Assessing the Reconstruction of Macro-molecular Assemblies: the Example of the Nuclear Pore Complex

Frédéric Cazals, Tom Dreyfus

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Assessing the Reconstruction of Macro-molecular Assemblies: the Example of the Nuclear Pore Complex

Frédéric Cazals — Tom Dreyfus

N° 7513 — version 2

version initiale February 2011 — version révisée March 2011

Thème BIO

ISSN 0249-6399 ISRN INRIA/RR--7513--FR+ENG

INRIA
SOPHIA ANTIPOLIS

Assessing the Reconstruction of Macro-molecular Assemblies: the Example of the Nuclear Pore Complex

Frédéric Cazals *, Tom Dreyfus †

Thème BIO — Systèmes biologiques
Projet ABS

Rapport de recherche n° 7513 — version 2 — version initiale February 2011 — version révisée March 2011 — 44 pages

Abstract: The reconstruction of large protein assemblies is a major challenge due to their plasticity and due to the flexibility of the proteins involved. An emerging trend to cope with these uncertainties consists of performing the reconstruction by integrating experimental data from several sources, a strategy recently used to propose qualitative reconstructions of the Nuclear Pore Complex. Yet, the absence of clearly identified canonical reconstructions and the lack of quantitative assessment with respect to the experimental data are detrimental to the mechanistic exploitation of the results.

To leverage such reconstructions, this work proposes a modeling framework inherently accommodating uncertainties, and allowing a precise assessment of the reconstructed models. We make three contributions. First, we introduce *toleranced models* to accommodate the positional and conformational uncertainties of protein instances within large assemblies. A toleranced model is a continuum of geometries whose distinct topologies can be enumerated, and mining stable complexes amidst this finite set hints at important structures in the assembly. Second, we present a panoply of tools to perform a multi-scale topological, geometric, and biochemical assessment of the complexes associated to a toleranced model, at the assembly and local levels. At the assembly level, we assess the prominence of contacts and the quality of the reconstruction, in particular w.r.t symmetries. At the local level, the complexes encountered in the toleranced model are used to confirm / question / suggest protein contacts within a known 3D template known at atomic resolution. Third, we apply our machinery to the NPC for which we (i) report prominent contacts uncovering sub-complexes of the NPC, (ii) explain the closure of the two rings involving 16 copies of the Y-complex, and (iii) develop a new 3D template for the T-complex.

These contributions should prove instrumental in enhancing the reconstruction of assemblies, and in selecting the models which best comply with experimental data.

Key-words: Proteins, macro-molecular complexes, structural biology, nuclear pore complex. Union of balls, curved Voronoi diagrams, curved α-shapes, stability, topological persistence, graph matching, maximum common sub-graphs.

* INRIA Sophia-Antipolis-Méditerranée, Algorithms-Biology-Structure; Frederic.Cazals@sophia.inria.fr
† INRIA Sophia-Antipolis-Méditerranée, Algorithms-Biology-Structure; Tom.Dreyfus@sophia.inria.fr

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Evaluation de la Reconstructions de Gros Assemblages Protéiques: l'Exemple du Pore Nucléaire

Résumé : La reconstruction de gros assemblages est un challenge majeur en raison de leur plasticité, mais aussi de la flexibilité des protéines impliquées. Une stratégie émergeante pour faire face à ces incertitudes consiste à intégrer des données expérimentales diverses, cette stratégie ayant fait ses preuves pour la reconstruction de modèles qualitatifs du pore nucléaire (NPC), qui est le plus gros complexe protéique connu à ce jour dans la cellule eucaryote. Néanmoins, l'absence d'une part de reconstructions canoniques et d'autre part d'évaluation quantitative de la cohérence des modèles produits avec les données expérimentales utilisées nuit à l'exploitation des résultats.

Pour améliorer ces reconstructions, ce travail propose un paradigme de modélisation prenant en compte de façon inhérente les incertitudes, et permettant par ailleurs une évaluation précise des modèles reconstruits. Les contributions présentées sont triples. Tout d'abord, nous introduisons les *modèles tolérancés* de façon à prendre en compte les incertitudes relatives à la position et à la forme des protéines dans un assemblage. Un modèle tolérancé est un continuum géométrique dont on peut énumérer toutes les topologies possibles, et les régions stables au sein de celles-ci met sont autant d'indices vers des parties potentiellement importantes de l'assemblage. Ensuite, nous présentons une panoplie d'outils d'analyse topologique, géométrique, et biochimique des complexes associés à un modèle tolérancé, à la fois au niveau global et local. Au niveau de l'assemblage, nous évaluons la prégnance des contacts et la qualité de la reconstruction, en particulier vis à vis des symétries. Au niveau local, les complexes observés sont utilisés pour confirmer / infirmer / suggérer de nouveaux contacts au sein d'un template 3D d'un sous-système. Enfin, nous appliquons ces outils au pore nucléaire, pour lequel nous (i) mettons en exergue des contacts prégnants relatifs à plusieurs sous-systèmes, (ii) étudions la fermeture des deux anneaux impliquant 16 copies du complexe Y, et (iii) développons un nouveau template 3D pour le T-complex.

De façon générale, nous pensons que ces travaux vont permettre d'une part d'améliorer la reconstruction de gros assemblages, et d'autre part de sélectionner les reconstructions montrant la plus forte cohérence avec les données expérimentales.

Mots-clés : Protéines, complexes macro-moléculaires, interfaces, pore nucleaire, biologie structurale. Union de boules, diagrammes de Voronoi courbes, α-shapes courbes, stabilité, persistence topologique, matching de graphes, sous-graphes maximaux.

1 Reconstructing Large Macro-molecular Assemblies

Reconstruction by data integration. Large protein assemblies such as the Nuclear Pore Complex (NPC), chaperonin cavities, the proteasome or ATP synthases, to name a few, are key to numerous biological functions. To improve our understanding of these functions, one would ideally like to build and animate atomic models of these molecular machines. However, this task is especially tough, due to their size and their plasticity, but also due to the flexibility of the proteins involved. In a sense, the modeling challenges arising in this context are different from those faced for binary docking, and also from those encountered for intermediate size complexes which are often amenable to a processing mixing (cryo-EM) image analysis and classical docking. To face these new challenges, an emerging paradigm is that of reconstruction by data integration [AFK+08]. In a nutshell, the strategy is reminiscent from NMR and consists of mixing experimental data from a variety of sources, so as to find out the model(s) best complying with the data. This strategy has been in particular used to propose plausible models of the Nuclear Pore Complex [ADV+07a], the largest assembly known to date in the eukaryotic cell, and consisting of 456 protein *instances* of 30 *types*.

Modeling with uncertainties and model assessment. Reconstruction by data integration requires three ingredients. First, a parametrized model must be adopted, typically a collection of balls to model a protein with pseudo-atoms. Second, as in NMR, a functional measuring the agreement between a model and the data must be chosen. In [ADV+07b], this functional is based upon *restraints*, namely penalties associated to the experimental data. Third, an optimization scheme must be selected. The design of restraints is notoriously challenging, due to the ambiguous nature and/or the noise level of the data. For example, Tandem Affinity Purification (TAP) gives access to a *pullout* i.e. a list of protein types which are known to interact with one tagged protein type, but no information on the number of complexes or on the stoichiometry of proteins types within a complex is provided. In cryo-EM, the envelope enclosing an assembly is often imprecisely defined, in particular in regions of low density. For immuno-EM labelling experiments, positional uncertainties arise from the microscope resolution.

These uncertainties coupled with the complexity of the functional being optimized, which in general is non convex, have two consequences. First, it is impossible to single out a unique reconstruction, and a set of plausible reconstructions must be considered. As an example, 1000 plausible models of the NPC were reported in [ADV+07b]. Interestingly, averaging the positions of all balls of a particular protein type across these models resulted in 30 so-called *probability density maps*, each such map encoding the probability of presence of a particular protein type at a particular location in the NPC. Second, the assessment of all models (individual and averaged) is non trivial. In particular, the lack of straightforward statistical analysis of the individual models and the absence of assessment for the averaged models are detrimental to the mechanistic exploitation of the reconstruction results. At this stage, such models therefore remain qualitative.

Contributions. This work tackles the difficulties just discussed, and proposes a modeling framework inherently accommodating uncertainties, and allowing a precise assessment of reconstructed models. We make three contributions.

First, we introduce *toleranced models* to accommodate the positional and conformational uncertainties of protein instances within large assemblies. A toleranced model is a collection of *toleranced balls*, each such ball consisting of two concentric balls called the *inner* and *outer* balls, respectively meant to encode high and low confidence regions. A toleranced model is a one-parameter family of shapes, since growing the radii of toleranced balls results in a continuum of nested geometries, whence accommodating the aforementioned uncertainties. In particular, it is possible to enumerate the finite set of topologies encountered along the growth process, each connected component associated to a given topology being a protein *complex*.

Second, we present a panoply of tools to perform a topological, geometric, and biochemical assessment of the complexes associated to a toleranced model, at the global and local levels. At the global level, a multi-scale investigation of the protein complexes involving selected protein types provides information on prominent contacts and on the overall quality of the reconstruction, which is especially useful in the presence of symmetries. At the local level, let a *template* be a 3D model meant to probe the protein complexes of the toleranced model. We confirm / question / suggest protein contacts of the template based on the prominent contacts seen in the toleranced model,and hint at missing and or ill-placed proteins. Note that these tools can naturally be used to run in-silico experiments aiming at testing hypothesis.

Third, we apply our machinery to the NPC, so as to bridge the gap between global yet qualitative models of the whole NPC, and atomic models of sub-complexes. Starting from a toleranced model derived from the probability density maps of Alber et al. [ADV+07a], we (i) report prominent contacts uncovering sub-complexes of the NPC, (ii) explain the closure of the two rings involving copies of the Y-complex, in the context of the work by Blobel et al [KB09] and Vetter et al [SSF+08], and (iii) develop a new template for the T-complex.

Mathematically, our framework elaborates on previous work in computational geometry, computational topology, and graph theory. Tracking the evolution of topological features associated with a collection of growing balls is the seminal contribution of affine α-shape [Ede92], which was later put in the context of Morse theory [GJ03] and topological persistence [CSEH05, CCS11]. In this context, the novelty of our work resides in the introduction of toleranced models, and in the ability to investigate the one-parameter family of shapes defined by the α-shape of an additively-multiplicatively weighted Voronoi diagram [CD10]. Also, the comparison of the contacts within a protein complex and a template is phrased in terms of Maximal Common Induced Sub-graph (MCIS) and Maximal Common Edge Sub-graph (MCES), which are enumeration problems admitting exact algorithms [CK05].

2 Theory

2.1 Toleranced Models of Proteins and Assemblies

Motivations. In molecular modeling, facing uncertainties on the shape of proteins and/or on their positions is commonplace. In the sequel, we develop a model accommodating uncertain models i.e. models featuring high and low confidence regions.

Toleranced models. We consider a protein assembly consisting of n *protein instances* from $p < n$ *protein types*. For example, the NPC consists of $n = 456$ instances of $p = 30$ protein types. Let a *toleranced ball* $\overline{B_i}(p_i; r_i^-, r_i^+)$ be a pair of concentric balls, the *inner* and *outer* balls, respectively, of radii $r_i^+ > r_i^-$. Inner and outer balls are respectively meant to encode high confidence regions and uncertain regions. In order to deal with balls of intermediate size, we introduce a parameter $\lambda > 0$ governing a *growth process* consisting of linearly interpolating and extrapolating the radii. That is, the *grown ball* $\overline{B_i}[\lambda]$ stands for the ball centered at p_i and of radius:

$$r_i(\lambda) = r_i^- + \lambda(r_i^+ - r_i^-). \tag{1}$$

Note that for $\lambda = 0$ (resp. $\lambda = 1$), the grown ball matches the inner (resp. outer) ball.

We define a *toleranced protein* $\overline{P_j}$ as a collection of toleranced balls, and a *toleranced assembly* as a collection of toleranced proteins. For a given value of λ, a protein of intermediate size is denoted $\overline{P_j}[\lambda]$, and $\mathcal{F}_\lambda$ denotes the domain corresponding to the union of growing balls, that is $\mathcal{F}_\lambda = \cup_i \overline{B_i}[\lambda] = \cup_j \overline{P_j}[\lambda]$. For a fixed λ, the topology of the domain $\mathcal{F}_\lambda$ is of utmost interest: a connected component of this domain is called a *complex*; the domain is called a *mixture* if it involves several complexes.

These notions are illustrated on Fig. 1. The construction of toleranced models depends on the uncertainties of the data processed, and the reader is referred to section 3.2 for toleranced models derived from probability density maps in the context of the NPC.

Curved Voronoi diagrams. We have assumed so far that the p protein types are tantamount. Assume now that these types belong to two groups, which we term the red and blue groups for the sake of exposure. This *bicolor* setting is meant to deal with models of large assemblies, where the red group will refer to the protein types involved in a TAP experiment or to those seen in a sub-complex.

To compute complexes and mixtures in the bicolor setting, we resort to the theory of curved Voronoi diagrams and α-shapes [CD10]. Intuitively, the growth process of Eq. (1) allows one partition the three-dimensional space of into so-called *Voronoi regions*, with one region V_i for each toleranced ball $\overline{B_i}$: a point p belongs to V_i if the growing ball $\overline{B_i}[\lambda]$ reaches point p before any ball $\overline{B_j}[\lambda] \neq \overline{B_i}[\lambda]$. A region V_i is bounded by curved bisectors defined by $\overline{B_i}$ and neighboring balls.

For a given ball $\overline{B_i}[\lambda]$, consider its *restriction* to its Voronoi region, that is the intersection $\overline{B_i}[\lambda] \cap V_i$. These restrictions naturally partition the domain $\mathcal{F}_\lambda$, and their connected components correspond to the aforementioned complexes. Moreover, we use the pairwise intersections between the restrictions involved in a complex C to define its *skeleton* graph G_C: its nodes are the toleranced proteins of C; an edge links $\overline{P_i}$ and $\overline{P_j}$ provided that there exists two intersecting restrictions, one from $\overline{P_i}$ and one from $\overline{P_j}$.

Stability analysis. Growing λ results in merges between complexes. The set of finite topologies [1] corresponding to this evolution can be represented in a directed acyclic graph called *Hasse diagram*, a special graph whose nodes are the complexes, with an edge joining (generically) two nodes when the complexes merge along the growth process. The origin (endpoint) of an edge therefore represents the birth (resp. death) of a complex C: at $\lambda = \lambda_b(C)$, the complex gets formed by a merge of two or more complexes; at $\lambda = \lambda_d(C)$, the complex dies by merging with at least another complex. Thus, the *lifetime* $s(C) = \lambda_d(C) - \lambda_b(C)$ provides a measure of the topological stability of the complex C. Also, the *ancestors* and *successors* of C are the complexes contained into and containing, respectively, the complex C. See Fig. 1(Bottom row) for an illustration.

In the bicolor setting, let T be the list of red protein types. A complex C of the Hasse diagram is made of instances whose types are in T. If each type of T is present exactly once in C, the complex C is termed an *isolated copy*. The number and the lifetime of isolated copies give a measure of the separability of the different copies of a complex involving all the types of T. Note that the intersection of the lifetime intervals of the different isolated copies may be empty.

2.2 Topological Assessment of Complexes

Motivations. From a topological standpoint, a complex C associated to a node of the Hasse diagram is characterized by its skeleton graph. In this section, we present tools to compare the skeleton graph of a complex C against that of a template T of C. Practically, T shall be a co-crystallized complex or a high-resolution model built in-silico, and the protein types in T identify the red proteins of the bicolor setting.

Comparing the skeleton graphs of a template and of a complex. We assume that all the types of the instances present in C are present in T, and formalize this comparison in terms of graph theory.

The skeleton graph G_C corresponds to a complex C whose nodes are protein instances i.e. each instance carries a unique identifying label. On the other hand, we assume that the nodes of G_t are protein types, so that a node of G_C (a protein instance) can be uniquely mapped to a node of G_t (a protein type). (This latter assumption is warranted by the fact that the templates of the NPC to be analyzed have at most one instance of each protein type.) Note also that the complex C may not feature instances of all the types found in the template T. We therefore denote $G_{t|C}$ the *restricted template* i.e. the graph obtained by removing from G_t all the nodes whose protein types are not found in the protein instances of G_C, and the edges incident on these nodes. To compare the graphs $G_{t|C}$ and G_C, we use the concept of *matching*.

A *matching* from $G_{t|C}$ to C maps vertices of $G_{t|C}$ (protein types of the template) to vertices of G_C (protein instances of the complex), and edges of $G_{t|C}$ (contacts within the template) to edges of G_t (contacts within the complex). Taking the template as reference, we assess a matching with its:

- *Matching protein type(s):* a protein type of $G_{t|C}$ with a corresponding instance in C.
- *Missing protein type(s):* a protein type of $G_{t|C}$ with no corresponding instance in C.
- *Matching contact(s):* a contact in $G_{t|C}$ with a counterpart in C.
- *Missing contact(s):* a contact in $G_{t|C}$ with no counterpart in C.
- *Extra contact(s):* a contact in C with no counterpart in $G_{t|C}$.

Computing matchings is tantamount to computing maximal cliques [CK05], and of particular interests are the matchings associated to the so-called Maximal Common Induced Sub-graph (MCIS) and Maximal Common Edge Sub-graph (MCES) of $G_{t|C}$ and G_C. See the supplemental section 7.3.

Perfect matchings. Along the growth process, we are interested in the complexes C which exhibit a perfect matching with the associated restricted template $G_{t|C}$, and which are maximal—there exists no perfect matching for the successors of C. Such complexes are easily obtained from the matchings provided by a MCIS calculation between the graphs G_C and $G_{t|C}$ in each node of the Hasse diagram.

[1] We track the evolution of connected components, but not that of higher order homology generators.

Alternate matchings. Consider a complex C such that there is no perfect matching for C or any of its successors. In that case, we are interested in maximizing the number of common contacts between G_C and $G_{t|C}$, which corresponds to a MCES calculation. To report such matchings, we proceed as follows. First, for each complex C which is a root of the Hasse diagram, we compute the MCES between G_C and $G_{t|C}$. Second, let A be a matching returned by the MCES calculation. We search the ancestor D of C involving the protein instances and contacts of C matched by A, and minimizing the number of extra contacts.

2.3 Geometric Assessment of Complexes

Motivations. For large values of λ, the topological assessment of section 2.2 may be satisfactory if all (most) of the contacts seen in the template are found. However, a misplaced protein may require a large value of λ to obtain the correct contacts. This motivates the following statistic.

Volume ratio. Estimating the volume $Vol_{ref}(P_i)$ of a protein instance P_i from its sequence [HGC94], let $Vol_{ref}(C) = \sum_{P_i \in C} Vol_{ref}(P_i)$ the reference volume of the complex C, estimated from its constituting instances. On the other hand, for a fixed λ, let $Vol_\lambda(C)$ be the volume of the complex C, defined as the sum of the volumes of the Voronoi restrictions [2] of its toleranced proteins. The following ratio, which should ideally be close to one, is used to make a geometric assessment:

$$r_\lambda(C) = Vol_\lambda(C)/Vol_{ref}(C). \tag{2}$$

2.4 Combining the Geometric, Topological and Biochemical Assessments

Mining contacts. As a global assessment, we investigate all pairwise contacts of protein types appearing in the Hasse diagram. In [ADV+07b, Fig. 10], recall that the *contact frequency* f_{ij} between two protein types p_i and p_j is defined as the fraction of structures in the ensemble (of size 1000) for which at least one contact between two protein instances of these types is observed. The toleranced model being a continuous geometric model, we define a *contact probability* analogous to the contact frequency. Having painted all the proteins types in red, consider the Hasse diagram for the range of λ-values $[0, \lambda_{\max}]$, as discussed in section 3.2. Consider two protein types p_i and p_j, and a stoichiometry $k \geq 1$. As soon as k pairwise contacts between distinct pairs of instances of these types are observed, say at $\lambda = \lambda(p_i, p_j)$, the contact probability $p_{ij}^{(k)}$ is set as $p_{ij}^{(k)} = 1 - \lambda(p_i, p_j)/\lambda_{\max}$; if the two types make strictly less than k contacts, then $p_{ij}^{(k)} = 0$. For a given probability b, the set of *k-significant contacts* $S_b^{(k)}$ is the set of contacts such that $p_{ij}^{(k)} \geq b$ and $p_{ij}^{(k+1)} < b$.

Global assessment w.r.t. a collection of types: stoichiometry, symmetry, stability. Assume that the red proteins are instances of types prescribed in a set T, called a *pullout*. The following parameters can be assessed.

- *Stoichiometry.* Analyzing the complexes of the Hasse diagram has several interests: first, one sees whether the set T corresponds to a single complex or to a mixture of complexes; second, one can spot the copies associated to the set T—see section 2.1; third, if T corresponds to a TAP experiment, one can check whether each complex contains the tagged protein.
- *Symmetry.* For an assembly with symmetries, one can compare the number of complexes with the expected number. For example, in the NPC, the multiplicity of selected complexes is expected to be 16.
- *Topological stability.* In section 2.1, the stability of a complex has been defined as the difference between its birth and death dates. This information is particularly relevant to know when a given complex collides with another one to form a larger complex. For an assembly involving a prescribed number of complexes, one expects the variation of the number of complexes as a function of λ to exhibit a plateau. Also, for an assembly with symmetries, the homogeneity of the model can be inferred from the stability of complexes featuring the same types, but located in different places.
- *Geometric accuracy.* A complex may involve the correct protein instances, but may have a loose geometry. Comparing its volume to that occupied by its constituting instances is the goal of the volume ratio of Eq. (2).

[2] In the bicolor setting, the volume of a red complex is defined from its constituting red restrictions in the CW Voronoi diagram. Practically, however, we add up the volumes of the restrictions in the power diagram, as explained in [CHL11].

Local assessment w.r.t. a 3D model: perfect and alternate matchings. Assume now that we wish to compare a complex C against a model T. Complex T may come from a crystal structure, or may have been designed in-silico. The geometric and topological assessments just presented naturally apply. However, a more precise picture is obtained using the tools presented in section 2.2, which allow a precise comparison between the two skeleton graphs, both in terms of common protein types and commons contacts.

3 Material and Methods

In this section, we briefly present the NPC and sub-systems of interest, and provide elementary algorithms to build toleranced models of this assembly, based on the probability density maps from [ADV$^+$07b].

3.1 Structure of the NPC and Sub-systems of Interest

NPC: overall structure. The NPC is the channel regulating the nucleo-cytoplasmic exchanges of eukaryotic cells. It is composed of eight symmetrical spokes, each of them divided in two symmetrical half spokes. A model of the NPC developed in [HSBH07] involves four functional concentric cylinders, namely (i) the channel cylinder, containing proteins having unstructured regions (filaments) regulating the active transport; (ii) the adapter cylinder, involving intermediate proteins between channel proteins and scaffold proteins; (iii) the coat cylinder, which defines the scaffold of the NPC; (iv) the pore membrane cylinder, anchoring the NPC into the nuclear membrane.

The Y-complex and related complexes. Each half spoke of the NPC restricted to the coat cylinder contains a heptamer called the Y-complex (Nup133, Nup84, Nup145C, Sec13, Nup120, Nup85 and Seh1). To describe two models of the NPC, illustrated on Fig. 2, we decompose it into sub-systems, namely [3] the Y_X-short-arm, the Y_X-long-arm, the Y_X-edge, the Y_X-tail, and the Y-arms, the Y-core, the Y-main, and Y-junction.

The model, from Blobel et al. [KB09], comes from a reconstruction involving single particle EM data, together with crystal structures of the Y_X-short-arm, the Y_X-long-arm, the Y_X-edge and Nup133. Using size-exclusion chromatography and analytical ultracentrifugation, the authors show that two opposite proteins of the Y-complex namely (Nup120, Nup133) interact in a head-to-tail fashion [SMD$^+$09], a contact motivating the embedding of copies of the Y-complex into the NPC in a ring-like fashion.

Using the same pairwise contacts together with those involved in the Y_X-tail, Brohawn et al. [BS09] propose an embedding of the Y-complex into the NPC where the Y_X-tail extremities point towards the cytoplasmic and nuclear hemispheres of their respective half-spokes. This proposal is motivated by homology considerations with coat vesicles and interactions with Nic96 sub-complexes. We shall investigate these models using the skeleton graphs of Fig. 2.

The T-complex and related complexes. The protein Nic96 is located in the adapter cylinder, and makes the T-*complex* complex with instances of Nsp1, Nup49 and Nup57. We split these proteins into the T-core i.e. (Nic-96, Nsp1) and the T-leg i.e. (Nup49, Nup57), see Fig. 3. Filaments of the latter three proteins are involved in the regulation of the traffic trough the NPC. We are not aware of any crystal structure of complexes involving two or more such proteins.

Contacts between proteins of the T-core were determined by purification experiments [GDH93]. Similarly, it has been shown that Nup57 binds Nsp1 and Nup49 independently [SHL$^+$97], which motivates the first skeleton graph $G_t(T)$ of the T-complex. A second skeleton graph $G_t(T\text{-comp})$ encodes all possible contacts. This model is warranted by in vitro binding assay experiments [SSF$^+$08] showing interactions between the filaments of the T-leg proteins with Nic96. Finally, the skeleton graph $G_t(T\text{-new})$ refers to $G_t(T\text{-comp})$ without the contact between Nup57 and Nic96. See Fig. 3.

3.2 Constructing Toleranced Models

The NPC model of [ADV$^+$07b] involves 30 types, whence 34 maps due to four duplicated types (Nup82, Nsp1, Nic96, Nup145N). The map of Gle1 being missing from `http://salilab.org/npc/`, we use the remaining 33 maps as input, for a total of 29 types. We build a toleranced model for each type and merge them to obtain a toleranced model of the whole NPC. (Note that tagging some types as red results in a bicolor toleranced model.) Processing a given map is a three-stage process.

[3] The subscript X in Y_X hints at a crystal structure, see section 7.2.

First, we allocate occupancy volumes to protein instances. This step consists of collecting voxels in such a way that the volume covered by these voxels matches the estimated volume of all instances, namely Vol_{ref} multiplied by the stoichiometry. These voxels are collected by a greedy region growing strategy, as explained in [ADV$^+$07b, Caption of Fig.9, page 691]. Second, we compute a canonical representation involving 18 toleranced balls for each instance. (In [ADV$^+$07b], at most 12 balls are used to represent a protein instance.) Consider an occupancy volume to be covered with 18 toleranced balls of identical radius. Using a principal components analysis, each volume is assigned one of the four canonical arrangements of Fig. 4, which correspond to shapes which are roughly isotropic, flat, semi-linear and linear. Finally, we set the inner and outer radii. For a given protein type, the inner radius is set so that the volume of the union of the 18 inner balls matches the estimated volume of the protein Vol_{ref}. Since the probability density maps of large proteins tend to be more accurate than those of small proteins, see the supplemental Fig. 9, we set the outer radius such that the discrepancy $r_i^+ - r_i^-$ is proportional to α/r_i^-:

$$r_i^+ = \frac{\alpha}{r_i^-} + r_i^-. \tag{3}$$

This formula entails that the Hasse diagram representing the evolution of skeleton graphs depends only on the inner radii $\{r_i^-\}$, but not on the parameter α. We arbitrarily set $\alpha = 10$ and compute the whole λ-complex of the toleranced model. To examine models with decent geometric accuracy, we stop when the volume ratio deteriorates, namely at $\lambda = \lambda_{\max}$ with $r_{\lambda_{\max}} \sim 5$. The reader is referred to the supplemental sections 7.1 and 7.2 for the mathematical details and for the assessment of toleranced models w.r.t. known crystal structures.

4 Results

In this section, we apply the global and local analysis of section 2.4 to the Y-complex and the T-complex. We aim at bridging the gap between copies of these complexes embedded in the toleranced model of the NPC, and those obtained by crystallography and modeling.

4.1 Contact Analysis

Contact frequencies versus contact probabilities. For $k = 1$, we compare the probability $p_{ij}^{(1)}$ to the contact frequency f_{ij}. As discussed in section 3.2, there are 29 protein types and 435 possible contacts, including homotipic ones. Using the two probabilities $0 \leq a < b \leq 1$, the 435 contact frequencies f_{ij} are sorted into three classes in [ADV$^+$07b]: $F_1 : f_{ij} \leq a$; $F_2 : a < f_{ij} < b$; $F_3 : b \leq f_{ij}$. Similarly, we segregate the contacts observed from the Hasse diagram into the three classes $P_1^{(1)} : p_{ij}^{(1)} \leq a; P_2^{(1)} : a < p_{ij}^{(1)} < b; P_3^{(1)} : b \leq p_{ij}^{(1)}$.

For $a = 0.25$, $b = 0.65$ and $\lambda_{\max} = 1$, the sizes of the classes are $\mid F_1 \mid = 325, \mid F_2 \mid = 79$ and $\mid F_3 \mid = 31$. Moreover, 93.5% of the contacts in F_3 belong to $P_3^{(1)}$, and 60.5% of the contacts in F_1 belong to $P_1^{(1)}$. The contact probability is more discriminative than the contact frequency since the maximum number of contacts in $P_2^{(1)}$ and F_2 are respectively of 53 and 79. For the more stringent values $a = 0.1$, $b = 0.9$ and $\lambda_{\max} = 1$, one has $\mid F_1 \mid = 220, \mid F_2 \mid = 196$ and $\mid F_3 \mid = 19$ contacts. Then, 79% of contacts in F_3 belong to $P_3^{(1)}$, 73% of contacts in F_1 belong to $P_1^{(1)}$, while the maximum number of contacts in $P_2^{(1)}$ is 95—to be compared to 196 contacts in F_2. See the supplemental Figs. 15. and 14.

We also use the values $a = 0.1$, $b = 0.9$ to report mismatches. A pair of types belonging to F_1 but $P_3^{(1)}$ is called *over-represented* in the toleranced model. The supplemental Table 2 lists the 23 over-represented pairs. Note that all these contacts are over-represented for $\lambda_{\max} \geq 0.21$, which clearly indicates that the corresponding contacts appear early in the growth process. Similarly, a pair belonging to F_3 but $P_1^{(1)}$ is termed *under-represented* in the toleranced model. The supplemental Table 3 lists such cases, which are under-represented for $\lambda_{\max} \leq 0.28$. The illustrations presented on the supplemental Figs. 16. and 17 clearly support our contact probability.

On k-significant contacts. To leverage the previous information, we now focus on pairs of types making up a prescribed number of contacts. For $\lambda_{\max} = 1$, we observe 183 contacts (over 435) in $S_{0.65}^{(k\geq 1)}$, but only 36 in $S_{0.65}^{(k\geq 16)}$. Building the graph whose edges correspond to types displaying at least 11 contacts uncovers sub-complexes of the NPC, including the Y-complex and T-complex. See the supplemental Fig. 6.

4.2 Y-complex Analysis

Stoichiometry, symmetry, stability. The evolution of complexes involving the seven types of the Y-complex is provided by the Hasse diagram on Fig. 5 (Top-Left). Out of sixteen expected copies of the Y-complex, eight are observed in the range $\lambda = 0$ ($r_\lambda = 0.86$) and $\lambda = 0.28$ ($r_\lambda = 2.14$). These correspond to the fat nodes on the Hasse diagram, one of them being singled out on Fig. 5 (Bottom-Left). The stability of these eight complexes is heterogeneous as their lifetimes span the range $s(C) = 0.02$ ($\Delta r_\lambda = 0.06$) and $s(C) = 0.46$ ($\Delta r_\lambda = 2.47$). Also, they do not coexists since the intersection of their lifetime intervals is empty. These observations show that contacts between protein instances belonging to several copies of the Y-complex can prevail over contacts within the isolated copies. The curve featuring the evolution of the number of complexes, see Fig. 5 (Middle-Left), does not exhibit any plateau, but its slope decreases beyond $\lambda = 0.31$ ($r_\lambda = 2.21$), and all complexes finally merge into two complexes defining the inner and outer ring of the NPC at $\lambda = 0.68$ ($r_\lambda = 3.67$). These two complexes do not merge before $\lambda_{\max}$, as collisions between their proteins are hindered by instances of the remaining 23 protein types. These two stable rings hint at correct head-to-tail contacts in the model of Blobel et al.

Perfect matchings. Perfect matchings reflect the largest sub-complexes of the Y-complex without any missing or extra contact w.r.t. the model—see the rows tagged with $G_t(Y)$ in the supplemental Table 8.

We first classify the matchings following their similarity with known sub-complexes of the Y-complex: 16 copies of the Y_X-tail (P_1, P_2), 12 of the Y_X-edge (five split in two subunits (P_3, P_4); seven entire units cf P_5), 12 of the Y_X-short-arm (P_6, P_7) and 16 of the Y_X-long-arm (P_8). In addition, we have four perfect matchings corresponding to the Y-core (P_9). The four remaining perfect matchings (P_{10}) have one matching protein type i.e. Sec13. Let us inspect these perfect matchings.

The sixteen perfect matchings involving the Y_X-tail (P_1, P_2) show that Nup133 and Nup84 are well positioned w.r.t. one another. But the absence of larger perfect matching shows that in the fourteen complexes under investigation in P_1, Nup133 makes an erroneous contact with Nup145C, a fact clearly related to its elongated shape – the canonical configurations of the different instances of Nup133 are either linear or semi-linear.
The 12 perfect matchings involving parts of the Y_X-edge show that there is an additional contact between Sec13 and Nup84 in the model for at least five copies of the Y_X-edge (P_3, P_4).
The contact between Y_X-short-arm and Nup85 appears in 16 perfect matchings (P_6, P_7, P_9), while the one between Y_X-short-arm and Nup145C appears in only five matchings (P_7, P_9). As illustrated by the supplemental Fig. 11, each copy of the Y-complex are split into two pieces. The 16 perfect matchings of the Y_X-long-arm (P_8) show a good relative position between Seh1 and Nup85. The same holds for the four protein of Y-core (P_9), as evidenced by the four matchings.
Finally, the matchings involving Sec13 involve protein instances without any valid contact. Their positioning appears as uncertain, a fact likely related to the small size of Sec13. (With a molecular weight of 33 kDa, Sec13 is the smallest one of the NPC.) As a matter of fact, no available data for the position of Sec13 is found in [ADV+07b, supplemental Table 7]. Interestingly, the volume ratios associated with these results are bounded by 2.57 (P_9).

Alternate matchings. Alternate matchings aim at maximizing the number of common contacts, and involve largest sub-complexes of the Y-complex. We first computed matchings with $G_t(Y)$, see the supplemental Table 9. We get 11 for (Y_X-tail,Y_X-edge) (A_1, A_2), and 11 for (Y_X-short-arm,Y_X-long-arm) (A_3). These matchings correspond to eleven copies of the Y-complex split in two sub-complexes. Together with the five matchings for the whole Y-complex (A_4, A_5), we get an overall stoichiometry of 16, as expected.

The number of extra contacts observed is bounded by seven for a maximum of fifteen. (Seven proteins make at most twenty one pairwise contacts, out of which six belong to the template.) Note that the only missing protein type in the five matchings involving Y-main is restricted to Sec13 in three of them. Note also that the volume ratio does not exceed 4.71 (A_3).

Further in-silico experiments. Sec13 having a poor relative position with respect to the proteins of the Y-complex, we investigated its relative position w.r.t. proteins of the Y_X-edge i.e. Nup84 and Nup145C. As shown in the supplemental section 7.6.1, the global analysis reveals 13 isolated copies of the Y_X-edge, thus witnessing a correct positioning of Sec13 w.r.t. to Nup84 and Nup145C.

To get results on the Y-complex not affected by the poor global positioning of Sec13, we removed Sec13 from the toleranced model and repeated the global and local analysis w.r.t. the template skeleton $G_t(Y)$ for the six remaining types of the Y-complex, see the supplemental section 7.6.2. The evolution of the number of

complexes reveals a plateau for 14 complexes, and the corresponding Hasse diagram singles out nine isolated copies of the Y-complex instead of eight with Sec13. No gain is observed for perfect and alternate matchings, though, a fact related to erroneous relative positions. See the supplemental Fig. 20 and the supplemental Tables 10 and 11.

Finally, having removed Sec13 from the toleranced model, we investigate the closure of the two rings involving the 16 copies of the Y-complex. Having grouped from the Hasse diagram the instances of the six protein types in 16 sets, we report—also from the Hasse diagram—the first event connecting two proteins of two distinct copies. These 16 events involve: nine contacts (Nup133, Nup85) appearing in between $\lambda = 0.09$ and $\lambda = 0.70$; five contacts (Nup84, Nup85) appearing in between $\lambda = 0.52$ and $\lambda = 0.69$; one contact (Nup133, Nup120) appearing at $\lambda = 0$; one contact (Nup84, Nup120) appearing at $\lambda = 0.06$. Note that Nup85 is represented in 14 of the 16 contacts. Furthermore the volume ratios associated to the λ values are smaller than 3.67 showing the consistency of the contacts. See the supplemental Fig. 5.

4.3 T-complex Analysis

As opposed to the Y-complex, no (sub-)complex of the T-complex has been crystallized. In the following, we therefore investigate the coherence between putative pairwise contacts in the T-complex, and the copies of the T-complex embedded in the toleranced model of the NPC.

Stoichiometry, symmetry, stability. The Hasse diagram at Fig. 5 (Top-Right) shows that the 16 copies—the expected number—of the T-complex get formed thanks to merges in-between $\lambda = 0$ ($r_\lambda = 0.72$) and $\lambda = 0.15$ ($r_\lambda = 1.24$). Their lifetimes are rather homogeneous since they vary in-between $s(C) = 0.10$ ($\Delta r_\lambda = 0$) and $s(C) = 0.33$ ($\Delta r_\lambda = 1.29$), and the copies coexist in-between $\lambda = 0.15$ and $\lambda = 0.22$. These results show that contacts inside a copy of the T-complex prevail over contacts between different copies of the T-complex. As shown on the red curve on Fig. 5 (Middle-Right), these copies are stable until $\lambda = 0.23$ ($r_\lambda = 1.68$). After this plateau, the number of complexes drops to eight at $\lambda = 0.33$ ($r_\lambda = 1.90$), and these do not merge before $\lambda_{\max}$. Thus, merges of copies of the T-complex within a spoke are privileged over merges across spokes.

Perfect matchings. As summarized in the supplemental Table 8, we computed the perfect matchings w.r.t. the skeleton graph $G_t(T)$ for all nodes of the Hasse diagram. One gets sixteen perfect matchings corresponding to the T-leg (P_{11}, P_{12}), fourteen to the T-core (P_{13}) and two to the entire T-complex (P_{14}). A careful inspection shows that all contacts of $G_t(T)$ are found in all copies of the T-complex. The low number of perfect matchings (2) owes to extra contacts, namely (Nup49 and Nsp1) and/or (Nup49 and Nic96) and/or (Nup57 and Nic96). Yet, we found Nup57 and Nic96 in 16 different perfect matchings (P_{13}, P_{14}), showing that these two types do not make any contact.

Alternate matchings. As seen from the supplemental Table 9, 18 alternate matchings of the entire T-complex are found (A_6), for a volume ratio less than 2.36. Further investigation shows two instances of Nup49, each interacting with two instances of Nup57 belonging to two copies of the T-complex, contribute to two extra matchings—whence 18 matchings and not 16. These spurious matching tough, are easily ruled out from the λ value for which contacts between Nup49 and Nup57 appear, since the second contact appears at $\lambda = 0.38$, after the last merge of complexes at $\lambda = 0.33$. The analysis of alternate matchings also exhibits at most two extra contacts between Nup49 and Nsp1, and between Nup49 and Nic96.

Further in-silico experiments. To single out frequent contacts not present in $G_t(T)$, we consider the complete skeleton graph $G_t(T\text{-comp})$. We obtain 18 perfect matchings corresponding to the T-leg (P_{15}, P_{16}) and 16 to the T-core (P_{17}, P_{18}, P_{19}). These matchings highlight two relevant contacts absent from the skeleton $G_t(T)$: Nup49 and Nsp1 obtained 16 times (P_{16}), and Nup49 and Nic96 obtained ten times (P_{18}). Adding these contacts to the skeleton graph $G_t(T)$ yields $G_t(T\text{-new})$. This new graph yields eight perfect matchings with T-leg (P_{20}, P_{21}), six to the T-core (P_{22}, P_{23}) and ten to the entire T-complex (P_{23}). We note that the number of perfect matchings with the entire T-complex node set moves from two for $G_t(T)$ to ten for $G_t(T\text{-new})$.

In terms of alternate matchings, $G_t(T\text{-new})$ yields 22 alternate matchings, containing in particular the ten perfect matchings already discussed—the matchings counted in the lines P_{23} and A_{10} are in one-to-one correspondence. The extra 12 matchings owe again to contacts between protein instances of different copies of the T-complex. These extra matchings involve at most two missing contacts corresponding to the contacts

added w.r.t. $G_t(T)$. Also, these matchings do not have any extra contact, except one corresponding to the a contact between Nup57 and Nic96, see line A_{11}.

5 Discussion and Outlook

Significant pairwise contacts. The stoichiometry-dependent contact probability defined from the toleranced model singles out prominent interactions and sheds light on the structure of the NPC. For a unit stoichiometry, this probability sharpens the contact frequency of Alber et al. While 63.4% of the pairs of the low and high contact frequency classes are found in the equivalent contact frequency classes, our probability is more discriminative and also exhibits a better coherence with the density maps. Also, focusing on pairs of types making a prescribed set of contacts results in a graph revealing sub-complexes of the NPC, including the Y-complex and the T-complex.

Y-complex and structure of the NPC. Our analysis highlights a good coherence between the global fuzzy models of Alber et al., and the local high-resolution models of Blobel et al., in particular regarding the two stable rings which support the *head-to-tail* embedding of the Y-complex. These rings come from the merge of the proteins involved in 16 copies of the Y-complex, eight of them being singled out by the Hasse diagram of the Y-complex. The remaining eight copies do not appear isolated, as contacts across copies prevail on contacts within the isolated copies. In any case, the merge events between the copies of the Y-complex involve Nup133, Nup85, Nup84 and Nup120, with a prominent role of Nup85.

Contact-wise, the coherence between the complexes observed and the skeletons of the atomic models (obtained by crystallography and modeling) is medium, as complexes of the toleranced model are often split. This in turn owes to the accuracy of the probability density maps. We note in passing that testing the *fence-like* organization of the proteins of the Y-complex [DMS$^+$08] is not possible, as this model requires 32 copies of each protein while the maps of Alber et al contain 16 of them.

A new model for the T-complex. The 16 stable copies of the T-complex evidenced by our toleranced model support current knowledge. Also, the fact that these copies eventually merge into eight complexes while exploring the toleranced model shows that copies of the T-complex are well separated across spokes, yet the two copies within a spoke lie next to one another. Our local analysis shows that all the contacts in the model of Schlaich et al [SHL$^+$97] are supported by our toleranced model. The same holds for all the contacts of the model of Vetter et al. [SSF$^+$08], but the contact between Nic96 and Nup57. In fact, the contacts between Nic96 and Nup57 and that between Nic96 and Nup49 deserve two comments. First, while both are supposed to involve unstructured filaments [SSF$^+$08], the second one only is supported by the toleranced model. Second, the contact frequencies in the analysis of Alber et al. [ADV$^+$07b, Supplemental, page 75, pullout #28] are of 0.42 for Nic96 and Nup57, and of 0.44 for Nic96 and Nup49, which again contrasts with the asymmetry observed in the toleranced model.

Methodology. Toleranced models allow a multi-scale investigation of the complexes involving a prescribed set of proteins, at the global and local levels. At the assembly level, they allow the detection discrepancies in-between regions of the assembly, which is especially useful in the presence of symmetries. At the local level, they allow the comparison of the toleranced model with a 3D template known or modeled at atomic resolution, so as to confirm / question / suggest protein contacts of the template based on the prominent contacts seen in the toleranced model. In the context of assembly reconstruction by data integration, these analysis will prove especially useful to qualify the completeness of the experimental data and their coherence, and to perform model selection. We use these analysis to hint at missing or ill-placed proteins.

May be most importantly, toleranced models allow testing hypothesis, by repainting or removing a given protein type, and permit investigating the interactions between any protein types—not just those obtained from TAP experiments. In particular, the analysis presented in this work for the Y-complex and the T-complex can be replicated at essentially no cost for any subset of protein type, which should provide incentives to run new experiments.

On a more general perspective, the results obtained in this work are based on toleranced models using elementary canonical representations of proteins. The design of more elaborate toleranced models will sharpen the statistics obtained. Also, toleranced models should prove useful in different contexts where uncertainties are faced, such as soft docking, handling anisotropic temperature factors in crystallography, or the processing of cryo EM maps.

References

[ADV+07a] Frank Alber, Svetlana Dokudovskaya, Liesbeth M. Veenhoff, Wenzhu Zhang, Julia Kipper, Damien Devos, Adisetyantari Suprapto, Orit Karni-Schmidt, Rosemary Williams, Brian T. Chait, Andrej Sali, and Michael P. Rout. The molecular architecture of the nuclear pore complex. *Nature*, 450(7170):695–701, Nov 2007.

[ADV+07b] Frank Alber, Svetlana Dokudovskaya, Liesbeth M. Veenhoff, Wenzhu Zhang, Julia Kipper, Damien Devos, Adisetyantari Suprapto, Orit Karni-Schmidt, Rosemary Williams, Brian T. Chait, and Michael P. Rout Andrej Sali. Determining the architectures of macromolecular assemblies. *Nature*, 450(7170):683–694, Nov 2007.

[AFK+08] Frank Alber, Friedrich Förster, Dmitry Korkin, Maya Topf, and Andrej Sali. Integrating diverse data for structure determination of macromolecular assemblies. *Ann. Rev. Biochem.*, 77:11.1–11.35, 2008.

[BLS+08] Stephen G. Brohawn, Nina C. Leksa, Eric D. Spear, Kanagalaghatta R. Rajashankar, and Thomas U. Schwartz. Structural evidence for common ancestry of the nuclear pore complex and vesicle coats. *Science*, 322:1369–1373, 2008.

[BS09] Stephen G. Brohawn and Thomas U. Schwartz. Molecular architecture of the Nup84–Nup145C–Sec13 edge element in the nuclear pore complex lattice. *Nature Structural & Molecular Biology*, pages 1173–1178, 2009.

[CCS11] Frédéric Cazals and David Cohen-Steiner. Reconstructing 3d compact sets. *Computational Geometry Theory and Applications*, 2011. To appear; available from `ftp://ftp-sop.inria.fr/abs/fcazals/papers/flowRecons1.pdf`.

[CD10] Frédéric Cazals and Tom Dreyfus. Multi-scale geometric modeling of ambiguous shapes with toleranced balls and compoundly weighted α-shapes. In B. Levy and O. Sorkine, editors, *Symposium on Geometry Processing*, Lyon, 2010. Also as INRIA Tech report 7306.

[CHL11] Frédéric Cazals, Kanhere Harshad, and Sebastien Loriot. Computing the volume of union of balls: a certified algorithm. *ACM Transactions on Mathematical Software*, 2011. To appear. Available as INRIA Tech report 7013.

[CK05] Frédéric Cazals and Chinmay Karande. An algorithm for reporting maximal c-cliques. *Theoretical Computer Science*, 349(3):484–490, 2005.

[CSEH05] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. In *ACM Symp. Comp. Geometry*, 2005.

[DMS+08] Erik W. Debler, Yingli Ma, Hyuk-Soo Seo, Kuo-Chiang Hsia, Thomas R. Noriega, Günter Blobel, and André Hoelz. A fence-like coat for the nuclear pore membrane. *Molecular Cell*, 32:815–826, 2008.

[Ede92] Herbert Edelsbrunner. Weighted alpha shapes. Technical Report UIUCDCS-R-92-1760, Dept. Comput. Sci., Univ. Illinois, Urbana, IL, 1992.

[GDH93] Paola Grandi, Valerie Doye, and Eduard C. Hurt. Purification of NSP1 reveals complex formation with'GLFG'nucleoporins and a novel nuclear pore protein NIC96. *The EMBO Journal*, 12(8):3061, 1993.

[GJ03] Joachim Giesen and Matthias John. The flow complex: A data structure for geometric modeling. In *ACM SODA*, 2003.

[HGC94] Yehouda Harpaz, Mark Gerstein, and Cyrus Chothia. Volume changes on protein folding. *Structure*, 2:641–649, 1994.

[HSBH07] Kuo-Chiang Hsia, Pete Stavropoulos, Günter Blobel, and André Hoelz. Architecture of a coat for the nuclear pore membrane. *Cell*, 131(7):1313–1326, 2007.

[JS07] Sandra Jeudy and Thomas U. Schwartz. Crystal structure of nucleoporin Nic96 reveals a novel, intricate helical domain architecture. *J. Biol. Chem.*, 282(1):34904–34912, 2007.

[KB09] Marti Kampmann and Günter Blobel. Three-dimensional structure and flexibility of a membrane-coating module of the nuclear pore complex. *Nature structural & molecular biology*, 16(7):782–788, 2009.

[Koc01] Ina Koch. Fundamental study: Enumerating all connected maximal common subgraphs in two graphs. *Theoretical Comp. Sc.*, 250(1-2):1–30, 2001.

[LBS09] Nina C. Leksa, Stephen G. Brohawn, and Thomas U. Schwartz. The structure of the scaffold nucleoporin nup120 reveals a new and unexpected domain architecture. *Structure*, 17:1082–1091, 2009.

[NHD+09] Vivien Nagy, Kuo-Chiang Hsia, Eric W. Debler, Marti Kampmann, Andrew M. Davenport, Günter Blobel, and André Hoelz. Structure of a trimeric nucleoporin complex reveals alternate oligomerization states. *Proceedings of the National Academy of Sciences*, 106(42):17693, 2009.

[OBSC00] Atsuyuki Okabe, Barry Boots, Kokochi Sugihara, and Sung N. Chiu. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams (2nd Ed.)*. Wiley, 2000.

[SHL+97] Nikolaus L. Schlaich, Markus Haner, Ariel Lustig, Ueli Aebi, and Eduard C. Hurt. In vitro reconstitution of a heterotrimeric nucleoporin complex consisting of recombinant Nsp1p, Nup49p, and Nup57p. *Molecular biology of the cell*, 8(1):33–46, 1997.

[SMD+04] Hyuk-Soo Seo, Yingli Ma, Erik W. Debler, Daniel Wacker, Stephan Kutik, Günter Blobel, and André Hoelz. Structural and functional analysis of nup133 domains reveals modular building blocks of the nuclear pore complex. *J. Cell Biol.*, 167:591–597, 2004.

[SMD+09] Hyuk-Soo Seo, Yingli Ma, Erik W. Debler, Daniel Wacker, Stephan Kutik, Günter Blobel, and André Hoelz. Structural and functional analysis of nup120 suggests ring formation of the nup84 complex. *PNAS*, pages 14281–14286, 2009.

[SSF+08] Nils Schrader, Philipp Stelter, Dirk Flemming, Ruth Kunze, Eduard Hurt, and Ingrid R. Vetter. Structural basis of the nic96 subcomplex organization in the nuclear pore channel. *Molecular cell*, 29(1):46–55, 2008.

[WS09] James R. R. Whittle and Thomas U. Schwartz. Architectural nucleoporins nup157/170 and nup133 are structurally related and descend from a second ancestral element. *J. Biol. Chem.*, 284:28442–28452, 2009.

6 Artwork

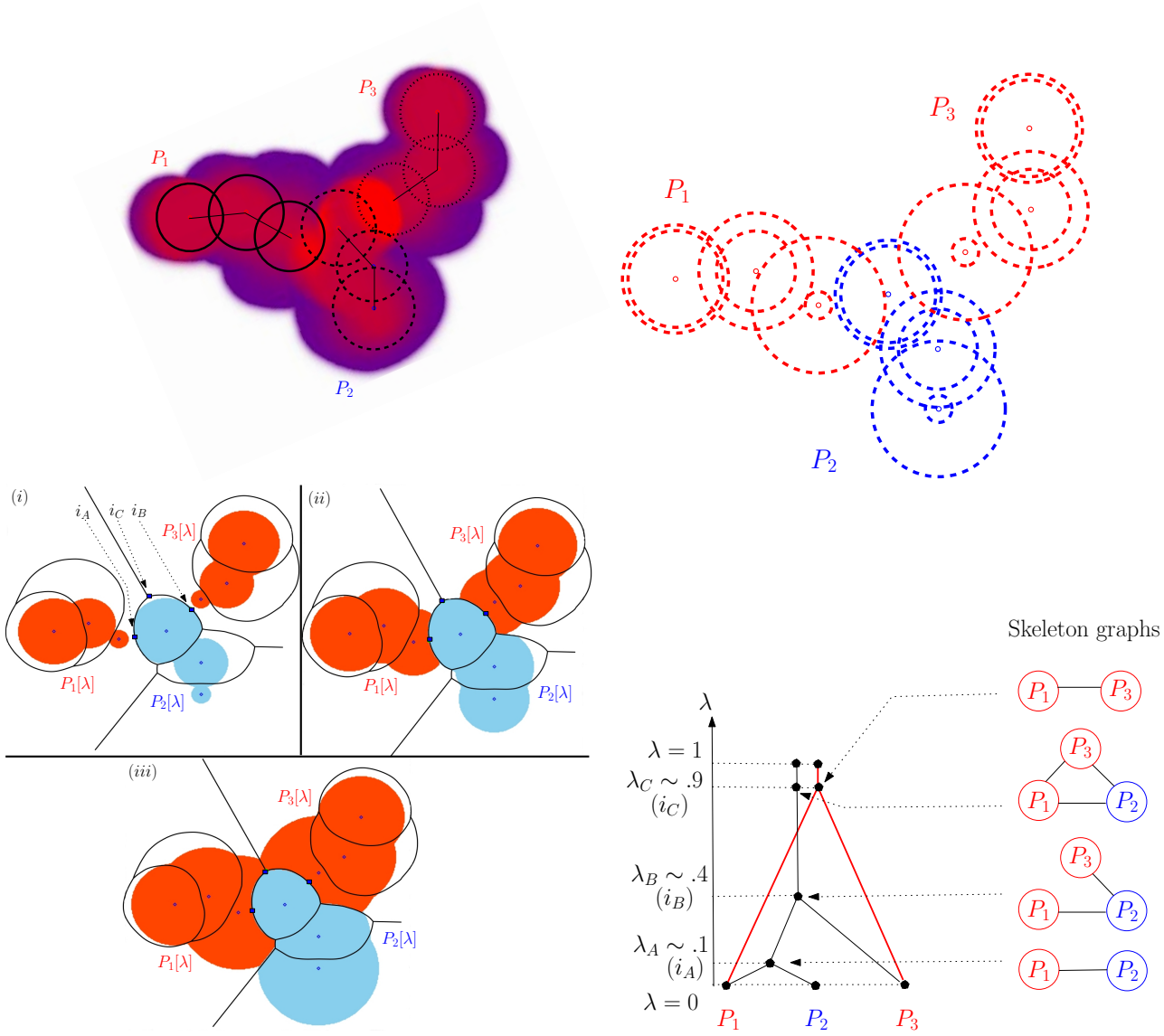


Figure 1: Tracking the interactions of three toleranced proteins of three toleranced balls each. **(Top left)** Three conformations of three flexible molecules, and a probability density map whose color indicates the probability of a given point to be covered by a random conformation of the ternary complex — from low (blue pixels) to high (red pixels) probabilities. **(Top right)** The associated bicolor toleranced model, with one blue and two red molecules. Each toleranced molecule consists of a set of pairs of concentric balls, the inner and outer balls. **(Bottom left)** Sub-figures (i,ii,iii) respectively show grown balls $\overline{B_i}[\lambda]$ for $\lambda = 0, 0.5, 1$. The region of the plane consisting of points first reached by a growing toleranced ball is the Voronoi region of this ball, represented by solid lines. Colored solid regions feature the *restrictions* i.e. the intersection of a growing ball and its Voronoi region. Along the growth process, the restrictions intersect in three points i_A, i_B, i_C, represented as blue squares. **(Bottom right)** Hasse diagrams encoding contacts between the protein instances. Black tree: all instances; red tree: red instances only.

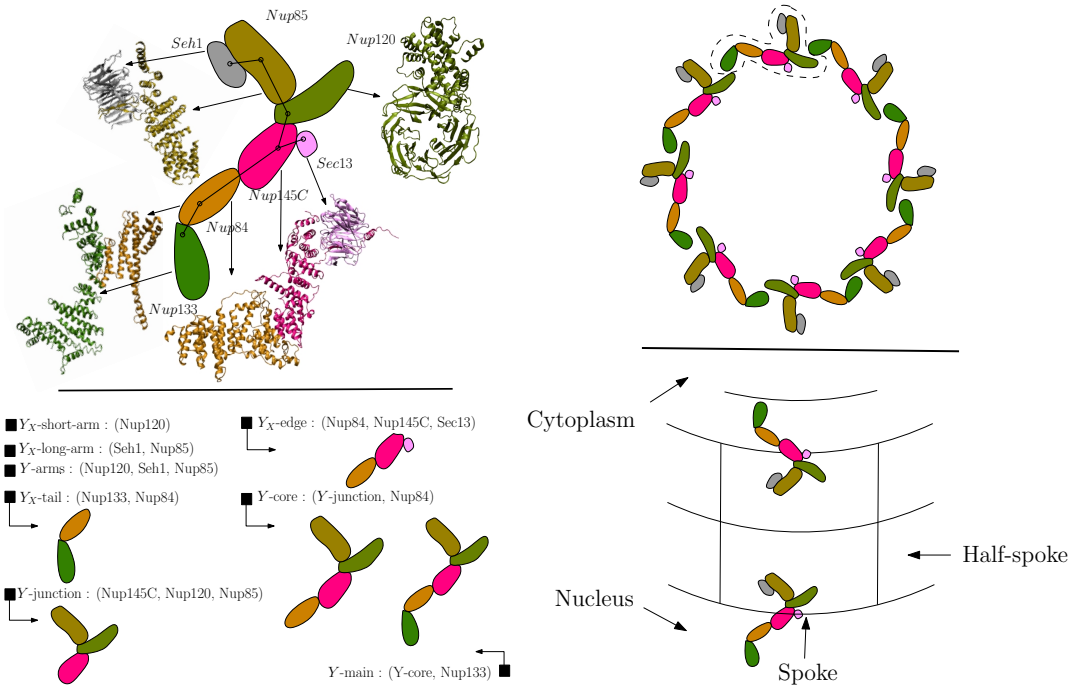


Figure 2: (**Top Left.**) Putative Y-complex model after [KB09]. It is composed of 7 proteins: Nup133 (light green), Nup84 (tan), Nup145C (red), Sec13 (pink), Nup120 (dark green), Nup85 (yellow) and Seh1 (gray). The skeleton graph $G_t(Y)$ of the Y-complex is represented in black solid lines. **(Bottom Left.)** Terminology used for sub-complexes of the Y-complex. **(Top right.)** Putative arrangement of Y-complexes, from [KB09]. Interactions between Nup133 and Nup120 account for one ring of head-to-tail Y-complexes in the cytoplasmic and nuclear hemispheres. **(Bottom right.)** Putative arrangement of Y-complexes in a spoke [BS09]. Each spoke of the NPC contains two Y-complexes with Y_X-tail pointing towards the cytoplasmic and nuclear hemispheres.

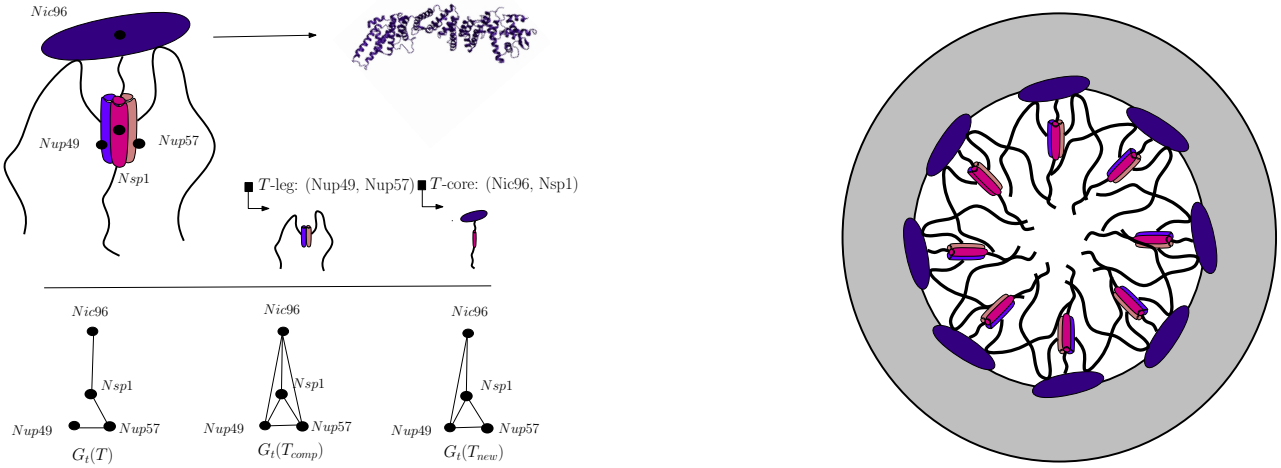


Figure 3: **(Top Left.)** The T-complex consists of Nic96 (dark blue), Nsp1 (magenta), Nup49 (light blue) and Nup57 (apricot). Filaments are non structured domains of Nsp1, Nup49 and Nup57. **(Bottom Left.)** The skeleton graphs for the T-complex. **(Right.)** The putative location of instances of the T-complex in the inner rim of the NPC.

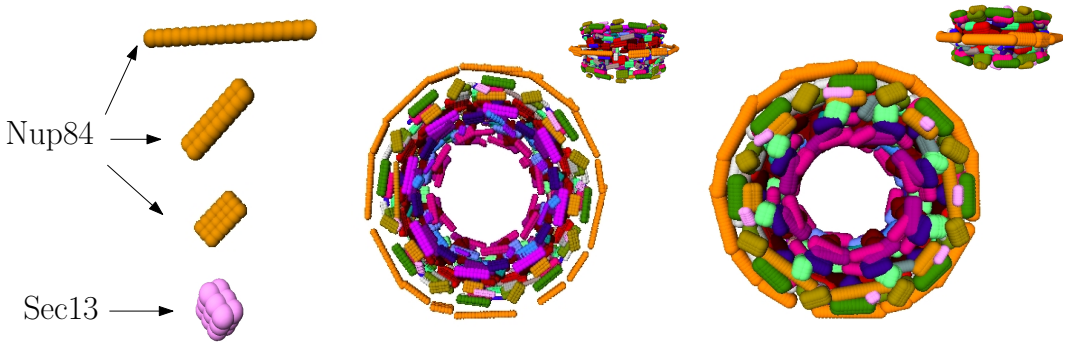
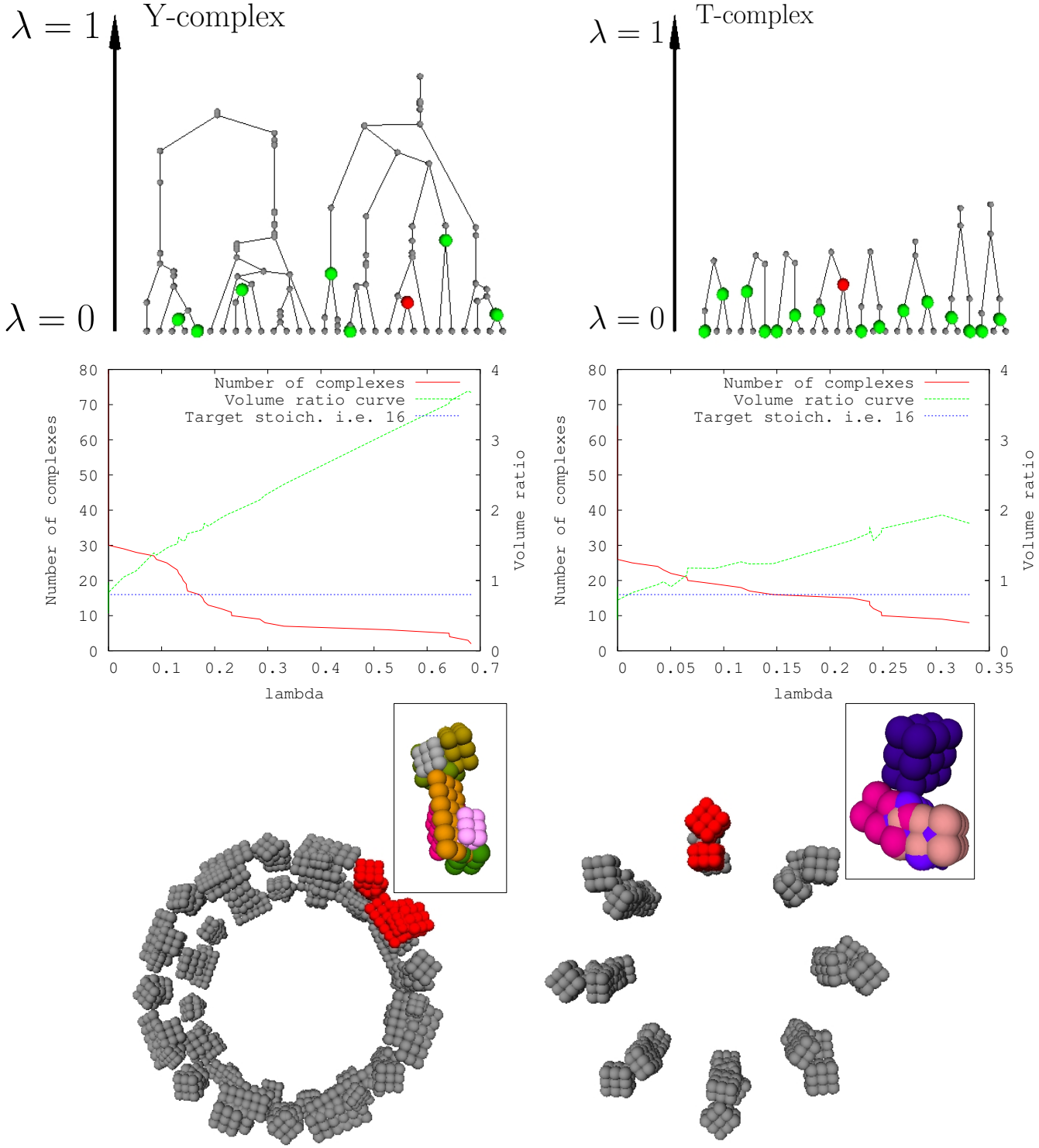


Figure 4: Toleranced Model of the whole NPC. **(Left.)** The four canonical configurations, 18 balls each, illustrated with protein types Nup84 and Sec13. **(Middle / Right.)** Views of the inner balls (middle, $\lambda = 0$), and outer balls (right, $\lambda = 1$).

Figure 5: Global assessment for the Y-complex (**Left column**) and the T-complex (**Right column**). **(Top row.)** The Hasse diagram representing the evolution of the connected components. Fat nodes correspond to isolated copies. **(Middle row.)** Evolution of the number of complexes and volume ratio r_λ as a function of λ. **(Bottom row.)** The complex corresponding to the red fat node of the Hasse diagram (main caption), with the protein instances highlighted with the color code of Fig. 2 (inset).



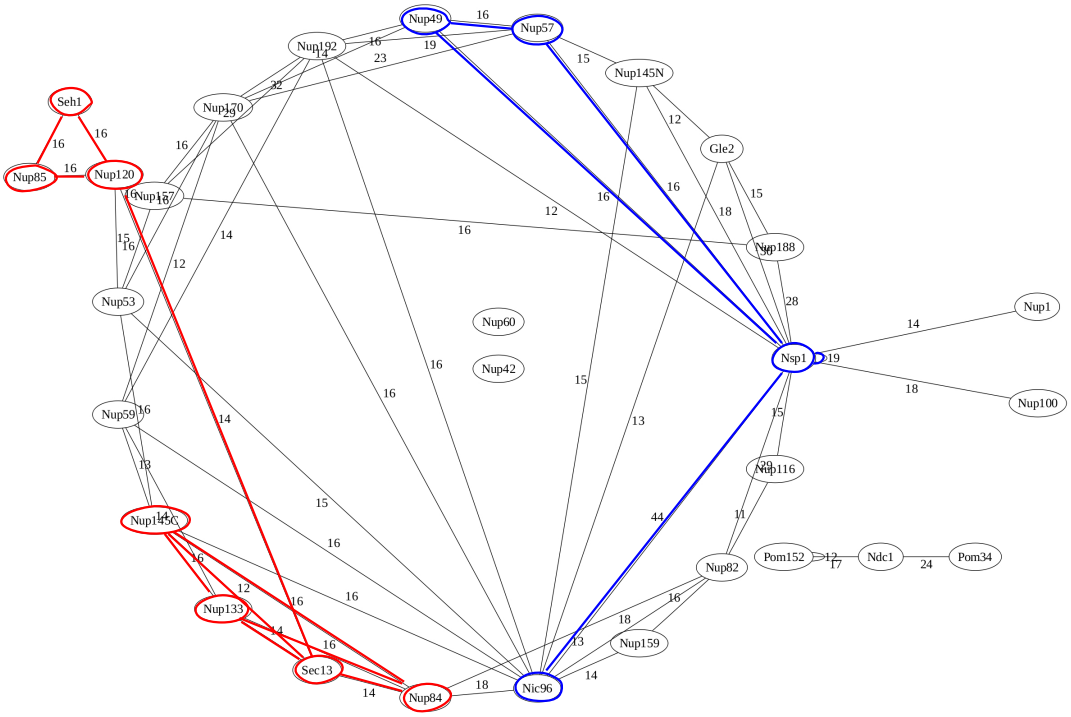


Figure 6: Graph of contacts in $S_{0.65}^{(k>10)}$ for $\lambda_{\max} = 1$—see section 2.4. The red and blue sub-graphs respectively correspond to the Y-complex and T-complex.

7 Supplemental

7.1 Curved α-shapes and Toleranced Models

The toleranced models introduced in section 2.1 use α-shapes associated with a compoundly weighted Voronoi diagram. The matter is actually quite subtle, and the following remarks are meant to intuitively clarify the meaning of the pairs of toleranced balls considered in the skeleton graphs. The reader is referred to [CD10] for the precise mathematical statements.

Toleranced balls: inner and outer radii versus interpolation and extrapolation. Intuitively speaking, toleranced models are best described in terms of inner and outer balls, the elementary geometric operation consisting of interpolating the radius between these radii. But the radius can also be extrapolated on both ends. In fact, in the theory of curved Voronoi diagrams [OBSC00], a toleranced ball whose radius is interpolated as

$$r_i(\lambda) = r_i^- + \lambda(r_i^+ - r_i^-) = || \, c_i p \, ||, \tag{4}$$

yields the following compoundly-weighted distance, see [CD10]:

$$\lambda(B_i, p) = \frac{1}{r_i^+ - r_i^-}(|| \, c_i p \, || - r_i^-). \tag{5}$$

Phrased differently, interpolating or extrapolating the radius is equivalent to varying the generalized distance of Eq. (5).

Intersecting balls versus intersecting restrictions. The pairs of balls reported correspond to balls whose restrictions intersect, and these pairs form a subset of all pairs of intersecting balls. (Recall that a restriction is the intersection between a ball and its Voronoi region.) In line with this comment, if ones keeps growing balls all the way to $\lambda = \infty$, one does not end up with all pairs of intersecting balls, but the pairs giving rise to a bisector in the Voronoi diagram defined by the balls. For a large enough value of λ, the pairs obtained are the abstract simplices of the dual complex of the Voronoi diagram.

Curved α-shapes. In the bicolor setting, resorting to curved α-shapes to identify complexes of a given color is actually compulsory. To see why, term the intersection point p between the red spheres bounding two balls $\overline{B_i}[\lambda]$ and $\overline{B_j}[\lambda]$ *pure* if p is not contained within a blue ball. Denoting $\lambda(B_i, p)$ the compoundly-weighted distance between a point p and the ball $\overline{B_i}$, pure intersections correspond to privileged binary contacts in the following sense: a pure contact point p is such that $\lambda(B_i, p) = \lambda(B_j, p) \leq \lambda(B_k, p), \forall k \neq i, j$. Contacts between restrictions retrieved from the curved α-complex have this property. On the other hand, contacts directly read from pairs of intersecting balls may not be pure. For example, on Fig. 1, the first intersection point between $\overline{P_1}$ and $\overline{P_3}$ is contained within a blue ball.

Setting inner and outer radii. Equation (3) provides a parametrization of the outer radius as a function of λ and r_i^-. Consider a collection of toleranced balls whose outer radii are set this way, that is $\{\overline{S_i}(c_i; r_i^-; r_i^+ = \frac{\alpha}{r_i^-} + r_i^-)\}$. Under the assumption $r_i^+ = \alpha/r_i^- + r_i^-$, the equation (5) becomes

$$\lambda(B_i, p) = \frac{r_i^-}{\alpha}(|| \, c_i p \, || - r_i^-). \tag{6}$$

If one equates two such equations to define a Voronoi bisector, that is $\lambda(B_i, p) = \lambda(B_j, p)$, the α cancel out. Phrased differently, the CW VD of the toleranced balls does not depend on α.

7.2 Toleranced Models: Assessment

7.2.1 On the Density Maps Used

The quality of the toleranced models built in section 3.2 depends on the accuracy of the probability density maps used. In the following, on a per-density map basis, we report statistics aiming at qualifying these maps, in particular regarding the number of connected components (c.c.) of voxels having a non null probability, and the volume of these c.c. with respect to their expected volume. (Following the terminology introduced in section 2.3, the reference volume $Vol_{ref}(P)$ of a protein P is the volume estimated from its sequence [HGC94].) We also report statistics aiming at assessing the geometric accuracy of our toleranced models, based on volume comparison with known crystal structures.

On the number of connected components. Ideally, the number of c.c. of a map should match the stoichiometry of the corresponding protein. But as illustrated on Fig. 7, this is not always the case. To further this observation, Fig. 8 displays the number of c.c. for each density map: this number is larger than / equal to / less than the stoichiometry in five / 19 / nine cases. The former case, such as Sec13 on Fig. 7, corresponds to ambiguous locations which induce multiple connected components per instance. The latter one, such as Nup170 on the same figure, occurs when multiple c.c. located nearby merge. This phenomenon is extreme for Pom152, since a single c.c. corresponding to a filled torus is observed.

On the volume of connected components. Assume that the density map of the protein type P contains say p c.c.. Denoting $Vol(cc_i)$ the volume of the ith c.c., consider the set of volume ratios

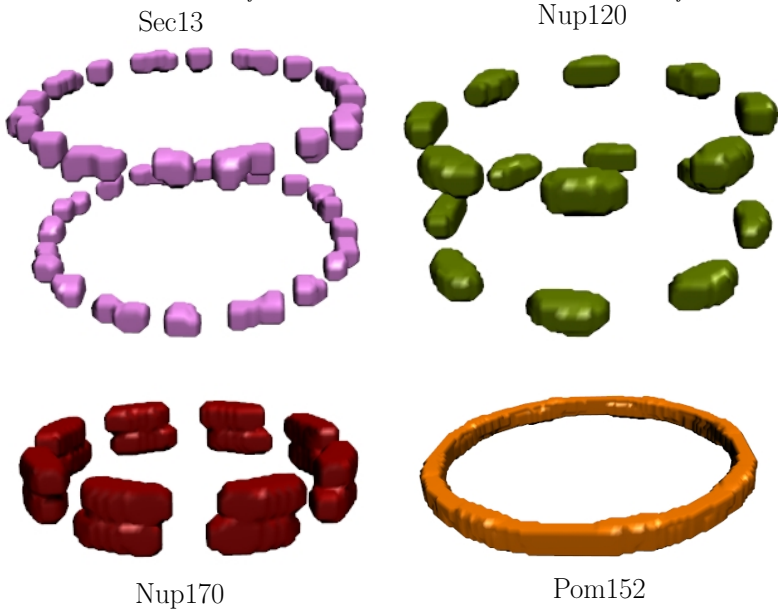
$$v_{cc_i} = Vol(cc_i)/Vol_{ref}(P), \text{ for } i = 1, \ldots, p. \tag{7}$$

The box plots [4] of the 19 density maps with correct stoichiometry [5] are drawn on Fig. 9. The volume ratio tend to decrease when the reference volume increases. In other words, the density maps of large proteins tend to exhibit less geometric uncertainties than those of small proteins.

[4] Recall that the box plot of a set of values is presented as follows. First, the rectangle displays three values, namely the first and third quartiles (small sides of the rectangle), and the median (bold line-segment inside the rectangle). Second, the whiskers extend to the extrema values of the plot, limited by 1.5 times the inter-quartile distance. Values below and above these thresholds are represented by circles.

[5] In this analysis, we restrict ourselves to maps which have the correct stoichiometry, since the meaning of c.c. in the remaining cases in unclear. For example, a c.c. within a plethoric map can be significant or can be insignificant. In theory, analysing the relative importance of c.c. in any map can be done using Morse theory and persistence theory, in a manner similar to the algorithms developed in [CCS11] in the context of Morse theory of the distance function. Yet, for general (density) maps, effective algorithms for Morse-Smale decompositions yet have to be developed.

Figure 7: Example density maps, all voxels with a non null density being displayed. The number of c.c. and the reference volumes estimated from the sequence are respectively: Sec13: $Vol_{ref} = 40.7nm^3$, 32 c.c.; Nup120: $Vol_{ref} = 149.8nm^3$, 16 c.c.; Nup170: $Vol_{ref} = 210.9nm^3$, 8 c.c.; Pom152: $Vol_{ref} = 188.4nm^3$, 1 c.c..



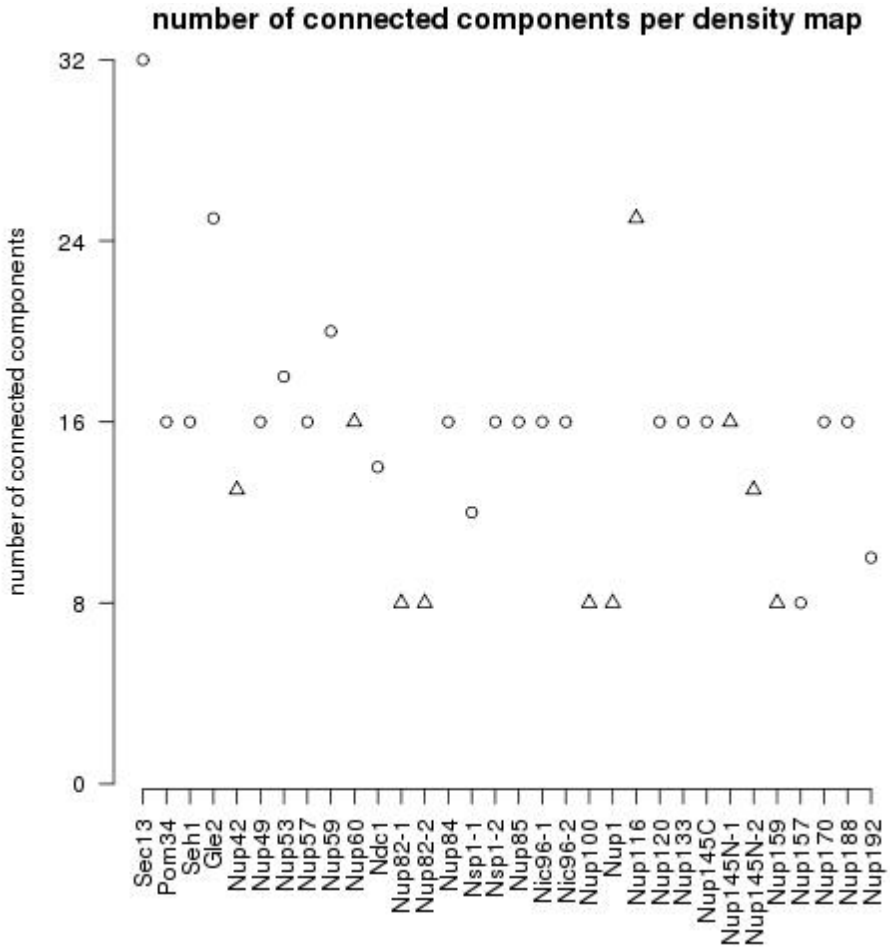


Figure 8: Number of connected components of voxels with non null density per density map—excepted for Pom152 which has a single c.c.. Indices along the x-axis correspond to the 32 maps, index by increasing reference volume. Disks corresponds to maps with a stoichiometry of 16, while triangles correspond to a stoichiometry of 8. A total of 19 maps exhibit the expected stoichiometry.

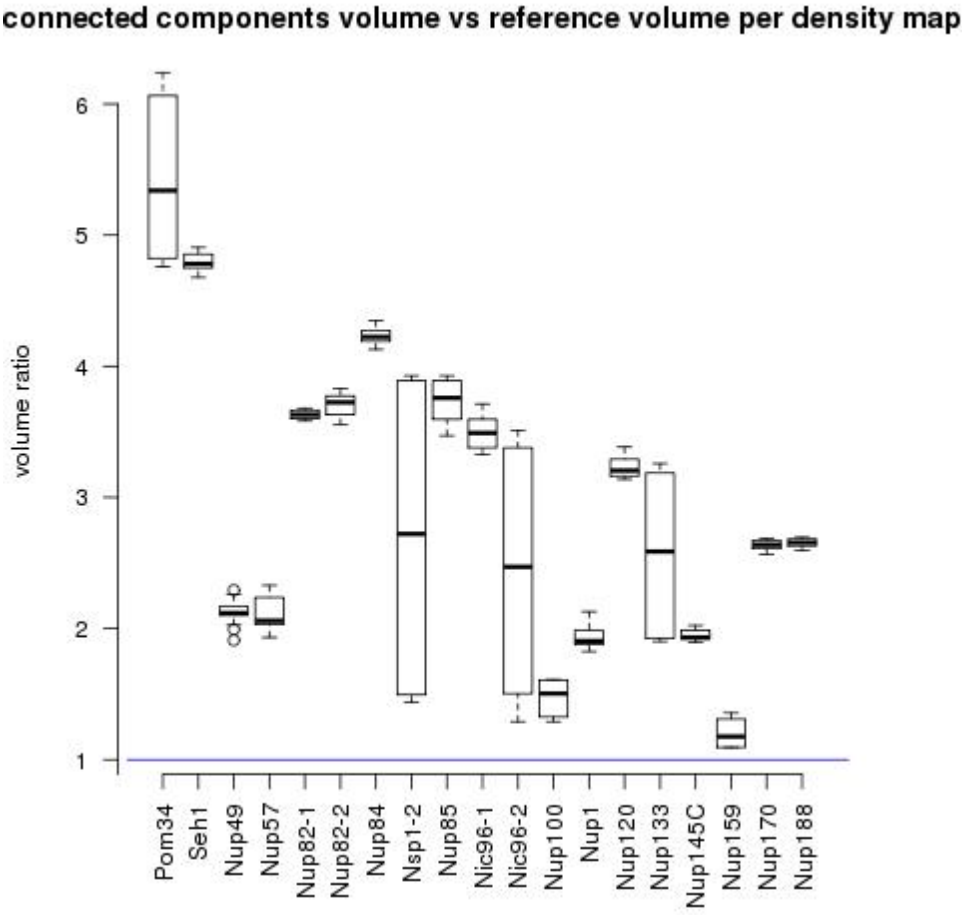


Figure 9: Box plots of the volume ratios v_{cc_i} of Eq. (7), for density maps with a number of c.c. matching the stoichiometry of the protein type.

7.2.2 Selecting Ambiguous Density Maps

The stoichiometry of the T-complex is 16, which requires 16 instances of Nic96 and Nsp1. On the other hand, there are 2 density maps for Nic96 (and Nsp1), each involving 16 instances. Since each map contains one protein instance per half-spoke of the NPC, out of the four possible pairs, (two options for Nic96 and two options for Nsp1) we select the pair producing the best results i.e maximizing global results. The four resulting Hasse diagrams are shown on Fig. 10. Only the Hasse diagram at the Bottom Right reveals 16 isolated copies of the T-complex, motivating the selection of the corresponding two density maps. Note that a calculation with the four density maps would have required selecting the relevant instances of Nic96 and Nsp1 within each half-spoke, an ill-posed problem.

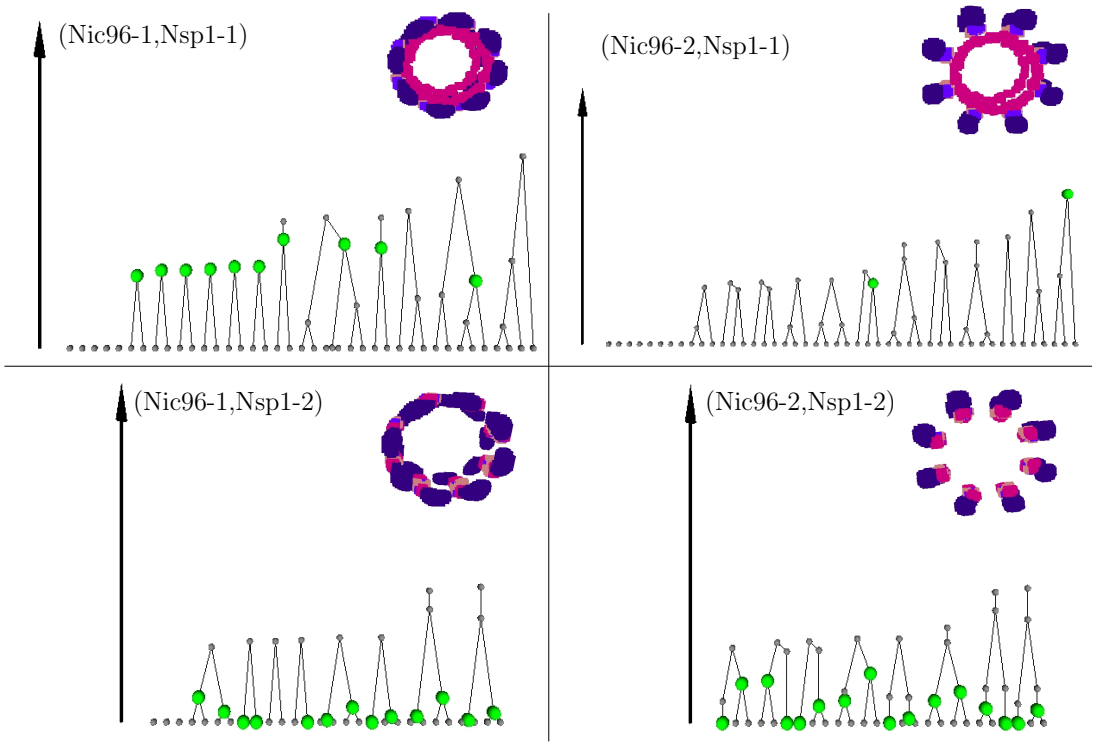


Figure 10: The four Hasse diagrams associated to the four density maps of Nic96 and Nsp1 — these two protein types define the T-core of the T-complex. The density maps used in each case are shown as inset. The two maps selected four our study are those of the Bottom-Right Figure, see section 4.3.

On the connectedness of copies of the Y-complex. As mentioned in section 4.2, each copy of the Y-complex is split into two components. That is, for each copy, there exists a value of the probability such that the level set surfaces of the maps restricted to this copy have two connected components, one including Nup145C and the other one Y_X-short-arm. See Fig. 11.

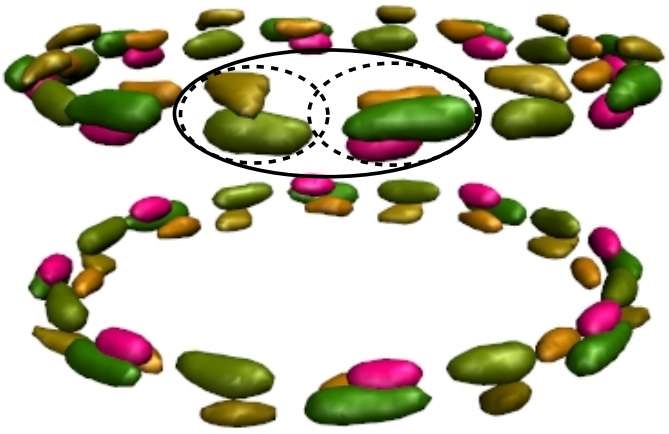


Figure 11: Union of the level-set surfaces from the density maps of protein types of Y-main, at intensity 0.5. The color codes are those of Fig. 2. The circled region illustrates the split of a Y-complex into two pieces (Nup133, Nup84, Nup145C) and (Nup120, Nup85).

7.2.3 Assessment of Toleranced Models

We now wish to assess the geometric accuracy of the toleranced model of section 3.2, by comparing sub-complexes with known crystal structures. To this end, given a crystal structure, term a sub-complex encountered along the growth process of *compliant* provided that it contains protein instances of the types found in the crystal structure. As seen from Table 1, we compare:

- Vol_{ref} The reference volume of these items computed from the sequence [HGC94], by adding up reference volumes on a per-residue basis.
- $V_{r=0}, V_{r=1.4}$ Consider the Van der Waals model of a known crystal structure. We compute the volume $V_{r=0}$ of this model, and the volume $V_{r=1.4}$ of the associated Solvent Accessible model, namely the model obtained by expanding the VdW radii of $1.4\mathring{A}$. These volumes are meant to provide references for the volumes of toleranced models.
- $V_{c,\lambda_{first}}, V_{c,\lambda_{last}}$ The volume of two *compliant* sub-complexes, namely the first one and the last one encountered along the growth process, respectively encountered at $\lambda = \lambda_{first}$ and $\lambda = \lambda_{last}$. Note that these complexes are spotted from the Hasse diagram. Following the volume ratio of Eq. (2), the volume of a compliant complex is computed as the sum of the volumes of its Voronoi restrictions in the power diagram [CHL11], using our software Vorlume, see `http://cgal.inria.fr/abs/Vorlume/`. These volumes, expressed in nm^3, are denoted $V_{c,\lambda_{first}}$ and $V_{c,\lambda_{last}}$.

Crystal structures versus reference Volumes. The upper left region of Table 1 compares the reference volume to $V_{r=0}$ and $V_{r=1.4}$. Regarding Van der Waals models, the ratio $Vol_{ref}/V_{r=0}$ lies in the range 0.33 - 0.49, respectively for the Y_X-edge and Nic96, showing that Van der Walls volumes underestimate the volume of globular proteins. On the other hand, excepted for Nup133 and the Y_X-tail, the ratio $Vol_{ref}/V_{r=1.4}$ lies in the range 0.65 - 1.02, values respectively attained for the Y_X-edge and Nic96. Thus, Solvent Accessible models on a per-atom basis provide a relatively good approximation of reference volumes estimated on a per-residue basis.

Reference volumes versus volumes of compliant complexes. As seen from the upper-right region of Table 1, excepted for three copies of Y_X-long-arm, all isolated copies of all sub-complexes appear at $\lambda = 0$ with a volume ratio varying in the range 0.77 - 0.97 for the Y_X-long-arm and the Y_X-short-arm. We note that these values are comparable to those of Solvent Accessible models. (As explained in section 3.2, the inner radius is set such that the volume ratio of an isolated protein for $\lambda = 0$ is equal to one. The values observed, which are less than one, are due to overlaps with other protein instances.)

The lower-right region of Table 1 reports these ratios for sub-complexes of Y-complex and T-complex with no known crystal structure. All isolated sub-complexes of the Y-complex appear with a volume ratio in the range 0.83 - 2.22 for the Y-arms and the Y-main. For the T-complex, though, isolated copies have a volume ratio in the range 0.17 - 1.49 for the T-leg and the T-core. The lower bound for the T-leg corresponds to a copy partially covered by the remaining toleranced proteins of the NPC, whence a considerably reduced volume.

Protein types	ref.	PDB id	Res ($\mathring{A}$)	$\frac{V_{r=0}}{Vol_{ref}}$	$\frac{V_{r=1.4}}{Vol_{ref}}$	Vol_{ref}	λ_{first}	$\frac{V_{c,\lambda_{first}}}{Vol_{ref}}$	λ_{last}	$\frac{V_{c,\lambda_{last}}}{Vol_{ref}}$
Y_X-edge	[NHD$^+$09]	3IKO	3.20	0.33	0.66	324.3	0	0.83	0	0.87
	[BS09]	3JRO	4.00	0.31	0.65	324.3	0	0.83	0	0.87
Y_X-long-arm	[DMS$^+$08]	3F3F	2.90	0.48	0.97	153.3	0	0.77	0.17	1.65
	[BLS$^+$08]	3EWE	3.50	0.37	0.79	153.3	0	0.77	0.17	1.65
Y_X-short-arm	[SMD$^+$09]	3F7F	2.60	0.45	0.91	149.8	0	0.89	0	0.97
	[SMD$^+$09]	3H7N	3.00	0.45	0.91	149.8	0	0.89	0	0.97
	[LBS09]	3HXR	3.00	0.41	0.85	149.8	0	0.89	0	0.97
Y_X-tail (homologous)	[WS09]	3I4R	3.53	0.23	0.51	269.9	0	0.79	0	0.89
Nup133 (N-terminal)	[SMD$^+$04]	1XKS	2.35	0.20	0.42	165.7	0	0.82	0	0.91
Nic96	[JS07]	2QX5	2.50	0.45	0.92	119.9	0	0.77	0	0.88
Nic96	[SSF$^+$08]	2RFO	2.60	0.49	1.02	119.9	0	0.77	0	0.88
Y-arms						302.1	0	0.83	0	0.93
Y-junction						434.6	0.04	1.14	0.58	2.07
Y-core						538.8	0	0.86	0.44	2.18
Y-main						704.5	0	0.86	0.44	2.22
Y-complex						793.1	0.04	1.11	0.21	1.85
T-leg						354.8	0	0.17	0	0.27
T-core						224.2	0	0.79	0.15	1.49
T-complex						579.0	0	0.48	0.15	0.78

Table 1: Comparison of volumes of selected proteins and sub-complexes : crystal structures versus toleranced models. **Top.** Crystal structures versus toleranced models of sub-complexes of the Y-complex and the T-complex. **Bottom.** Toleranced models of interesting sub-complexes of the NPC.

7.3 Maximal Common Edge/Induced Sub-graphs

In section 2.2, we presented tools to compare the skeleton graphs of a complex and of a template. In this section, we formally present the notions of matching and common sub-graphs.

7.3.1 Matchings

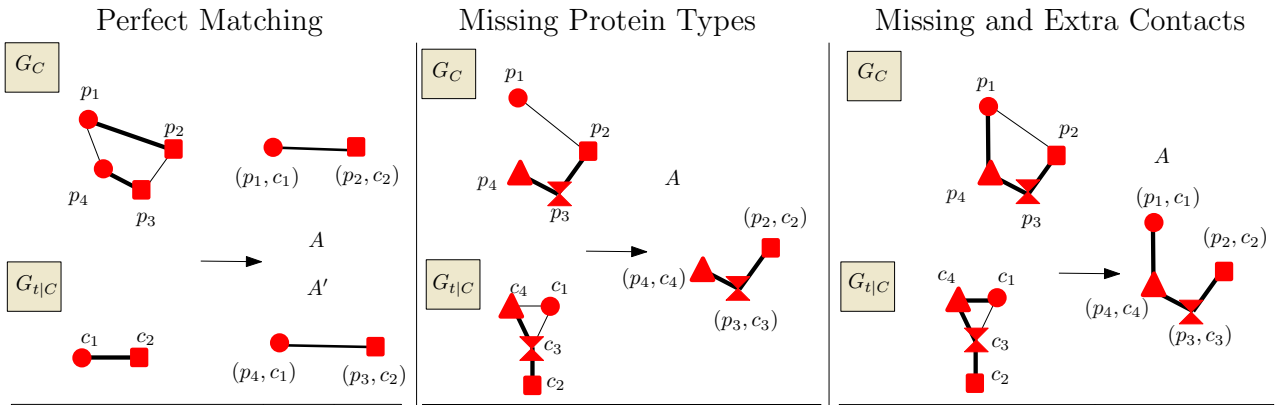
A *matching* A from $G_{t|C}$ to C maps (i) vertices of $G_{t|C}$ (protein types of the template) to vertices of G_C (protein instances of the complex), and (ii) edges of $G_{t|C}$ (contacts within the template) to edges of G_t (contacts within the complex). The inverse of map A, from G_t to $G_{t|C}$, is denoted A^{-1}. Taking the template as reference, we assess a matching with the following five categories, illustrated on Fig. 12:

- *Matching protein type(s):* a protein type of the restricted template with a corresponding instance in the complex:
$$V^{\sim}(G_{t|C}; G_C; A) = \{(u,v) \in V[G_{t|C}] \times V[G_C], A(u) = v\} \tag{8}$$
- *Missing protein type(s):* a protein type of the restricted template with no corresponding instance in the complex:
$$V^{-}(G_{t|C}; G_C; A) = \{u \in V[G_{t|C}], A(u) = \emptyset\} \tag{9}$$
- *Matching contact(s):* a contact in the restricted template with a counterpart in the complex:
$$E^{\sim}(G_{t|C}; G_C; A) = \{(u_1,u_2),(v_1,v_2) \in E[G_{t|C}] \times E[G_C], A(u_1,u_2) = (v_1,v_2)\} \tag{10}$$
- *Missing contact(s):* a contact in the restricted template with no counterpart in the complex:
$$E^{-}(G_{t|C}; G_C; A) = \{(u,v) \in E[G_{t|C}], A(u) \neq \emptyset, A(v) \neq \emptyset, A(u,v) = \emptyset\} \tag{11}$$
- *Extra contact(s):* a contact in the complex with no counterpart in the restricted template:
$$E^{+}(G_{t|C}; G_C; A) = \{(u,v) \in E[G_C], A^{-1}(u) \neq \emptyset, A^{-1}(v) \neq \emptyset, A^{-1}(u,v) = \emptyset\} \tag{12}$$

Using these sets, the *signature* of the matching A is defined by:

$$S(G_{t|C}; G_C; A) = \{V^{\sim}, V^{-}, E^{\sim}, E^{-}, E^{+}\}. \tag{13}$$

Note in particular that the matching is called *perfect* provided that the three sets V^-, E^-, E^+ are empty, in which case $G_{t|C}$ is isomorphic to an induced sub-graph of G_C.



Matchings	A	A'
Matching Protein Types	(p_1, c_1) (p_2, c_2)	(p_3, c_2) (p_4, c_1)
Matching Contacts	$(c_1, c_2) \leftrightarrow (p_1, p_2)$	$(c_1, c_2) \leftrightarrow (p_4, p_3)$
Missing Protein Types	$\emptyset$	$\emptyset$
Missing Contacts	$\emptyset$	$\emptyset$
Extra Contacts	$\emptyset$	$\emptyset$

Matchings	A
Matching Protein Types	(p_2, c_2) (p_3, c_3) (p_4, c_4)
Matching Contacts	$(c_2, c_3) \leftrightarrow (p_2, p_3)$ $(c_4, c_3) \leftrightarrow (p_4, p_3)$
Missing Protein Types	c_1
Missing Contacts	$\emptyset$
Extra Contacts	$\emptyset$

Matchings	A
Matching Protein Types	(p_1, c_1) (p_2, c_2) (p_3, c_3) (p_4, c_4)
Matching Contacts	$(c_2, c_3) \leftrightarrow (p_2, p_3)$ $(c_4, c_3) \leftrightarrow (p_4, p_3)$ $(c_4, c_1) \leftrightarrow (p_4, p_1)$
Missing Protein Types	$\emptyset$
Missing Contacts	(c_1, c_3)
Extra Contacts	(p_1, p_2)

Figure 12: Comparing the skeleton graphs G_C of a complex C and $G_{t|C}$ of a template t restricted to C. The match between a protein instance of G_C and a type of $G_{t|C}$ is materialized by an identical geometric shape (disk, square, triangle, hourglass). Matched contacts corresponds to bold edges. The adjectives matching/missing/extra qualify G_C w.r.t. $G_{t|C}$. **(Left.)** A Maximal Common Induced Sub-graph calculation yields two perfect matchings. **(Middle.)** A Maximal Common Edge Sub-graph calculation yields a matching with one missing protein type. **(Right.)** A Maximal Common Edge Sub-graph calculation yields a matching with missing and extra contacts.

7.3.2 Computing matchings from Maximal Common Induced/Edge Sub-graph

Definitions. Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two undirected labelled graphs.

Definition. 1. *A* **Maximal Common Edge Sub-graph (MCES)** *of G_1 and G_2 is a graph H that is isomorphic to sub-graphs G'_1 of G_1 and G'_2 of G_2, such that there is no other Common Edge Sub-graph H' of G_1 and G_2 containing H.*

Definition. 2. *An* **induced sub-graph** *G' of G is a sub-graph of G such that for all pairs of vertices (u, v) of G', (u, v) is an edge of G' iff it is an edge of G.*

Definition. 3. *A* **Maximal Common Induced Sub-graph (MCIS)** *of G_1 and G_2 is a graph H that is isomorphic to* induced sub-graphs *G'_1 of G_1 and G'_2 of G_2, such that there is no other Common Induced Sub-graph H' of G_1 and G_2 containing H.*

These notions are illustrated on Fig. 13. Notice in particular that a MCES or MCIS calculation yields in general several matchings.

Algorithms. The calculation of all maximal common sub-graphs of two graphs G_1 and G_2 is equivalent to the enumeration of all maximal cliques of a so-called product graph [Koc01], a problem for which exact algorithms were proposed in [CK05]. In fact, there are two kind of product graphs:

- the *edge product graph*, from which we generate Maximal Common Edge Sub-graphs (MCES). Each node of the *edge product graph* is associated to a pair of edges $(e_1 \in E[G_1], e_2 \in E[G_2])$, and there is an edge between two nodes (e_1, e_2) and (f_1, f_2) iff (e_1, f_1) and (e_2, f_2) are incident together to a same vertex, or are not incident together to a same vertex.
- the *vertex product graph*, from which we generate Maximal Common Induced Sub-graphs (MCIS). Each node of the *vertex product graph* is associated to a pair of vertices $(u_1 \in V[G_1], u_2 \in V[G_2])$, and there is an edge between two nodes (u_1, u_2) and (v_1, v_2) iff (u_1, v_1) and (u_2, v_2) are neighbors together, or are not neighbors together.

Note that the definition of product graphs is purely topological. But in our setting, a protein type is associated to each vertex of graphs G_1 and G_2, and we only match two vertices provided that they carry the same protein type. Similarly, matching two edges requires the agreement of their vertices. As an example, consider Fig. 13 and assume that a matches x, that b matches y and that c matches z. Under these hypothesis, there is a single MCES and two MCIS.

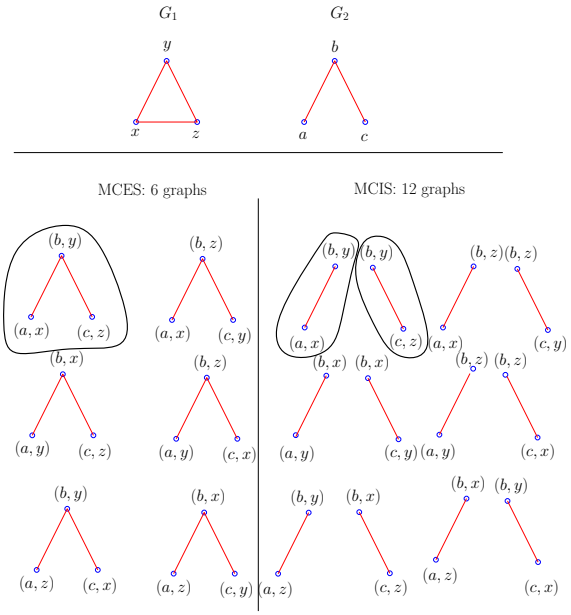


Figure 13: **Top.** Two labelled graphs G_1 and G_2. **Bottom Left.** Maximal Common Edge Sub-graph of G_1 and G_2 (MCES). **Bottom Right.**: Maximal Common Induced Sub-graph (MCIS) of G_1 and G_2. If we impose a correspondence between labels $((a, x), (b, y), (c, z))$, there is one MCES and there are two MCIS—the circled graphs.

7.4 Results for Contact Analysis

7.4.1 Over and Under-represented pairs in the toleranced model

Section 4.1 compares the classes of low/medium/high contact frequencies and probabilities.

For each class $F_{i,i=1,2,3}$, the Figs. 14 and 15 present the variation of the cardinality of the classes $P_i^{(1)}, i = 1, 2, 3$ while varying $\lambda_{\max}$ in $[0, 1]$. Note that increasing $\lambda_{\max}$ yields a monotonic increase (resp. decrease) of the cardinality of the class $P_3^{(1)}$ (resp. $P_1^{(1)}$).

The over-represented and under-represented pairs of types (in the toleranced model) are listed in Tables 2 and 3. Finally, the improved coherence between our contact probability and the density maps is illustrated by the examples of Fig. 16 and 17.

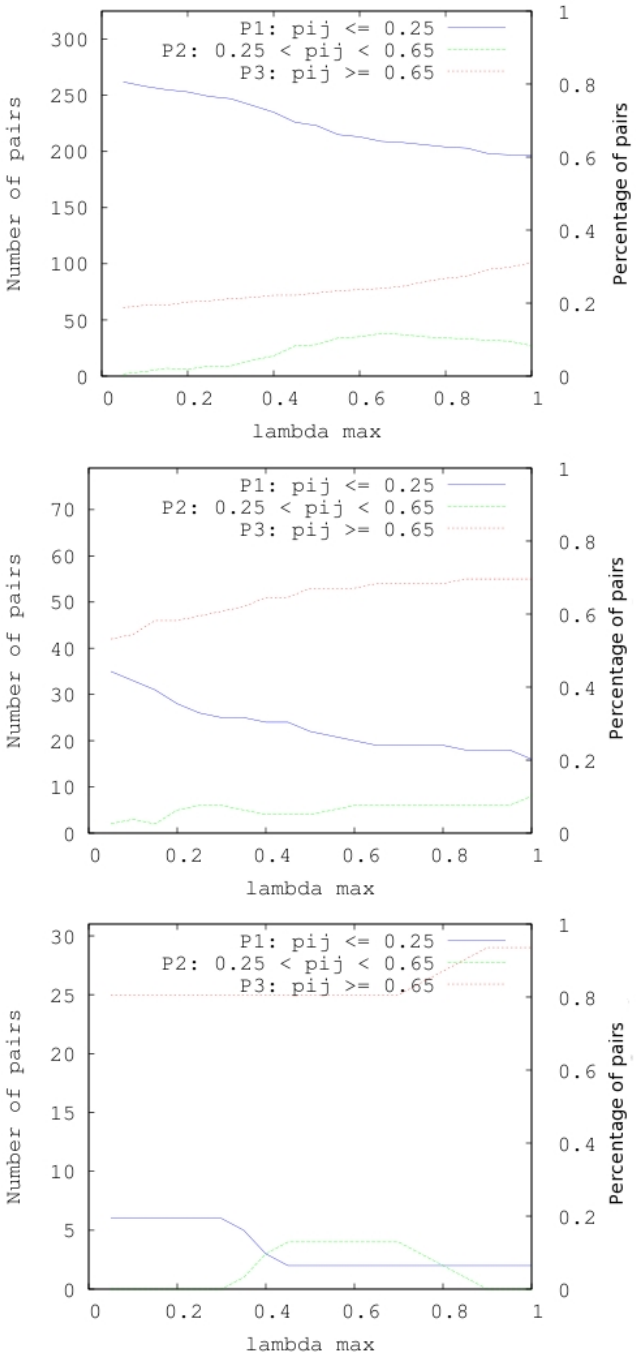


Figure 14: Partitioning the set of pairs of protein types with a prescribed contact frequency f_{ij}, from [ADV$^+$07b], into the three classes of increasing contact probability $P_i^{(1)}, i = 1, 2, 3$. **Top:** low frequencies F_1 i.e. $f_{ij} \leq a$; **Middle:** medium frequencies F_2 i.e. $a < f_{ij} < b$; **Bottom:** high frequencies F_3 i.e. $b \leq f_{ij}$. Following [ADV$^+$07b], the thresholds are $a = 0.25$ and $b = 0.65$.

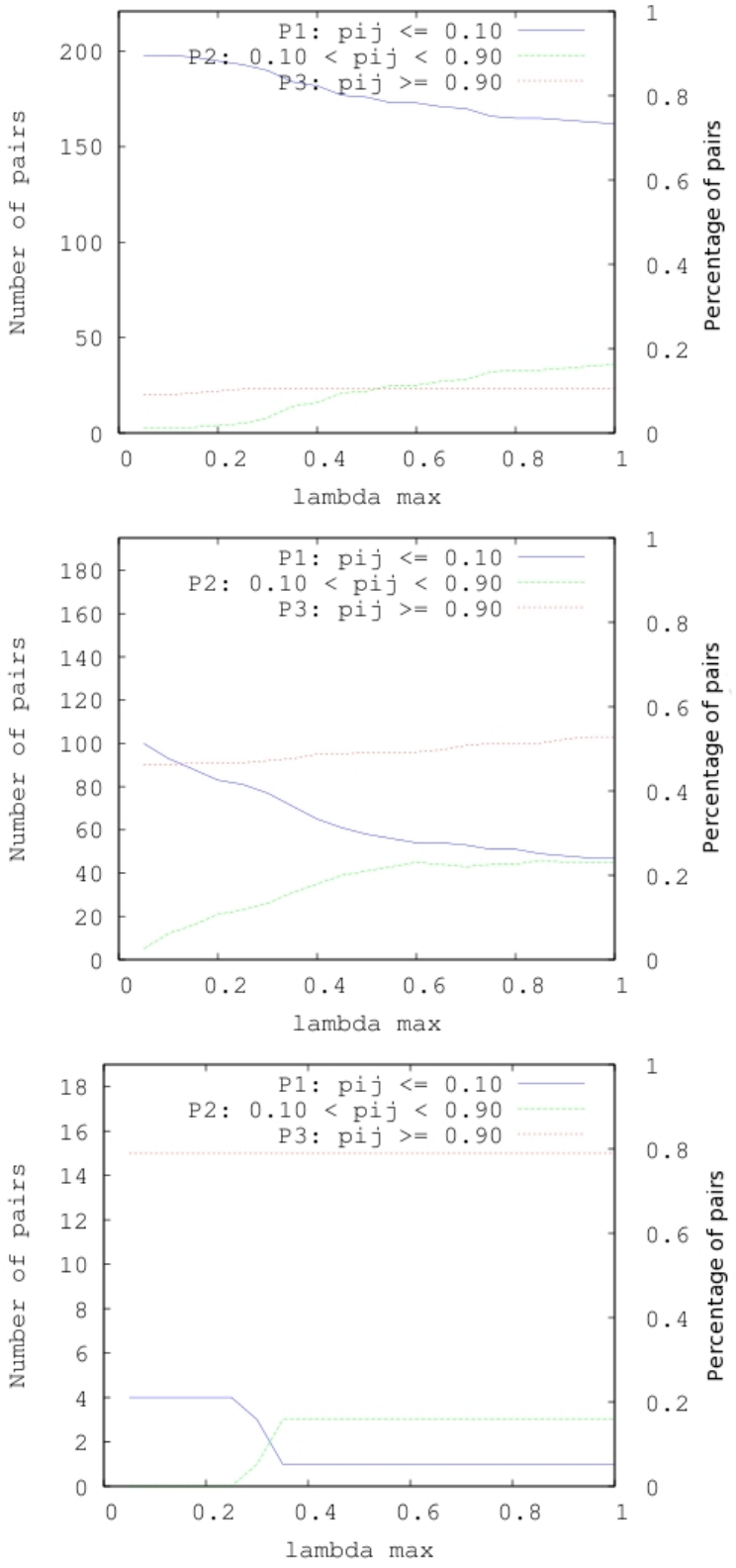


Figure 15: Partitioning the set of pairs of protein types with a prescribed contact frequency f_{ij}, from [ADV$^+$07b], into the three classes of increasing contact probability $P_i^{(1)}, i = 1, 2, 3$. **Top:** low frequencies F_1 i.e. $f_{ij} \leq a$; **Middle:** medium frequencies F_2 i.e. $a < f_{ij} < b$; **Bottom:** high frequencies F_3 i.e. $b \leq f_{ij}$. Thresholds are $a = 0.1$ and $b = 0.9$.

Contact	f_{ij}	$p_{ij}^{(1)}$	$\lambda_{\max}$
Nup59 Nup59	0	1	0
Pom34 Pom34	0.02	1	0
Nsp1 Nsp1	0.02	1	0
Nup60 Nup145N	0.03	1	0
Nup60 Pom34	0.03	1	0
Nup145N Nup49	0.04	1	0
Nup1 Nup145N	0.05	1	0
Nup60 Ndc1	0.06	1	0
Nup84 Nup60	0.07	1	0
Nsp1 Nup145N	0.07	1	0
Nup145C Nup60	0.08	1	0
Sec13 Nup159	0.08	1	0
Nsp1 Nup60	0.08	1	0
Nup49 Nup116	0.08	1	0
Nup57 Nup145N	0.08	1	0
Nsp1 Nup42	0.09	1	0
Nup60 Nup59	0.09	1	0
Nup42 Nup116	0.09	1	0
Nup57 Nup116	0.09	1	0
Sec13 Nup145N	0.1	1	0
Nup59 Pom34	0.03	0.9	0.15
Seh1 Nup60	0.06	0.9	0.18
Gle2 Nup57	0.08	0.9	0.21

Table 2: Over-represented pairs of types in the toleranced model for $a = 0.1$ and $b = 0.9$—that is pairs in $P_3^{(1)}$ and F_1. The last column is the smallest $\lambda_{\max}$ value for which the contact is over-represented—the smaller the value the more significant the contact.

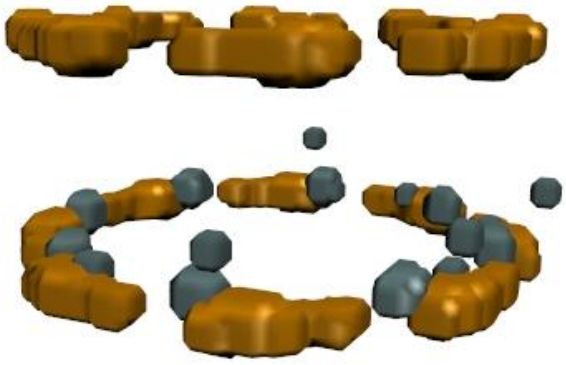


Figure 16: An example over-represented pair. The overlapping density maps of Nup84 (stoichiometry: 16) and Nup60 (stoichiometry: 8), from `http://salilab.org/npc/`, visualized with VMD. Their contact frequency from [ADV$^+$07b] is $f_{ij} = 0.07$, while the contact probability from the toleranced model is $p_{ij}^{(1)} = 1$.

Contacts	f_{ij}	$p_{ij}^{(1)}$	$\lambda_{\max}$
Nup192 Pom152	0.98	0	1
Nup170 Ndc1	0.91	0.1	0.35
Nup188 Nic96	1	0.1	0.32
Pom152 Pom34	1	0.1	0.28

Table 3: Under-represented pairs of types for $a = 0.1$ and $b = 0.9$, i.e. pairs in $P_1^{(1)}$ and F_3. The last column is the largest $\lambda_{\max}$ value for which the contact is under-represented—the larger the value the less significant the contact.

Figure 17: An example under-represented pair. The disjoint density maps of Nup192 (stoichiometry: 16) and Pom152 (stoichiometry: 16), from `http://salilab.org/npc/`, visualized with VMD. Their contact frequency from [ADV+07b] is $f_{ij} = 0.98$, while the contact probability from the toleranced model is $p_{ij}^{(1)} = 0$.

7.4.2 On k-significant contacts

As explained in section 4.1, for a given $\lambda_{\max}$ and two fixed probabilities $0 \leq a < b \leq 1$, the contacts observed in the Hasse diagram are partitioned into the classes $P_i^{(k)}, i = 1, 2, 3$. The variation of the cardinality of these classes with $\lambda_{\max}$ and k is displayed on Fig. 18. We note that the curves are just shifted when $\lambda_{\max}$ varies. In the following, we consider $\lambda_{\max} = 1$ (solid lines).

The red and blue curves show that the contact probability to have k instances of the contacts between two protein types decreases when k increases. The green curves show the discriminant property of the contact probability since less than 40 pairs of protein types are in $P_2^{(1)}$, and green curves tend to decrease when k increases.

The partition of all contacts into classes$S_{0.65}^{(k)}$ can be found in Tables 4, 5, 6 and 7.

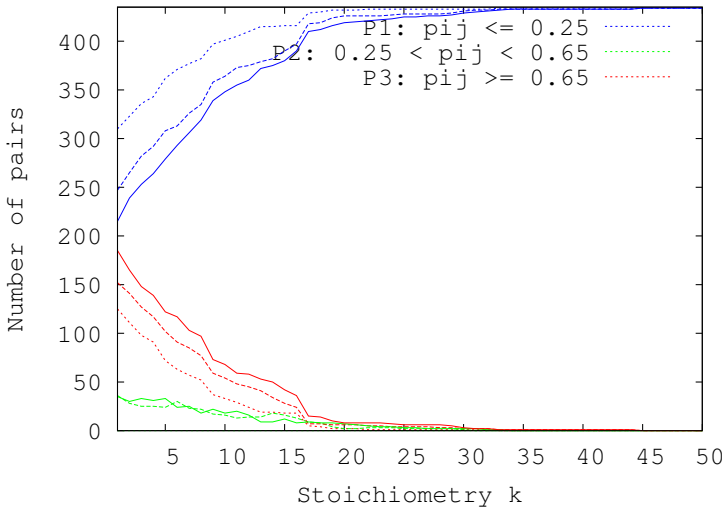


Figure 18: Cardinality of the classes $P_1^{(k)}, P_2^{(k)}, P_3^{(k)}$ as a function of k. **Dotted lines** for $\lambda_{\max} = 0$, **dashed lines** for $\lambda_{\max} = 0.5$ and **solid lines** for $\lambda_{\max} = 1$.

Contact type	k	$p_{ij}^{(k)}$
Nic96 Nsp1	44	0.84
Nup192 Nup170	32	1.00
Gle2 Nsp1	30	0.80
Nup192 Nup157	29	0.68
Nup82 Nsp1	29	0.76
Nup188 Nsp1	28	0.69
Ndc1 Pom34	24	0.72
Nup170 Nup57	23	0.71
Nup192 Nup57	19	0.72
Nsp1 Nsp1	19	0.68
Nic96 Nup82	18	0.89
Nsp1 Nup100	18	0.72
Nup84 Nic96	18	0.65
Nsp1 Nup145N	18	0.69
Pom152 Ndc1	17	0.72
Nsp1 Nup49	16	1.00
Nup57 Nup49	16	1.00
Nup192 Nic96	16	1.00
Nup120 Nup85	16	1.00
Nup120 Seh1	16	1.00
Nup133 Nup84	16	1.00
Nup133 Nup145C	16	1.00
Nup170 Nic96	16	1.00
Nup84 Nup145C	16	1.00
Nup145C Nic96	16	1.00
Nup82 Nup159	16	1.00
Nsp1 Nup57	16	1.00
Nic96 Nup59	16	0.95
Nup192 Nup49	16	0.84
Nup85 Seh1	16	0.83
Nup170 Nup53	16	0.80
Nup188 Nup157	16	0.79
Nup145C Nup53	16	0.79
Nup170 Nup157	16	0.78
Nup157 Nup53	16	0.77
Nup157 Nup120	16	0.72

Table 4: Pairs in $S_{0.65}^{(k\geq 16)}$ for $\lambda_{\max} = 1$. Refer to section 2.4 for the definition of $S_b^{(k)}$.

Contact type	k	$p_{ij}^{(k)}$
Nup188 Gle2	15	0.90
Nic96 Nup145N	15	0.88
Nup120 Nup53	15	0.81
Nic96 Nup53	15	0.81
Nsp1 Nup116	15	0.75
Nup57 Nup145N	15	0.66
Nup192 Nup59	14	0.81
Nup133 Sec13	14	0.81
Nup84 Sec13	14	0.79
Nic96 Nup159	14	0.72
Nup120 Sec13	14	0.71
Nup1 Nsp1	14	0.69
Nup133 Nup59	14	0.67
Nup170 Nup49	14	0.65
Nup145C Nup59	13	1.00
Gle2 Nic96	13	0.93
Nup84 Nup82	13	0.87
Nup145C Sec13	12	1.00
Nup170 Nup59	12	0.89
Gle2 Nup145N	12	0.85
Nup192 Nsp1	12	0.71
Pom152 Pom152	12	0.68
Nup82 Nup116	11	0.68

Table 5: Pairs in $S_{0.65}^{(11\leq k\leq 15)}$ for $\lambda_{\max} = 1$. Refer to section 2.4 for the definition of $S_b^{(k)}$.

Contact type	k	$p_{ij}^{(k)}$	Contact type	k	$p_{ij}^{(k)}$
Nup82 Nup100	10	1.00	Nup188 Nup1	8	0.74
Nup82 Nup42	10	0.91	Seh1 Nup145N	8	0.71
Nup157 Seh1	10	0.85	Nup188 Nic96	8	0.68
Nup188 Nup145N	10	0.74	Nup59 Ndc1	8	0.68
Nic96 Nup49	10	0.72	Nup84 Seh1	8	0.67
Nup192 Nup145N	10	0.70	Nup57 Nup57	8	0.67
Nup120 Nup84	10	0.67	Nup59 Pom34	8	0.66
Nup192 Nup188	10	0.66	Seh1 Gle2	8	0.66
Nup85 Sec13	10	0.66	Nic96 Nup60	8	0.66
Gle2 Nup82	9	0.86	Nup157 Nup59	7	1.00
Nsp1 Nup42	9	0.79	Nup159 Nup100	7	0.98
Nup170 Nup145N	9	0.76	Nup120 Nup59	7	0.86
Nup59 Nup53	9	0.73	Gle2 Nup100	7	0.72
Gle2 Nup53	9	0.65	Nup1 Nup145N	7	0.69
Nup188 Nup100	8	1.00	Nup120 Nup145N	7	0.66
Nup170 Nup170	8	1.00	Nsp1 Nup159	6	0.92
Nup84 Nup159	8	1.00	Nup116 Nup100	6	0.92
Nic96 Nup1	8	1.00	Pom34 Pom34	6	0.76
Nup82 Nup82	8	1.00	Nic96 Ndc1	6	0.76
Nic96 Nup42	8	1.00	Nup192 Nup60	6	0.75
Nup42 Nup100	8	1.00	Nic96 Nup116	6	0.73
Nup145N Nup49	8	1.00	Nup133 Nic96	6	0.72
Nic96 Nup100	8	0.96	Nup145C Nup60	6	0.70
Nup42 Nup159	8	0.90	Seh1 Sec13	6	0.69
Nup84 Nup145N	8	0.86	Nup60 Ndc1	6	0.69
Nup133 Nup85	8	0.82	Pom152 Pom34	6	0.68
Nup53 Ndc1	8	0.80	Nup60 Nup145N	6	0.68
Nup157 Nup157	8	0.78	Nup120 Nup159	6	0.67
Nup145C Nup159	8	0.77	Nup120 Gle2	6	0.66

Table 6: Pairs in $S_{0.65}^{(6\leq k\leq 10)}$ for $\lambda_{\max} = 1$. Refer to section 2.4 for the definition of $S_b^{(k)}$.

Contact type	k	$p_{ij}^{(k)}$	Contact type	k	$p_{ij}^{(k)}$
Nup188 Nup116	5	1.00	Nup85 Nup84	2	1.00
Gle2 Nup116	5	1.00	Nup133 Nup120	2	0.97
Nup60 Nup59	5	0.99	Nup120 Nup145C	2	0.95
Nup188 Nup49	5	0.84	Gle2 Nup57	2	0.94
Sec13 Nup159	5	0.65	Nup84 Nup116	2	0.93
Nup84 Nup60	4	1.00	Nup192 Nup116	2	0.89
Nup159 Nup116	4	1.00	Nup157 Nup60	2	0.82
Nup57 Nup116	4	1.00	Nup120 Nup60	2	0.80
Nup49 Nup116	4	1.00	Nup1 Nup60	2	0.73
Nup170 Nup60	4	0.98	Seh1 Nup53	2	0.73
Nup85 Nup145N	4	0.98	Nup85 Nup60	2	0.69
Nup60 Pom34	4	0.95	Nup145C Ndc1	2	0.69
Nup42 Nup116	4	0.93	Nup170 Ndc1	2	0.68
Nup192 Nup53	4	0.87	Nup170 Pom34	2	0.67
Nup85 Nup159	4	0.85	Nic96 Pom34	2	0.67
Seh1 Nup159	4	0.82	Gle2 Nup42	1	1.00
Nup145C Nup145N	4	0.79	Nup170 Nup116	1	0.84
Nup157 Gle2	4	0.68	Nup60 Nup53	1	0.82
Nup157 Ndc1	4	0.67	Nic96 Nup57	1	0.79
Nup84 Nup59	4	0.66	Nup57 Nup100	1	0.77
Nup84 Gle2	4	0.66	Nup188 Nup42	1	0.74
Nsp1 Nup60	4	0.66	Sec13 Nup82	1	0.72
Sec13 Nup145N	3	0.94	Nup49 Nup100	1	0.71
Seh1 Nup60	3	0.89	Seh1 Nup82	1	0.71
Nup188 Nup57	3	0.83	Nup133 Nup60	1	0.70
Gle2 Nup1	3	0.82	Nup84 Nsp1	1	0.69
Nup170 Nsp1	3	0.74	Gle2 Nup60	1	0.69
Nup53 Pom34	3	0.71	Nup60 Nup49	1	0.69
Gle2 Nup159	3	0.69	Nsp1 Nup53	1	0.67
Nup53 Nup145N	3	0.66	Nup145C Nup82	1	0.67
Sec13 Nup53	3	0.65	Nup188 Nup60	1	0.66
Gle2 Nup49	2	1.00	Nup170 Gle2	1	0.66
Nup59 Nup59	2	1.00	Nup145C Nsp1	1	0.65

Table 7: Pairs in $S_{0.65}^{(1\leq k\leq 5)}$ for $\lambda_{\max} = 1$. Refer to section 2.4 for the definition of $S_b^{(k)}$.

7.5 Results for Perfect and Alternate matchings

Template; tag	#	$V^{\sim}$	$\min r_\lambda$	$\max r_\lambda$
$G_t(Y)$;P_1	14	Y_X-tail	0.77	0.90
$G_t(Y)$;P_2	2	(Y_X-tail,Nup145C)	0.85	0.88
$G_t(Y)$;P_3	5	(Nup145C,Nup84)	0.81	0.88
$G_t(Y)$;P_4	5	(Nup145C,Sec13)	0.81	0.86
$G_t(Y)$;P_5	7	Y_X-edge	0.78	0.88
$G_t(Y)$;P_6	11	(Y_X-short-arm,Nup85)	0.88	0.91
$G_t(Y)$;P_7	1	Y-junction	1.78	1.78
$G_t(Y)$;P_8	16	(Y_X-long-arm)	0.77	1.63
$G_t(Y)$;P_9	4	Y-core	1.15	2.57
$G_t(Y)$;P_{10}	4	Sec13	0.58	0.69
$G_t(T)$;P_{11}	10	T-leg	0.57	0.75
$G_t(T)$;P_{12}	6	(T-leg,Nsp1)	0.61	0.72
$G_t(T)$;P_{13}	14	(T-core,Nup57)	0.74	1.37
$G_t(T)$;P_{14}	2	T-complex	1.79	2.28
$G_t(T$-comp);P_{15}	2	T-leg	2.42	2.79
$G_t(T$-comp);P_{16}	16	(T-leg,Nsp1)	0.61	0.81
$G_t(T$-comp);P_{17}	5	T-core	0.79	1.46
$G_t(T$-comp);P_{18}	10	(T-core,Nup49)	0.97	1.76
$G_t(T$-comp);P_{19}	1	(T-core,Nup57)	1.91	1.91
$G_t(T$-new);P_{20}	6	(T-leg,Nsp1)	0.61	0.81
$G_t(T$-new);P_{21}	2	(T-leg,Nic96)	2.13	2.25
$G_t(T$-new);P_{22}	6	(T-core,Nup57)	0.78	1.37
$G_t(T$-new);P_{23}	10	T-complex	0.98	1.73

Table 8: Perfect matchings for the templates $G_t(Y)$, $G_t(T)$, $G_t(T$-comp) and $G_t(T$-new). Each matching is identified by a tag (P_i) referenced in the text. The columns read as follows: $V^{\sim}$: protein types involved in the matching; #: number of identical matchings; $\min r_\lambda(C)$ and $\max r_\lambda(C)$: min and max volume ratios amidst identical matchings.

Template; tag	#	$V^{\sim}$	$\mid V^{-} \mid$	$\mid E^{\sim} \mid$	$\mid E^{-} \mid$	min $\mid E^{+} \mid$	max $\mid E^{+} \mid$	$\min r_\lambda$	$\max r_\lambda$
$G_t(Y)$;A_1	1	(Y_X-tail,Nup145C)	4	2	0	1	1	3.43	3.43
$G_t(Y)$;A_2	10	(Y_X-tail,Y_X-edge)	3	3	0	3	3	3.61	4.32
$G_t(Y)$;A_3	11	Y-arms	4	2	0	1	1	0.94	4.71
$G_t(Y)$;A_4	3	(Y-main,Seh1)	1	5	0	2	6	1.13	2.96
$G_t(Y)$;A_5	2	Y-complex	0	6	0	7	7	3.37	3.47
$G_t(T)$;A_6	18	T-complex	0	3	0	0	2	0.83	2.36
$G_t(T\text{-comp})$;A_7	11	T-complex	0	5	1	0	0	0.98	1.85
$G_t(T\text{-comp})$;A_8	7	T-complex	0	4	2	0	0	1.17	2.22
$G_t(T\text{-comp})$;A_9	4	T-complex	0	3	3	0	0	1.80	2.28
$G_t(T\text{-new})$;A_{10}	10	T-complex	0	5	0	0	0	0.98	1.73
$G_t(T\text{-new})$;A_{11}	8	T-complex	0	4	1	0	1	1.17	2.26
$G_t(T\text{-new})$;A_{12}	4	T-complex	0	3	2	0	0	1.79	2.29

Table 9: Alternate matchings for the templates $G_t(Y), G_t(T), G_t(T\text{-comp})$ and $G_t(T\text{-new})$. Each matching is identified by a tag (A_i) referenced in the text. The columns read as follows: #: number of identical matchings; $V^{\sim}$: protein types involved in the matching; $|V^{-}|, |E^{\sim}|, |E^{-}|, \min|E^{+}|, \max|E^{+}|$: size of the sets involved in the signature of the matching—min and max taken amidst all identical matchings; $\min r_\lambda$ and $\max r_\lambda$: min and max volume ratios amidst identical matchings.

7.6 Further in-silico Experiments

7.6.1 Location of Sec13 Relatively to the Y_X-edge

In section 4.2, we mentioned the satisfactory positioning of Sec13 w.r.t. proteins of the Y_X-edge. This claim is warranted by the Hasse diagram shown on Fig. 19 (Top), which reveals 13 isolated copies of the Y_X-edge. All these isolated copies appear at $\lambda = 0$ ($r_\lambda = 0.87$) and have a lifetime varying in-between $s(C) = 0.21$ ($\Delta r_\lambda = 2.40$) and $s(C) = \lambda_{\max}$ ($\Delta r_\lambda = 18.66$). They coexist until $\lambda = 0.21$, witnessing the close positioning of Sec13 with respect to Nup84 and Nup145C in all these isolated copies, as in the crystal structure of the Y_X-edge.

Also, one has 18 complexes at $\lambda = 0.21$ ($r_\lambda = 3.24$), which is larger than the expected stoichiometry. These 18 complexes do not merge for larger values of λ as shown on Fig. 19 (Middle plot).

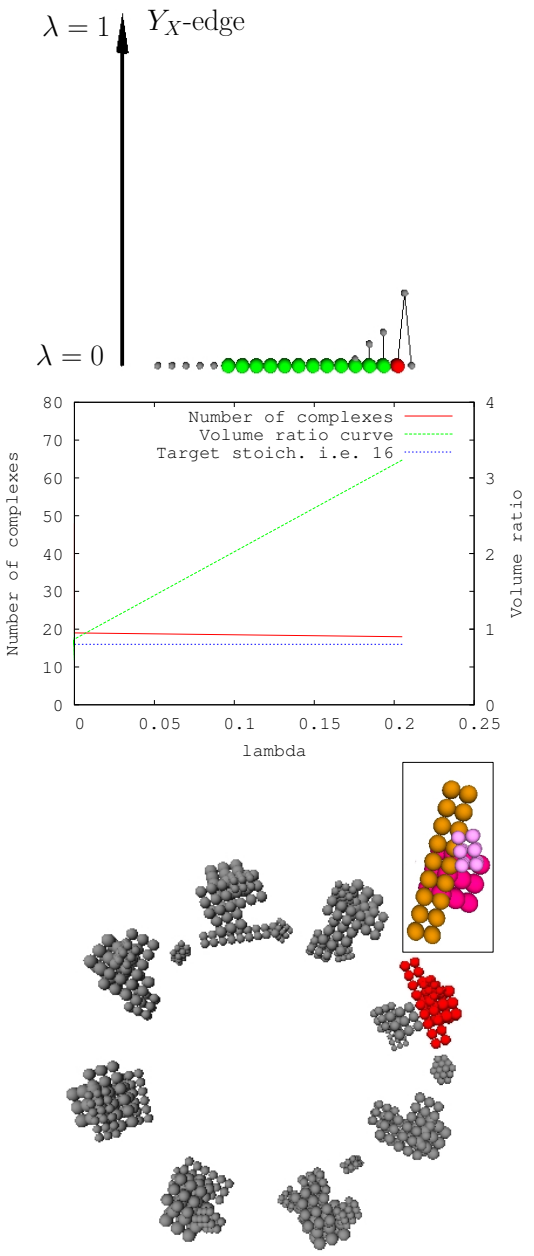


Figure 19: Evolution of the number of complexes associated to the pullout of the Y_X-edge. Compare to Fig. 5.

7.6.2 Removing Sec13 From the Toleranced Model

The global analysis of the Y-complex without Sec13 has been discussed in Section 4.2, and is illustrated on Fig. 20. Note that the lifetimes of the isolated copies of the Y-complex without Sec13 vary in-between $s(C) = 0.03$ ($\Delta r_\lambda = 0.18$) and $s(C) = 0.24$ ($\Delta r_\lambda = 0.94$), showing that the isolated copies are less stable than those found for the whole Y-complex. Not surprisingly, the intersection of the lifetime intervals is empty—as for the entire Y-complex. The local analysis is represented in Tables 10 and 11. We classify the perfect and alternate matchings of $G_t(Y)$ following the entries of Tables 8 and 9. The perfect matchings of Table 10 do not reveal a significant improvement: the perfect matchings observed involve more protein types (e.g Y-core in P_9), but require a larger volume ratio. The same holds for alternate matchings, see Table 11.

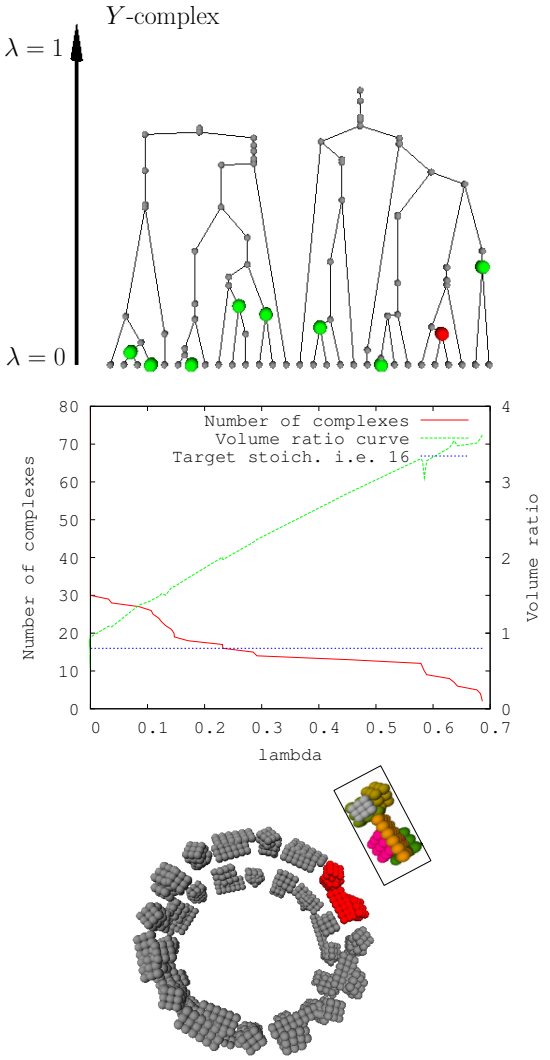


Figure 20: Evolution of the number of complexes associated to the pullout of the Y-complex. Sec13 was removed from the toleranced model. Compare to Fig. 5.

Template; tag	#	$V^{\sim}$	$\min r_\lambda$	$\max r_\lambda$
$G_t(Y)$;P_1	13	Y_X-tail	0.77	0.90
$G_t(Y)$;P_2	3	(Y_X-tail,Nup145C)	0.85	0.87
$G_t(Y)$;P_3, P_4, P_5	7	(Nup145C,Nup84)	0.81	0.88
$G_t(Y)$;P_6	9	(Y_X-short-arm,Nup85)	0.88	0.91
$G_t(Y)$;P_7	1	Y-junction	2.26	2.26
$G_t(Y)$;P_8	16	(Y_X-long-arm)	0.77	1.54
$G_t(Y)$;P_9	6	Y-core	1.10	3.05

Table 10: Perfect matchings of $G_t(Y)$. Sec13 was removed from the toleranced model. The tags P_i match those used in Table 8.

Template; tag	#	$V^{\sim}$	$\mid V^{-} \mid$	$\mid E^{\sim} \mid$	$\mid E^{-} \mid$	$\max \mid E^{+} \mid$	$\min \mid E^{+} \mid$	$\min r_\lambda$	$\max r_\lambda$
$G_t(Y)$;A_1, A_2	9	(Y_X-tail,Nup145C)	4	2	0	1	1	3.03	3.82
$G_t(Y)$;A_3	9	(Y-arms)	4	2	0	1	1	0.94	4.79
$G_t(Y)$;A_4, A_5	7	(Y-main,Seh1)	1	5	0	2	6	1.09	3.45

Table 11: Alternate matchings of $G_t(Y)$. Sec13 was removed from the toleranced model. The tags A_i match those used in Table 9.

7.6.3 Painting Nup133 in Blue Reveals Specific Interactions

In section 4.2, having removed Sec13 from the toleranced model, we analyzed the incidence of repainting Nup133 in blue i.e. we investigated the pullout of the Y-complex without Nup133. As shown on Fig. 21, there are 14 isolated copies of the Y-complex, which is more than the 12 isolated copies obtained in the global analysis of the Y-complex on Fig. 20. The stability of these complexes is heterogeneous as their lifetimes span the range $s(C) = 0.01$ ($\Delta r_\lambda = 0$) and $s(C) = 1.71$ ($\Delta r_\lambda = 10.68$). The intersection of lifetime intervals is empty, a property already observed for the entire Y-complex. However, as shown on the red curve of Fig. 21, there are six complexes that do not merge after $\lambda = 0.80$ ($r_\lambda = 3.62$), which is larger than the two complexes found at the top the Hasse diagram on Fig. 20. Note that if we remove Nup133 from $G_t(Y)$, the template skeleton remains connected. The only possible way to increase the number of complexes that do not merge on the red curve on Fig. 21 is to break the connection between several copies of the Y-complex.

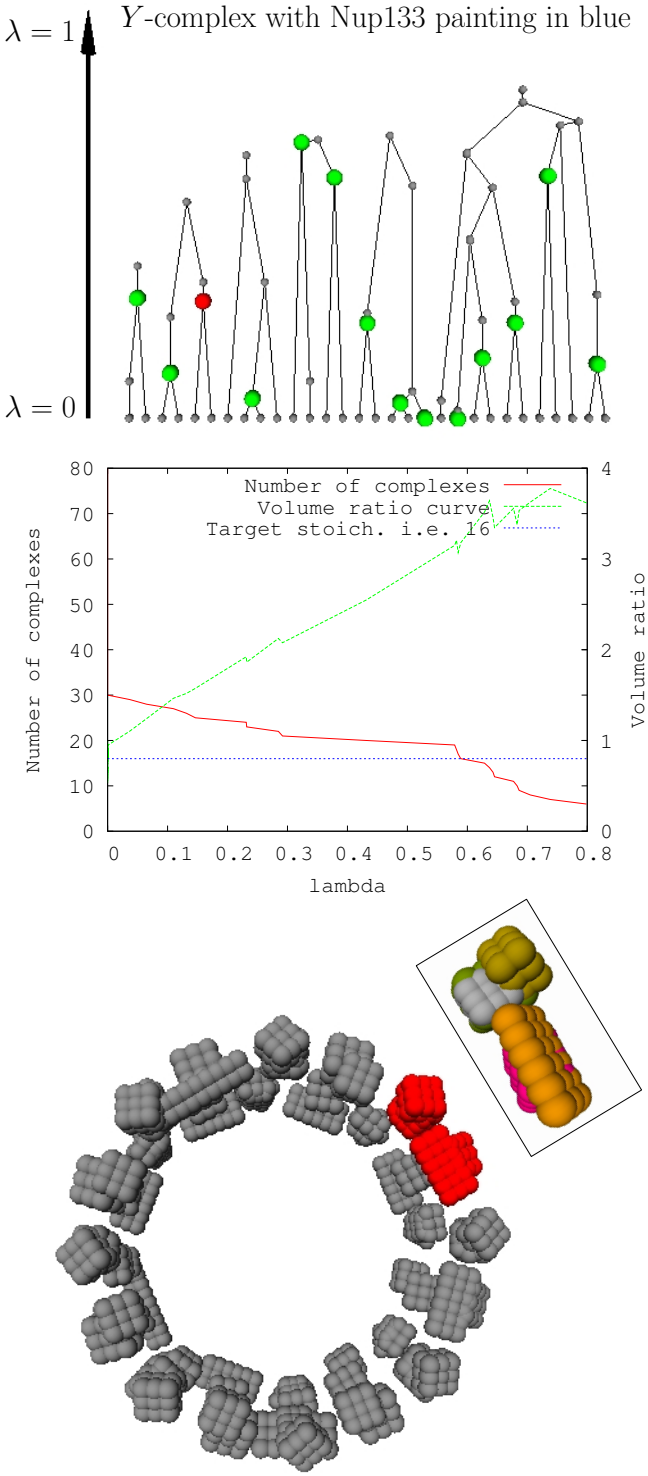


Figure 21: Evolution of number of complexes associated to the pullout of the Y-complex without Nup133. Sec13 was removed from the toleranced model. Compare to Fig. 5.

Contents

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399