Helsinki University of Technology Laboratory of Computational Engineering Publications

Teknillisen korkeakoulun Laskennallisen tekniikan laboratorion julkaisuja

Espoo 2007 REPORT B62

Statistical Analysis and Modeling for Biomolecular Structures

Vibhor Kumar

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Electrical and Communications Engineering, Helsinki University of Technology, for public examination and debate in Auditorium S1 at Helsinki University of Technology (Espoo, Finland) on the 17th of August, 2007, at 12 noon.

Helsinki University of Technology
Department of Electrical and Communications Engineering
Laboratory of Computational Engineering

Teknillinen korkeakoulu
Sähkö- ja tietoliikennetekniikan osasto
Laskennallisen tekniikan laboratorio

Distribution:
Helsinki University of Technology
Laboratory of Computational Engineering
P. O. Box 9203
FIN-02015
FINLAND

Online in PDF format: http://lib.tkk.fi/Diss/2007/isbn9789512288915/

Tel. +358 9 451 4845
Fax. +358 9 451 4830
http://www.lce.hut.fi

E-mail: vkumar@cc.hut.fi

ISBN-13 978-951-22-8890-8 (printed)
ISBN-13 978-951-22-8891-5 (PDF)
ISSN 1455-0474

Picaset Oy
Helsinki 2007

ABSTRACT OF DOCTORAL DISSERTATION	HELSINKI UNIVERSITY OF TECHNOLOGY P.O. BOX 1000, FI-02015 TKK http://www.tkk.fi

Author Vibhor Kumar

Name of the dissertation

STATISTICAL ANALYSIS AND MODELING FOR BIOMOLECULAR STRUCTURES

Manuscript submitted 12/3/2007	Manuscript revised 15.6.2007

Date of the defence 17.8.2007

□ Monograph	☒ Article dissertation (summary + original articles)

Department	Department of Electrical and Communication Engineering
Laboratory	Laboratory of Computational Engineering
Field of research	Computational systems biology
Opponent(s)	Prof. Olli Yli-Harja
Supervisor	Prof. Kimmo Kaski
Instructor	Dr. Jukka Heikkonen and Dr Peter Engelhardt

Abstract

Most of the recent studies on biomolecules address their three dimensional structure since it is closely related to their functions in a biological system. Determination of structure of biomolecules can be done by using various methods, which rely on data from various experimental instruments or on computational approaches to previously obtained data or datasets. Single particle reconstruction using electron microscopic images of macromolecules has proven resource-wise to be useful and affordable for determining their molecular structure in increasing details.

The main goal of this thesis is to contribute to the single particle reconstruction methodology, by adding a process of denoising in the analysis of the cryo-electron microscopic images. First, the denoising methods are briefly surveyed and their efficiencies for filtering cryo-electron microscopic images are evaluated. In this thesis, the focus has been set to information theoretic minimum description length (MDL) principle for coding efficiently the essential part of the signal. This approach can also be applied to reduce noise in signals and here it is used to develop a novel denoising method for cryo-electron microscopic images. An existing denoising method has been modified to suit the given problem in single particle reconstruction. In addition, a more general denoising method has been developed, discovering a novel way to find model class by using the MDL principle. This method was then thoroughly tested and compared with co-existing methods in order to evaluate the utility of denoising in single particle reconstruction.

A secondary goal in the research for this thesis deals with studying protein oligomerisation, using computational approaches. The focus has been to recognize interacting residues in proteins for oligomerization and to model the interaction site for hantavirus N-protein. In order to unravel the interaction structure, the approach has been to understand the phenomenon of protein folding towards quaternary structure.

Keywords MDL denoising, cryo-EM, biomolecular structure

ISBN (printed) 978-951-22-8890-8	ISSN (printed) 1455-0474
ISBN (pdf) 978-951-22-8891-5	ISSN (pdf)
Language English	Number of pages 68p + app. 56p

Publisher Laboratory of Computational Engineering

Print distribution Laboratory of Computational Engineering

☒ The dissertation can be read at http://lib.tkk.fi/Diss/2007/isbn9789512288915

Preface

This thesis was completed in the Computational Systems Biology group of Laboratory of Computational Engineering, Helsinki University of Technology, to conclude my Doctor of Science degree. The enthusiasm and jubilant moments I shared with other researchers, while working on this thesis, will remain with me as precious memories.

I am grateful to my supervisor Professor Kimmo Kaski for his help in proper documentation of all my works and for providing an excellent environment for this work to take place. I had the privilege to have discussions and work with Dr. Jorma Rissanen and I had wonderful experience and enlightment due to his intelligent suggestions. I am grateful to Docents Jukka Heikkonen and Peter Engelhardt for showing me the path to proceed forward. Without their friendly company it would have been hard for me to stay in Finland and complete this work. I would also like to thank my collaborators Dr Pasi Kaukinen and Docent Alexander Plyusnin for working with me to solve a challenging puzzle in the field of virology.

I would also like to thank other people in our laboratory, namely, Dr. Sami Brandt, Tapio Nieminen, Janne Ojanen, Ville Lilja, Danai Laksameethanasan and Dr Mark Anderman. I am thankful to Eeva Lampinen and Kaija Virolainen for patiently taking care of so many practical matters that have come up throughout the years. I would also like to express my gratitude to my other collaborators Eija Heikkilä and Docent Sarah Butcher. Although our common work is not included in this thesis, they have significantly influenced my doctoral studies. Finally, I dedicate this thesis to the memory of my mother, late Mrs. Asha Sinha.

Vibhor Kumar

List of publications

This thesis consists of a general introduction and the following five publications:

(I) Vibhor Kumar, Jukka Heikkonen, Peter Engelhardt and Kimmo Kaski, Robust filtering and particle picking in micrograph images towards 3D reconstruction of purified proteins with cryo-electron microscopy, *Journal of Structural Biology*, 145(1-2): 41−51, 2004.

(II) Pasi Kaukinen, Vibhor Kumar, Kirsi Tulimäki, Peter Engelhardt, Antti Vaheri and Alexander Plyusnin, Oligomerization of hantavirus N protein: C-terminal α-helices interact to form a shared hydrophobic space, *Journal of Virology*, 78(24): 13669−13677, 2004.

(III) Vibhor Kumar, Jukka Heikkonen, Jorma Rissanen and Kimmo Kaski, Minimum description length denoising with histogram models, *IEEE Transactions on Signal Processing,* 54(8): 2922−2928, 2006.

(IV) Vibhor Kumar and Jukka Heikkonen, Denoising with flexible histogram models on minimum description length principle, in B. Enyedi and T. Farhinger, editors, *Proceedings of 13th International conference on Systems, Signals and Image Processing (IWSSIP-2006)*, pp: 167−172, Budapest, Hungary, 2006.

(V) Agne Alminaite, Vera Halttunen, Vibhor Kumar, Antti Vaheri, Liisa Holm and Alexander Plyusnin, Oligomerization of hantavirus nucleocapsid protein: analysis of the N-terminal coiled-coil domain, *Journal of Virology*, 80(18): 9073−9081, 2006.

In addition there is one more supporting publication that is not included in this thesis:

(S1) Alexander Plyusnin, Vibhor Kumar, Olli Vapalahti and Peter Engelhardt, Nucleocapsid protein of hantaviruses (*Bunyaviridae*): Structure and functions, in R. H. Cheng, editor, *Structure Based Study of Viral Replication*, World Scientific Publishing Co., 2007, ISBN: 978-981-270-405-4.

Throughout the overview these publications are referred by the corresponding numerals.

Author’s contribution

The research reported in this thesis is a result of collaboration between myself and the other authors of the included publications. I initiated the research lines leading to Publications I, III and IV. I had a central role in developing the ideas for Publications II and V. I have implemented all computer programs for Publications I, III and IV. I have analyzed all the data and performed simulations in all publications except Publication V. I am an equal contributor to Publication II and I have contributed significantly to the writing about computational methodologies in Publications I, II, III and IV. I have also been actively involved in developing the models presented in Publication V and verified its analytical results.

Contents

Abbreviations

Å	Ångström
3D	three dimensional
2D	two dimensional
CA	correspondence analysis
cryo-EM	cryo-electron microscope
C-terminal	polypeptide chain terminal (end) with a free carboxyl group
CTF	contrast transfer function
DCT	discrete cosine transform
DWT	discrete wavelet transform
EM	electron microscope
ln	natural log (to base e)
L(x)	code length of x
MAE	mean absolute error
MDL	minimum description length
ML	maximum likelihood
MSE	mean square error
NML	normalized maximum likelihood
N-terminal	polypeptide chain terminal (end) with a free amino group
NMR	nuclear magnetic resonance
PCA	principle component analysis
PDB	protein data bank
RNA	ribonucleic acid
SNR	signal to noise ratio
SPR	single particle reconstruction
RNA	ribonucleic acid
TEM	transmission electron microscope
$\Gamma(k)$	gamma function

Chapter 1

Introduction

The recent development in deciphering the sequences of human genome and other organisms leaves us still quite far from fully understanding the structure, function, and response these sequences generate at the system level. Such one-dimensional genome sequences project themselves to higher dimensional spaces as larger scale structures of proteins, cell structures, metabolic networks, and their interactions leading to the vast complexity of living things. In addition to the variable complexity between different levels of biological constituents and more macroscopically between species there is individual variability within species. All of this has recently been a driving force in developing affordable and automated 'high throughput' methods to study gene sequences and networks [1], [2], protein and macromolecular structures [3], [4], metabolites [5], and various complexes of them. Such a system level and often holistic approach to biology has coined (or revived) a new concept of systems biology [6], which - to be more medically target oriented like the 'father' of system biology Leroy Hood – is seen as 'predictive, preventive and personalized medicine'.

In order to understand complex biological systems and their behavior in terms of their structure and function requires in many cases the integration of experimental and computational research, in which sense the aim of research of systems biology is the same with that of structural biology [7]. The focus of structural biology is microscopically oriented since it is mainly interested in the architecture and shape of biological macromolecules like proteins and what causes them to have the structures they have. This is of great importance in biology since the three-dimensional (3D) shapes of macromolecules are responsible for most of the functions of a cell. The shape of a molecule is its "tertiary structure", which depends in a complicated way on the molecule's basic composition, or "primary structure." Then as the quotation by Harrison [7] states 'biology rests on structural observations' there is a great need for new research methodologies and their development for faster and more accurate analysis of biological systems. Even in case of much more limited task for studying protein structures and their interactions – being one of the main goals of this thesis – is so broad that using a single approach or method is not sufficient to get holistic or detailed picture of the system.

In addition to these challenges many of the experimental methods do not have sufficient measurement resolution, for example because biomolecules are so small that even with the most advanced light microscopes it is impossible to study their details. To overcome this difficulty, methods with higher resolution could be used but also methods involving measurements on vast numbers of identical molecules at the same time, like it is done in X-ray crystallography, NMR, and electron microscopy (EM). These methods are most often used to study the static "native states" of macromolecules but they are also varyingly used to study nascent or denatured molecules. However, it may turn out that resolution-wise even this may not be sufficient for deciphering the details of biomolecules. Thus there is need to combine various approaches, experimental and computational, to get more detailed images of macromolecular assemblies of interest. This thesis aims to do just this by combining the experimental cryo-electron microscopy (cryo-EM) imaging and tomography method with novel computational signal and image processing methods.

1.1 Background for biomolecular structure determination

In order to determine the structure of a biomolecule, especially in case of proteins and protein complexes, one can use either a direct approach of an experiment-based sampling and its data processing or indirect computational modeling-based approach. In the former approach the data can be collected by using various experimental techniques like X-ray crystallography, NMR, and EM, each with its own characteristics and thus setting limitations to the method [4] in usability. In the later approach one uses already known structures and thus the obvious limitation is the availability of known structures. Hence the task for methodological development is still very challenging as is evident by the fact that the structures of numerous known and unknown protein still remain to be solved [4]. Next the three main experimental techniques are briefly described but putting emphasis to EM due to it being the main method for generating data in this thesis.

One of the most commonly used experimental techniques to determine protein structure is X-ray crystallography which is based on X-ray diffraction pattern data, obtained from crystalline samples [8]. The pattern is influenced by electron density of the sample crystal. The patterns of amplitudes and phases in the diffraction data are used to calculate an electron-density map of the building units of the crystal. The resolution of the electron-density map calculated from the diffraction data is dependent upon quality of the protein crystal. However, the time needed for protein crystallization and the availability of appropriate particle accelerators used for X-ray generation, has discouraged many researcher to use it. The other discouraging factor is that the determined structure can not be trusted fully due to lack of natural conditions for proteins in crystals. The most stringent limitation of the use of X-ray crystallography is that some proteins can not be crystallized or finding the right conditions for crystallizing the protein is itself a tedious task.

Another experimental technique to determine protein structure is NMR (Nuclear Magnetic Resonance) which can produce chemically and structurally specific spectroscopic data of proteins. NMR-based protein structure determination [9] is currently possible both in aqueous solution and solid state and it exploits the fact that each nucleus in the molecule experiences a distinct chemical environment and thus shows a characteristic chemical shift when an external magnetic field is applied. The resulting conformation-dependent dispersion of the chemical shifts is measured and used to derive structural constraints for NMR-structure calculation. However, the limitations of NMR spectroscopy as the result of low inherent sensitivity [4] of the technique, limits its use for getting high resolution structure of certain proteins, usually not more than approximately in mass 40 kDa [10].

The third experimental technique to determine protein structures is EM, which is the main method for generating micrographic image data in this thesis. In its original form of transmission electron microscopy (TEM) [11] technique, a high voltage electron beam, controlled by magnetic lenses is focused through a thin layer of a prepared sample on a thin film on a supporting grid, to form a highly magnified image of the specimen. Since the wavelength of electrons is at sub-Ångström (Å) level the practical achievable resolution of an image could even be 1 Å for example with a 300kV TEM. In ordinary TEM the sample has to be dehydrated to avoid the evaporation of water in the vacuum conditions of the microscope column. However, to keep samples in hydrated native state and to avoid evaporation in the microscope, the technique of using vitrified samples has proven successful. This technique is termed "cryo-electron microscopy", in which the sample used in the TEM is rapidly frozen to cryogenic temperatures (of around $-196\ ^{o}C$) in order to avoid crystallization of the water present, i.e. to form amorphous ice [11], [12]. The macromolecules in the vitrified sample in the holes of carbon film do not touch the adhering surface so they have less chances of deformation of molecular 3D configuration. In the cryo-EM the vitrified samples are kept frozen by using liquid nitrogen (or even liquid helium) that protects them also from radiation damage [12]. With cryo-EM it is also possible to obtain images of trapped samples in different dynamic states like viral fusion, different states of cell components and interaction of protein assemblies [13]. This is not usually possible by using X-ray or NMR based methods. One can use the cryo-EM images to make 3D reconstructions of biomolecules by applying single-particle reconstruction (SPR) or electron crystallography techniques.

The SPR [12] approach using EM image projections of proteins is emerging as an affordable and rapid way to solve protein structures in their native state. It has aspects of electron tomography [13] where a 3D reconstruction of sample is created from tilted 2D images, but SPR method has potential to provide even near-atomic resolution using hundreds of thousands of projection images. However, since SPR uses cryo-EM images it is not free from problems. It has limitations in different steps of reconstruction and processing the EM images. On the other hand, the electron crystallography method exploits similar features as X-ray crystallography, that the periodic ordering of proteins in two dimensions as well as along one-dimensional helices can be used to determine important structural features [11]. The periodically ordered protein structures in a two dimensional (2D) plane are referred to as 2D crystals. In comparison to 3D crystals, the 2D crystals are easier to produce for certain kinds of proteins and can also give high-resolution structures with EM techniques. However, the methods involved in this are not necessarily rapid, for example, it took years to obtain the 2D crystals for NhaA protein that were used to generate its 3D reconstruction [14].

As mentioned above, a totally different approach to determine protein structure is to use entirely computational techniques by utilizing the knowledge of known protein structures. This approach relies on molecular modeling methods, sequence-pattern matching and optimization methods. One of the methods using purely computational approach is called protein threading [15] where the target sequence is threaded through the backbone structures of a collection of template protein structures to show sequence similarity partially or fully. Recently such methods have been used to support or complement the experiment-based methods described above [7].

The structure of protein complexes and interaction sites in them also belongs to the realm of structure determination of biomolecules. In this for example, finding protein interaction structure to high resolution is a very challenging task by using solely the X-ray crystallography method. To do it properly one would need to have crystals of two or more proteins in a tight complex. One of the structure determination methods is protein threading which would work only for one protein not for the interacting structures. Also, it is well known that the protein folding and interactions involve forces at quantum level, in which case the use of conventional laboratory methods or structure determination is insufficient to understand the real forces involved. It is now quite well accepted that integrated approaches that combine biochemical and computational analyses with reconstructed structures have increasingly important roles in providing the framework for understanding the structure and function of biomolecules.

1.2 Purpose of the research

The main research goal in this thesis is to find and develop an efficient way of 3D-structure determination of proteins and biological complexes by using cryo-EM SPR methods and by modeling the interaction structures of protein. A schematic diagram of different methodologies used in this thesis is shown in Figure 1.1.

The SPR method for making 3D reconstruction of proteins and biological complexes has shown promising results for some very symmetric structures [16]. However, high-resolution reconstruction using cryo-EM images for asymmetric biological structures still poses a formidable challenge. Various factors responsible for limiting the resolution are dependent upon contrast transfer function (CTF), noise and alignment techniques [12]. The CTF is an image aberration oscillating function generated in cryo-EM due to different factors like acceleration voltage for electrons, spherical aberration of magnetic lenses and defocus values of sample grid [12]. At large focus level CTF has more oscillations in high spatial-frequency domain causing disruption and loss of high resolution information [11], [12]. In images taken close to focus, high-resolution information dominates the low resolution information so the noisy particles are no longer discernible against the background [12], hence causing problems in alignment. Hence there is a need for denoising methods to highlight the features of samples without loosing its high frequency details. Several studies for CTF correction and alignment techniques and their development have been adopted as integral parts of SPR method. However, noise removal from the 2D cryo-EM images still needs considerable attention as it can improve the efficiency of other steps in SPR

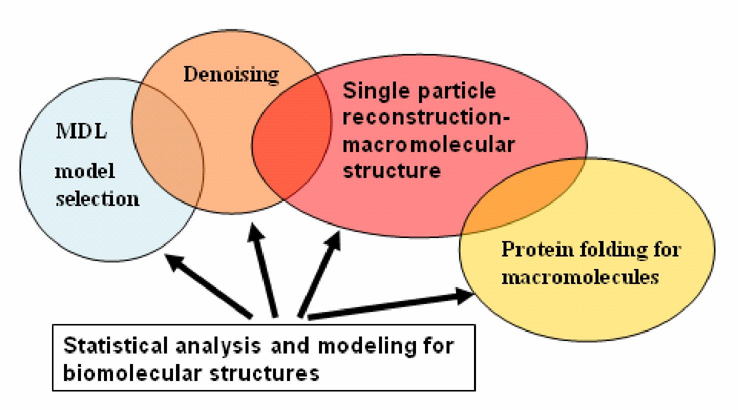
Denoising
Single particle reconstruction-macromolecular structure
MDL
model selection
Protein folding for macromolecules
Statistical analysis and modeling for biomolecular structures

Figure 1.1: Schematic diagram representing different scientific methods used in this thesis for augmenting current understanding and developing methodologies involved in solving structures of biomolecules, ranging from minimum description length (MDL) principle to protein folding. Overlapping regions describe the common grounds of research between different methods.

method tremendously. The state of the art in SPR is such that its methods have matured so much that if they are supported by an optimized denoising method, there could be a paradigm shift in the field of structure determination of biomolecules. Inspired by such speculations, the research done here is to analyze the noise in cryo-EM micrograph images and to develop methods for removing noise to enable better picking and alignment of particles from cryo-EM micrograph images. This leads to very significant improvements in high resolution structure determination of proteins and their complexes. The denoising methods are developed theoretically such that they can be applied in other fields as well, since noise has similar statistical properties in many applications. The concept followed here for denoising is based on the minimum description length (MDL) principle [17] which is strongly connected to information theory and statistics through the interdependencies between code length and probability distributions. For the denoised micrographs, some new particle picking methods have also been introduced in Publication I as discussed in Chapter 3.

The other prominent research question considered in this thesis is how to predict protein structures involved in the quaternary structure of the proteins and how to validate such predictions? The quest for doing such a prediction leads to study of forces and patterns involved in protein interactions. Here, the study has been done with hantavirus nucleocapsid protein whose structure determination through crystallization is still under progress. In our study we have not only predicted its interaction structure but also explained the role of amino acids and different forces involved in the interaction. Similar approaches are needed for many proteins and complexes like lipoproteins and pathogenic viral proteins which need to be studied in order to advance to the stage of targeted drug development, being one of the key ideas of systems biology. The fulfillment of this promise is near since it is soon possible to get insights into the structures of the targeted proteins even at the level of individual amino-acid residues.

Chapter 2

Minimum description length based denoising

In signal processing in general and specifically in 2D image processing, one of the main tasks is to reduce noise to improve signal to noise ratio (SNR). The search for efficient signal and image denoising and image restoration methods is important for many applications ranging from the field of movies [18] to the field of deep oil exploration [19]. It is a valid challenge crossing the borders of statistics, filter design and functional analysis. Denoising is also an important step in cryo-EM SPR in order to avoid false solutions and to achieve faster convergence and higher resolution of reconstructed 3D structures of biomolecular samples.

A search across the literature frames a view that many researchers have been trying to increase SNR in cryo-EM related data but the detailed survey reveals the fact that only few of them have been trying to tackle this problem by treating each 2D image individually [20], [21]. Most of them deal with the problem of denoising the reconstructed 3D volume [22], [23], [24], [25], which is an average of original images. The averaging based denoising method proposed by Mielikäinen and Ravantti [26], uses information from many images and relies heavily on finding similarity between them. Such averaging procedures using alignment partially or fully are already embedded in the steps of SPR. Hence the average-based methods do not provide true additional benefit to SPR. For SPR the denoising of individual 2D images has to be done before the steps of alignment or 3D reconstruction to improve their efficiency to have full utility of increasing SNR. In order to denoise 2D cryo-EM images, the first problem to be tackled is to define noise and recognize sources of noise. In a pioneer research work Doerschuk [27] has quoted that noise due to image recording process in cryo-EM images has a Poisson distribution due to electron counting statistics. However one has to be aware of that unlike TEM images, noise in cryo-EM images is not only due to image recording process but also due to the vitrified buffered water solution (amorphous ice) around sample particles. Several other investigators have also tried to handle cryo-EM noise by assuming a distribution for it. The assumption of noise distribution being Gaussian has been commonly used [27], [28], but no real proof has been presented. In fact, noise in cryo-EM images circumvents randomness of electron counting statistics, inelastic collisions of electrons with amorphous ice [29] and its varying thickness [30]. It also seems to be overlapped with CTF aberration [12] and as discussed in Chapter 3, the calculation of a clear separation line between CTF aberration and noise is not a trivial task.

To understand different approaches for denoising and their potential to filter cryo-EM images this chapter gives an overview of denoising methods. Later it explains the MDL principles with its application to improve SNR. The approaches by other researchers to denoise cryo-EM related data, are mentioned with an overview of denoising methods.

2.1 Denoising methods

To have an understanding about noise present in the observed data y, it can often be written in following mathematical form

$$y = A \otimes x + \varepsilon \, , \tag{2.1}$$

where A is the point spread function (PSF) distorting the original data x by convolving with it. The second additive component ε is noise which is the unwanted random component. Here the PSF matrix (A) is without loss of generality, taken as unity and data matrices y and x and noise ε are all taken as 2D matrices. Conventionally noise has been considered as independent and identically distributed (i.i.d) for the sake of simplicity. On the basis of the central limit theorem most methods claim noise to have Gaussian distribution, but this assumption has helped researchers only to some extent in real world applications. Noise can be generated due to the source of the data itself or due to the observation equipments. Every denoising method has its own assumption about the source and statistical property of the noise.

Currently, most of the proposed image denoising methods can be classified according to their original concepts as local smoothing filters or transform domain filters or as combination of them. Local smoothing methods are generally believed to be based on isotropic linear filtering, anisotropic filtering or total variation minimization, which consider neighborhood of the pixels according to their spatial location. Some local smoothing methods can also be based on the neighborhood of a pixel according to the gray level values such as in case of Susan filter [31]. The transform domain filters use basis vectors to project the image in to another frame in order to separate the image as a set of non-noisy coefficients and noisy coefficients.

Spatial filters

In isotropic linear filtering [32] one does the smoothing of the image by convolving it with the Gaussian kernel of type

$$G(x) = \frac{1}{(2\pi h^2)} e^{-|x|^2 / 2h^2} \, . \tag{2.2}$$

Here, the noise reduction exploits the fact that the neighborhood window of size h involved in the smoothing is wide enough, so that the noise is reduced by averaging. The Gaussian convolution works optimally on smooth and regular data but performs poorly on other parts of the image, like edges or textures, where the Laplacian gradient of the image is large, producing a blurred image.

On the other hand an anisotropic filter [32] avoids the blurring effect of the Gaussian envelope by convolving the image y at pixel (i, j) only in the direction orthogonal to the image gradient $D_y = (\partial y/\partial i, \partial y/\partial j)$. In this case the Gaussian kernel is similar to isotropic linear filter except at the location where the gradient $D_y \neq 0$. The approach of such filter is an extension to the work presented by Perona and Malik [33] where the properties of the physical process of diffusion have been used to enhance edges and the sharp feature of the image. The noise at pixel (i, j) according to anisotropic filter is

$$\varepsilon(i,j) = -\frac{1}{2}h^2 \, |D_y| \, curv(y(i,j)) \, , \tag{2.3}$$

when $D_y \neq 0$ and $curv(y(i,j))$ denotes the curvature, i.e. the negative inverse of the radius of curvature of the level curve passing through pixel (i, j). There are many versions of anisotropic filters, achieving an asymptotic estimate equivalent to the one in equation (2.3), one of which is the median filter [34], [35]. The method proposed recently by Fernandez and Li [25] for denoising 3D cryo-tomograms, uses such an approach to enhance features in 3D electron density maps. Their method is based on the hybrid of edge enhancing diffusion and coherence-enhancing diffusion (CED) approaches first introduced by Frangakis and Hegerl [23]. It also includes a 3D plane-enhancing CED diffusion mode, which enhances surface-like or plane-like local structures in tomograms.

A different kind of spatial filter called the 'total variation' [32] minimization method assumes the original non-noisy image as a collection of connected objects having smooth contours and edges such that the image variance is less inside the objects but increases across the boundaries. For image data y defined in space Ω, the restoration problem using total variation minimization can be written as the solution of following form

$$TVF_{\lambda}(x) = \min_{x}\{TV(x) + \lambda(\frac{1}{2}\|x - y\|^2)\} \, , \tag{2.4}$$

where the regularization function $TV(x) = \sum_{\Omega} |\nabla x|$ represents the edge information in the image. The Lagrange multiplier λ works as trade-off between the smoothness and the sharpness terms. The noise at pixel (i, j) is defined as

$$\varepsilon(i,j) = -\frac{1}{2\lambda} curv(TVF_{\lambda}(x(i,j))) \, . \tag{2.5}$$

Similar to the anisotropic case, edges are maintained if their curvature is small. However, detailed non-connected boundaries and textures could be smoothed out if λ is too small.

There are also other neighborhood filters [32] which take into account grey level values to define neighboring pixels. They are fully non-local algorithms, since pixels belonging to the whole image are used for the estimation of the pixel at location (i, j). One such filter, Susan filter [31] proposes a closed formula for the denoised image intensity $x(i,j)$ at pixel (i,j) that can be derived from a noisy image y, as

$$x(i,j) = \frac{1}{C(i,j)} \int_{B_\rho(i,j)} y(i',j') e^{\frac{-|y(i',j')-y(i,j)|^2}{h^2}} di', dj', \tag{2.6}$$

where $B_\rho(i,j)$ is a circle of radius ρ around the centre (i,j) and

$$C(i,j) = \int_{B_\rho(i,j)} e^{\frac{-|y(i',j')-y(i,j)|^2}{h^2}} di', dj' \tag{2.7}$$

is a normalization factor at pixel (i,j).

A similar method with modification of the neighborhood filtering approach is bilateral filtering. This filter is worth mentioning here because in a recent publication Jiang et al. [20] have claimed their bilateral filtering method to be working efficiently on EM 2D images and 3D volumes. They have shown that bilateral filtering improves the feature detection in 3D volumes and they demonstrated it using a helix-hunter [36] method for the computational identification of helices in P8 protein of rice dwarf virus structure with resolution around 8 Å. In principle their formula for denoised image intensity $x(i,j)$ at pixel (i,j) that can be derived from noisy image y, can be written in the following form

$$x(i,j) = \int (e^{\frac{-((i'-i)^2+(j'-j)^2)}{2\sigma^2}})(e^{\frac{-(y(i',j')-y(i,j))^2}{h^2}}) y(i',j') di', dj', \tag{2.8}$$

which is quite similar to the equation (2.6), except that instead of taking a circular window $B_\rho(i,j)$ it uses a Gaussian function with standard deviation σ to control the smoothing region in the image. The other term with parameter h controls the discrimination between true features and noise with support from the assumption that larger pixel intensity value variations are mainly due to true features while smaller pixel intensity value variations are due to noise.

The spatial filters assume the difference only between sharp boundaries and plain objects; they do not pick patterns or special features. Most of the spatial filters need an external parameter, which could be the variance estimate of noise or the number of iterations needed for cleaning the noise. The local spatial filters always leave artifacts in the denoised image. For instance, the total variation method tends to find spurious edges in the noise, shown in Figure 2.1(g) and similarly bilateral filtering creates unwanted spurious local artifacts if the smoothing region is small. Picking pattern and highlighting some desired informative parts in images is not possible using

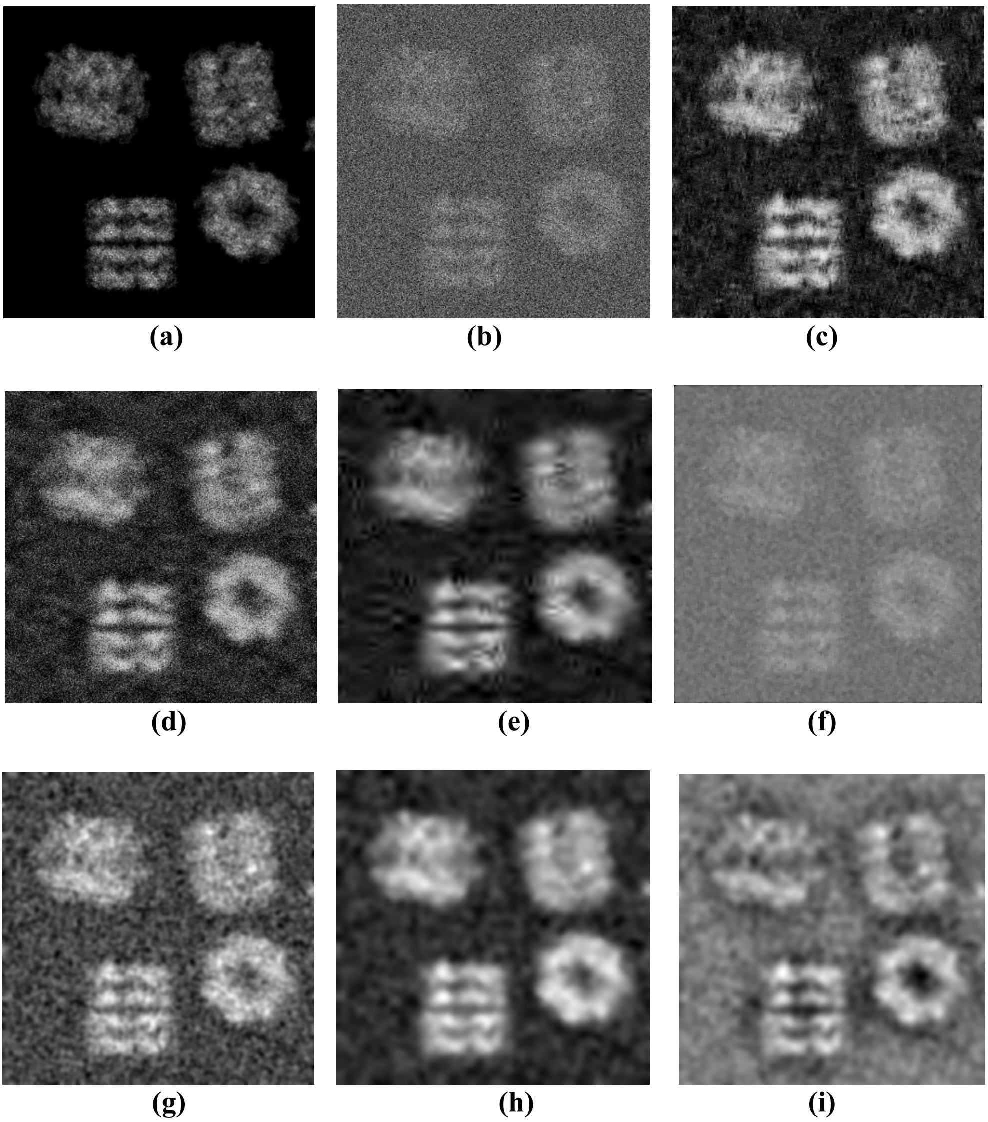

Figure 2.1: Nearly equal success of different methods in removing simulated Gaussian noise added to image having projections of the prokaryotic chaperone protein GroEL 3D volume (from protein data bank [4]). (a) original image without noise, (b) noisy image with Gaussian noise, (c) the denoised image using modified MDL-histo of Publication IV and wavelet level-wise weighting method proposed in Publication I, (d) denoised image using NML based MDL denoising method and wavelet level-wise weighting, (e) denoised image using wavelet shrinkage using BayesShrink [44] based soft thresholding, (f) filtered image using median filtering [35], (g) image denoised using total variation method, (h) denoised image by Bilateral filtering [20], (i) band-pass filtered image.

spatial filters. On the other hand, the transform domain filters have the capacity to pick patterns and features by the use of basis vectors suitable for the feature and the pattern in the image. A collection of results obtained by using different methods for simulated Gaussian noise as shown in Figure 2.1, gives an overview of their performance. The results shown in Figure 2.1 (c), (d), (e) and (i), have been produced using transform domain filters.

Transform domain filters

Most of the transform domain filters work such that initially the image is transformed into coefficients. Some transform coefficients are then thresholded to zero or rescaled down to produce a set of non-noisy coefficients which are then inverted back to obtain the denoised image. One such Fourier-domain filter is Fourier-Wiener filter [37] where the Fourier basis weights global image characteristics more than local ones and create periodic pattern as artifacts in the denoised image. To avoid this effect, other basis vectors are used to take into account more local features; some of such transforms are: wavelet [37], local discrete cosine transform (DCT) [37], bandlet [38] and curvelet [39] transforms.

The wavelet transform is an example of a transform which has adapted representation for localized features as well as for long range smoothing functions. They have advantages over traditional Fourier methods in analyzing physical situations, where the signal contains discontinuities and sharp spikes. Wavelet transform is done by decomposing the signal by convolving it with wavelet function φ(t) (i.e. the mother wavelet) for each level. [40] The wavelet function φ(t) is translated and scaled according to equation

$$\varphi_{s,l}(t) = \frac{1}{\sqrt{2^s}}\varphi(\frac{t-l}{2^s}) \tag{2.9}$$

where s is the wavelet level and l is the location for which wavelet coefficient is being calculated. The complex conjugate of the translated and scaled wavelet function is then correlated with the data $x(t)$ to get wavelet coefficient at location l and level s as

$$w(s,l) = \int_{-\infty}^{\infty} x(t)\varphi_{s,l}^{*}(t)dt\,. \tag{2.10}$$

In order to avoid the infinite number of levels needed, a compact support scaling function ψ(t) is used after last level to represent the rest of the information in the signal [40]. A wavelet function satisfies the orthogonality property such that its inner product with itself is unity and inner product with other functions obtained after dilating or shifting it, is zero [40]. Several wavelet functions have been derived to highlight different aspects of data. One of the wavelet function φ(t) called Mexican Hat function is represented in equation (1) of Publication I. Wavelet transformations can be classified in to two major types as continuous wavelet transform (CWT) and discrete wavelet transform (DWT). The CWT is computed just as explained above i.e. changing the scale of the wavelet function and convolving with the signal at every level. In the discrete case, filters of different cut-off frequencies are convolved with signal after scaling the signal to different levels. For DWT the signal is itself scaled at every level and convolved with the scaled wavelet function. The scaled and translated wavelet function for DWT can be demonstrated as

$$\varphi_{s,l}(t) = \frac{1}{\sqrt{2^s}} \varphi(2^{-s} t - l) \tag{2.11}$$

which is almost similar to the one in equation (2.9) except that location in signal is not scaled as signal itself is scaled for level s. DWT results in lesser and non-redundant coefficients and is quite useful in applications like data compression and feature extraction [37]. However, CWT which produces redundant coefficients has its own applications like time-frequency analysis and denoising [40]. The discussion in thesis applies to both types of wavelet transform, however notice that the orthogonality criteria is not necessary for derivation of method proposed in Publications III and IV.

The lossy compression [37] based approach for denoising using wavelets has been studied by many researchers in different ways. Most of them choose one threshold for one wavelet basis vector level coefficients and threshold coefficients below the threshold. The pioneer proposal given in a seminal work of Donoho and Johnstone, has been followed by series of other proposals by the same researchers with different thresholding criteria called the universal threshold [41], SureShrink [42], VisuShrink and RiskShrink [41]. These works were based on obtaining near-optimal properties in the minimax sense by minimizing the worst case risk over large class of signals using nonlinear shrinkage estimators with the help of thresholds. The risk associated with threshold rule to achieve estimate $\hat{x}$ is described as

$$R(x,\hat{x}) = E\{\| x - \hat{x} \|^2\} \tag{2.12}$$

where the expectation is calculated over the noise distribution. An estimate that minimizes the risk can be written as $\hat{x} = inf_{\hat{x}} sup_x R(\hat{x}, x)$ [41]. The optimal operator would attenuate all noisy coefficients in order to minimize the risk but Donoho and Johnstone explained a simple kind of thresholding criteria of 0 or 1, which can help in minimizing the worst case risk. The universal threshold λ_U, derived by them as optimal in the minimax sense is given by

$$\lambda_U = \hat{\sigma}\sqrt{2\log(N)}, \tag{2.13}$$

where $\hat{\sigma}$ is the estimate of standard deviation of noise and N is the number of data points. The standard deviation of noise is usually estimated by the median absolute deviation (MAD) estimate

$$\hat{\sigma} = median(| c_1^{N_1} |) / 0.6745 , \tag{2.14}$$

where, $| c_1^{N_1} |$ is the vector of absolute values of the wavelet coefficients at the finest detail level with N_1 coefficients and 0.6745 being a normalization factor due to the Gaussian noise assumption. Anticipating the weakness of such an approach, Donoho and Johnstone came up with new approaches for finding thresholds, like Sureshrink, Visushrink [42], [43]. Wavelet denoising has been attempted using various Bayesian approaches; one of them is adaptive thresholding using generalized Gaussian prior, known as BayesShrink [44]. Recently another class of threshold rules has stemmed from information theoretic ideas related to minimizing the description length and complexity distribution [17], [45] and this we will discuss in next section.

2.2 Minimum description length principle

In MDL based denoising technique it is assumed that the complexity of noise is greater than the complexity of the smooth data. Another way to describe this is that MDL principle assumes noise as the incompressible part of data and meaningful information-bearing signal as the compressible part, given the model class. The model class serves simply as a language by which the properties we wish to learn can be described. The modeling approach to select the best model and model class, can be formalized in different ways. One such way can be explained in terms of complexity. The concept of complexity was first described by Solomonoff [45] but the Kolmogorov complexity has quite often been referenced [46]. The Kolmogorov complexity of a binary string $X^n = x_1, x_2, ...x_n$, is defined as the length of its shortest description $p_U(X^n)$ in the language of a universal computer *U* generating the input string [45],

$$K_U(X^n) = \min | P_U(X^n) |. \tag{2.15}$$

Kolmogorov complexity gives a universal model which can also be represented with a probability term [46] as

$$P_K(X^n) = C 2^{-K_U(X^n)} , \tag{2.16}$$

where, $C = 1 / \sum_{y \in B} 2^{-K_U(y)}$ is a finite normalization term calculated over the set of all binary strings *B*. The universal distribution $P_K(X^n)$ in equation (2.16) has the property to represent any computational probability distribution $Q(X^n)$ so that

$$P_K(X^n) > AQ(X^n) , \tag{2.17}$$

for some constant A. The major drawback of Kolmogorov complexity measure is its non-computability, so in practical applications it has always to be approximated [45]. The current version of MDL principle has evolved around this algorithmic complexity theory. In MDL theory instead of working on all possible programs to find the complexity, a model class is defined such that it is restricted to be a set of parametric probabilistic models.

The Kolgomorov complexity can be presented in terms of conditional complexity $K(x \mid y)$ as the length of the shortest program that generates string x from another string y. It can be shown [46] that

$$K(x^n, y^n) < K(x^n) + K(y^n \mid x^n). \tag{2.18}$$

To describe sequence x^n we can do it in two parts as $K(A) + \log|A|$, where first part describes optimally a set A with the number of bits given by the Kolmogorov complexity $K(A)$ and the second part describes x^n in A with $\log|A|$ bits, with $|A|$ denoting the number of elements in A.

Another concept with similar idea, called stochastic complexity [47], is also used as a founding terminology for explaining MDL and it can be defined as the shortest code length for the data given a probabilistic model class. In order to understand further and avoid non-computability problem, consider a family of probability models $M_\gamma = f(x^n, \theta^k, \gamma)$, where $\theta^k = \theta_1, ..., \theta_k$ are parameters and γ is the structure index containing information about the subset of included parameters. To encode x^n, θ^k, γ we need a prior for the parameters and the real valued parameters must be quantized to get finite code length for them. For coding prior distribution, hyper-prior distributions or hyper-parameters [48] may be needed and so on. To truncate this chain of hyper-priors, a constraint can be used such that the highest prior included can be calculated without adding it to the total code length. A fundamental construct in the theory is a universal distribution $\hat{f}(x^n; M_\gamma)$, which has no other parameters than the structure index that has the information about the included parameters. The code length *L(a)* for any object *a* defined by a probability *P(a)* is taken as $L(a) = \log 1/P(a)$ [47]. This implies that with a universal model and a code length for the structure index $L(\gamma)$ the optimum index $\hat{\gamma}(x^n)$ can be searched for or estimated by using the following minimization

$$\min_{\gamma}\{\log(1/\hat{f}(x^n; M_\gamma)) + L(\gamma)\}. \tag{2.19}$$

Two-part MDL

As shown before, to encode a data sequence x^n, the MDL principle needs generally two main parts, which can be stated as follows; given the data D and the model M, the total code length *L(D,M)* can be written as

$$L(D,M) = L(D|M) + L(M), \tag{2.20}$$

where *L(M)* is the length of the description of the model and *L(D|M)* is the length in bits of the description of the data encoded with the help of model *M.* The explanation behind this code length based model selection is simple, see Figure 2.2. If the quality of fitness of the model to represent the data is improved by making the model more complex, the code length *L(M)* needed to encode the model increases. The minimization of the total code length then finds the balance between model complexity and accuracy.

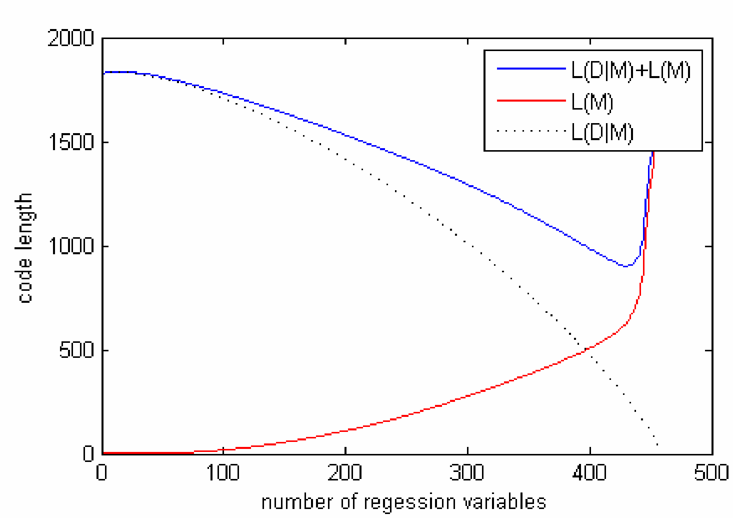

Figure 2.2: The plot of code length with respect to number of regression variables (model). *L(M)* is the code length of the regression model and *L(D|M)* is code length of the data with given model M represented by number of regression variable in its structure index.

Supported by the discussions above, it can be shown that minimizing the two- part code length

$$L(x^n, \theta^k) = -\log P(x^n; \theta^k) + L(\theta^k) \tag{2.21}$$

gives a statistical fit and minimized code length [17], [45] for the model (the second term). It is equivalent to finding the maximum likelihood (ML) parameters $\hat{\theta}^k$. Since the real-valued parameters must be quantized, the resulting precision δ should also be taken into account. In order to achieve the shortest code length the precision of ML parameters must be optimized. A previous work by Rissanen [49] shows that the optimized truncation precision $\hat{\delta}_j$ is of the order of $1/\sqrt{n}$, when $-\log P(x^n; \theta^k)$ grows proportionally to n. With optimal precision the minimization of the total code length, given the model class is equivalent to minimizing the expression

$$L(x^n, \theta^k) = -\log P(x^n; \theta^k) + \frac{k}{2}\log(n). \tag{2.22}$$

For a parametric density function $q(x^n) = f(x^n, \theta^k(x^n), \gamma)$ the ML code length can be used to judge the level of coding performance by the regret

$$\Re(q) = -\log q(x^n) - (-\log f(x^n; \hat{\theta}^k(x^n), \gamma)), \tag{2.23}$$

and as a measure of optimal model solution of the minimax [47] problem

$$\min_q[\max_{x^n \in X} \Re(q)] = \min_q[\max_{x^n \in X}\{-\log q(x^n) - (-\log f(x^n; \hat{\theta}^k(x^n), \gamma))\}]. \tag{2.24}$$

The unique solution to this minimax problem is the normalized maximum likelihood (NML) density function [47], as a universal model:

$$\hat{f}_\gamma(x^n) = \frac{f(x^n; \hat{\theta}(x^n), \gamma)}{C_\gamma{}^n}, \tag{2.25}$$

where

$$C_\gamma{}^n = \int_X f(x^n; \hat{\theta}(x^n)) dx^n . \tag{2.26}$$

The logarithm of the denominator $\ln(C_\gamma{}^n)$ is a measure of 'parametric complexity' of the model class. Since this must be bounded, it may be necessary to restrict the range X of the integration by hyper parameters. For the Gaussian family, Rissanen [50] describes an elegant renormalization scheme, where the hyper parameters [48], [50] defining the range of the data are optimized and a second normalization is performed such that the resulting code is complete. This renormalized NML can be used for model selection in linear regression and denoising. In denoising, the MDL model selection is performed by considering each subset of the regression coefficients as a model class and minimizing the stochastic complexity $-\log \hat{f}_\gamma(x^n)$ of the data given the model class.

Linear regression

In order to look at the linear regression problem in more detail, we will deal with the data of type $(y_t, x_{1t}, x_{2t}, ..., x_{lt})$ for $t = 1, 2, ..., n,$ for which we wish to find the influence of regression variables x_{it} on the corresponding values y_t of the regression variable $y.$ To find the number of regression variables that are most important we must be able to compare the performance of subsets of different sizes. The regression based linear model can be written as

$$y_t = \beta' x_t + \varepsilon_t = \sum_{i \in \gamma} \beta_i x_{it} + \varepsilon_t , \tag{2.27}$$

where, $\gamma = \{i_1,...,i_k\}$ denotes a subset of the indices of the regression variables and ε_t denotes independent Gaussian process of zero mean and variance $\tau = \sigma^2$. For such a model the observed data $y^n = y_1,...,y_n$ are also normally distributed with the density function

$$f(y^n;\gamma,\beta,\tau) = \frac{1}{(2\pi\tau)^{n/2}} e^{-\frac{1}{2\tau}\Sigma\ (y_t - \beta' x_t)^2} . \tag{2.28}$$

Taking the $X'_\gamma = \{x_{it}; i \in \gamma\}$ as the $k \times n$ matrix defined by the values of the regression variables with indices in γ and taking $Z_\gamma = X'_\gamma X_\gamma = n\Sigma_\gamma$, the ML solution for the parameters is given by

$$\hat{\beta}(y^n) = Z_\gamma^{-1} X'_\gamma y^n \tag{2.29}$$

$$\hat{\tau}(y^n) = \frac{1}{n}\sum_t (y_t - \hat{\beta}'(y^n)x_t)^2 . \tag{2.30}$$

The derivation of NML density function for the given model class, is not a trivial task and has been attempted using different approaches [47], [51]. An exact formula has been derived by Rissanen [50] to avoid unbounded complexity as follows

$$\hat{f}(y^n;\gamma,\tau_0,R) = -\frac{f(y^n;\gamma,\hat{\beta}(y^n),\hat{\tau}(y^n))}{\int_{Y(\tau_0,R)} f(z^n;\gamma,\hat{\beta}(z^n),\hat{\tau}(z^n))dz^n} , \tag{2.31}$$

where $Y(\tau_0,R)$ represents the set of all the data value y^n such that

$$Y(\tau_0,R) = \{z^n : \hat{\tau}(z^n) \geq \tau_0, \hat{\beta}'(y^n)\Sigma\hat{\beta}(y^n) \leq R\} . \tag{2.32}$$

The renormalization process [50] yields another NML density of limited support for each pair of parameters (R,τ_0), which can be written as

$$\hat{f}(y^n;\gamma) = -\frac{\hat{f}(y^n;\gamma,\hat{\tau}(y^n),\hat{R}(y^n))}{\int_{Y(\tau_0,R)} \hat{f}(z^n;\gamma,\hat{\tau}(z^n),\hat{R}(z^n))dz^n} . \tag{2.33}$$

After setting $\tau_0 = \hat{\tau}$ and varying $\hat{\tau}$ over the range $[\tau_1,\tau_2]$ and setting $R = \hat{R}$ and varying $\hat{R}$ over the range $[R_1,R_2]$ and integrating the function $\hat{f}(y^n;\gamma,\tau_0,R)$ it can be shown that the negative logarithm of $\hat{f}(y^n;\gamma)$ is given by

$$-\ln \hat{f}(y;\gamma) = \frac{n-k}{2}\ln\hat{\tau} + \frac{k}{2}\ln\hat{R} - \ln\Gamma(\frac{n-k}{2}) - \ln\Gamma(\frac{k}{2}) + \frac{n}{2}\ln(n\pi) + \ln[\ln\frac{\tau_2}{\tau_1}\ln\frac{R_2}{R_1}] \tag{2.34}$$

The last term in this equation can be ignored if the value of $\hat{\tau}$ and $\hat{R}$ do not vary too much as a function γ. For detail derivation of this NML criteria, see the references [50], [51]. After applying Stirling's approximation to the $\Gamma-$ function we get the NML criterion for $0<k\leq n$ as:

$$\min_{\gamma}\{\frac{n-k}{2}\ln\hat{\tau}+\frac{k}{2}\ln\hat{R}+\left(\frac{n-k-1}{2}\right)\ln\frac{1}{n-k}-\left(\frac{k-1}{2}\right)\ln k+\ln k+\ln[1+\ln(1+k)]\}, \tag{2.35}$$

where k denotes the number of elements of γ. The last two terms represent the code length needed to code integer k which can be ignored for usual regression problems as their value is relatively too small. This criterion has been used in various applications and also in modified forms.

2.3 MDL denoising and wavelet shrinkage

The linear regression based approach presented above can be extended to regression matrix-based transformations of observed data. The regression matrix can represent many transforms like wavelet or discrete cosine transform [37]. Taking example of wavelet transform which is an orthonormal transform, the transformation can be written as

$$c=Wy \tag{2.36}$$

$$y=W^{T}c \tag{2.37}$$

with y describing the data matrix and c describing the wavelet coefficient matrix and W representing the wavelet transformation matrix. If the subsets of the regression variables are taken as model class then the approach of MDL is to find the model which gives minimum code length when used to code the data. In other words, the problem is to choose the subset γ of the basis vectors or the corresponding parameters in such a way that the stochastic complexity of the given data is minimized. The denoised signal can be written as

$$\hat{x}=W^{T}c_{\gamma}(y), \tag{2.38}$$

where W^{T} is the regression matrix and $c_{\gamma}(y)$ is the vector of coefficients corresponding to the optimal model specified by γ. In $c_{\gamma}(y)$ only the selected coefficients are present and it represents the ML estimate $\hat{\beta}$. Other ML estimate $\hat{\tau}$ in equation (2.35) is now given by

$$\hat{\tau}=\frac{\sum_{t=1}^{n}(y_t-\hat{x}_t)^2}{n}=\frac{c'c-\hat{c}'\hat{c}}{n}, \tag{2.39}$$

where $\hat{c}$ is the vector of non-selected coefficients [50]. The criterion for finding the best subset γ is then equivalent to

$$\min_{\gamma}\{\frac{n-k}{2}\ln\frac{c'c-\hat{c}'\hat{c}}{n-k}+\frac{k}{2}\ln\frac{\hat{c}'\hat{c}}{k}+\frac{1}{2}\ln(k(n-k))+\ln k+\ln(\ln(k))]\}. \quad (2.40)$$

Using Parseval's equality, Rissanen [50] has proved that the sum of the squared deviations $\hat{\tau}$ is minimized by the k largest coefficients in absolute value. This can be achieved by sorting the coefficients according to their absolute value and start checking for minimization from the largest value. The sorting would lead to loss of location information of the coefficients, so the locations have to be coded [52]. This leads to the code length to be minimized as

$$\min_{\gamma}\{\frac{n-k}{2}\ln\frac{c'c-\hat{c}'\hat{c}}{n-k}+\frac{k}{2}\ln\frac{\hat{c}'\hat{c}}{k}+\frac{1}{2}\ln(k(n-k))+\ln\binom{n}{k}+\ln k+\ln(\ln(k))]\}, \quad (2.41)$$

which after ignoring small valued terms, can be written as

$$\min_{\gamma}\{\frac{n-k}{2}\ln\frac{c'c-\hat{c}'\hat{c}}{(n-k)^3}+\frac{k}{2}\ln\frac{\hat{c}'\hat{c}}{k^3}+\frac{1}{2}\ln(k(n-k))\}. \quad (2.42)$$

The term k determines the threshold λ for the coefficients, such that all coefficients larger than λ in absolute value are only selected and other coefficients are set to zero. The minimization criteria mentioned above need the regression matrices to be orthonormal. It is quite similar to using threshold by other proposed methods to truncate wavelet coefficients to achieve denoised data, but in most of the other methods the prior information about noise variance should be known.

Such approaches to pick one threshold as the boundary between noise coefficients and data coefficients could perform satisfactorily for low variance of noise. However, when the statistical properties of noise get very similar to that of data or when the distribution of noise is unknown, such single threshold method would be insufficient to demarcate noise from data. But, the question arises how to model the noise if we cannot know or assume any distribution? The answer to this lies in approach of coding the data without assuming any distribution of the data. Using such non-parametric coding and dividing the code in two parts would lead to their automatic separation in to two parts, where one part would have more complexity than the other. The part of the signal with more complexity can be demarcated as noise. The approach used in Publication III follows this policy and employs histogram based coding of the wavelet coefficients. Histograms at each level of the data and for the common noise can be changed to fit the distribution of the data and noise in a prudent manner. This kind of approach automatically fit to the distributions of data and noise, whatever they look like. Instead of using one threshold the selection of bins is done such that

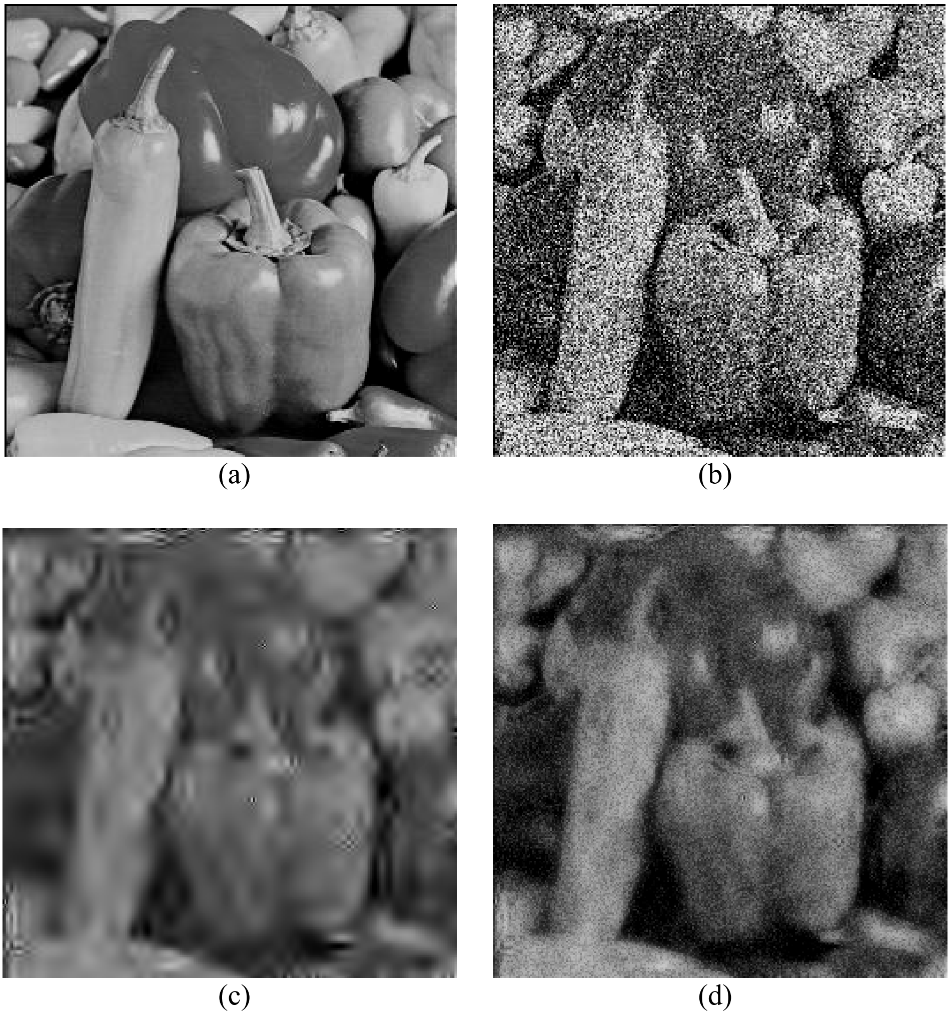

Figure 2.3: Difference between traditional soft thresholding [30] and the proposed wavelet level wise weighing and shrinkage method in Publication I using NML criteria to find the threshold. (a) The original image, (b) noisy image with Gaussian noise, (c) the traditional soft thresholding result [30], (d) denoised image using wavelet shrinking proposed in Publication I.

distribution of coefficients among the data histograms and noise histogram leads to minimum code length. The approach has been demonstrated with equal bin width histograms in Publication III, as well as variable bin histograms in Publication IV. The simulation results have already shown that such method is able to distribute the coefficients among noise and data histograms quite effectively. After the introduction of improvements, like gradual iterative picking and greedy approaches in Publication IV, this method outperformed other wavelet thresholding methods like Bayesthresh [44] and universal threshold [41], in term of image denoising.

There are several possible methods to pick non-noisy data coefficients from the coefficients of original noisy data, in transform domain. However, the assumed non-noisy coefficients may not always represent only the required data. The non-picked coefficients usually retain some features of non-noisy data, such as in wavelet domain the edges and sharp feature coefficients have chances to get categorized as noise. Especially when the noise variance is high the resulting denoised image by most of the thresholding methods are either left with remnant noise or loose lot of genuine information in the image. To avoid such attenuation of denoised image, a new method was developed and presented in Publication I, which calculates the shrinkage factor for each coefficients based on wavelet level-wise weighing. The purpose of this approach is to do denoising and enhance the features of samples without loosing their high frequency information from the cryo-EM images. Thus it has the capacity to enhance the samples in images taken at low defocus value as well as retrieve their high frequency details. Such a method based on code length principle would have a valid explanation as the non-noisy data tends to need less code length. This approach assigns weight to different basis vector layers or wavelet levels on the basis of compressibility. Instead of setting the coefficients below the determined threshold to zero, they are scaled to smaller values and the different wavelet level coefficients are weighted according to the number of their coefficients above the threshold.

2.4 Comparison of methods

The shrinking method proposed in Publication I, seems similar to the soft thresholding scheme proposed by Donoho and Johnstone [42], but in actuality both methods are quite opposite of each other. The traditional soft thresholding approach [42] only helps in avoiding artifacts and adjusting contrast in the resulting denoised images but not in retrieving back the genuine information lost with the noisy coefficients. The difference can easily be seen from the results shown in Figure 2.3. Such an approach is needed for high variance noise of unknown distribution in cryo-EM images, as most of the denoising methods tend to either create a lot of artifacts or completely smooth out the genuine information in the image. This kind of behavior of traditional denoising methods can be explained by the fact that most of them are developed for Gaussian noise and have been tested for simulated easy noise. On observing Figure 2.1, it becomes clear that all the methods tested gave nearly equal performance for Gaussian noise.

The results shown with original cryo-EM images in Figure 2.4 give completely different judgment for different methods. Spatial filters produced sharp and edgy denoised images with some artifacts, while wavelet based denoising methods produced noiseless and little blurred images. The spatial filters simply can not find difference between particle projections and noisy neighboring vitrified water. As shown in Figure 2.4 (e) and (f), bilateral filter [20] and total variation filter take high pixel intensity locations as genuine data and enhance them even if they are actually lying outside the

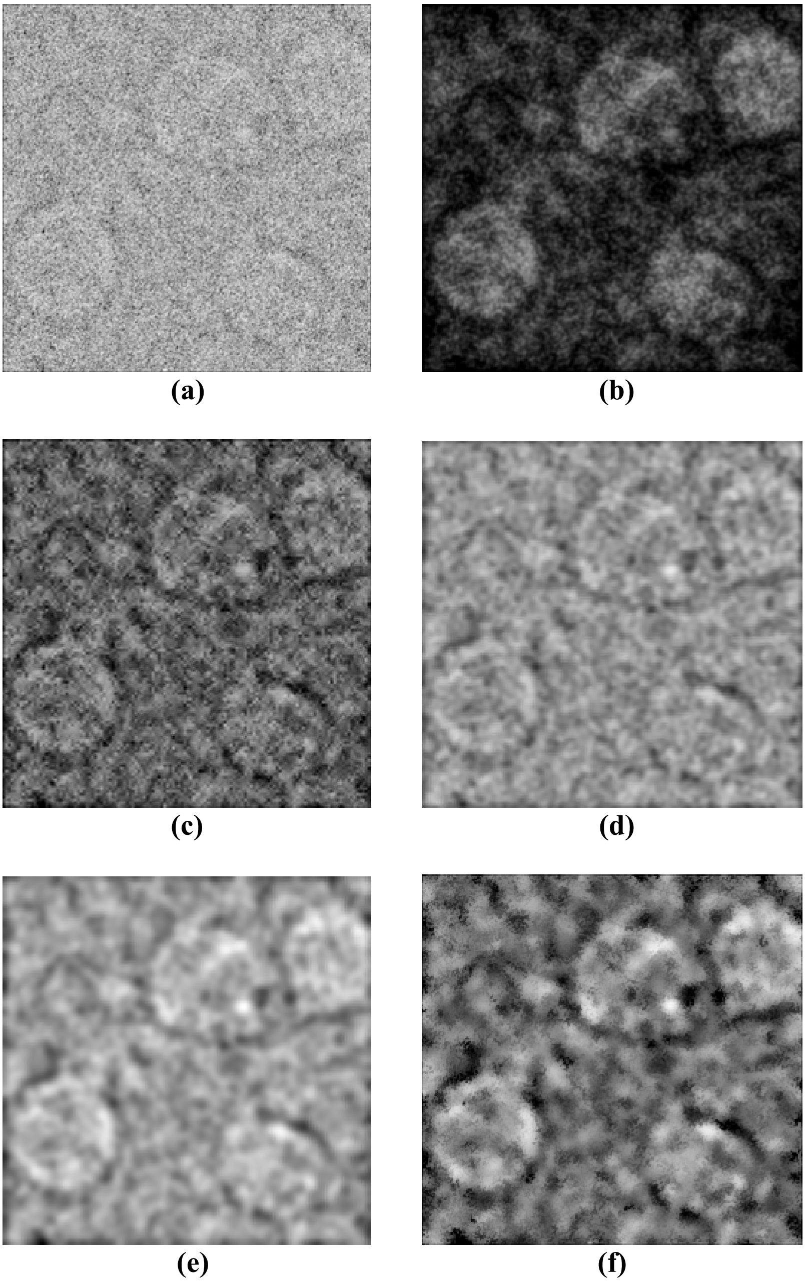

Figure 2.4: Comparison of different denoising methods on real data. (a) original micrograph image of LDL particles (by courtesy of Sarah Butcher, University of Helsinki and Kati Öörni, Wihuri Research Institute), (b) denoised image using NML based MDL denoising method and wavelet level-wise weighting, (c) the denoised image using modified MDL-histo of Publication IV and wavelet level-wise weighing method proposed in Publication I, (d) band-pass filtered image, (e) image denoised using total variation method, (f) denoised image by bilateral filtering.

sample particles. Even frequency domain band-pass filtering (Figure 2.4 d) may give better results than spatial filter as shown in Figures 2.4 (e) and (f).

Chapter 3

Statistical analysis for single particle reconstruction

The SPR method is based on the assumption that the projections used to reconstruct a 3D structure have been made from structurally congruent particles. At the atomic level these particles are dynamic but at nano-resolution level they can be regarded as static. Since the noise in TEM and cryo-EM images enhances the differences among projections, the success of the steps involved in the SPR method depends on their capacity to overcome these differences. For this purpose, it is often necessary to use several thousand projections (particles) for reconstructing 3D object from its 2D projections. Once particles have been picked and centered, they need to be classified and aligned for the reconstruction. These steps provide information on the relationships between different particles. The information can then be used towards putting together a 3D volume. Figure 3.1 gives a schematic representation of the procedures in SPR. This chapter gives an overview of the steps involved in SPR, including particle picking, classification, alignment, CTF correction, 3D reconstruction and model generation. It also presents the link between the noise and the efficiency of some steps and shows results of additional analysis for this thesis through included figures.

3.1 Particle picking

For small amount of sample, it is trivial to select or pick the particle projections from few micrographs. The procedure for picking thousands of particles from several micrographs is a tedious task, which emphasizes the need for automatic picking method. Automatic particle picking procedures generally use correlation peak finding algorithms to select particles [53], in which an average model of a particle is generated and correlated across the whole micrograph. The correlation value between the model and the micrograph, gives the probability of particle presence and the area with high correlation value is selected and boxed out into a different set. Another approach used to select barely visible particles from close-to-focus and noisy micrograph, is to pick visible particle locations from the micrograph image of the same position in the same sample but far from focus. This can be done if the relative orientation of the two micrographs is known [12]. There are also other methods which take into account the local variance [54] in the image or the local histogram in the image boxes as proposed in Publication I while some other methods employ different transforms to select the assumed geometrically shaped particle picked [55], [56]. The prior assumption of shape of projections to pick particle works for symmetrical particles but not for all the particles. Given the noise in the micrograph the correlation values can be false, so most of the methods have a refinement procedure to drop the picked particles which are false-positives. These refinement procedures may themselves serve as independent picking methods as proposed in Publication I. One of the refinement method employs Gabor filter that posses optimal representation property in both spatial and frequency domain with orientation. The Gabor kernel function given in equation (13) of Publication I is based on the two-dimensional Gabor function proposed by Daugman to model the spatial summation properties of simple cells in the visual cortex [57]. It has a single direction sinusoidal part controlled by frequency and a global Gaussian blurring. The function has both real and imaginary part, however only one part is used to convolve it with the boxed images to obtain coefficients in different orientations. The corrected version of real part of the function for direction θ at location (x, y) would be

$$\psi_{\sigma,f,\gamma,\theta}(x, y) = \cos(x')\exp\left(-\frac{x'^2 + \gamma y'^2}{2\sigma^2}\right), \tag{3.1}$$

where σ is the standard deviation for Gaussian blur and γ is the elongation parameter. The corrected versions of x' and y' are such that $x' = (x\cos(\theta) + y\sin(\theta))f$ and $y' = (-x\sin(\theta) + y\cos(\theta))f$ for frequency f . As mentioned in Publication I, the array of the sum of the coefficients of one image is then convolved with similar array of other image to get the similarity index.

3.2 Classification

Doing classification and making class averages of cryo-EM images is a traditional way to reduce noise and simplify the reconstruction process. The picked projection images of the sample particles can be classified using various methods that are efficient with high SNR, when a model representation (reference images) for the data is available [58]. To achieve the model representations from the image data, a commonly used method is multivariate data analysis [59] based on either principle component analysis (PCA) or correspondence analysis (CA). While PCA is done on the basis of Euclidean distances, the CA method [60] is dependent on the statistical χ^2 distances between the relative frequencies. The relative frequencies are obtained after the data in each row and column are normalized such that they add up to one. The rows contain the images and columns contain the pixel values. Thus correspondence analysis [60] technique does statistical analysis of such two way tables in order to find important representative features among several images [58]. After using data reduction technique such as CA or PCA to make reference images, it is possible to decide the classes of all particles The reference-based classification can be done on the basis of features or direct pixel. information of the images. Classification methods have evolved with different approaches such as hard or fuzzy classification approach [61], partitional-hierarchical ascendant merging [62] and self-organizing approaches of classification. Although classification is meant to increase the SNR in images, it may also produce worse results due to its failure to deal with high variance noise and images could be assigned wrong classes producing false projection averages and hence lead to false 3D volumes. When the image set is heterogeneous, the multi-reference classification of the entire set of images can also be done after reference-free alignment or after having projection from different 3D volumes. The class averages made by multi-reference classification can be used to make multiple 3D volumes. The classification process is followed by further alignment and making class averages.

3.3 Alignment

Usually, picking algorithms can put the picked particles in the centre of the selected area (boxes). Though the need for centering of projections in boxes for asymmetrical particles has been questioned, researchers still do fine adjustments for centering to increase the efficiency of other steps of SPR. The centering methods employ different approaches to find the centre of particles. Some make pixel value based weighted average of spatial locations while other use only intensity peaks. In the later approach, image is translated so that number of intensity peaks near the centre is maximum. For finding orientations, the alignment method heavily relies on template based cross-correlation rather than the pair-wise features matching and optimization like computer vision techniques. In contrast to other fields like computer tomography [63] and computer vision, the sequence of the projection images taken, does not exist for SPR. A possible matching among each pair of images from a set of several thousands of images could be quite tedious task to do, so the alignment methods use common templates for calculating cross-correlation with particle images. The templates are generally the projections of the 3D volume reconstruction made during previous iteration or using initial model generation methods. The mode of orientation representation vary for different methods and users as some use Euler angles θ,φ,ω format while others use vector format [64] representation. Different methods use different transformed space to calculate correlation coefficients to align the particles. PFT2 [65] software uses polar Fourier transform to align the particles and has been proved to be quite effective for aligning icosahedral viruses [66], [67]. Sinogram [12] based alignment have also been used to align noisy images. Conventional cross-correlation functions are ‘squared’ correlation functions which can bias towards the low frequency components in the data. In order to avoid this biasing, mutual cross-correlation function can be used for alignment in final iteration of processing to focus on the high resolution data components [12]. One projection can be assigned either one or many orientations based on choice and complexity of data. The assigned orientations are used to make the 3D reconstructions which can be further used for tem template based matching and alignment in an iterative manner, see Figure 3.1.

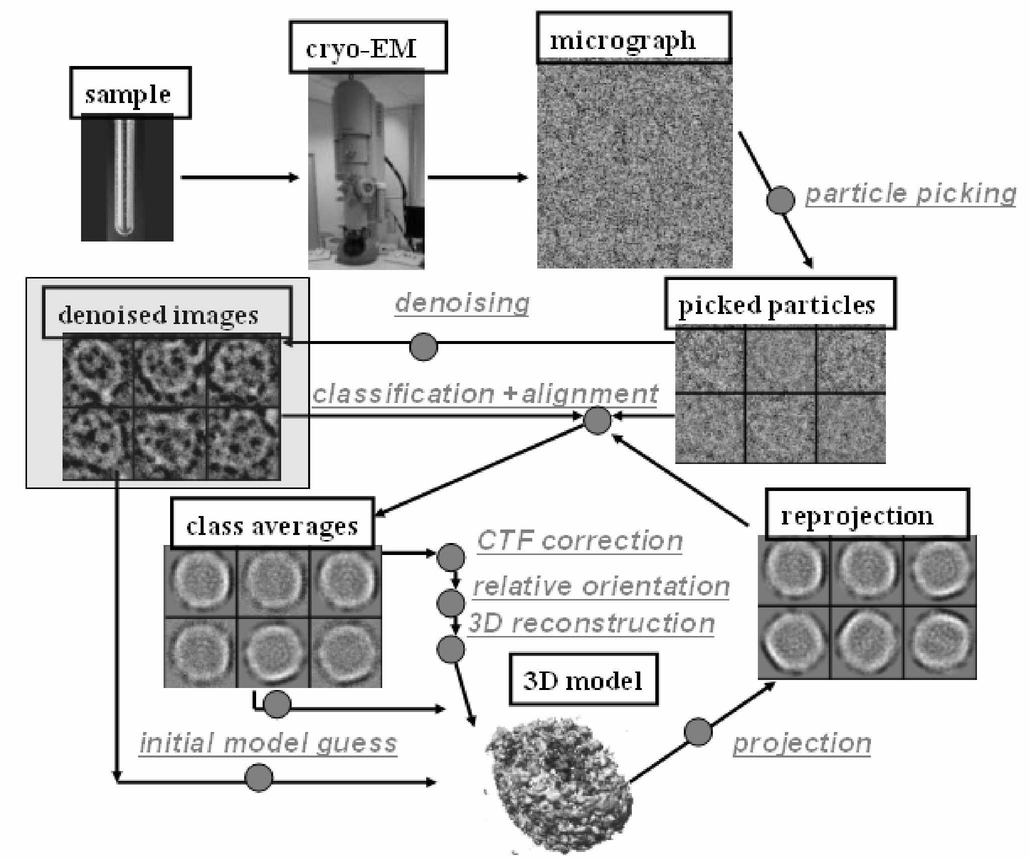

Figure 3.1: The schematic diagram of the process steps and cycle involved in single particle reconstruction. The process names are underlined and are in grey color. The grey box represents the set of process involving denoised image. (cryo-EM images by courtesy of Sarah Butcher, University of Helsinki)

3.4 Contrast transfer function correction

The data measured by a TEM and cryo-EM can be presented in following function form

$$I(s,\Phi) = C(s)E(s)F(s,\Phi) + N(s,\Phi), \tag{3.2}$$

where $I(s,\Phi)$ is the measured data, $C(s)$ is the CTF, $E(s)$ is the envelope function and $N(s,\Phi)$ represents random noise and $F(s,\Phi)$ is the projection of the 3D structure in orientation Φ. The knowledge of $C(s)$ and $E(s)$ is important for estimating $F(s,\Phi)$. Since noise $N(s,\Phi)$ properties are unknown, it can not be subtracted directly from the images. To determine $C(s)$ and $E(s)$, examination of the rotationally averaged power spectra of projection image, helps effectively [58]. *C(s)*, *E(s)* and *N(s)* can all be parameterized based on simulated theoretical models and comparison of the modeled power spectra with the original average power spectra. The mean noise level then can be reduced by averaging, when the actual CTF correction is done while generating class averages from aligned 2D single particle images or while making 3D reconstruction. The CTF based averaging method weights different images in Fourier space and the weights vary for different spatial frequencies depending on the quantity of information they have.

The CTF correction technique described above makes suppositions for models which may not be always correct and lead to suboptimal estimation of parameters. To solve this problem, few methods [35], [62] incorporate Wiener filter into the correction equations, but accurate Wiener filtering needs an accurate measure of the spectral SNR in the final averaged image. Wiener filter can then do the correction as such

$$\bar{F}(s,\Phi) = \frac{H^*(s)I(s,\Phi)}{|H(S)|^2 + (\frac{1}{SNR})} \tag{3.3}$$

where $H(s) = C(s)E(s)$ represents the CTF determined for one micrograph and $H^*(s)$ is the complex conjugate of $H(s)$.

A relative SNR measure may not be adequate for optimal Weiner filtering in the case of noisy TEM and cryo-EM images. To get a more optimal measure, the CTF parameter determination can be performed using structure factor [68] if it can be obtained. An approximate structure factor not only helps in estimating CTF parameters but also provides more accurate estimation of the SNR in class averages. The Wiener filter can then be applied to the class-averages. While making the 3D reconstruction, the additional averaging can lead to higher SNR but if the CTF correction is not done optimally for cryo-EM images, the 3D reconstruction may be lacking detailed information about some structural details. Better alignment and reconstructions can be achieved when the estimation of CTF parameters is improved iteratively at every cycle of the reconstruction process [69].

3.5 Three dimensional reconstruction

After the orientation determination for particles the 3D reconstruction from a set of 2D projections or class average or projections, can be done using the principle first explained by Radon [70]. This is the backprojection method which is the inverse of the projection method. In Fourier space the backprojection corresponds to simple combining of central sections associated with projections. Different variants of backprojection method such as weighted backprojection and filtered backprojection have been able to produce better result.

Theoretically, the projection of 3D object is assumed as a summation along rays that extend from minus infinity $-\infty$ to plus infinity $+\infty$ and which represent an infinitely thin central section in Fourier domain. In reality the 3D reconstruction extends only over maximum finite distance D. This causes the central section in the Fourier transform to be a slab with thickness of (1/D) rather than an infinitely thin plane, which lead to overlap of the Fourier slabs in the low frequency region causing in to a blurred 3D reconstruction [12]. Also as the denoising has tendency to suppress high frequency information, the resulting 3D reconstruction from denoised image may tend to be blurred. To avoid this situation one can use original noisy images for making 3D reconstruction based on orientations, determined for corresponding denoised images. Such 3D reconstructions are called here as 'hybrid 3D reconstruction'. Figure 3.2 shows the results of reconstruction using the hybrid method. To avoid the blurring problem, high pass filtering of the projections can also be used just like in some fields of computer tomography. Instead of using single approximate filter for all images it is possible to have a specific filter computed for each image used in the reconstruction [12]. Besides this, the high frequency regions in the 3D Fourier reconstruction have

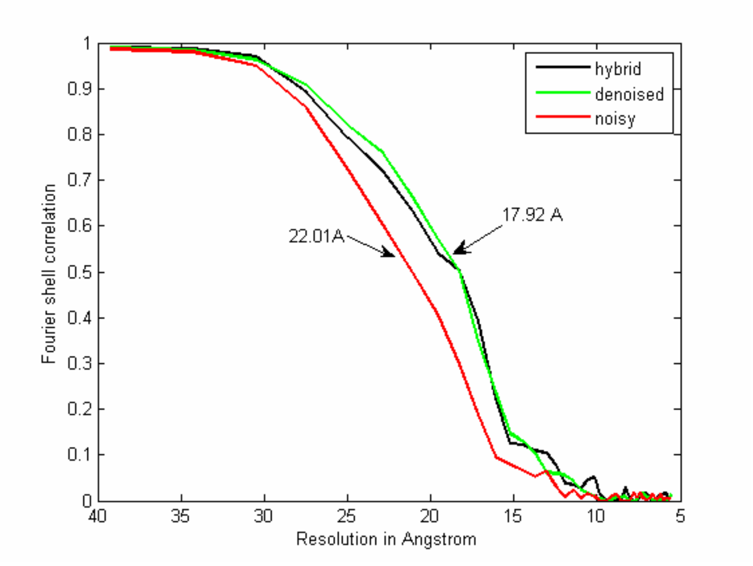

Figure 3.2: The Fourier shell correlation [58] plot to show resolution achieved by using single particle reconstruction for 10,000 projections taken from 3D volume of the prokaryotic chaperone protein GroEL Protein X-ray structure. Noise of high variance was added to projections which were randomly translated. The noisy images were later denoised and single particle reconstruction was performed using EMAN [19], assuming no symmetry and running for 9 iterations. The hybrid method used denoised images only for classification and noisy images to make class averages and 3D reconstruction. The resolution achieved with denoised images and hybrid approach was 17.92 Å and resolution achieved by processing only original noisy images is 22.01Å.

gaps between Fourier 2D projection slices, causing missing values of the high resolution information in the resulting 3D reconstruction. Methods like MEM (Maximum Entropy Method) [71] and POCS (Projection onto Convex Sets) [72] for filling the missing 3D information seem to improve the high frequency details but their results of recovered 3D volumes could still be objected and termed as too much extrapolation of data. The original estimate of 3D volume by backprojection method can also be further refined by using methods such as ART (Algebraic Reconstruction Technique) [73], which is more practical in theoretical explanation for restoration of high resolution information.

3.6 Generation of initial 3D volume

For an inverse problem like SPR, in which the source of data is modeled and found iteratively, starting from an initial model, the properties of initial model can effectively influence final results. When alignment has not been done and orientations of 2D projections are unknown, making approximately correct initial 3D reconstruction is extremely important. A commonly employed method for this purpose, is the common line method [12] which tries to find orientations for two projections p_i and p_j of a density map F onto planes corresponding to orientations θ_i and θ_j, respectively by finding common line on the two projections. All one-dimensional projections of F onto a line passing through the origin in the plane corresponding to the orientation θ_j can be computed and found in other projection p_i; this collection of such projections is also called the sinogram of p_i.

For noiseless projections the pair wise common lines of three projections can be used to determine uniquely, the relative orientations of the projections in 3D space. The calculation of this is trivial using some arithmetic operations. However, the projection images produced by cryo-EM are extremely noisy and so to find one-dimensional projections that are equal becomes a computationally heavy task. It explains the necessity of increasing SNR in the images to be aligned even if some resolution in the images is lost. An example simulation has been shown in Figure 3.3, where it can be clearly noticed that noisy images could lead to wrong structures with common line methods. The same noisy 2D images after denoising gives structures in Figure 3.3(e), more similar to original 3D structure shown in Figure 3.3(b). For highly symmetrical structures like icosahedral viruses and protein oligomers, the combination of denoising and common line method could help in avoiding spurious results caused by inefficiencies of other steps in the SPR.

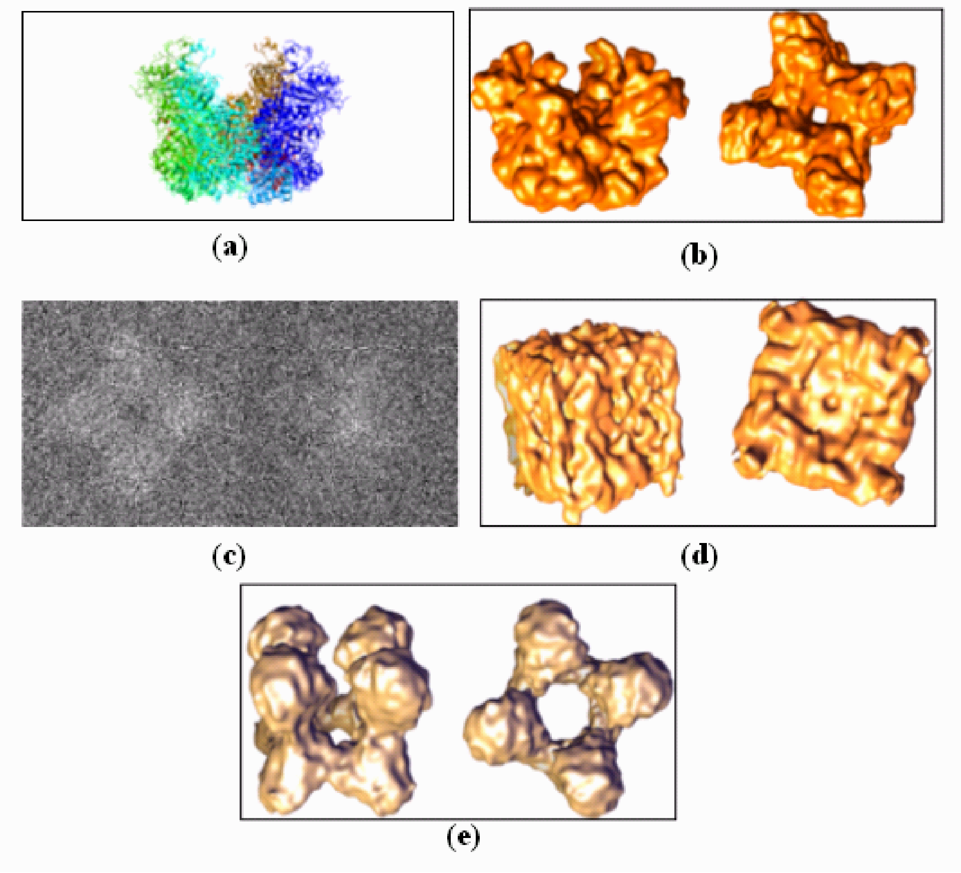

Figure 3.3: The difference in the results of common line method. UCSF Chimera [74] was used for 3D visualization. (a) an oxidoreductase protein structure, PDB [75] code 1AA1[1], (b) the side and top views of 3D volume obtained after converting pdb structure to 3D volume at sampling rate of 1.4 Å /voxel, (c) two of the 351 projections made from the 3D volume and added with flat band Gaussian noise, (d) the side and top views of 3D volume reconstructed using common line method for the noisy projection images and assuming a four fold symmetry an axis, (e) the side and top views of 3D volume reconstructed using common line method for denoised projection images and assuming four fold symmetry around an axis.

Chapter 4

Computer modeling and validation

A protein molecule is a linear polymer composed of amino acids which form a chain, known as its ‘primary structure’. In a protein sequence, amino acids are joined together by peptide bonds between their amine and carboxyl groups. The terminal of a protein sequence with amino acid with a free amine group is called ‘N-terminal’ of the protein. Similarly the terminal of a protein with a free carboxyl group is called its ‘C-terminal’. The 3D shape of a protein in its native state is critical for its molecular function. The short-range forces at the atomic level give rise to the local regular configuration of protein sequence, called ‘secondary structure’. The protein's polypeptide chain with secondary structure folds further in 3D space to form more complex structures known as ‘super-secondary structure’ and ‘tertiary structure’. A protein molecule can interact with other protein molecules either after completely folding to its own tertiary structure or by folding together with the polypeptide chains of other protein molecules. The structure of a complex of many protein molecules is called ‘quaternary structure’.

Solving the protein structure is mainly to determine the location of atoms of amino acids of proteins in 3D space. The 3D structure of proteins and protein complexes obtained by experimental methods are mostly docked with the predicted molecular structure of the proteins. Sometimes the configurations of the amino acid residues are predicted from the details of the 3D reconstruction obtained for the protein. At the same time, the characterization of protein interactions is done, in order to find answers for problems ranging from rational drug design [76] to analysis of metabolic [5] and signal transduction networks. As the number of experimentally determined structures for protein and protein complexes is still quite small, methods for computational prediction of protein-protein interaction sites are becoming increasingly important. An important research field included in the thesis is to characterize the interaction location and interaction pattern involved in the oligomerization of hantavirus N-protein presented in Publications II and V. Since the structure of N-protein is not yet known, the approach followed here is unconventional and it reveals the potential of a different perspective about protein folding for quaternary structure.

The overwhelming emphasis on determining tertiary structures, given by the current popular opinion about protein folding, might be blocking the study of quaternary structures of many proteins. The approach proposed in this chapter is to view the phenomenon of protein folding from another perspective and not to stress on the tertiary structure. In order to understand the protein structure and to model it in a proper way, the basic forces responsible for the folding and the stability of the tertiary and quaternary structure of the protein have to be understood properly. Therefore, this chapter explains such forces by outlining the role of different amino acids for stability of protein and protein complexes. Later in this chapter, insight in to the steps involved in the prediction of proteins and protein interaction structures is given as well as the explanation of the approach used for the candidate protein mentioned.

4.1 Protein structure forces and interaction site prediction

Most of the forces present in protein structures are local and they influence the folding of proteins. Though the origin of these forces can be explained in detail by quantum mechanics, a brief description about them would serve as a background to start with. Some of the basic forces influencing protein structure are briefly presented here, but for more details see [77], [78].

In proteins the basic covalent bond, formed by covalently sharing an electron between two atoms, connects atoms of an amino acid. Besides being in amide bonds they are also responsible for cysteine side chains interactions, which provide stability to protein structure, folded properly or even improperly [77]. While the covalent bonding produces strong short-range forces, electrostatic interactions among amino acids bearing electrical charges can be long-range. Due to their long range strength they have an important role in the formation of higher order structures like tertiary and quaternary structures [77]. Some of the atoms have a partially negative charge which they try to reduce by sharing a partially positively-charged hydrogen atom leading to formation of hydrogen bond. It leads to the secondary structure formation in proteins as most of the hydrogen bonds are present between positively-charged amide and negatively charged carboxyl groups in alpha helices and beta sheets. A major force responsible for the proper configuration of protein structures [77], [78] is due to hydrophobic bonding. They pack hydrophobic side chains in order to shield them from interactions with polar water molecules. These side chains at protein surfaces involve hydrophobic bonding with neighboring proteins.

Yet another form of force, called van der Waals force, is due to weak electrostatic interactions, among transient dipoles of electron clouds around the pair of atoms. Van der Waals forces can be either attractive or repulsive depending on the distance between the two atoms [79]. These forces provide an important component for protein structure because of the close proximity of atoms with each other. Other types of interactions, which are caused by aromatic amino acids, are cation-π interactions [80] and aromatic-aromatic interactions [81]. A cation-π interaction occurs when an electron cloud of an aromatic ring interacts with a positively-charged cation close to it. Also many aromatic side chains (phenylaline, tyrosine and tryptophan), found in proteins are involved in aromatic-aromatic pairings [81]. Simulation of such forces, suggest that interacting aromatic rings are generally perpendicular to each other such that the π-electron clouds do not come in contact with each other and hydrogen atom at the end of one ring can interact with the π-electrons of another ring, e.g. see Figure 4.1. The effect of aromatic amino acids for protein folding and stability can be understood by the fact that adding aromatic pairs and aromatic clusters in a protein increases its thermal stability [81], [82]. These forces have influenced the research done for Publications II and V.

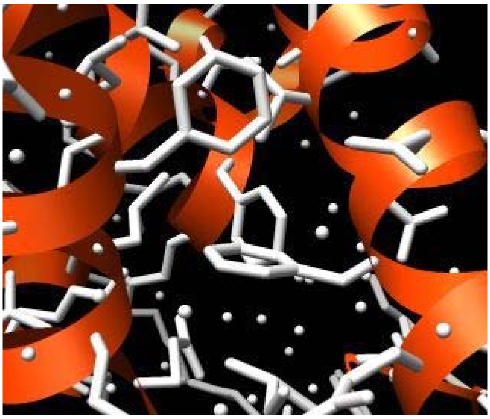

Figure 4.1: The orientations of benzene rings of phenylaline amino acids relative to each other while having aromatic-aromatic interaction in bacteriophage T4 lysozyme protein. All the three rings are perpendicular to their neighbors. The visualization tool UCSF Chimera [74] was used to produce the figure from the lysozyme structure (PDB code 112L).

Protein complexes have different properties depending upon their constituents and the function they perform. Cyclic oligomers and dimers tend to have different complementary electrostatic surfaces at their monomer interfaces [83]. The protein complexes formed by different proteins involved in catalysis, are mostly results of interaction among their polar amino acids. Based on these known features of protein interaction investigators have claimed to be able to predict quaternary structures from primary sequence information [84], [85] using computational methods only. However, the task at hand to predict parts of protein molecules involved in interaction, still remains as a challenge. For some proteins, even if their structures are known, it may not be a trivial to find their interaction sites. This can be understood from the survey research for community wide experiment on Critical Assessment of Predicted Interactions (CAPRI) [86]. Researchers have used different computational approaches and different properties of the amino acid chain in order to predict interaction sites [87], [88], [89], [90]. A convincing method proposed by Jones and Thornton [91], for predicting interaction site on known protein structures, uses the knowledge of interaction sites from other protein structure patches. In their approach, residue patches defined on the surfaces of isolated proteins are analyzed for six parameters termed as salvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and accessible surface area. Such an approach can also be used for a protein using its secondary structure, even if the overall folding of the protein is not known. The attempt done in Publication II is based on such an approach, combined with modeling of the interaction structure.

4.2 Systematic approach to predict interaction site

As mentioned above one of the tasks in this thesis was to characterize interaction of hantavirus N-protein monomers of sequence length of 421 amino acids. For such proteins, given a set of acting forces, the question is how to proceed such that the interacting amino acids and the forces involved can be determined and the structure of the protein interaction can be predicted. Taking an example of a dimer forming protein composed of 450 amino acids and assuming that 5 amino acids of a monomer binds with 7 amino acids of another monomer, the number of possible combinations of interacting amino acids would exceed 10^{20}. Hence there has to be reasonable way for reducing the search space for active amino acids, for example relying on bioinformatics approach. The first step in such a systematic approach would be to align all the amino acid sequences of the protein and closely related proteins across all the species. For this purpose multiple sequence alignment could be used to align peptide positions believed to be homologous, applying various algorithms. The algorithms for multiple-sequence alignment use different strategies like dynamic programming [92], probabilistic hidden Markov model [93], genetic algorithm [94] and progressive alignment with clustering [95]. The sequence alignment technique, which gives priority to finding motif [96], is quite useful here as it finds blocks of conserved locations. The homologous and conserved amino acid locations give hint about the functional units of the protein and the interaction site could be one of such locations on the protein [97]. This exploits the evolutionary conservation behavior of proteins to remain functional. The next step would be to look for domains and motifs, which are well known for interaction of proteins. This approach works quite well with eukaryotic proteins but it is hard to find well-known domains in virus proteins. If domains are found, the structure of such domains could be determined from the available protein structure data bank. The next stage would be to try to dock different domains using some available methods for protein ligand docking or protein-protein docking [98], [99]. Such an attempt was done in order to find potential interacting partners for type 1 PDZ domain. We performed docking with different known and unknown interacting partner protein sequences using program GOLD [98]. One such docking result about a known interaction is shown in Figure 4.2.

Chances of getting known domain structures is rare and even if domains are found the proteins should be checked for interaction after chopping the conserved regions. In most of the cases the probable interacting side (N terminal or C terminal) of the protein can be found by such an approach. Another procedure would be to predict the secondary structure of the protein using the available tools like PsiPred [100], Sam-T99 [101], JUFO [102], Profsec [103] and JPRED [104]. These tools use different algorithm for predicting secondary structures. PsiPred [100] uses neural network with the position specific scoring matrices generated by Psi-BLAST [105] which combine statistically significant results produced during sequence alignment. Sam-T99 [101] uses a hidden markov model iteratively to analyze sequence homologies and predict

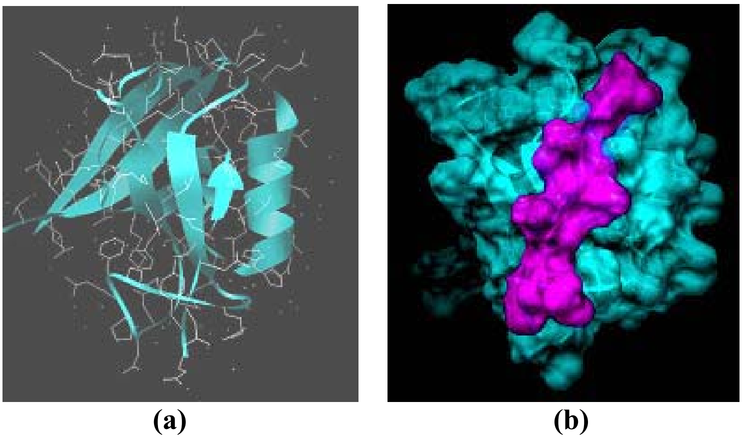

Figure 4.2: (a) The structure of PDZ domain of the erbin protein, (b) the van der Waal surface shown for protein rebind, PDZ domain docked with ERBB2 C-terminal amino acids, here the docking was performed using program GOLD [98], [99]. The ERBB2 C-terminal amino acids chain was used as a flexible ligand and no prior information about docking location was given. The figures were made using visualisation software UCSF Chimera [74].

secondary structure. JUFO [102] and PROFsec [103] servers both use neural network, in their own way. Combining results from different prediction servers helps in avoiding ambiguities, especially about the boundaries of α-helices. Therefore some servers like JPRED [104] combine results of different secondary structure prediction methods to deliver results. One can combine the predicted secondary structure with the information about the hydrophobic property of the peptides to predict the interaction sites. The hydrophobicity plot for proteins gives a measure of burial of different amino acids. For non-membrane and non-amphipatic proteins the less hydrophobic sites have higher chances of being exposed and hence higher chances of participation in heterotypic or enzymatic interaction. The flexibility of the protein hints about the clusters in protein 3D structure as these globular clusters would behave as individual components and the protein can be chopped on those flexible locations without disrupting the structure. The chopped portions of protein can be checked for interaction. Figure 4.3 shows the flexibility and hydrophobicity of hantavirus N-protein. For homotypic interaction the experiments for chopping few amino acids from either end or both ends of the proteins would give a strong hint about interaction mode, as homotypic interaction often occur by exchange of N-terminal or the C-terminal secondary structures. For homotypic interactions if there is any region of high hydrophobicity near any of the terminals of the protein, then the experiments with mutations at those locations could be given high priority.

After predicting the secondary structure of the proteins the search can be done to find the patterns in amino acid sequence for some well known super-secondary structures. Methods like Rosseta [107] and 3D-PSSM [108] would be quite helpful in finding those patterns. Rossetta method, combines fragments of known protein structures with a Monte Carlo strategy to obtain a feasible native protein

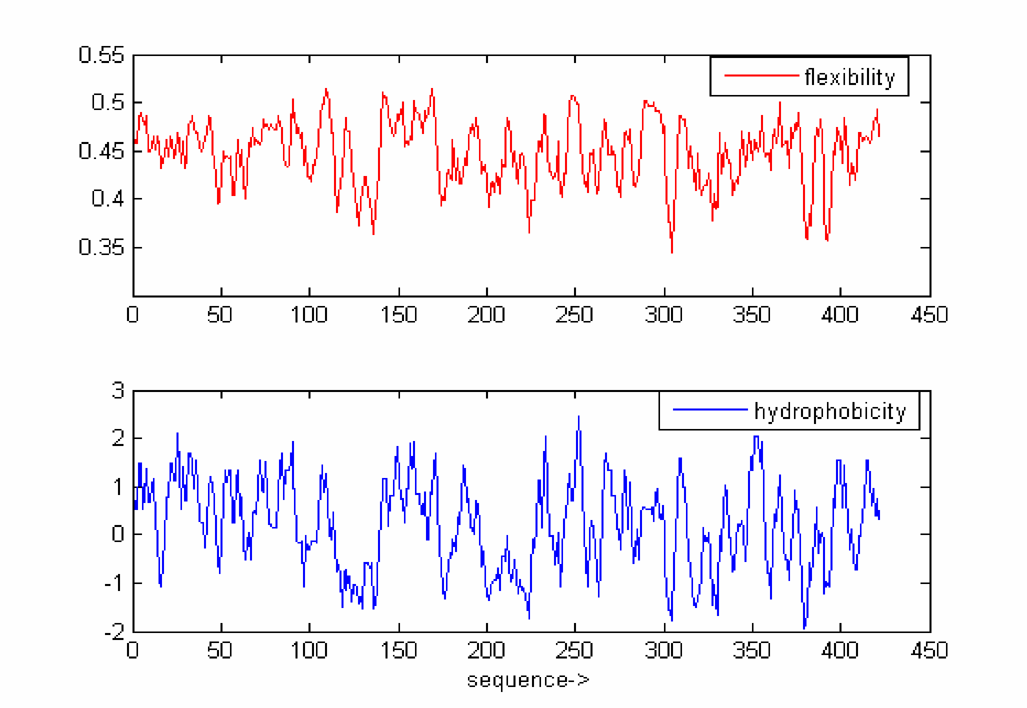

Figure 4.3: The flexibility and hydrophobicity plots for Tula-virus N-protein sequence, made using MATLAB toolbox. The hydrophobicity was calculated using method proposed by Hopp and Woods [106].

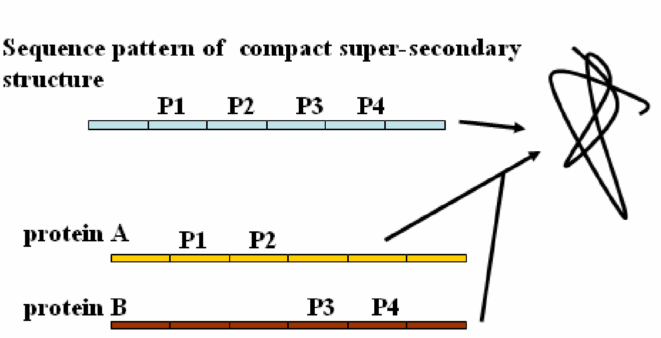

Figure 4.4: The conceptual representation of interacting sequences of different proteins to form a super-secondary structure. The patterns P1, P2, P3 and P4 of a super-secondary structure may be only in one protein (top-cyan) or across many protein sequences (down-brown and yellow).

structure for a sequence. Recently it has demonstrated high accuracy for short contiguous peptides of 60-80 amino acid residues [107]. The 3D-PSSM method attempts to predict 3D structure using its own library of known protein structures onto which the sequence is mapped (threaded) and score calculated using various criteria including the secondary structure match, solvation potential and PSSM scores [108].

For interaction site prediction, one has to extend the search for pattern of super-secondary structure beyond a single protein sequence. These patterns can be present in a single protein sequence or across more than one sequences giving hints about their possible interaction sites. This has been demonstrated in Figure 4.4. It represents how a mutual folding of sequences of two proteins to form a super-secondary structure leads to protein interaction. For a good example see Publication V, where there is a description for interaction of N-terminal sequences of different N-protein monomers to coiled-coil structure.

4.3 Structure prediction for protein interaction

Since the first proposal about protein secondary structures (α-helices and β-sheets) by Pauling and Corey [109], [110] and first description of a full protein structure solved by Kendrew and colleagues [111], there have been many developments and now researchers are endeavoring to determine interaction structure among proteins. The accuracy of the secondary-structure prediction has constantly been improving and some methods claim the accuracy to be more than 75 percent [104]. The predicted secondary structure not only enables to know the polarity or charge distribution of the amino acids in 3D space but also hints at the potential of different parts of the protein to form super-secondary structures. Thus after the prediction of secondary structure, the folding of protein into super-secondary structure can be done by finding some known pattern. Such patterns for super-secondary structure in the protein sequences, would lead to formation of linkages among corresponding secondary structures in the same protein or in different proteins. Some examples of such super-secondary structures are coiled-coil helices, beta barrels and helix-turn-helix [112]. For a non-membrane protein the search for amino acid patterns for forming super-secondary structures with amino acids from other protein molecules, could be possible if their tendency of burial into the hydrophobic core is estimated.

Various computational searches in this field [113], [114], [115] have made classifications of different structural domains found in the available structures in PDB [75]. Domains are semi-independent 3D sub-units that are compact and may fold independently [114], [116], [117], [118]. An increasing number of methods and databases address the problem of identifying patterns of structural domains in sequences [116].

Recently Jinfeng Liu and Burkhard Rost [117] developed a method to chop proteins into domain like fragments. Using their chopping methodology, they claimed that 70 percent of all dissected proteins contained more than one domain like fragment and three-fourth of all domains appeared shorter than 210 residues. This

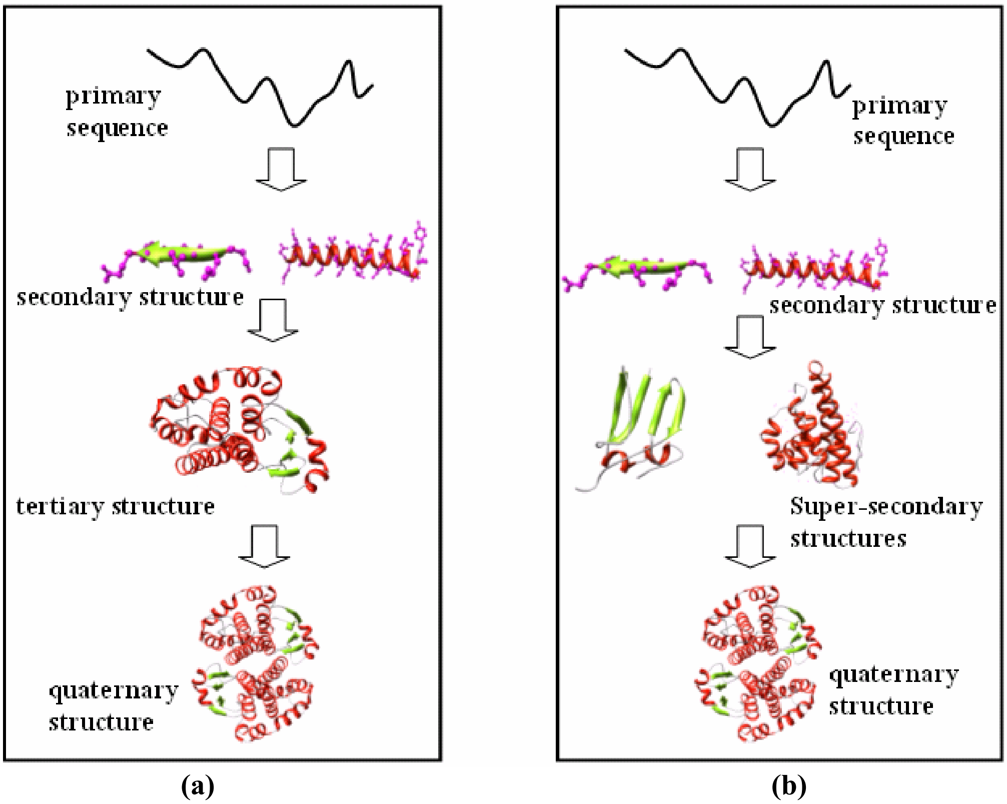

Figure 4.5: Two perspectives of protein folding towards quaternary structure starting from primary sequence.

represents the importance of the localized compact structures in comparison to tertiary structure of individual protein in macro-molecular complexes. These small domains can be called as super-secondary structures and can be used directly to predict quaternary structures of protein complexes. The trend of giving priority in protein structure determination is shifting from determining protein tertiary structures towards modeling quaternary structure of proteins. An argument is that, with localized folding, giving emphasis on tertiary structure is not worthwhile. Such a plausible argument supports the work done in Publication II and V. The 3D structure of N-protein trimer, obtained by SPR method gave a low resolution structural description of quaternary structure for N-protein oligomers. After pondering over two perspectives for protein folding for quaternary structures as shown in Figure 4.5 and having other experimental results [119], it was clear that the folding of different portions of N-protein monomer is more localized and independent from other distant peptides of their sequences. If we consider the conservation of quaternary structures [120] of similar sized N-proteins of negative stranded RNA viruses, then the structures published for N-protein of two similar viruses, Rabies virus [121] and vesicular stomatitis virus [122] and results in

Publication II together with the publication S1 all support our conclusion about localized folding of N-protein of hantavirus.

Modeling interacting surfaces

A survey across the PDB database [75] for quaternary structures leads to knowledge of some classes of commonly occurring super-secondary patterns of protein interaction. The pattern which came out to be as most dominant for the C-terminal interaction among N-protein monomers is 'domain swapping'. The flexibility of the region around such amino acids is also an important factor as those amino acids most often need to protrude out to perform swapping of secondary or super-secondary structures. Simulations for domain swapping presented in [123], address the fact that model with the flexible hinge loop has a lower energy barrier from the one in which the hinge dihedral energy barrier must be crossed to swap with neighboring protein.

Overall, using the approach proposed above could lead to more than one potential interaction structures. The models of each possible interaction structure could be made by using available tools like Insight-II (Accelerys, Inc.) and Swiss-model [124]. The calculation of energy in the interaction structure would give a good hint of the most probable interaction structure. It should not only have potential to exist but also feasible to be formed by crossing of low energy barrier by the polypeptide chains involved, see Publication II. The modeled interaction structures can then be checked by using mutations. Even though experiments done with mutants could give some quite strong hint about the importance of different amino acids, they cannot confirm the role of amino acids in interaction. The solution to this problem can be achieved by performing simulation based mutations and further improvement of the simulation model based on mutant experiments.

The presence of tryptophan in the proximity of binding site always hints to its important role in the binding of proteins, as tryptophan has the capacity to interact with different kinds of forces involving aromatic-aromatic interaction, hydrogen binding and cation-π interaction. In order to check the possibility of tryptophan role, the modeled secondary structure of N-protein was made to dock such that tryptophans could juxtapose to their potential binding amino acids of the other interacting C-terminal polypeptide chain. The structural constraints filtered out many possible interactions of tryptophan, for C terminal interaction of the N-protein. The remaining and possible candidate amino acids were then checked by mutating those sites in the protein and checking for the interaction. While doing simulations for C-terminal interaction of N-protein it became quite clear that the initiation of N-protein interaction is done by other parts in the N-protein monomer, while C-terminals only stabilizes this interaction. The candidate polypeptide for initiation of N-protein interaction at its N-terminal was then modeled and its structure was proposed in Publication V.

Chapter 5

Summary of results

The challenge of improving signal to noise ratio in various imaging problems has been tackled using a statistical modeling approach in Publications I, III and IV. The methods followed here are based on defining noise by the MDL principle, such that noise is that part of the data which is random and is incompressible. Under these assumptions, coding the data in two parts, the part with higher code length represents noise. The principle of MDL has its background in information theory and basically deals with the coding of the data using the statistical properties of the data. This principle can be applied efficiently in many fields to choose or recognize a model to represent the system or data. Each application tends to have its own way of implementing this principle but they all share the similar notion of choosing the best model using the code-length principle. Some applications use it on data after transforming it using basis vectors, regression matrices or eigenvectors, while others apply it directly on the data.

A completely novel way of choosing the best model after transforming the data has been proposed in Publication III and further developed in Publication IV. It is different from normal approaches that are derived from the mean squared error (MSE) criterion or other conventional methods to minimize theoretical error while estimating the parameters. Examples of these conventional methods include the maximum likelihood approach and minimum-mean-square error-based methods. The histogram coding based model selection proposed in Publication III is completely nonparametric and free from any assumption concerning the statistical distribution of the data. Although it has been demonstrated with wavelet coefficients in order to denoise noisy signals, the general implication of this approach goes beyond denoising. The basic theoretical method published in Publications III and IV can also be used for other statistical signal processing problems in its original form or in an extended form. Such applications could include non-parametric expectation maximization or pattern recognition and classification.

For cryo-EM images the properties of noise and noise source are still being studied [29], [30] as it crucially affects the results of the SPR method. However, without proper conclusion about noise properties, assumptions of parametric distributions of the noise are completely illogical. However a basic property of noise is that it cannot be defined properly because it is random. The methods published in Publication I, III and IV exploit this property of the noise in denoising and enhancing cryo-EM images. In Publication I the method of enhancing the desirable part of the data without even knowing what part is desired, exploits the code-length principle. The weighting of different basis vectors based on their complexity can be compared to choosing best transforms to represent the data, but it is not exactly the same. Approaches where researchers tried to find the best wavelet basis vector family, have been previously described [44], [125], [126]. According to the method published in Publication I, first the search for the best basis vector family is done and then the non-noisy coefficients are chosen after applying the chosen basis vector family, using the MDL principle. After these steps, the priority of different level basis vectors among the chosen basis vector family can be determined using the code-length criterion. Although the initial demonstration of the proposed method has been shown with wavelet transform basis vectors, other similar transforms can also be use in order to filter and enhance different features. For example, the use of curvelet [39] and edgelet [127] basis vectors with the proposed approach would help in removing noise without loosing details of sharp features.

In publication I, we proposed some denoising based methods to automatically pick the particle projection from the cryo-EM micrograph images. There is no assumption of the particle shape in these methods and they promise to be picking particles despite high variance of noise or thick vitrified layers. Each of these methods can be individually used for picking particles but in Publication I we propose to use them in sequence. A model is needed for some of the proposed methods so that it is possible to avoid picking impurities in the micrograph, even if they could not be separated during the process of purification of proteins for sample preparation. The Gabor filter based method to compare different projections can also be used for rotation-invariant classification of images, using multi-resolution features and has been demonstrated in Publication I.

The proposed denoising approach was applied to real and simulated cryo-EM data, in order to check the benefits in SPR method. The test samples were real cryo-EM data of hantavirus N-protein and simulated data of GroEL particles. As expected, SPR method with denoised images gave higher resolution and had faster convergence in comparison with noisy images, but there is also a possibility of making 3D electron density map using ‘hybrid methods’. For obtaining these 3D maps the orientation determination can be done with denoised images and the 3D maps can be reconstructed by using original noisy images. If the prerequisite of high SNR is satisfied in the original cryo-EM images, simple template based classification can help in further reduction of noise and making reliable class averages and 3D volumes. If the SNR is low in original noisy images then a simple classification process without denoising will not be entirely reliable. In such situations, the 3D volumes produced from those class averages of noisy images have either less resolution or they are completely different from the original 3D structure and can result in misleading or wrong conclusions about the structure and function of the biomolecule. The simulations done with denoised projection images showed that the 3D volumes generated after processing denoised images were always more similar to the original 3D volumes.

In Publication II a proposal for structure was given for a protein which is sticky and has not yet been crystallized. The approach to model super-secondary structures of the C-terminal and N-terminal polypeptides of hantavirus N-protein has been explained in Publications II and V. This is a simple example of combining electron microscopic 3D data with modeling. Here, a detailed study on the interaction of the protein to form homo dimer and homo trimer was done. The modeling of C-terminal of N-protein was done by starting from the secondary structure prediction to side chain addition, followed by the modeling of the interaction. The 3D structure made from EM images for the N–protein trimer was taken as reference for the model. In order to model the amino acid forces involved, detailed experiments were planned so that determination of the actual forces could be done without ignoring other forces. Arriving at the conclusion proposed in Publication II was not a trivial task as there could be many possible structures to choose from. The design of initial set of models for binding of amino acids was done, using the constraint of side-chain hindrances, forces and amino acid charges. The secondary structure prediction for C-terminal polypeptide of N-protein reduced the search space, which was still large enough and could need few more years to verify them using other instrumental methods. The 3D charge and hydrophobic distribution on the modeled secondary structure demonstrated the conserved functional pockets in the N-protein monomer. The prediction of binding sites and binding mode leads to the prediction of other functional sites in N-protein. Thus it became fairly trivial to predict where RNA would fit in the N-protein oligomers after having a 3D reconstruction and knowledge about monomer's binding pattern. The proposal given for RNA binding location in N-protein in Publication II was one of the first for hantavirus. It has recently been confirmed by the X-ray structure for N-protein oligomers of two similar viruses, namely Rabies virus [121] and vesicular stomatitis virus [122] and in the publication S1, where a detail model of N-protein has been demonstrated.

Through the series of works presented in this thesis, there has been a serious attempt to contribute to the fundamental concepts in three fields. First, in the field of signal processing, a novel MDL-based model selection approach has been proposed. Second, SPR method from cryo-EM images, has been added with one extra process of denoising for improving results. Finally, a different perspective towards protein folding has come to light due to the approach followed in publications II and V.

References

[1] S. A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, New York, 1993, ISBN10: 0-19-507951-5.

[2] L. Hallivuori, Vibhor Kumar and J. Heikkonen, Using Kalman filters to model gene regulatory networks, in: J. Rousu, S. Kaski and E. Ukkonen, editors, Probabilistic Modeling and Machine Learning in Structural and Systems Biology (PMSB 2006), International workshop proceedings, Finland, 2006.

[3] F. H. C. Crick and J. D. Watson, Structure of Small Viruses, *Nature*, 177: 473−475, 1956.

[4] H. L. Liu and J. P. Hsu, Recent developments in structural proteomics for protein structure determination, *Proteomics*, 5(8): 2056−2068, 2005.

[5] J. J. Gerrits, J. Dijkstra and J. France, Description of a model integrating protein and energy metabolism in preruminant calves, *The Journal of Nutrition*, 127(6): 1229−1242, 1997.

[6] F. J. Bruggeman and H. V. Westerhoff, The nature of systems biology, *TRENDS in Microbiology*, 15(1): 45−50, 2006.

[7] S. C. Harrison, Whither structural biology, *Nature structural and Molecular Biology*, 11(1): 12−15, 2004.

[8] J. Drenth, *Principles of Protein X-Ray Crystallography,* Springer-Verlag Inc. NY, 1999, ISBN 0-387-98587-5.

[9] K. Wuthrich, Protein structure determination in solution by NMR spectroscopy, *Journal of Biological Chemistry*, 265(36): 22059−22062, 1990.

[10] M. Kainosho, T. Torizawa, Y. Iwashita, T. Terauchi, A. Mei Ono and P. Guntert, Optimal isotope labelling for NMR protein structure determinations, *Nature,* 440: 52−57, 2006.

[11] H. Stahlberg, *Transmission Electron Microscopy of Proteins and Protein Complexes*, Habilitation (Biophysics), University of Basel, Switzerland, 2002.

[12] V. Heel, B. Gowen, R. Matadeen, E. V. Orlova, R. Finn, T. Pape, D. Cohen, H. Stark, R. Schmidt, M. Schatz and A. Patwardhan, Single-particle electron cryo-microscopy: towards atomic resolution, *Quarterly Reviews of Biophysics,* 33 (4): 307−369, 2000.

[13] S. Subramaniam and J. L. S. Milne, Three-dimensional electron micrscopy at moecular resolution, *Annual Review of Biophysics and Biomolecular Structure*, 33: 141−155, 2004.

[14] K. A. Williams, Three-dimensional structure of the ion-coupled transport protein NhaA, *Nature*, 403: 112−115, 2000.

[15] D. T. Jones and C. Hadley, Threading methods for protein structure prediction, in *Bioinformatics: Sequence, Structure and Databanks*, Editors D. Higgins and W. R. Taylor, pp: 1−13, Springer-Verlag, Heidelberg, 2004.

[16] Z. H. Zhou, M. L. Baker, W. Jiang, M. Dougherty, J. Jakana, G. Dong, G. Lu and W. Chiu, Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus, *Nature Structural and Molecular Biology,* 8: 868−873, 2001.

[17] J. Rissanen, Modeling by Shortest Data Description, *Automatica*, 14: 465−471, 1978.

[18] A. Buades, B. Coll and J. M. Morel, Denoising image sequences does not require motion estimation, AVSS-2005, IEEE Conference, Como, Italy, 2005.

[19] M. N. Nabighian, M. E. Ander, V. J. S. Grauch, R. O. Hansen, T. R. LaFehr, Y. Li, W. C. Pearson, J. W. Peirce, J. D. Phillips and M. E. Ruder, Historical development of the gravity method in exploration, *Geophysics*, 70(6): 63−89, 2005.

[20] W. Jiang, M. L. Baker, Q. Wu, C. Bajaj and W. Chiua, Applications of a bilateral denoising filter in biological electron microscopy, *Journal of Structural Biology,* 144: 114−122, 2003.

[21] C.O.S. Sorzanoa, E. Ortiza, M. Lópezc and J. Rodrigod, Improved Bayesian image denoising based on wavelets with applications to electron microscopy, *Pattern Recognition*, 39: 1205−1213, 2006.

[22] W. C. Moss, S. Haase, J. M. Lyle, D. A. Agard and J. W. Sedat, A novel 3D wavelet-based filter for visualizing features in noisy biological data, *Journal of Microscopy*, 219(2): 43−49, 2005.

[23] A. S. Frangakis, A. Stoschek and R. Hegerl, Wavelet transform filtering and nonlinear anisotropic diffusion assessed for signal reconstruction performance on multidimensional biomedical data, *IEEE Transactions on Biomedical Engineering* ,48(2): 213−222, 2001.

[24] T. Deschamps, R. Malladi and I. Ravve, Fast Evolution of Image Manifolds and Application to Filtering and Segmentation in 3D Medical Images, *IEEE Transactions on Visualization and Computer Graphics*, 10(5): 525−535, 2004.

[25] J. J. Fernandeza and S. Li, An improved algorithm for anisotropic nonlinear diffusion for denoising cryo-tomograms, *Journal of Structural Biology,* 144: 152−161, 2003.

[26] T. Mielikäinen and J. Ravantti, Sinogram denoising of Cryo-electron microscopy images, *Lecture notes in Computer Science*, 3483: 1251−1261, 2005.

[27] P. C. Doerschuk, Ab Initio reconstruction and experimental design for cryo-electron microscopy, *IEEE Transactions on Information Theory*, 46(5):1714−1729, 2000.

[28] A. L. Saad, Orientation determination by wavelets matching for 3D reconstruction of very noisy electron microscopic virus images, *BMC Structural Biology*, doi:10.1186/1472-6807-5-5, 2005.

[29] I. Angert, C. Burmester, C. Dinges, H. Rose and R. R. Schröder, Elastic and inelastic scattering cross-sections of amorphous layers of carbon and vitrified ice, *Ultramicroscopy*, 63: 181−192, 1996.

[30] B. Feja and U. Aebi, Determination of the inelastic mean free path of electrons in vitrified ice layers for on-line thickness measurements by zero-loss imaging, *Journal of Microscopy*, 193(1): 15−19, 1999.

[31] S. M. Smith and J. M. Brady, Susan - a new approach to low level image processing, International, *Journal of Computer Vision*, 23(1): 45−78, 1997.

[32] A. Baudes, B. Coll and J. M. Morel, A review of image denoising algorithms, with a new one, *Multiscale Model Simulation,* 4(2): 490−530, 2005.

[33] P. Perona and J. Malik, Scale space and edge detection using anisotropic diffusion, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 12: 629−639, 1990.

[34] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall Information and System Sciences, Prentice Hall, 1989.

[35] S. J. Ludtke, P. R. Baldwin and W. Chiu, EMAN: semi-automated software for high-resolution single-particle reconstructions, *Journal of Structural Biology,* 128: 82−97, 1999.

[36] W. Jiang, M. L. Baker, S. J. Ludtke and W. Chiu, Bridging the information gap: computational tools for intermediate resolution structure interpretation, *Journal of Molecular Biology,* 308: 1033−1044, 2001.

[37] R. C. Gonzales and R. E. Woods, *Digital Image Processing*, Second Edition, Prentice Hall, 2002.

[38] E. Le Pennec and S. Mallat, Geometrical image compression with bandlets, *Proceedings of the SPIE 2003*, 5150: 1273−1286, 2003.

[39] J. Starck, E. Candµes and D. Donoho, The curvelet transform for image denoising, *IEEE Transactions on Image Processing,* 11: 670−684, 2000.

[40] S. Mallat, *A Wavelet Tour for Signal Processing*, Academic Press, 1999.

[41] D. L. Donoho and I. M. Johnstone, Ideal spatial adaptation via wavelet shrinkage, *Biometrika*, 81: 425−455, 1994.

[42] D. L. Donoho and I. M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, *Journal of American Statistical Association,* 90(432): 1200−1224, 1994.

[43] D. L. Donoho, Denoising by soft-thresholding, *IEEE Transactions on Signal Processing*, 41(3): 613−627, 1995.

[44] S. G. Chang, B. Yu and M. Vetterli, Adaptive wavelet thresholding for image denoising and compression, *IEEE Transactions on Image Processing*, 9: 1532−1546, 2000.

[45] P. D. Grünwald, I. J. Myung and M. A. Pitt, *Advances in Minimum Description Length, Theory and Applications,* MIT Press, 2005, ISBN-10: 0-262-07262-9.

[46] P. D. Grünwald and P. M. B. Vitányi, Kolmogorov complexity and information theory with an interpretation in terms of questions and answers, *Journal of Logic, Language and Information,* 12: 497−529, 2003.

[47] J. Rissanen, Fisher information and stochastic complexity, *IEEE Transactions on Information Theory*, 42: 48−54, 1996.

[48] M. H. Hansen and Bin Yu, Model selection and the principle of minimum description length, *Journal of American Statistical Association*, 96(454): 746−774, 2001.

[49] J. Rissanen, A universal prior for integers and estimation by minimum description length, *The Annals of Statistics*, 11(2): 416−431, 1983.

[50] J. Rissanen, MDL denoising, *IEEE Transactions on Information Theory,* 46: 2537–2543, 2000.

[51] B. Dom, MDL estimation for small sample sizes and its application to linear regression, IBM Research Report RJ10030, June 13, 1996.

[52] T. Roos, P. Myllymäki and J. Rissanen, MDL denoising revisited, Submitted to *IEEE Transactions on Information Theory*, September 2006.

[53] B. K Rath and J. Frank, Fast automatic particle picking from cryo-electron micrographs using a locally normalized cross-correlation function: a case study, *Journal of Structural Biology*, 145(1-2): 84−90, 2004.

[54] J. R. Plaisier, R. I. Koning, H. K. Koerten, M. van Heel and J.P. Abrahams, TYSON: Robust searching, sorting and selecting of single particles in electron micrographs, *Journal of Structural Biology,* 145: 76−83, 2004.

[55] Z. Yu and C. Bajaj, Detecting circular and rectangular particles based on geometric feature detection in electron micrographs, *Journal of Structural Biology,* 145: 168−180, 2004.

[56] Y. Zhu, B. Carragher, F. Mouche and C. S. Potter, Automatic particle detection through efficient Hough transforms, *IEEE Transcations on Medical Imagining,* 22(9): 1053−1062, 2003.

[57] J. G. Daugman, Uncertainty relations for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters, *Journal of the Optical Society of America A*, 2: 1160−1169, 1985.

[58] J. Frank, *Three-dimensional Electron Microscopy of Macromolecular Assemblies*, Oxford University Press, 2006.

[59] M. van Heel and J. Frank, Use of multivariate statistics in analysing the images of biological macro-molecules, *Ultramicroscopy,* 6: 187−194, 1981.

[60] L. Lebart, A. Maurineau and K. M. Warwick, *Multivariate Descriptive Statistical analysis*, Wiley, New York, N.Y. , 1984.

[61] S. Bodjanova, Hard and fuzzy classification within the framework of hierarchical and optimization clustering, *Kybernetika*, 28: 81−84, 1992.

[62] J. Frank., M. Radermacher, P. Penczek, J. Zhu, Y. Li, M. Ladjadj and A. Leith, SPIDER and WEB: processing and visualization of images 3D electron microscopy and related fields, *Journal of Structural Biology,* 116: 190−199, 1996.

[63] S. Brandt, J. Heikkonen and P. Engelhardt, Automatic alignment of transmission electron microscope tilt-series without fiducial markers, *Journal of Structural Biology*, 136: 201−213, 2001.

[64] J. B. Heymann, M. Chagoyen and D. M. Belnap, Common conventions for interchange and archiving of three-dimensional electron microscopy information in structural biology, *Journal of Structural Biology*, 151: 196−207, 2005.

[65] T. S. Baker and R. H. Cheng, A model-based approach for determining orientations of biological macromolecules imaged by cryoelectron microscopy, *Journal of Structural Biology,* 116: 120–130, 1996.

[66] R. H. Cheng and L. Hammer, *Conformational Proteomics of Macromolecular Architecture*, World Scientific Printers (S) Pte Ltd, Singapore, 2004, ISBN 981-238-614-9.

[67] P. Natarajan, G. C. Lander, C. M. Shepherd, V. S. Reddy, C. L. Brooks and J. E. Johnson, Exploring icosahedral virus structures with VIPER, *Nature Reviews Microbiology*, 3(10): 809−817, 2005.

[68] S. J. Ludtke, J. Jakana, J. Song, D. T. Chuang and W. Chiu, A 11.5 Å single particle reconstruction of GroEL using EMAN, *Journal of Molecular Biology,* 314: 253−262, 2001.

[69] C. O. S. Sorzano1, R Marabini, G T Herman, Y Censor and J M Carazo, Transfer function restoration in 3D electron microscopy via iterative data refinement, *Physics in Medicine and Biology,* 49, 509−522, 2004.

[70] J. Radon, Uber die Bestimmung von Funktionen durch ihre Integralwerte langs gewisser Manningfaltigkeiten, Beritchte uber die Verhandlungen der Königlich Sachsischen gesellschaft der wissenschaftenzu Leipzig, *Math Physics Klasse,* 69: 262−277, 1917.

[71] U. Skoglund, L. G. Öfverstedt, R. M. Burnett and G. Bricogne, Maximum-entropy three-dimensional reconstruction with deconvolution of the contrast transfer function: A test application with adenovirus, *Journal of Structural Biology*, 117(3): 173−188, 1996.

[72] R. Lenz, 3-D Reconstruction with a projection onto convex sets algorithm, *Optics Communications,* 57: 21−25, 1986.

[73] R. Marabini, G. T. Herman and J. M. Carazo, 3D reconstruction in electron microscopy using ART with smooth spherically symmetric volume elements (blobs), *Ultramicroscopy*, 72(1-2): 53−65, 1998.

[74] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt,E. C. Meng and T. E. Ferrin, UCSF Chimera - A visualization system for exploratory research and analysis, *Journal of Computational Chemistry*, 25(13): 1605−1612, 2004.

[75] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I.N. Shindyalov and P. E. Bourne: The protein data bank, *Nucleic Acids Research*, 28: 235−242, 2000.

[76] K. A. Thiel, Structure-aided drug design's next generation, *Nature Biotechnology*, 22: 513−519, 2004.

[77] K. A. Dill, Dominant forces in protein folding, *Biochemistry*, 29(31): 7133−5155, 1990.

[78] J. W. Ponder and D. Case, Force fields for protein simulations, *Advances in Protein Chemistry*, 66: 27−85, 2003.

[79] D. Leckhood, Measuring the forces that control protein interactions, *Annual Review Biophysical and Biomolecular Structure*, 29: 1−26, 2000.

[80] S. Mecozzi, A. P. West and D. A. Dougherty, Cation-pi interactions in aromatics of biological and medicinal interest: electrostatic potential surfaces as a useful qualitative guide, *Proceedings of National Academy of Science*, 93: 10566−10571, 1996.

[81] S. K. Burley and G. A. Petsko, Aromatic-aromatic interaction: a mechanism of protein structure stabilization, *Science*, 229: 23–28, 1985.

[82] L. Serrano, M. Bycroft and A. R. Fersht, Aromatic-aromatic interactions and protein stability investigation by double-mutant cycles, *Journal of Molecular Biology*, 218: 465−475, 1991.

[83] Y. Tsuchiya, K. Kinoshita and H. Nakamura, Analyses of homo-oligomer interfaces of proteins from the complementarity of molecular surface, electrostatic potential and hydrophobicity, *Protein Engineering Design and Selection*, 19(9):421−429, 2006.

[84] R. Garian, Prediction of quaternary structure from primary structure, *Bioinformatics*, 17(6), 551-556, 2001.

[85] D. Casciari, M. Seeber and F. Fanelli, Quaternary structure predictions of transmembrane proteins starting from the monomer: a docking-based approach, *BMC Bioinformatics*, 7: 340, 2006.

[86] S. Vajda, Classification of protein complexes based on docking difficulty, *PROTEINS: Structure, Function and Bioinformatics*, 60:176−180, 2005.

[87] C. Yan, V. Honavar and D. Dobbs, Identification of interface residues in protease-inhibitor and antigen-antibody complexes: a support vector machine approach, *Neural Computing and Application*, 13: 123−129, 2004

[88] A. S. Aytuna, A. Gursoy and O. Keskin, Prediction of protein–protein interactions by combining structure and sequence conservation in protein interfaces, *Bioinformatics*, 21(12): 2850−2855, 2005.

[89] A. Koike and T. Takagi, Prediction of protein-protein interaction sites using support vector machines, *Protein Engineering Design and Selection*, 17(2): 165−173, 2004.

[90] P. Fariselli, F. Pazos, A. Valencia and R. Casadio, Prediction of protein–protein interaction sites in heterocomplexes with neural networks, *European Journal of Biochemistry*, 269: 1356−1361, 2002.

[91] S. Jones and J. M. Thorton, Prediction of protein-protein interaction sites using patch analysis, *Journal of Molecular Biology,* 272: 133−143, 1997.

[92] L. Wang and T. Jiang, On the complexity of multiple sequence alignment, *Journal of Computational Biology*, 1: 337−348, 1994.

[93] R. Durbin, S. Eddy, A. Krogh and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge University Press, 1998.

[94] C. Notredame and D. G. Higgins, SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Research*, 24(8): 1515−1524, 1996.

[95] J. D. Thompson, D. G. Higgins and T. J. Gibson, CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice, *Nucleic Acids Research*, 22: 4673−4680, 1994.

[96] S. Henikoff, J. G. Henikoff, W. J. Alford and S. Pietrokovski, Automated construction and graphical presentation of protein blocks from unaligned sequences, *Gene*, 163: 17−26, 1995.

[97] O. Lichtarge, H. R. Bourne and F. E. Cohen, An evolutionary trace method defines binding surfaces common to protein families, *Journal of Molecular Biology,* 257: 342−358, 1996.

[98] E. K. Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo and I. A. Vakser, Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques, *Proceedings of National Academy of Science*, 89(6): 2195−2199, 1992.

[99] J. J. Gray, S. Moughon, C. Wang, O. S. Furman, B. Kuhlman, C. A. Rohl and D. Baker, Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations, *Journal of Molecular Biology,* 331(1): 281−299, 2003.

[100] L. J. McGuffin, K. Bryson and D. T. Jones, The PSIPRED protein structure prediction server, *Bioinformatics,* 16: 404−405, 2000.

[101] K. Karplus, C. Barrett and R. Hughey, Hidden Markov models for detecting remote protein homologies, *Bioinformatics*, 14: 846−856, 1998.

[102] J. Meiler, M. Mueller, A. Zeidler and F. Schmaeschke, JUFO: Secondary structure prediction for proteins, *2002, http://www.jens-meiler.de/.*

[103] B. Rost and J. Liu, The PredictProtein server, *Nucleic Acids Research,* 31: 3300−3304, 2003.

[104] J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay and G. J. Barton, Jpred: A consensus secondary structure prediction server, *Bioinformatics,* 14: 892−893, 1998.

[105] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, 25(17): 3389−3402, 1997.

[106]T.P. Hopp and K.R. Woods, Prediction of protein antigenic determinants from amino acid sequences, *Proceedings of National Academy of Science*, USA, 78: 3824−3828, 1981.

[107] D. E. Kim, D. Chivian and D. Baker, Protein structure prediction and analysis using the Robetta server, *Nucleic Acids Research*, 32: 526−531, 2004.

[108] L. A. Kelley, R. M. MacCallum and M. J. E. Sternberg, Enhanced Genome Annotation using Structural Profiles in the Program 3D-PSSM, *Journal of Molecular Biology* 299(2): 501−522, 2000.

[109] L. Pauling and R.B Corey. Configurations of polypeptide chains with favored orientations of the polypeptide around single bonds: two pleated sheets, *Proceedings of National Academy of Science*, 37: 729−740, 1951.

[110] L. Pauling and R. B. Corey, Two hydrogen-bonded spiral configurations of the polypeptide chain, *Journal of American Chemistry Society*, 72: 5349, 1950.

[111] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff and D. C. Phillips, A three-dimensional model of the myoglobin molecule obtained by X-ray analysis, *Nature*, 181: 662, 1958.

[112] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells and J. M. Thornton, CATH - A hierarchic classification of protein domain structures, *Structure,* 5(8): 1093−1108, 1997.

[113] G. D. Rose, Hierarchic organization of domains in globular proteins, *Journal of Molecular Biology*, 134: 447−470, 1979.

[114] C. A. Orengo and J. M. Thornton, Protein families and their evolution - A structural perspective, *Annual Review of Biochemistry*, 74: 867−900, 2005.

[115] M. Tress, C. H. Tai, G. Wang, L. Ezkurdia, G. Lopez, A. Valencia, B. Lee and R. L. Dunbrack, Domain definition and target classification for CASP6, *Protein: Structure, Function and Bioinformatics,* 61(S7): 8−18, 2005.

[116] R. Janice, Folding and association of proteins, *Progress in Biophysical and Molecular Biology*, 49: 117−137, 1987.

[117] J. Liu and B. Roost, CHOP proteins into structural domain-like fragments, *Proteins*, 55(3): 678−688, 2004.

[118] L. Holm and C. Sander, Parser for protein folding units, *Proteins*, 19: 256−268, 1994.

[119] P. Kaukauna, A. Vacherie and A. Plunging, Mapping of the regions involved in phonotypic interactions of Tula hantavirus N protein, *Journal of Virology,* 77: 10910−10916, 2003.

[120] M. S. Formicari, G. Parisi and J. Echave, Quaternary structure constraints on evolutionary sequence divergence, *Molecular Biology and Evolution*, doi:10.1093/molbev/msl181, 2006.

[121] A. A. V. Albertan, K. W. Amy, M. Thaddeus, B. G. R. Raymond, MR. Cedric, Segue, W. Winfred and W. H. R. Rob, Crystal structure of the Rabies virus nucleoprotein-RNA complex, *Science*, 313: 360−363, 2006.

[122] J. G. Todd, Z. Xing, W. W. Gail and L. Ming, Structure of the vesicular stomatitis virus nucleoprotein-RNA Complex, *Science*, 313: 363−366, 2006.

[123] S. Yang, S. S. Chow, Y. Levy, M. S. Cheung, H. Levine, P. G. Wooliness and J. N. On chic, Domain swapping is a consequence of minimal frustration, *Proceedings of National Academy of Science*, 101(38): 13786−13791, 2004.

[124] N. Guex and M. C. Peitsch, SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling. *Electrophoresis,* 18: 2714−2723, 1997.

[125] S. E. Ferrando and L. A. Kolasa, Averages of best wavelet basis estimates for denoising, *Journal of Computational and Applied Mathematics*, 136(1): 357−367, 2001.

[126] F. Luisier, T. Blu, B. Forster and M. Unser, Which wavelet bases are the best for image denoising, in *Wavelets XI.* edited by M. Papadakis, A. F. Laine and A. Michael, *Proceedings of the SPIE*, 5914: 100−111, 2005.

[127] D. L. Donoho, H. Xiaohming, M. A. Unser, A. Aldroubi and A. F. Laine, Combined image representation using edgelets and wavelets, in *Wavelet applications in signal and image processing VII*, SPIE proceedings series, Denver CO, pp: 19−23, July 1999.