structure of human platelet-derived growth factor BB. EMBO J. *11*, 3921–3926.

Pandit, J., Bohm, A., Jancarik, J., Halenbeck, R., Koths, K., and Kim, S.H. (1992). Three-dimensional structure of dimeric human recombinant macrophage colony-stimulating factor. Science *258*, 1358–1362.

Petoukhov, M.V., and Svergun, D.I. (2005). Global rigid body modeling of macromolecular complexes against small-angle scattering data. Biophys. J. *89*, 1237–1250.

Roessle, M., Round, A.R., Franke, D., Moritz, S., Huchler, R., Fritsche, M., Malthan, D., Klaering, R., and Svergun, D.I. (2008). Automated sample-changing robot for solution scattering experiments at the EMBL Hamburg SAXS station X33. J. Appl. Crystallogr. *41*, 913–917.

Roussel, M.F., Downing, J.R., Rettenmier, C.W., and Sherr, C.J. (1988). A point mutation in the extracellular domain of the human CSF-1 receptor (c-fms proto-oncogene product) activates its transforming potential. Cell *55*, 979–988.

Ruch, C., Skiniotis, G., Steinmetz, M.O., Walz, T., and Ballmer-Hofer, K. (2007). Structure of a VEGF-VEGF receptor complex determined by electron microscopy. Nat. Struct. Mol. Biol. *14*, 249–250.

Savvides, S.N., Boone, T., and Andrew Karplus, P. (2000). Flt3 ligand structure and unexpected commonalities of helical bundles and cystine knots. Nat. Struct. Biol. *7*, 486–491.

Schubert, C., Schalk-Hihi, C., Struble, G.T., Ma, H.C., Petrounia, I.P., Brandt, B., Deckman, I.C., Patch, R.J., Player, M.R., Spurlino, J.C., and Springer, B.A. (2007). Crystal structure of the tyrosine kinase domain of colony-stimulating factor-1 receptor (cFMS) in complex with two inhibitors. J. Biol. Chem. *282*, 4094–4101.

Shaikh, T.R., Gao, H.X., Baxter, W.T., Asturias, F.J., Boisset, N., Leith, A., and Frank, J. (2008). SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. Nat. Protoc. *3*, 1941–1974.

Shim, A.H., Liu, H., Focia, P.J., Chen, X., Lin, P.C., and He, X. (2010). Structures of a platelet-derived growth factor/propeptide complex and a platelet-derived growth factor/receptor complex. Proc. Natl. Acad. Sci. USA *107*, 11307–11312.

Siebert, X., and Navaza, J. (2009). UROX 2.0: an interactive tool for fitting atomic models into electron-microscopy reconstructions. Acta Crystallogr. D Biol. Crystallogr. *65*, 651–658.

Streeter, P.R., Minster, N.I., Kahn, L.E., Hood, W.F., Vickery, L.E., Thurman, T.L., Monahan, J.B., Welply, J.K., McKearn, J.P., and Woulfe, S.L. (2001). Progenipoietins: biological characterization of a family of dual agonists of fetal liver tyrosine kinase-3 and the granulocyte colony-stimulating factor receptor. Exp. Hematol. *29*, 41–50.

Svergun, D., Barberato, C., and Koch, M.H.J. (1995). CRYSOL—a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. J. Appl. Crystallogr. *28*, 768–773.

Svergun, D.I. (1992). Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria. J. Appl. Crystallogr. *25*, 495–503.

Svergun, D.I. (1999). Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. Biophys. J. *76*, 2879–2886.

Svergun, D.I., Petoukhov, M.V., and Koch, M.H.J. (2001). Determination of domain structure of proteins from X-ray solution scattering. Biophys. J. *80*, 2946–2953.

van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R., and Schatz, M. (1996). A new generation of the IMAGIC image processing system. J. Struct. Biol. *116*, 17–24.

Verstraete, K., Koch, S., Ertugrul, S., Vandenberghe, I., Aerts, M., Vandriessche, G., Thiede, C., and Savvides, S.N. (2009). Efficient production of bioactive recombinant human Flt3 ligand in *E. coli*. Protein J. *28*, 57–65.

Verstraete, K., Remmerie, B., Elegheert, J., Lintermans, B., Haegerman, G., Vanhoenacker, P., Van Craenenbroeck, K., and Savvides, S.N. (2011a). Inducible production of recombinant human Flt3 ectodomain variants in mammalian cells and preliminary crystallographic analysis of Flt3 ligand-receptor complexes. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. *67*, 325–331.

Verstraete, K., Vandriessche, G., Januar, M., Elegheert, J., Shkumatov, A.V., Desfosses, A., Van Craenenbroeck, K., Svergun, D.I., Gutsche, I., Vergauwen, B., and Savvides, S.N. (2011b). Structural insights into the extracellular assembly of the hematopoietic Flt3 signaling complex. Blood *118*, 60–68.

Volkov, V.V., and Svergun, D.I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. J. Appl. Crystallogr. *36*, 860–864.

Walter, M., Lucet, I.S., Patel, O., Broughton, S.E., Bamert, R., Williams, N.K., Fantino, E., Wilks, A.F., and Rossjohn, J. (2007). The 2.7 A crystal structure of the autoinhibited human c-Fms kinase domain. J. Mol. Biol. *367*, 839–847.

Wei, S., Nandi, S., Chitu, V., Yeung, Y.G., Yu, W., Huang, M., Williams, L.T., Lin, H., and Stanley, E.R. (2010). Functional overlap but differential expression of CSF-1 and IL-34 in their CSF-1 receptor-mediated regulation of myeloid cells. J. Leukoc. Biol. *88*, 495–505.

Wiesmann, C., Fuh, G., Christinger, H.W., Eigenbrot, C., Wells, J.A., and de Vos, A.M. (1997). Crystal structure at 1.7 A resolution of VEGF in complex with domain 2 of the Flt-1 receptor. Cell *91*, 695–704.

Wu, X., Milne, J.L., Borgnia, M.J., Rostapshov, A.V., Subramaniam, S., and Brooks, B.R. (2003). A core-weighted fitting method for docking atomic structures into low-resolution maps: application to cryo-electron microscopy. J. Struct. Biol. *141*, 63–76.

Wu, X.W., and Brooks, B.R. (2003). Self-guided Langevin dynamics simulation method. Chem. Phys. Lett. *381*, 512–518.

Yang, Y., Yuzawa, S., and Schlessinger, J. (2008). Contacts between membrane proximal regions of the PDGF receptor ectodomain are required for receptor activation but not for receptor dimerization. Proc. Natl. Acad. Sci. USA *105*, 7681–7686.

Yang, Y., Xie, P., Opatowsky, Y., and Schlessinger, J. (2010). Direct contacts between extracellular membrane-proximal domains are required for VEGF receptor activation and cell signaling. Proc. Natl. Acad. Sci. USA *107*, 1906–1911.

Yuzawa, S., Opatowsky, Y., Zhang, Z., Mandiyan, V., Lax, I., and Schlessinger, J. (2007). Structural basis for activation of the receptor tyrosine kinase KIT by stem cell factor. Cell *130*, 323–334.

Zhang, Z., Zhang, R., Joachimiak, A., Schlessinger, J., and Kong, X.P. (2000). Crystal structure of human stem cell factor: implication for stem cell factor receptor dimerization and activation. Proc. Natl. Acad. Sci. USA *97*, 7732–7737.

Bibliography

Adrian, M, J Dubochet, J Lepault, and AW McDowall. 1984. “Cryo-electron Microscopy of Viruses.” Nature 308. http://www.nature.com/nature/journal/v308/n5954/abs/308032a0.html.

Albertini, Aurélie a V, Cedric R Clapier, Amy K Wernimont, Guy Schoehn, Winfried Weissenhorn, and Rob W H Ruigrok. 2006. “Isolation and Crystallization of a Unique Size Category of Recombinant Rabies Virus Nucleoprotein-RNA Rings.” Journal of Structural Biology 158 (1) (April): 129–33. doi:10.1016/j.jsb.2006.10.011. http://www.ncbi.nlm.nih.gov/pubmed/17126031.

Albertini, Aurélie a V, Amy K Wernimont, Tadeusz Muziol, Raimond B G Ravelli, Cedric R Clapier, Guy Schoehn, Winfried Weissenhorn, and Rob W H Ruigrok. 2006. “Crystal Structure of the Rabies Virus nucleoprotein-RNA Complex.” Science (New York, N.Y.) 313 (5785) (July 21): 360–3. doi:10.1126/science.1125280. http://www.ncbi.nlm.nih.gov/pubmed/16778023.

Amos, L a, and a Klug. 1975. “Three-dimensional Image Reconstructions of the Contractile Tail of T4 Bacteriophage.” Journal of Molecular Biology 99 (1) (November 25): 51–64. http://www.ncbi.nlm.nih.gov/pubmed/1206701.

Avinoam, Ori, Karen Fridman, Clari Valansi, Inbal Abutbul, Tzviya Zeev-Ben-Mordehai, Ulrike E Maurer, Amir Sapir, et al. 2011. “Conserved Eukaryotic Fusogens Can Fuse Viral Envelopes to Cells.” Science (New York, N.Y.) 332 (6029) (April 29): 589–92. doi:10.1126/science.1202333. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3084904&tool=pmcentrez&rendertype=abstract.

Baker, TS, J Drak, and M Bina. 1989. “The Capsid of Small Papova Viruses Contains 72 Pentameric Capsomeres: Direct Evidence from Cryo-electron-microscopy of Simian Virus 40.” Biophysical Journal 55 (2) (February): 243–53. doi:10.1016/S0006-3495(89)82799-7. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1330465&tool=pmcentrez&rendertype=abstract.

Bamford, C. H., W. E. Hanby, and F. Happey. 1951. “The Structure of Synthetic Polypeptides. I. X-Ray Investigation.” Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 205 (1080) (January 22): 30–47. doi:10.1098/rspa.1951.0015. http://rspa.royalsocietypublishing.org/cgi/content/abstract/205/1080/30.

Baudin, F, C Bach, S Cusack, and R W Ruigrok. 1994. “Structure of Influenza Virus RNP. I. Influenza Virus Nucleoprotein Melts Secondary Structure in Panhandle RNA and Exposes the Bases to the Solvent.” The EMBO Journal 13 (13) (July 1): 3158–65. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=395207&tool=pmcentrez&rendertype=abstract.

Beroukhim, Rameen, and Nigel Unwin. 1997. “Distortion Correction of Tubular Crystals: Improvements in the Acetylcholine Receptor Structure.” Ultramicroscopy 70 (1-2) (December): 57–81. http://dx.doi.org/10.1016/S0304-3991(97)00070-3.

Bharat, Tanmay, Norman E Davey, Pavel Ulbrich, James D Riches, Alex de Marco, Michaela Rumlova, Carsten Sachse, Tomas Ruml, and John A G Briggs. 2012. “Structure of the Immature Retroviral Capsid at 8 Å Resolution by Cryo-electron Microscopy.” Nature 487 (7407) (June 3): 385–9. doi:10.1038/nature11169. http://www.ncbi.nlm.nih.gov/pubmed/22722831.

Bharat, Tanmay, Takeshi Noda, James D Riches, Verena Kraehling, Larissa Kolesnikova, Stephan Becker, Yoshihiro Kawaoka, and John a G Briggs. 2012. “Structural Dissection of Ebola Virus and Its Assembly

Determinants Using Cryo-electron Tomography." Proceedings of the National Academy of Sciences of the United States of America 109 (11) (March 13): 4275–80. doi:10.1073/pnas.1120453109. http://www.ncbi.nlm.nih.gov/pubmed/22371572.

Bharat, Tanmay, James D Riches, Larissa Kolesnikova, Sonja Welsch, Verena Krähling, Norman Davey, Marie-Laure Parsy, Stephan Becker, and John a G Briggs. 2011. "Cryo-electron Tomography of Marburg Virus Particles and Their Morphogenesis Within Infected Cells." PLoS Biology 9 (11) (November): e1001196. doi:10.1371/journal.pbio.1001196. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3217011&tool=pmcentrez&rendertype=abstract.

Bhella, David, Adam Ralph, Lindsay B Murphy, and Robert P Yeo. 2002. "Significant Differences in Nucleocapsid Morphology Within the Paramyxoviridae." The Journal of General Virology 83 (Pt 8) (August): 1831–9. http://www.ncbi.nlm.nih.gov/pubmed/12124447.

Bhella, David, Adam Ralph, and Robert Paul Yeo. 2004. "Conformational Flexibility in Recombinant Measles Virus Nucleocapsids Visualised by Cryo-negative Stain Electron Microscopy and Real-space Helical Reconstruction." Journal of Molecular Biology 340 (2) (July 2): 319–31. doi:10.1016/j.jmb.2004.05.015. http://www.ncbi.nlm.nih.gov/pubmed/15201055.

Bluemke, D A, B Carragher, and R Josephs. 1988. "The Reconstruction of Helical Particles with Variable Pitch." Ultramicroscopy 26 (3): 255–270. http://www.sciencedirect.com/science/article/pii/0304399188902264.

Blumberg, B M, C Giorgi, K Rose, and D Kolakofsky. 1984. "Preparation and Analysis of the Nucleocapsid Proteins of Vesicular Stomatitis Virus and Sendai Virus, and Analysis of the Sendai Virus leader-NP Gene Region." The Journal of General Virology 65 (Pt 4) (April): 769–79. http://www.ncbi.nlm.nih.gov/pubmed/6323622.

Bourhis, Jean-Marie, Kenth Johansson, Véronique Receveur-Bréchot, Christopher J Oldfield, Keith A Dunker, Bruno Canard, and Sonia Longhi. 2004. "The C-terminal Domain of Measles Virus Nucleoprotein Belongs to the Class of Intrinsically Disordered Proteins That Fold Upon Binding to Their Physiological Partner." Virus Research 99 (2) (February): 157–67. http://www.ncbi.nlm.nih.gov/pubmed/14749181.

Bremer, A, C Henn, A Engel, W Baumeister, and U Aebi. 1992. "Has Negative Staining Still a Place in Biomacromolecular Electron Microscopy?" Ultramicroscopy 46 (1-4) (October): 85–111. doi:10.1016/0304-3991(92)90008-8. http://www.ncbi.nlm.nih.gov/pubmed/1481278.

Brenner, S, and RW Horne. 1959. "A Negative Staining Method for High Resolution Electron Microscopy of Viruses." Biochimica Et Biophysica Acta 34: 3–10. http://www.ncbi.nlm.nih.gov/pubmed/13804200.

Brüggeller, P, and E Mayer. 1980. "Complete Vitrification in Pure Liquid Water and Dilute Aqueous Solutions." Nature 288. http://www.nature.com/nature/journal/v288/n5791/abs/288569a0.html.

Burgess, Stan a, Matt L Walker, Kavitha Thirumurugan, John Trinick, and Peter J Knight. 2004. "Use of Negative Stain and Single-particle Image Processing to Explore Dynamic Properties of Flexible Macromolecules." Journal of Structural Biology 147 (3) (September): 247–58. doi:10.1016/j.jsb.2004.04.004. http://www.ncbi.nlm.nih.gov/pubmed/15450294.

Calain, P, and L Roux. 1993. "The Rule of Six, a Basic Feature for Efficient Replication of Sendai Virus Defective Interfering RNA." Journal of Virology 67 (8) (August): 4822–30. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=237869&tool=pmcentrez&rendertype=abstract.

Chen, Zhiqiang, Todd J Green, Ming Luo, and Huilin Li. 2004. "Visualizing the RNA Molecule in the Bacterially Expressed Vesicular Stomatitis Virus nucleoprotein-RNA Complex." Structure (London, *England : 1993)* 12 (2) (February): 227–35. doi:10.1016/j.str.2004.01.001. http://www.ncbi.nlm.nih.gov/pubmed/14962383.

Chrétien, D, J M Kenney, S D Fuller, and R H Wade. 1996. “Determination of Microtubule Polarity by Cryo-electron Microscopy.” Structure 4: 1031–10410.

Clare, Daniel K, and Elena V Orlova. 2010. “4.6A Cryo-EM Reconstruction of Tobacco Mosaic Virus from Images Recorded at 300 keV on a 4k x 4k CCD Camera.” Journal of Structural Biology 171 (3) (September): 303–8. doi:10.1016/j.jsb.2010.06.011. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2939825&tool=pmcentrez&rendertype=abstract.

Cochran, W., F. H. Crick, and V. Vand. 1952. “The Structure of Synthetic Polypeptides. I. The Transform of Atoms on a Helix.” Acta Crystallographica 5 (5) (September 10): 581–586. doi:10.1107/S0365110X52001635. http://scripts.iucr.org/cgi-bin/paper?S0365110X52001635.

Couturier, Marie, Matt Buccellato, Stéphanie Costanzo, Jean-Marie Bourhis, Yaoling Shu, Magali Nicaise, Michel Desmadril, Christophe Flaudrops, Sonia Longhi, and Michael Oglesbee. 2010. “High Affinity Binding Between Hsp70 and the C-terminal Domain of the Measles Virus Nucleoprotein Requires an Hsp40 Co-chaperone.” J*ournal of Molecular Recognition : JMR* 23 (3) (January): 301–15. doi:10.1002/jmr.982. http://www.ncbi.nlm.nih.gov/pubmed/19718689.

Crowther, R A, and A Klug. 1975. “Structural Analysis Of Assemblies By Image Reconstruction From Electron Micrographs.” Annual Review of Biochemistry 44: 161–182.

Crowther, R A, R Padrón, and R Craig. 1985. “Arrangement of the Heads of Myosin in Relaxed Thick Filaments from Tarantula Muscle.” Journal of Molecular Biology 184 (3) (August 5): 429–39. http://www.ncbi.nlm.nih.gov/pubmed/4046022.

Crowther, R. A., D. J. DeRosier, and A. Klug. 1970. “The Reconstruction of a Three-Dimensional Structure from Projections and Its Application to Electron Microscopy.” Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 317 (1530) (June 23): 319–340. doi:10.1098/rspa.1970.0119. http://rspa.royalsocietypublishing.org/cgi/content/abstract/317/1530/319.

Curran, J, T Pelet, and D Kolakofsky. 1994. “An Acidic Activation-like Domain of the Sendai Virus P Protein Is Required for RNA Synthesis and Encapsidation.” Virology 202 (2) (August 1): 875–84. doi:10.1006/viro.1994.1409. http://www.ncbi.nlm.nih.gov/pubmed/8030249.

DeRosier, D J, and A Klug. 1968. “Reconstruction of Three Dimensional Structures from Electron Micrographs.” Nature 217: 130–134.

DeRosier, D J, and P B Moore. 1970. “Reconstruction of Three-dimensional Images from Electron Micrographs of Structures with Helical Symmetry.” Journal of Molecular Biology 52 (2) (September 14): 355–69. http://www.ncbi.nlm.nih.gov/pubmed/5485914.

Deckert, Jochen, Klaus Hartmuth, Daniel Boehringer, Nastaran Behzadnia, Cindy L Will, Berthold Kastner, Holger Stark, Henning Urlaub, and Reinhard Lu. 2006. “Protein Composition and Electron Microscopy Structure of Affinity-Purified Human Spliceosomal B Complexes Isolated Under Physiological Conditions” 26 (14): 5528–5543. doi:10.1128/MCB.00582-06.

Desfosses, Ambroise, Gaël Goret, Leandro Farias Estrozi, Rob W H Ruigrok, and Irina Gutsche. 2011. “Nucleoprotein-RNA Orientation in the Measles Virus Nucleocapsid by Three-dimensional Electron Microscopy.” Journal of Virology 85 (3) (February): 1391–5. doi:10.1128/JVI.01459-10. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3020520&tool=pmcentrez&rendertype=abstract.

Dubochet, J, M Adrian, J J Chang, J C Homo, J Lepault, a W McDowall, and P Schultz. 1988. “Cryo-electron Microscopy of Vitrified Specimens.” Quarterly Reviews of Biophysics 21 (2) (May): 129–228. http://www.ncbi.nlm.nih.gov/pubmed/3043536.

Dubochet, J, JJ Chang, R Freeman, J Lepault, and a W McDowall. 1982. “Frozen Aqueous Suspensions.” Ultramicroscopy 10: 55–61. http://www.sciencedirect.com/science/article/pii/0304399182901875.

Dubochet, J., and A.W. McDowall. 1981. “VITRIFICATION OF PURE WATER FOR ELECTRON MICROSCOPY.” Journal of Microscopy 124 (3) (December 2): 3–4. doi:10.1111/j.1365-2818.1981.tb02483.x. http://doi.wiley.com/10.1111/j.1365-2818.1981.tb02483.x.

Effantin, Grégory, Aurélien Dordor, Virginie Sandrin, Nicolas Martinelli, Wesley I Sundquist, Guy Schoehn, and Winfried Weissenhorn. 2012. “ESCRT-III CHMP2A and CHMP3 Form Variable Helical Polymers in Vitro and Act Synergistically During HIV-1 Budding.” Cellular Microbiology (October 10). doi:10.1111/cmi.12041. http://www.ncbi.nlm.nih.gov/pubmed/23051622.

Egelman, E H. 2000. “A Robust Algorithm for the Reconstruction of Helical Filaments Using Single-particle Methods.” Ultramicroscopy 85 (4) (December): 225–34. http://www.ncbi.nlm.nih.gov/pubmed/11125866.

———. 2007. “Single-particle Reconstruction from EM Images of Helical Filaments.” Current Opinion in Structural Biology 17 (5) (October): 556–61. doi:10.1016/j.sbi.2007.07.006. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2443787&tool=pmcentrez&rendertype=abstract.

Egelman, E H, and D J DeRosier. 1992. “Image Analysis Shows That Variations in Actin Crossover Spacings Are Random, Not Compensatory.” Biophysical Journal 63 (5) (November): 1299–305. doi:10.1016/S0006-3495(92)81716-2. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1261433&tool=pmcentrez&rendertype=abstract.

Egelman, E H, S S Wu, M Amrein, a Portner, and G Murti. 1989. “The Sendai Virus Nucleocapsid Exists in at Least Four Different Helical States.” Journal of Virology 63 (5) (May): 2233–43. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=250641&tool=pmcentrez&rendertype=abstract.

Egelman, E. H. 2007. “The Iterative Helical Real Space Reconstruction Method: Surmounting the Problems Posed by Real Polymers.” Journal of Structural Biology 157 (1): 83–94. http://www.ncbi.nlm.nih.gov/pubmed/16919474.

Egelman, Edward H. 2010. Reconstruction of Helical Filaments and Tubes. Methods in Enzymology. 1st ed. Vol. 482. Elsevier Inc. doi:10.1016/S0076-6879(10)82006-3. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245864&tool=pmcentrez&rendertype=abstract.

Elad, Nadav, Daniel K Clare, Helen R Saibil, and Elena V Orlova. 2008. “Detection and Separation of Heterogeneity in Molecular Complexes by Statistical Analysis of Their Two-dimensional Projections.” Journal of Structural Biology 162 (1) (April): 108–20. doi:10.1016/j.jsb.2007.11.007. http://www.ncbi.nlm.nih.gov/pubmed/18166488.

Fooks, a R, J R Stephenson, a Warnes, a B Dowsett, B K Rima, and G W Wilkinson. 1993. “Measles Virus Nucleocapsid Protein Expressed in Insect Cells Assembles into Nucleocapsid-like Structures.” The Journal of General Virology 74 (Pt 7) (July): 1439–44. http://www.ncbi.nlm.nih.gov/pubmed/8336125.

Frank, J, P Penczek, R Grassucci, and S Srivastava. 1991. “Three-dimensional Reconstruction of the 70S Escherichia Coli Ribosome in Ice: The Distribution of Ribosomal RNA.” The Journal of Cell Biology 115 (3) (November): 597–605. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2289182&tool=pmcentrez&rendertype=abstract.

Frank, Joachim. 1975. “Averaging of Low Exposure Electron Micrographs of Non-periodic Objects.” Ultramicroscopy 1: 159–162.

Frank, Joachim, W Goldfarb, D Eisenberg, and T S Baker. 1978. “Reconstruction of Glutamine Synthetase Using Computer Averaging.” Ultramicroscopy 3: 283–290.

Frank, Joachim, and Marin van Heel. 1982. “Correspondence Analysis of Aligned Images of Biological Particles.” Journal of Molecular Biology 161 (1) (October 15): 134–7. http://www.ncbi.nlm.nih.gov/pubmed/7154073.

Frank, Joachim, M Radermacher, P Penczek, J Zhu, Y Li, M Ladjadj, and A Leith. 1996. “SPIDER and WEB: Processing and Visualization of Images in 3D Electron Microscopy and Related Fields.” Journal of Structural Biology 116 (1): 190–9. doi:10.1006/jsbi.1996.0030. http://www.ncbi.nlm.nih.gov/pubmed/8742743.

Frank, Joachim, Adirana Verschoor, and Miloslav Boublik. 1981. “Computer Averaging of Electron Micrographs of 40S Ribosomal Subunits.” Science 214: 1353–1355.

Franklin, R. E., and R. G. Gosling. 1953. “Evidence for 2-chain Helix in Crystalline Structure of Sodium Deoxyribonucleate.” Nature 172: 156–157.

Fujii, Takashi, Atsuko H Iwane, Toshio Yanagida, and Keiichi Namba. 2010. “Direct Visualization of Secondary Structures of F-actin by Electron Cryomicroscopy.” Nature 467 (7316) (October 7): 724–8. doi:10.1038/nature09372. http://www.ncbi.nlm.nih.gov/pubmed/20844487.

Fujii, Takashi, Takayuki Kato, and Keiichi Namba. 2009a. “Specific Arrangement of Alpha-helical Coiled Coils in the Core Domain of the Bacterial Flagellar Hook for the Universal Joint Function.” Structure (London, *England : 1993)* 17 (11) (November 11): 1485–93. http://dx.doi.org/10.1016/j.str.2009.08.017.

———. 2009b. “Specific Arrangement of Alpha-helical Coiled Coils in the Core Domain of the Bacterial Flagellar Hook for the Universal Joint Function.” *Structure (London, England : 1993)* 17 (11) (November 11): 1485–93. doi:10.1016/j.str.2009.08.017. http://www.ncbi.nlm.nih.gov/pubmed/19913483.

Gao, Haixiao, Christian M T Spahn, Robert a Grassucci, and Joachim Frank. 2002. “An Assay for Local Quality in Cryo-electron Micrographs of Single Particles.” Ultramicroscopy 93 (2) (November): 169–78. http://www.ncbi.nlm.nih.gov/pubmed/12425594.

Ge, Peng, Jun Tsao, Stan Schein, Todd J Green, Ming Luo, and Z Hong Zhou. 2010. “Cryo-EM Model of the Bullet-shaped Vesicular Stomatitis Virus.” Science (New York, N.Y.) 327 (5966) (February 5): 689–93. doi:10.1126/science.1181766. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2892700&tool=pmcentrez&rendertype=abstract.

Ge, Peng, and Z Hong Zhou. 2011. “Hydrogen-bonding Networks and RNA Bases Revealed by Cryo Electron Microscopy Suggest a Triggering Mechanism for Calcium Switches.” Proceedings of the National Academy of Sciences of the United States of America 108 (23) (June 7): 9637–42. doi:10.1073/pnas.1018104108. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3111329&tool=pmcentrez&rendertype=abstract.

Green, Todd J, and Ming Luo. 2009. “Structure of the Vesicular Stomatitis Virus Nucleocapsid in Complex with the Nucleocapsid-binding Domain of the Small Polymerase Cofactor, P.” Proceedings of the National Academy of Sciences of the United States of America 106 (28) (July 14): 11713–8. doi:10.1073/pnas.0903228106. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2710649&tool=pmcentrez&rendertype=abstract.

Green, Todd J, Xin Zhang, Gail W Wertz, and Ming Luo. 2006. “Structure of the Vesicular Stomatitis Virus nucleoprotein-RNA Complex.” Science 313 (5785) (July 21): 357–60. doi:10.1126/science.1126953. http://www.ncbi.nlm.nih.gov/pubmed/16778022.

Grigorieff, Nikolaus. 2000. “Resolution Measurement in Structures Derived from Single Particles.” Acta Crystallographica Section D Biological Crystallography 56 (10) (October 1): 1270–1277. doi:10.1107/S0907444900009549. http://scripts.iucr.org/cgi-bin/paper?S0907444900009549.

Hall, Cecil E. 1955. “Electron Densitometry of Stained Virus Particles.” The Journal of Biophysical and Biochemical Cytology 1 (I).

Hanson, Jean, and J. Lowy. 1963. “The Structure of F-actin and of Actin Filaments Isolated from Muscle.” Journal of Molecular Biology 6 (1) (January): 46–IN5. doi:10.1016/S0022-2836(63)80081-9. http://linkinghub.elsevier.com/retrieve/pii/S0022283663800819.

Harauz, G, and Marin van Heel. 1986. “Exact Filters for General Geometry Three Dimensional Reconstruction.” Optik 73: 146. http://65.54.113.26/Publication/3064609/exact-filters-for-general-geometry-three-dimensional-reconstruction.

van Heel, M., G. Harauz, E. V. Orlova, R. Schmidt, and M. Schatz. 1996. “A New Generation of the IMAGIC Image Processing System.” Journal of Structural Biology 116 (1): 17–24. doi:10.1006/jsbi.1996.0004. http://www.ncbi.nlm.nih.gov/pubmed/8742718.

van Heel, Marin, and Joachim Frank. 1981. “Use Of Multivariate Statistics In Analysing The Images Of Biological Macromolecules.” Ultramicroscopy 6: 187–194.

van Heel, Marin, and Michael Schatz. 2005. “Fourier Shell Correlation Threshold Criteria.” Journal of Structural Biology 151 (3) (September): 250–62. http://www.ncbi.nlm.nih.gov/pubmed/16125414.

Heggeness, M. H. 1980. “Conformation of the Helical Nucleocapsids of Paramyxoviruses and Vesicular Stomatitis Virus: Reversible Coiling and Uncoiling Induced by Changes in Salt Concentration.” Proceedings of the National Academy of Sciences 77 (5) (May 1): 2631–2635. doi:10.1073/pnas.77.5.2631. http://www.pnas.org/cgi/doi/10.1073/pnas.77.5.2631.

Heggeness, Michael H., P.R. Smith, Ismo Ulmanen, Robert M. Krug, and Purnell W. Choppin. 1982. “Studies on the Helical Nucleocapsid of Influenza Virus.” Virology 118 (2) (April): 466–470. doi:10.1016/0042-6822(82)90367-1. http://dx.doi.org/10.1016/0042-6822(82)90367-1.

Hodgkinson, J L, C Peters, S a Kuznetsov, and W Steffen. 2005. “Three-dimensional Reconstruction of the Dynactin Complex by Single-particle Image Analysis.” Proceedings of the National Academy of Sciences of the United States of America 102 (10) (March 8): 3667–72. doi:10.1073/pnas.0409506102. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=553325&tool=pmcentrez&rendertype=abstract.

Hohn, Michael, Grant Tang, Grant Goodyear, P R Baldwin, Zhong Huang, Pawel A Penczek, Chao Yang, Robert M Glaeser, Paul D Adams, and Steven J Ludtke. 2007. “SPARX, a New Environment for Cryo-EM Image Processing.” Journal of Structural Biology 157 (1) (January): 47–55. http://www.ncbi.nlm.nih.gov/pubmed/16931051.

Holmes, Kenneth C, Isabel Angert, F Jon Kull, and Werner Jahn. 2003. “Electron Cryo-microscopy Shows How Strong Binding of Myosin to Actin Releases Nucleotide”: 423–427. doi:10.1038/nature01927.1.

Horne, RW, JM Hobart, and IP Ronchetti. 1975. “Application of the Negative Staining-carbon Technique to the Study of Virus Particles and Their Components by Electron Microscopy.” Micron 5: 233–261. http://www.sciencedirect.com/science/article/pii/0047720674900016.

Houben, Klaartje, Dominique Marion, Nicolas Tarbouriech, Rob W H Ruigrok, and Laurence Blanchard. 2007. “Interaction of the C-terminal Domains of Sendai Virus N and P Proteins: Comparison of Polymerase-nucleocapsid Interactions Within the Paramyxovirus Family.” Journal of Virology 81 (13) (July): 6807–16. doi:10.1128/JVI.00338-07. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1933331&tool=pmcentrez&rendertype=abstract.

Iseni, F, a Barge, F Baudin, D Blondel, and R W Ruigrok. 1998. “Characterization of Rabies Virus Nucleocapsids and Recombinant Nucleocapsid-like Structures.” The Journal of General Virology 79 (Pt 12 (December): 2909–19. http://www.ncbi.nlm.nih.gov/pubmed/9880004.

Iseni, F, F Baudin, D Blondel, and R W Ruigrok. 2000. “Structure of the RNA Inside the Vesicular Stomatitis Virus Nucleocapsid.” Rna New York Ny 6 (2): 270–281. http://www.ncbi.nlm.nih.gov/pubmed/10688365.

Jeng, T.-W., R.A. A Crowther, G. Stubbs, W Chiul, and W. Chiu. 1989. “Visualization of Alpha-helices in Tobacco Mosaic Virus by Cryo-electron Microscopy.” Journal of Molecular Biology 205 (1) (January): 251–257. doi:10.1016/0022-2836(89)90379-3. http://dx.doi.org/10.1016/0022-2836(89)90379-3.

Jensen, Malene Ringkjøbing, Guillaume Communie, Euripedes Almeida Ribeiro, Nicolas Martinez, Ambroise Desfosses, Loïc Salmon, Luca Mollica, et al. 2011. “Intrinsic Disorder in Measles Virus Nucleocapsids.” Proceedings of the National Academy of Sciences of the United States of America 108 (24) (June 14): 9839–44. doi:10.1073/pnas.1103270108. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3116414&tool=pmcentrez&rendertype=abstract.

Johansson, Kenth, Jean-Marie Bourhis, Valerie Campanacci, Christian Cambillau, Bruno Canard, and Sonia Longhi. 2003. “Crystal Structure of the Measles Virus Phosphoprotein Domain Responsible for the Induced Folding of the C-terminal Domain of the Nucleoprotein.” The Journal of Biological Chemistry 278 (45) (November 7): 44567–73. doi:10.1074/jbc.M308745200. http://www.ncbi.nlm.nih.gov/pubmed/12944395.

Karlin, D. 2003. “Structural Disorder and Modular Organization in Paramyxovirinae N and P.” Journal of General Virology 84 (12) (December 1): 3239–3252. doi:10.1099/vir.0.19451-0. http://vir.sgmjournals.org/cgi/doi/10.1099/vir.0.19451-0.

Kiselev, N A, F Y Lerner, and N B Livanova. 1971. “Electron Microscopy of Muscle Phosphorylase B.” Journal of Molecular Biology 62 (3) (December 28): 537–49. http://www.ncbi.nlm.nih.gov/pubmed/5167561.

Klaholz, Bruno P, Alexander G Myasnikov, and Marin Van Heel. 2004. “Visualization of Release Factor 3 on the Ribosome During Termination of Protein Synthesis” 427 (February): 3–6.

Klug, A, F. H. Crick, and H. W. Wyckoff. 1958. “Diffraction by Helical Structures.” Acta Crystallographica 11: 199.

Klug, A, and D J DeRosier. 1966. “Optical Filtering of Electron Micrographs : Reconstruction of One-Sided Images.” Nature 212: 29–32.

Kolakofsky, Daniel, Laurent Roux, Dominique Garcin, and Rob W H Ruigrok. 2005. “Paramyxovirus mRNA Editing, the ‘Rule of Six’ and Error Catastrophe: a Hypothesis.” The Journal of General Virology 86 (Pt 7) (July): 1869–77. doi:10.1099/vir.0.80986-0. http://www.ncbi.nlm.nih.gov/pubmed/15958664.

Korkhov, VM, and Carsten Sachse. 2010. “Three-dimensional Structure of TspO by Electron Cryomicroscopy of Helical Crystals.” Structure 18 (6): 677–687. http://www.sciencedirect.com/science/article/pii/S0969212610001073.

Korkhov, Vladimir M, Carsten Sachse, Judith M Short, and Christopher G Tate. 2010. “Three-dimensional Structure of TspO by Electron Cryomicroscopy of Helical Crystals.” *Structure (London, England : 1993)* 18 (6) (June 9): 677–87. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2911597&tool=pmcentrez&rendertype=abstract.

Lata, Suman, Guy Schoehn, Ankur Jain, Ricardo Pires, Jacob Piehler, Heinrich G Gottlinger, and Winfried Weissenhorn. 2008. “Helical Structures of ESCRT-III Are Disassembled by VPS4.” Science (New York, N.Y.) 321 (5894) (September 5): 1354–7. doi:10.1126/science.1161070.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2758909&tool=pmcentrez&rendertype=abstract.

Lee, Seunghee, Peter C Doerschuk, and John E Johnson. 2011. “Multiclass Maximum-likelihood Symmetry Determination and Motif Reconstruction of 3-D Helical Objects from Projection Images for Electron Microscopy.” *IEEE Transactions on Image Processing : a Publication of the IEEE Signal Processing* Society 20 (7) (July): 1962–76. doi:10.1109/TIP.2011.2107329. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3142268&tool=pmcentrez&rendertype=abstract.

Lepault, J, and K Leonard. 1985. “Three-dimensional Structure of Unstained, Frozen-hydrated Extended Tails of Bacteriophage T4.” Journal of Molecular Biology 182 (3) (April 5): 431–41. http://www.ncbi.nlm.nih.gov/pubmed/4009713.

Lepault, J, I Petitpas, I Erk, J Navaza, D Bigot, M Dona, P Vachette, J Cohen, and F a Rey. 2001. “Structural Polymorphism of the Major Capsid Protein of Rotavirus.” The EMBO Journal 20 (7) (April 2): 1498–507. doi:10.1093/emboj/20.7.1498. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=145494&tool=pmcentrez&rendertype=abstract.

Li, Huilin, David J. DeRosier, William V. Nicholson, Eva Nogales, and Kenneth H. Downing. 2002. “Microtubule Structure at 8 Å Resolution.” Structure 10 (10) (October): 1317–1328. doi:10.1016/S0969-2126(02)00827-4. http://dx.doi.org/10.1016/S0969-2126(02)00827-4.

Longhi, Sonia, Véronique Receveur-Bréchot, David Karlin, Kenth Johansson, Hervé Darbon, David Bhella, Robert Yeo, Stéphanie Finet, and Bruno Canard. 2003. “The C-terminal Domain of the Measles Virus Nucleoprotein Is Intrinsically Disordered and Folds Upon Binding to the C-terminal Moiety of the Phosphoprotein.” The Journal of Biological Chemistry 278 (20) (May 16): 18638–48. doi:10.1074/jbc.M300518200. http://www.ncbi.nlm.nih.gov/pubmed/12621042.

Low, Harry H, Carsten Sachse, Linda a Amos, and Jan Löwe. 2009. “Structure of a Bacterial Dynamin-like Protein Lipid Tube Provides a Mechanism for Assembly and Membrane Curving.” Cell 139 (7) (December 24): 1342–52. doi:10.1016/j.cell.2009.11.003. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2862293&tool=pmcentrez&rendertype=abstract.

Maki-Yonekura, Saori, Koji Yonekura, and Keiichi Namba. 2010. “Conformational Change of Flagellin for Polymorphic Supercoiling of the Flagellar Filament.” Nature Structural & Molecular Biology 17 (4) (April): 417–22. doi:10.1038/nsmb.1774. http://www.ncbi.nlm.nih.gov/pubmed/20228803.

Mandelkow, EM, and R Schultheiss. 1986. “On the Surface Lattice of Microtubules: Helix Starts, Protofilament Number, Seam, and Handedness.” *The Journal of Cell ...* 102 (March 1986): 1067–1073. http://jcb.rupress.org/content/102/3/1067.abstract.

Marabini, R, and J M Carazo. 1994. “Pattern Recognition and Classification of Images of Biological Macromolecules Using Artificial Neural Networks.” Biophysical Journal 66 (6) (June): 1804–14. doi:10.1016/S0006-3495(94)80974-9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1275906&tool=pmcentrez&rendertype=abstract.

Markham, Roy, Simon Frey, and G.J. Hills. 1963. “Methods for the Enhancement of Image Detail and Accentuation of Structure in Electron Microscopy.” Virology 20 (1) (May): 88–102. doi:10.1016/0042-6822(63)90143-0. http://linkinghub.elsevier.com/retrieve/pii/0042682263901430.

Markham, Roy, J. H. Hitchborn, G. J. Hills, and Simon Frey. 1964. “The Anatomy of the Tobacco Mosaic Virus.” Virology 22: 342–359.

Metlagel, Zoltan, Yayoi S Kikkawa, and Masahide Kikkawa. 2007. “Ruby-Helix: An Implementation of Helical Image Processing Based on Object-oriented Scripting Language.” Journal of Structural Biology 157 (1) (January): 95–105. doi:10.1016/j.jsb.2006.07.015. http://www.ncbi.nlm.nih.gov/pubmed/16996276.

Miyazawa, A, Y Fujiyoshi, M Stowell, and N Unwin. 1999. “Nicotinic Acetylcholine Receptor at 4.6 A Resolution: Transverse Tunnels in the Channel Wall.” Journal of Molecular Biology 288 (4) (May 14): 765–86. http://www.ncbi.nlm.nih.gov/pubmed/10329178.

Miyazawa, Atsuo, Yoshinori Fujiyoshi, and Nigel Unwin. 2003. “Structure and Gating Mechanism of the Acetylcholine Receptor Pore.” Nature 423 (6943) (June 26): 949–55. doi:10.1038/nature01748. http://dx.doi.org/10.1038/nature01748.

Moore, P B, H E Huxley, and D J DeRosier. 1970. “Three-dimensional Reconstruction of F-actin, Thin Filaments and Decorated Thin Filaments.” Journal of Molecular Biology 50 (2) (June 14): 279–95. http://www.ncbi.nlm.nih.gov/pubmed/7108961.

Moss, William J, and Diane E Griffin. 2006. “Global Measles Elimination.” Nature Reviews. Microbiology 4 (12) (December): 900–8. doi:10.1038/nrmicro1550. http://www.ncbi.nlm.nih.gov/pubmed/17088933.

Nakai, T., and A. F. Howatson. 1968. “The Fine Structure of Vesicular Stomatitis Virus.” Virology 35: 268–281.

Narita, a, T Yasunaga, T Ishikawa, K Mayanagi, and T Wakabayashi. 2001. “Ca(2+)-induced Switching of Troponin and Tropomyosin on Actin Filaments as Revealed by Electron Cryo-microscopy.” Journal of Molecular Biology 308 (2) (April 27): 241–61. doi:10.1006/jmbi.2001.4598. http://www.ncbi.nlm.nih.gov/pubmed/11327765.

Navaza, Jorge. 2003. “On the Three-dimensional Reconstruction of Icosahedral Particles.” Journal of Structural Biology 144 (1-2) (October): 13–23. http://dx.doi.org/10.1016/j.jsb.2003.09.007.

Newcomb, W W, and J C Brown. 1981. “Role of the Vesicular Stomatitis Virus Matrix Protein in Maintaining the Viral Nucleocapsid in the Condensed Form Found in Native Virions.” Journal of Virology 39 (1) (July): 295–9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=171289&tool=pmcentrez&rendertype=abstract.

Ogura, Toshihiko, Kenji Iwasaki, and Chikara Sato. 2003. “Topology Representing Network Enables Highly Accurate Classification of Protein Images Taken by Cryo Electron-microscope Without Masking.” Journal of Structural Biology 143 (3) (September): 185–200. doi:10.1016/j.jsb.2003.08.005. http://linkinghub.elsevier.com/retrieve/pii/S104784770300145X.

Ohi, Melanie, Ying Li, Yifan Cheng, and Thomas Walz. 2004. “Negative Staining and Image Classification - Powerful Tools in Modern Electron Microscopy.” Biological Procedures Online 6 (1) (January): 23–34. doi:10.1251/bpo70. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=389902&tool=pmcentrez&rendertype=abstract.

Owen, C H, D G Morgan, and D J DeRosier. 1996. “Image Analysis of Helical Objects: The Brandeis Helical Package.” Journal of Structural Biology 116 (1) (January): 167–75. http://dx.doi.org/10.1006/jsbi.1996.0027.

Parent, Kristin N, Christina T Deedas, Edward H Egelman, Sherwood R Casjens, Timothy S Baker, and Carolyn M Teschke. 2012. “Stepwise Molecular Display Utilizing Icosahedral and Helical Complexes of Phage Coat and Decoration Proteins in the Development of Robust Nanoscale Display Vehicles.” Biomaterials 33 (22) (August): 5628–37. doi:10.1016/j.biomaterials.2012.04.026. http://dx.doi.org/10.1016/j.biomaterials.2012.04.026.

Parent, Kristin N, Robert S Sinkovits, Margaret M Suhanovsky, Carolyn M Teschke, Edward H Egelman, and Timothy S Baker. 2010. “Cryo-reconstructions of P22 Polyheads Suggest That Phage Assembly Is

Nucleated by Trimeric Interactions Among Coat Proteins." Physical Biology 7 (4) (January): 045004. doi:10.1088/1478-3975/7/4/045004. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3202341&tool=pmcentrez&rendertype=abstract.

Paul, Danielle, Ardan Patwardhan, John M Squire, and Edward P Morris. 2004. "Single Particle Analysis of Filamentous and Highly Elongated Macromolecular Assemblies." Journal of Structural Biology 148 (2) (November): 236–50. doi:10.1016/j.jsb.2004.05.004. http://www.ncbi.nlm.nih.gov/pubmed/15477103.

Pauling, L, Robert B. Corey, and H.R. R Branson. 1951. "The Structure of Proteins; Two Hydrogen-bonded Helical Configurations of the Polypeptide Chain." Proceedings of the National Academy of Sciences of the United States of America 37 (4) (April): 205–11. http://www.pnas.org/content/37/4/205.short.

Penczek, P, M Radermacher, and J Frank. 1992. "Three-dimensional Reconstruction of Single Particles Embedded in Ice." Ultramicroscopy 40 (1) (January): 33–53. http://www.ncbi.nlm.nih.gov/pubmed/1580010.

Penczek, PA, Jun Zhu, and Joachim Frank. 1996. "A Common-lines Based Method for Determining Orientations for N> 3 Particle Projections Simultaneously." Ultramicroscopy 63: 205–218. http://www.sciencedirect.com/science/article/pii/030439919600037X.

Pomfret, Andrew J, William J Rice, and David L Stokes. 2007. "Application of the Iterative Helical Real-space Reconstruction Method to Large Membranous Tubular Crystals of P-type ATPases." Journal of Structural Biology 157 (1) (January): 106–16. doi:10.1016/j.jsb.2006.05.012. http://dx.doi.org/10.1016/j.jsb.2006.05.012.

Radermacher, M. 1988. "Three-dimensional Reconstruction of Single Particles from Random and Nonrandom Tilt Series." Journal of Electron Microscopy Technique 9 (4) (August): 359–94. doi:10.1002/jemt.1060090405. http://www.ncbi.nlm.nih.gov/pubmed/3058896.

Radermacher, M, T Ruiz, H Wieczorek, and G Grüber. 2001. "The Structure of the V(1)-ATPase Determined by Three-dimensional Electron Microscopy of Single Particles." Journal of Structural Biology 135 (1) (July): 26–37. doi:10.1006/jsbi.2001.4395. http://www.ncbi.nlm.nih.gov/pubmed/11562163.

Radermacher, M, T Wagenknecht, A Verschoor, and Joachim Frank. 1987. "Three-dimensional Structure of the Large Ribosomal Subunit from Escherichia Coli." The EMBO Journal 6 (4) (April): 1107–14. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=553509&tool=pmcentrez&rendertype=abstract.

Ramey, Vincent H, Hong-Wei Wang, and Eva Nogales. 2009. "Ab Initio Reconstruction of Helical Samples with Heterogeneity, Disorder and Coexisting Symmetries." Journal of Structural Biology 167 (2) (August): 97–105. doi:10.1016/j.jsb.2009.05.002. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2739800&tool=pmcentrez&rendertype=abstract.

Resch, Guenter P, Kenneth N Goldie, Andreas Hoenger, and J Victor Small. 2002. "Pure F-actin Networks Are Distorted and Branched by Steps in the Critical-point Drying Method." Journal of Structural Biology 137 (3) (March): 305–12. http://www.ncbi.nlm.nih.gov/pubmed/12096898.

Rudolph, M, I KRAUS, A DICKMANNS, M EICKMANN, W GARTEN, and R FICNER. 2003. "Crystal Structure of the Borna Disease Virus Nucleoprotein." Structure 11 (10) (October 1): 1219–1226. doi:10.1016/j.str.2003.08.011. http://www.cell.com/structure/fulltext/S0969-2126(03)00200-4.

Ruigrok, Rob W H, Thibaut Crépin, and Dan Kolakofsky. 2011. "Nucleoproteins and Nucleocapsids of Negative-strand RNA Viruses." Current Opinion in Microbiology 14 (4) (August): 504–10. doi:10.1016/j.mib.2011.07.011. http://www.ncbi.nlm.nih.gov/pubmed/21824806.

Sachse, Carsten, James Z Chen, Pierre-Damien Coureux, M Elizabeth Stroupe, Marcus Fändrich, and Nikolaus Grigorieff. 2007. "High-resolution Electron Microscopy of Helical Specimens: a Fresh Look at Tobacco Mosaic Virus." Journal of Molecular Biology 371 (3) (August 17): 812–35. doi:10.1016/j.jmb.2007.05.088. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2025690&tool=pmcentrez&rendertype=abstract.

Salunke, D M, D L Caspar, and R L Garcea. 1989. "Polymorphism in the Assembly of Polyomavirus Capsid Protein VP1." Biophysical Journal 56 (5) (November): 887–900. doi:10.1016/S0006-3495(89)82735-3. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1280588&tool=pmcentrez&rendertype=abstract.

Samatey, Fadel A, Hideyuki Matsunami, Katsumi Imada, Shigehiro Nagashima, Tanvir R Shaikh, Dennis R Thomas, James Z Chen, David J Derosier, Akio Kitao, and Keiichi Namba. 2004. "Structure of the Bacterial Flagellar Hook and Implication for the Molecular Universal Joint Mechanism." Nature 431 (7012) (October 28): 1062–8. doi:10.1038/nature02997. http://www.ncbi.nlm.nih.gov/pubmed/15510139.

Sander, B., M.M. Golas, and H. Stark. 2003. "Automatic CTF Correction for Single Particles Based Upon Multivariate Statistical Analysis of Individual Power Spectra." Journal of Structural Biology 142 (3) (June): 392–401. doi:10.1016/S1047-8477(03)00072-8. http://linkinghub.elsevier.com/retrieve/pii/S1047847703000728.

Schoehn, Guy, Frédéric Iseni, Manos Mavrakis, Danielle Blondel, and Rob W. H. Ruigrok. 2001. "Structure of Recombinant Rabies Virus nucleoprotein-RNA Complex and Identification of the Phosphoprotein Binding Site." Journal of Virology 75 (1): 490–498. doi:10.1128/JVI.75.1.490. http://jvi.asm.org/content/75/1/490.short.

Schoehn, Guy, Manos Mavrakis, Aurélie Albertini, Richard Wade, Andreas Hoenger, and Rob W H Ruigrok. 2004. "The 12 A Structure of Trypsin-treated Measles Virus N-RNA." Journal of Molecular Biology 339 (2) (May 28): 301–12. doi:10.1016/j.jmb.2004.03.073. http://www.ncbi.nlm.nih.gov/pubmed/15136034.

Serysheva, Irina I., Elena V Orlova, Wah Chiu, Michael B Sherman, Susan L. Hamilton, and Marin van Heel. 1995. "Electron Cryomicroscopy and Angular Reconstitution Used to Visualize the Skeletal Muscle Calcium Release Channel." Structural Biology 2 (1): 18–24.

Spehner, D, A Kirn, and R Drillien. 1991. "Assembly of Nucleocapsidlike Structures in Animal Cells Infected with a Vaccinia Virus Recombinant Encoding the Measles Virus Nucleoprotein." Journal of Virology 65 (11) (November): 6296–300. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=250336&tool=pmcentrez&rendertype=abstract.

Stewart, M, and R W Kensler. 1986. "Arrangement of Myosin Heads in Relaxed Thick Filaments from Frog Skeletal Muscle." Journal of Molecular Biology 192 (4) (December 20): 831–51. http://www.ncbi.nlm.nih.gov/pubmed/3495665.

Tang, Guang, Liwei Peng, Philip R Baldwin, Deepinder S Mann, Wen Jiang, Ian Rees, and Steven J Ludtke. 2007. "EMAN2: An Extensible Image Processing Suite for Electron Microscopy." Journal of Structural Biology 157 (1) (January): 38–46. http://www.ncbi.nlm.nih.gov/pubmed/16859925.

Tawar, Rajiv G, Stéphane Duquerroy, Clemens Vonrhein, Paloma F Varela, Laurence Damier-Piolle, Nathalie Castagné, Kirsty MacLellan, et al. 2009. "Crystal Structure of a Nucleocapsid-like nucleoprotein-RNA Complex of Respiratory Syncytial Virus." Science (New York, N.Y.) 326 (5957) (November 27): 1279–83. doi:10.1126/science.1177634. http://www.ncbi.nlm.nih.gov/pubmed/19965480.

Taylor, K A, and R M Glaeser. 1974. "Electron Diffraction of Frozen, Hydrated Protein Crystals." Science (New York, N.Y.) 186 (4168) (December 13): 1036–7. http://www.ncbi.nlm.nih.gov/pubmed/4469695.

Thuku, R Ndoria, Brandon W Weber, Arvind Varsani, and B Trevor Sewell. 2007. “Post-translational Cleavage of Recombinantly Expressed Nitrilase from Rhodococcus Rhodochrous J1 Yields a Stable, Active Helical Form.” The FEBS Journal 274 (8) (April): 2099–108. http://www.ncbi.nlm.nih.gov/pubmed/17371547.

Toyoshima, Chikashi. 2000. “Structure Determination of Tubular Crystals of Membrane Proteins . I . Indexing of Di ! Raction Patterns” 84.

Trachtenberg, Shlomo, Vitold E Galkin, and Edward H Egelman. 2005. “Refining the Structure of the Halobacterium Salinarum Flagellar Filament Using the Iterative Helical Real Space Reconstruction Method: Insights into Polymorphism.” Journal of Molecular Biology 346 (3): 665–676. http://www.ncbi.nlm.nih.gov/pubmed/15713454.

Trinick, J, J Cooper, J Seymour, and E H Egelman. 1986. “Cryo-electron Microscopy and Three-dimensional Reconstruction of Actin Filaments.” Journal of Microscopy 141 (Pt 3) (March): 349–60. http://www.ncbi.nlm.nih.gov/pubmed/3701854.

Unwin, N. 1993. “Nicotinic Acetylcholine Receptor at 9 Resolution.” Journal of Molecular Biology 229: 1101–1124. http://www2.mrc-lmb.cam.ac.uk/groups/nu/jmb93.pdf.

Unwin, Nigel, and Yoshinori Fujiyoshi. 2012. “Gating Movement of Acetylcholine Receptor Caught by Plunge-freezing.” Journal of Molecular Biology 422 (5) (October 5): 617–34. doi:10.1016/j.jmb.2012.07.010. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3443390&tool=pmcentrez&rendertype=abstract.

Unwin, P N, and A. Klug. 1974. “Electron Microscopy of the Stacked Disk Aggregate of Tobacco Mosaic Virus Protein.” Journal of Molecular Biology 87: 641–656.

Unwin, P.N.T., and R. Henderson. 1975. “Molecular Structure Determination by Electron Microscopy of Unstained Crystalline Specimens.” Journal of Molecular Biology 94 (3) (May 25): 425–40. http://www.ncbi.nlm.nih.gov/pubmed/1236957.

Vigers, G P, R A Crowther, and B M Pearse. 1986. “Location of the 100 Kd-50 Kd Accessory Proteins in Clathrin Coats.” The EMBO Journal 5 (9) (September): 2079–85. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1167085&tool=pmcentrez&rendertype=abstract.

Vogel, R H, S W Provencher, C H von Bonsdorff, M Adrian, and J Dubochet. 1986. “Envelope Structure of Semliki Forest Virus Reconstructed from Cryo-electron Micrographs.” Nature 320: 533–5. doi:10.1038/320533a0. http://www.ncbi.nlm.nih.gov/pubmed/3960136.

Wakabayashi, T, H E Huxley, L a Amos, and a Klug. 1975. “Three-dimensional Image Reconstruction of Actin-tropomyosin Complex and Actin-tropomyosin-troponin T-troponin I Complex.” Journal of Molecular Biology 93 (4) (April 25): 477–97. http://www.ncbi.nlm.nih.gov/pubmed/1142432.

Wang, HW, and E. Nogales. 2005. “An Iterative Fourier-Bessel Algorithm for Reconstruction of Helical Structures with Severe Bessel Overlap.” Journal of Structural Biology 149 (1) (January): 65–78. doi:10.1016/j.jsb.2004.08.006. http://www.ncbi.nlm.nih.gov/pubmed/15629658.

Wang, Ying A, Xiong Yu, Calvin Yip, Natalie C Strynadka, and Edward H Egelman. 2006. “Structural Polymorphism in Bacterial EspA Filaments Revealed by cryo-EM and an Improved Approach to Helical Reconstruction.” *Structure (London, England : 1993)* 14 (7) (July): 1189–96. doi:10.1016/j.str.2006.05.018. http://www.ncbi.nlm.nih.gov/pubmed/16843900.

Wang, Ying a, Xiong Yu, Stacy Overman, Masamichi Tsuboi, George J Thomas, and Edward H Egelman. 2006. “The Structure of a Filamentous Bacteriophage.” Journal of Molecular Biology 361 (2) (August 11): 209–15. doi:10.1016/j.jmb.2006.06.027. http://www.ncbi.nlm.nih.gov/pubmed/16843489.

Ward, Andrew, Michael F Moody, Brian Sheehan, Ronald a Milligan, and Bridget Carragher. 2003. “Windex: a Toolset for Indexing Helices.” Journal of Structural Biology 144 (1-2) (October): 172–183. doi:10.1016/j.jsb.2003.10.008. http://linkinghub.elsevier.com/retrieve/pii/S1047847703002193.

Watson, J. D., and F. H. C. Crick. 1953. “A Structure for Deoxyribose Nucleic Acid.” Nature 171: 737–738. http://eduardbardaji.com/DOCENCIA/prodnat/nucleicacids03.doc.

White, Helen E, Helen R Saibil, Athanasios Ignatiou, and Elena V Orlova. 2004. “Recognition and Separation of Single Particles with Size Variation by Statistical Analysis of Their Images”: 453–460. doi:10.1016/j.jmb.2003.12.015.

Whittaker, M, B O Carragher, and R A Milligan. 1995. “PHOELIX: a Package for Semi-automated Helical Reconstruction.” Ultramicroscopy 58 (3-4): 245–259. http://www.sciencedirect.com/science/article/pii/0304399195000578.

Wilkins, MH, AR Stokes, and HR Wilson. 1953. “Molecular Structure of Deoxypentose Nucleic Acids.” Nature 171: 738–740. http://www.ncbi.nlm.nih.gov/pubmed/12569936.

Williams, Robley C., and Ralph W. G. Wyckoff. 1944. “The Thickness of Electron Microscopic Objects.” Journal of Applied Physics 15 (10) (October 1): 712. doi:10.1063/1.1707376. http://link.aip.org/link/?JAPIAU/15/712/1.

Yamaguchi, Tomohiro, Takashi Fujii, Yoshito Abe, Teruhisa Hirai, Dongchon Kang, Keiichi Namba, Naotaka Hamasaki, and Kaoru Mitsuoka. 2010. “Helical Image Reconstruction of the Outward-open Human Erythrocyte Band 3 Membrane Domain in Tubular Crystals.” Journal of Structural Biology 169 (3) (March): 406–12. doi:10.1016/j.jsb.2009.12.009. http://www.ncbi.nlm.nih.gov/pubmed/20005958.

Yasunaga, T, and T Wakabayashi. 1996. “Extensible and Object-oriented System Eos Supplies a New Environment for Image Analysis of Electron Micrographs of Macromolecules.” Journal of Structural Biology 116 (1) (January): 155–60. http://dx.doi.org/10.1006/jsbi.1996.0025.

Yonekura, Koji, Saori Maki-Yonekura, and Keiichi Namba. 2003. “Complete Atomic Model of the Bacterial Flagellar Filament by Electron Cryomicroscopy.” Nature 424 (6949) (August 7): 643–50. doi:10.1038/nature01830. http://www.ncbi.nlm.nih.gov/pubmed/12904785.

———. 2005. “Building the Atomic Model for the Bacterial Flagellar Filament by Electron Cryomicroscopy and Image Analysis.” *Structure (London, England : 1993)* 13 (3) (March): 407–12. doi:10.1016/j.str.2005.02.003. http://www.ncbi.nlm.nih.gov/pubmed/15766542.

Thesis Summary

Flexible helical protein polymers exemplified by actin filaments, microtubules and bacterial flagella are ubiquitous in biology. Due to their size and intrinsic irregularities, the structure of these polymers cannot be solved by X-ray crystallography. Since half a century, three-dimensional (3D) reconstruction from two-dimensional (2D) Electron Microscopy (EM) images appears as a method of choice to solve the structure of large helical polymers. However, depending on the degree of flexibility of the analyzed helices, the 3D reconstruction process can still be a daunting task. For the most regular helices, the classical reciprocal space-based Fourier-Bessel approach can allow both to determine the helical symmetry and to calculate 3D structures. For more flexible structures, recent "single-particle" approaches consist in segmentation of long irregular helices into short (i.e. locally more regular) segments and their processing as asymmetrical objects with defined symmetry-imposed constraints (Egelman, 2000; Sachse et al., 2007). However, two major difficulties remain: the heterogeneous data must be sorted into homogeneous populations and the helical symmetry for each population has to be determined. In the presented work, we explored various single-particle approaches, developed new analysis methods, and implemented most of them into a user-friendly processing pipeline. The target biological objects were helical nucleocapsids of two negative strand RNA viruses, Measles (MeV) and Vesicular Stomatitis Virus (VSV ; the prototype for Rabies), the latter being particularly flexible in terms of helical parameters (diameter, number of subunits per turn). Nucleocapsids are formed by the viral genomic RNA coated by the nucleoprotein and serve as a template for viral replication and transcription. To overcome the heterogeneity problem, we used 2D classification, described general processing protocols and applications for helical segments, and introduced a new classification method based on the power spectra of the images. The determination of helical symmetry(ies) was addressed by a novel approach relying on ab initio exhaustive search of helical parameters whereby we start from a single 2D image, reconstruct as many 3D structures as parameters to test by cropping the image and assigning views to the obtained segments, and calculate the cross-correlation (CC) of the reprojection of the 3D model with the initial image. Applied to artificial data sets, the method was effectively able to detect a maximum of CC for the true symmetry parameters, but also showed intrinsic ambiguities of helical symmetry determination on which we extensively comment. Altogether, the result of this method-oriented work allowed us to address several biological questions. First, the 3D reconstruction by negative stain EM of two forms of nucleocapsids of MeV coupled to a docking of a homologous crystal structure enabled us to determine the orientation of the nucleoprotein and of the RNA in the nucleocapsids. Secondly, we assessed the structure of in vitro formed nucleocapsids of VSV and showed that assemblies close to the native viral nucleocapsids can be formed in the absence of any other viral proteins, thus providing new insights into the assembly of this virus. As a perspective of this work, our pipeline of flexible helical analysis is being extended and successfully used for other projects.

Les biopolymères hélicoïdaux flexibles sont ubiquitaires dans le monde biologique. Du fait de leur taille et de leur irrégularité, leur structure ne peut pas être résolue par cristallographie aux rayons X. Depuis un demi-siècle, la reconstruction 3D à partir d'images 2D obtenues par microscopie électronique (ME) s'est imposée comme une méthode de choix pour résoudre les structures de polymères hélicoïdaux. Cependant, en fonction du degré de flexibilité des hélices, le processus de reconstruction peut s'avérer être une tâche délicate. Pour les hélices les plus régulières, la méthode classique basée sur les méthodes de Fourier-Bessel permettent en même temps de déterminer les paramètres hélicoïdaux et de calculer des reconstructions 3D. Pour les structures plus flexibles, des approches récemment développées consistent à segmenter les long hélices en courts segments, localement plus régulier, et les traiter comme des particules isolées, tout en ajoutant des contraintes basées sur la symétrie (Egelman, 2000; Sachse et al., 2007). Deux difficultés majeures subsistent : les données, hétérogènes, doivent être séparées en sous-ensembles plus homogènes, et la symétrie doit être déterminée pour chaque sous-ensemble. Dans le travail présenté, pour résoudre ces problèmes, nous avons exploré différentes méthodes de particules isolées, développé de nouvelles approches, et implémenté la plupart dans une suite de modules de traitement d'image orientée utilisateur. Les objets biologiques étudiés ont été les nucléocapsides hélicoïdales et flexibles des virus de la Rougeole (MeV) et de la stomatite vésiculaire (VSV). Les nucléocapsides sont constituées du génome viral (ARN simple brin), couvert par la nucléoprotéine, et servent de matrices pour la transcription et la réplication virale. Pour palier a l'hétérogénéité des données, nous avons utilisé la classification 2D, décrit des protocoles de traitement et leur application aux segments hélicoïdaux, et introduit une nouvelle méthode de classification basée sur le spectre de puissance des images. Pour la détermination des paramètres de symétrie, nous proposons un approche nouvelle, ab initio, se basant sur une recherche quasi-exhaustive des paramètres et dans laquelle l'information de départ est une simple image 2D. Cette méthode a été testée sur des données artificielles et a montré qu'elle permet d'obtenir un score localement maximum pour les paramètres réels, même si sur un champ plus large, plusieurs solutions apparaissent possibles, montrant ainsi l'ambigüité intrinsèque de la détermination des paramètres de symétrie hélicoïdale sur une image 2D, que nous caractérisons et commentons en détail. Dans l'ensemble, les résultats de cette thèse orientée méthodes nous ont permis de répondre a plusieurs questions biologiques. Premièrement, les reconstructions 3D obtenues par coloration négative de deux formes de nucléocapsides de MeV associées au recalage d'un structure cristallographique d'une protéine homologue nous ont permis de déterminer l'orientation de la nucléoprotéine et de l'ARN viral dans les nucléocapsides. Deuxièmement, nous avons résolu la structure de nucléocapsides de VSV reconstituées in vitro et avons montré que des assemblages proches de ceux trouves dans le virus natif peuvent être formés en l'absence de toute autre protéine virale, apportant un nouveau regard sur l'assemblage de ce virus. En perspective de ce travail, notre suite de modules de traitement d'image adaptés aux hélices flexibles est maintenant étendue et utilisée avec succès pour d'autres projets.