Challenge: Image restoration/Superresolution for Single Particle Analysis

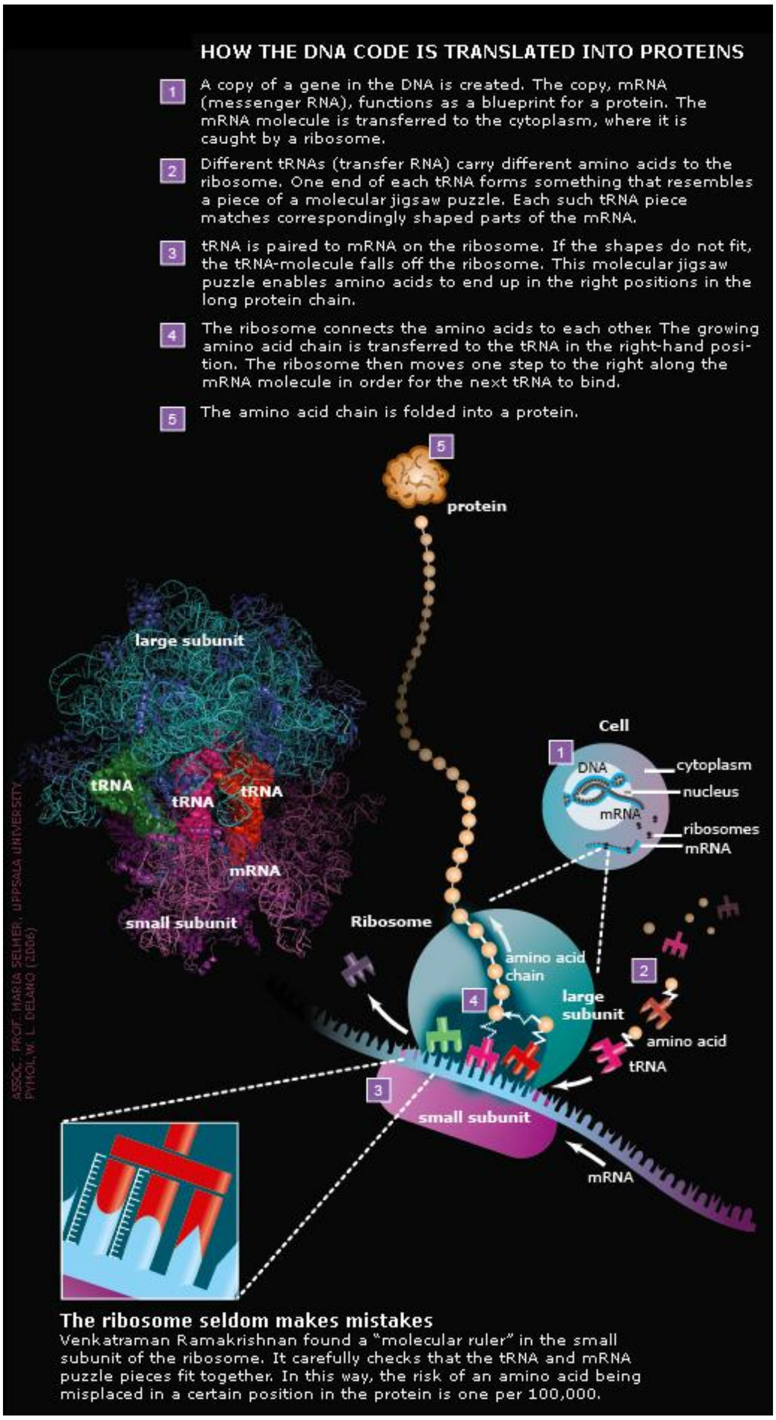
1. Significance and context

We may think of the eukaryotic cell as a complex automaton able of performing thousands of different functions and capable of adapting to its environment to perpetuate life through its lineage. These functions comprise its own nutrition and growing (metabolic functions), maintaining a given shape and internal organization (structural functions), communicating with other cells (signaling), moving in a medium (motility) or staying fixed at a place (adhesion), defending from external attacks (immunological functions), reproducing (replication), …

These functions are primarily performed by proteins, protein-derived compounds, RNA structures and/or a combination of them. In general, we refer to this set as macromolecular complexes, and their study has been called Structural Proteomics [Sali2003]. We may think of them as complex nanomachines carrying out a very specific biochemical function in a coordinated manner with other hundreds of thousands biochemical reactions concurrently taking place in the same cell (see Fig. 1).

Knowing the shape of the nanomachine and how it moves allows us to infer how it performs its functions within the cell (physiological conditions), and how it fails to perform them when it is “broken” or it is interfered (pathological or therapeutic conditions). It also allows us to design drugs that can block specific complexes, therefore, blocking the biochemical reaction and biochemical pathway taking place. It is estimated that all drugs in the market are currently interacting with about 300 different molecular (human or pathogen) targets [Overington2006].

Considering only the human genome, it contains about 35,000 genes and it has been estimated to produce up to 500,000 different proteins [Young2009]. The structure of 21,000 of these proteins has been resolved (about 4% of the total number of proteins) and publicly deposited at the Protein Data Bank (http://www.rcsb.org), although there is a high degree of redundancy in this database (only 30% of these 21,000 structures are unique, the rest are small variations of the formers). Solving a structure amounts to determining the location of each one of the structure atoms. X-ray crystallography and Nuclear Magnetic Resonance (NMR) are two most common experimental techniques to achieve these high-resolution structures. Experimental uncertainty translates into an uncertainty in the location of the atoms which is commonly referred to as resolution. Standard resolution for X-ray and NMR ranges between 1.5 and 3 Å (1 Å=10^{-10} meters; for instance, the Van der Waals radius of a carbon atom is 1.7 Å). Currently, there are large international consortia to address a massive determination of macromolecular structures by these techniques. In general, these efforts are referred to as structural genomics and they are aimed at cutting down the cost of structure determination [Chandonia2006].



HOW THE DNA CODE IS TRANSLATED INTO PROTEINS

1 A copy of a gene in the DNA is created. The copy, mRNA (messenger RNA), functions as a blueprint for a protein. The mRNA molecule is transferred to the cytoplasm, where it is caught by a ribosome.

2 Different tRNAs (transfer RNA) carry different amino acids to the ribosome. One end of each tRNA forms something that resembles a piece of a molecular jigsaw puzzle. Each such tRNA piece matches correspondingly shaped parts of the mRNA.

3 tRNA is paired to mRNA on the ribosome. If the shapes do not fit, the tRNA-molecule falls off the ribosome. This molecular jigsaw puzzle enables amino acids to end up in the right positions in the long protein chain.

4 The ribosome connects the amino acids to each other. The growing amino acid chain is transferred to the tRNA in the right-hand position. The ribosome then moves one step to the right along the mRNA molecule in order for the next tRNA to bind.

5 The amino acid chain is folded into a protein.

The ribosome seldom makes mistakes

Venkatraman Ramakrishnan found a "molecular ruler" in the small subunit of the ribosome. It carefully checks that the tRNA and mRNA puzzle pieces fit together. In this way, the risk of an amino acid being misplaced in a certain position in the protein is one per 100,000.

Fig. 1. (From the Nobel Prize summary) 2009 Nobel Prize in Chemistry was given to the studies of the structure and function of the ribosome (the machine in charge of translating messenger RNA into proteins). Watch http://www.youtube.com/watch?v=Jml8CFBWcDs to see the ribosome in action, and http://www.youtube.com/watch?v=RedO6rLNQ2o to see how different antibiotics interfere with the bacterial ribosomes.

However, not all macromolecular complexes are amenable to X-ray and NMR since not all of them crystallize (needed by X-ray diffraction), crystallization is a rather non-physiological condition, large macromolecular complexes cannot be resolved by NMR, and both techniques require a relatively high macromolecular concentration. Electron Microscopy (EM) is a complementary experimental technique that addresses these limitations [Henderson2004]: it is capable of looking at structures in nearly physiological state, it does not require high concentrations, and can naturally handle large complexes (this is particularly interesting since many proteins may participate in a single molecular tool). On the other hand, Electron Microscopy resolution is considered rather low compared to X-ray and NMR, in the range between 20 Å and 4 Å (with a current world record of 3.1 Å, http://www.ebi.ac.uk/pdbe-srv/emsearch/atlas/5256_summary.html).

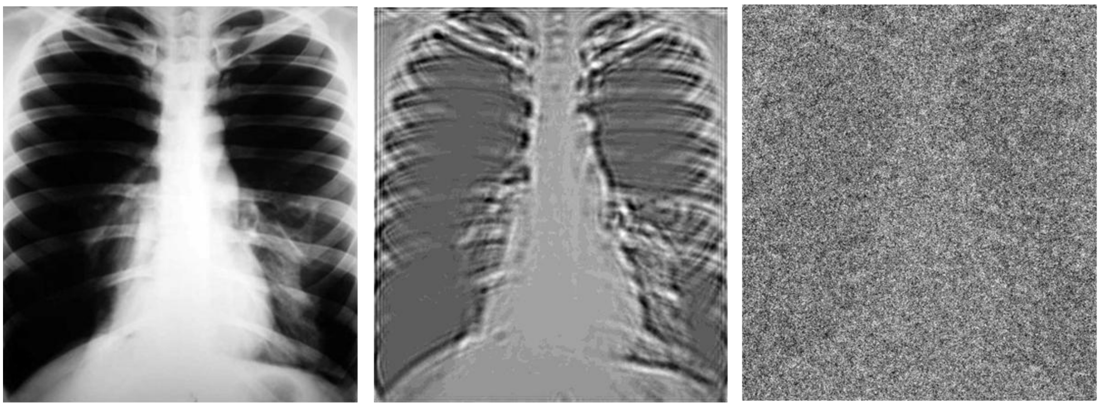
Resolution limits in Electron Microscopy mainly come from three sources [Henderson2004, Zhang2011]:

- Intrinsic macromolecular deformations: EM structures are determined by averaging hundreds of thousands of supposedly identical molecules. The problem is that this hypothesis may not be totally true and there might be subtle differences among the different copies of the macromolecule. This is actually the most limiting factor to high-resolution (in fact, those EM structures below 4 Å in resolution are highly symmetrical with very low possibilities of movement). However, this is also an advantage of the technique since it can visualize naturally occurring movements, giving hints on the way the macromolecule performs its function.
- Microscope aberrations: the Electron Microscope, as any other imaging device, introduces optical aberrations that distort the image (Fig. 2). Additionally, the electron beam damages the sample (by transferring energy to it, burning) and, consequently, the total electron dose must be kept low so that the sample keeps as much structural information as possible. This low dose constraint results in poorly contrasted images.
- Sample preparation: to be visualized in the Electron Microscope, macromolecules have to be embedded in amorphous ice at liquid nitrogen temperature (about -200°C). Ice thickness is not uniform along the sample and the contrast between the macromolecule and the ice is rather low. Acquired images contain information from the complex being reconstructed and information from the ice (which is seen as noise). Therefore, EM images have very low Signal-to-Noise ratio (below 1/10) and very low contrast (see Fig. 3)

Nevertheless, Electron Microscopy is gaining more and more resolution thanks to the following advances:

- Electron microscopes: the manufacturing technology of electron microscopes is progressing very rapidly (http://www.fei.com/resources/nanocenter/document-repository.aspx). In the recent years there have been great advances in the automation of images minimizing the electron dose needed, mechanical stability at cryo-temperature has been greatly increased, accelerating voltage has been raised to 300kV and its energy spread diminished, spherical aberration correctors have been incorporated, and new direct electron detectors have been put in-place (which has had a big impact on the reduction of the electron dose).
- Image processing: new and more powerful image processing algorithms have been put forward so that microscope aberrations, noise and structural heterogeneity are better handled [Sorzano2012]. The use of these new algorithms has made possible to

increase by two orders of magnitude the number of images that can be processed over the last 15 years (currently, there have been structural studies with several million images).

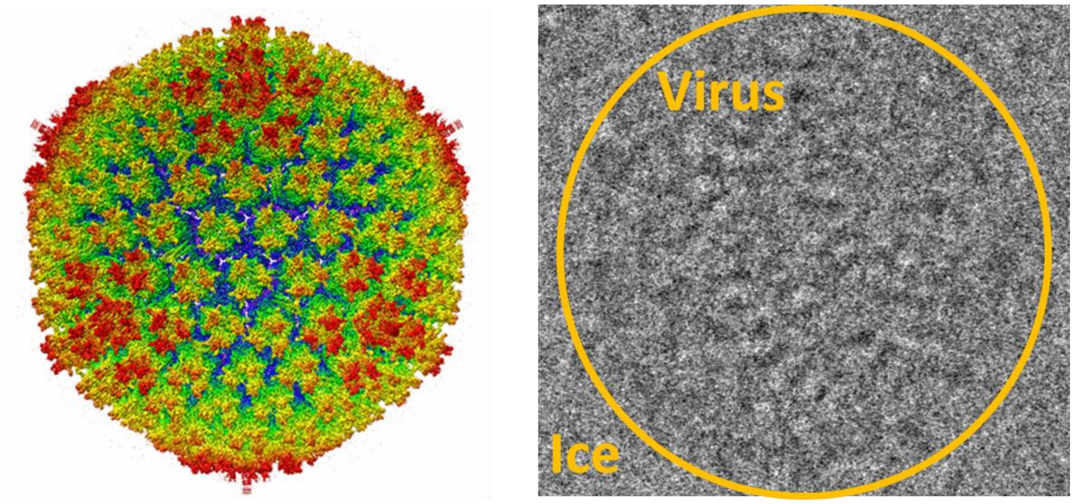


Ideal projection image

Distorted projection image

Distorted and noisy projection image

Fig. 2. Left: Ideal projection image (a 3D structure has been collapsed in 2D by projection). Middle: The Electron Microscope introduces optical distortions. Right: The recorded image is a noisy observation of the already distorted image.



Human adenovirus type 5 Virus (3D)

Electron Microscopy image (2D) of the virus embedded in ice

Fig. 3. Left: Example of macromolecular structure (Human adenovirus type 5) and its projection image acquired by an Electron Microscope (right).

2. Image restoration challenge

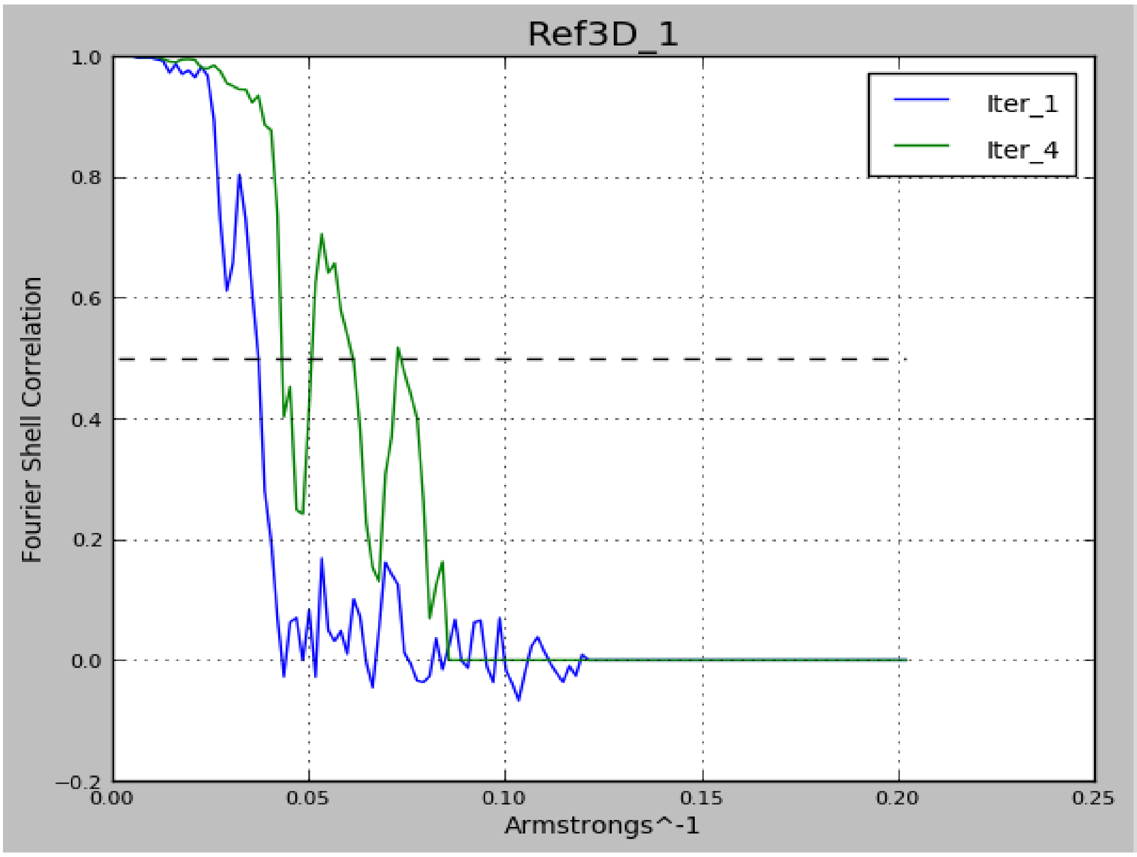
Within this framework there are important technological challenges that must be addressed by means of the efficient use of computational resources, new algorithmic developments in the field of image processing and pattern recognition, making these algorithms available to final users through the design of comprehensive software packages.

Hypothesis

The current resolution limit imposed by the experimental setting can be pushed forward by applying superresolution or image restoration techniques as has already been done in other scientific fields. The existence of macromolecular structures known at atomic resolution (Protein Data Bank: http://www.rcsb.org) provides lots of a priori information than can be used to enhance resolution. In order to exploit this information you may use any of the PDB to voxel density converters freely available as xmipp_volume_from_pdb, Spider cp from pdb, Eman procpdb, Situs pdb2vol, Bsoft bsf, or using any other program you may know. You may filter the volume to any desired frequency using xmipp_transform_filter, Spider ff, Eman e2filtertool, Bsoft bfilter, or again, any other tool of your own. To convert the format of the volumes you may use xmipp_image_convert, Eman e2proc3d, Bsoft bconvert, em2em, or any other tool.

Challenge

Several (in the order of a hundred) structures will be provided to participants at different resolutions. Challengers are supposed to design and apply **image restoration, superresolution, resolution enhancement or any other technique** that successfully recovers information not kept by the experimental setup. To check the quality of the resolution enhancement, the enhanced structure will be compared to the same structure at a resolution of 1 Å using the Fourier Shell Correlation [Grigorieff2000]. The Figure of Merit of each submission will be the average for all structures of the difference between the FSC of the original (low-resolution) structure and the high-resolution structure, and the FSC of the enhanced structure and the high-resolution structure. The figure below shows a typical Fourier Shell Correlation curve with an increase of resolution due to the iterative nature of the reconstruction process



The Figure of Merit for this challenge will be the área between the two curves. Note that this is calculated for 1 volume, and the challenge prize is on the average of the 120 volumes.

Expected outcome

Having access to a faithful, higher resolution structure will allow structural biologists to further understand the biological mechanisms underlying physiological and pathological functioning of these macromolecular machines.

The awarded teams will be required to give the source code and corresponding scripts so that results can be reproduced and verified.

Presentation of results

The best results will be invited to give a talk at IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2014 at Florence, Italy, and a special issue of *** will be devoted to publish selected algorithms.

Data availability and Submission of results

Test data is available from http://i2pc.cnb.csic.es/3dembenchmark starting on **October 1st**, 2013. Challengers may upload their results on this test data from **November, 1st**. Uploaded results will be automatically assessed so that challengers have immediate feedback on their performance. The resulting volumes have to be uploaded as a .tar.gz with the same filename, directory structure and data format as they were downloaded.

Challenge data is available on **January 15th**, 2014. **Challenge results** can be sent till **February, 28th**, 2014. Then, all results will be assessed and the challenge winners will be announced by **March, 21th**.

Key Dates

October 1st, 2013: Test data is available
November 1st, 2013: Results on test data can be submitted
December 2nd, 2013: Team registration to join the SP Cup competition
January 15th, 2014: Challenge data is available and results can be submitted
February 28th, 2014: Challenge results and paper submission closes
March 21st, 2014: Challenge winners are announced
May 4th-9th, 2014: Presentation of results at ICASSP 2014

3. Data generation, image restoration and superresolution

Test and challenge data has been generated mimicking the image formation process in Electron Microscopy:

- Four atomic structures have been downloaded from the PDB
- They were converted into voxel gray densities using xmipp_volume_from_pdb_ at 1 Angstrom/pixel.
- For each structure, 50,000 projections were generated at random orientations using xmipp_phantom_project.
- Noise was added to a Signal-to-Noise Ratio of 0.1, 0.2 and 0.4 and 34 different contrast transfer functions (CTFs) were applied (acceleration voltage 300kV, defocus range 1.8-2.1 µm, spherical aberration 2.26 mm) using xmipp_phatom_simulate_microscope. The CTFs are the same ones as the ones of the Bovine Papilloma Virus [Wolf2010].
- For each structure and SNR, 5 random subsets of different sizes were extracted from the 50,000 projections with the following distribution

SNR	Subset size
0.1	500, 1000, 5000, 10000
0.2	10000
0.4	10000

  and for each one of them we applied a projection matching with CTF correction [Scheres2010]. The resulting volume was low pass filtered to the resolution estimated by the protocol. A total of 120 volumes at different resolutions.

Note that the limitation of resolution is not imposed by the latest low-pass filtering that acts only as a “certifier” of the resolution loss, but by the low SNR, the alignment errors induced by this low SNR, the envelope of the Contrast Transfer Function that vanishes at a frequency of about 5 Angstroms, the fact of isotropically correcting a transfer function that is anisotropic, the lack of projections, interpolation errors, …

The loss of resolution in the volume does not correspond to the standard linear model

$$V_{observed} = HV_{ideal} + N$$

although, undeniably, a linear model is a first approximation to the non-linear relationship underneath. In any case, the convolution kernel H would have to be estimated from the data.

Superresolution algorithms needing several realizations of the observation may use the 5 repetitions within each SNR and subset size.

Finally, algorithms relying on *a priori* information may use standard image statistics known from natural images or exploit the content of the Protein Data Bank (http://www.rcsb.org/pdb) in which more than 90,000 structures are known at atomic resolution.

4. Directory structure and data formats

The different volumes are organized according to the different SNR and subset sizes

SNR_XX/SubsetSize_YYYYY/WWWW_ZZ.mrc

XX is the Signal-to-Noise Ratio and takes values 01, 02, 04 standing for 0.1, 0.2 and 0.4 respectively. YYYYY is the number of projections used for the 3D reconstruction. WWWW is the label identifying the molecule (four different molecules: aaaa, bbbb, cccc, and dddd). ZZ is the realization number (5 different realizations corresponding to 5 different random subsets whose size is YYYYY).

Files are written in MRC file format (specifications). From MATLAB you may use the function xmipp_read that is provided with Xmipp 3.1 or the MATLAB API provided by OMERO. Alternatively, volumes can be directly read in memory knowing that they follow a header+raw data structure with the parameters below:

Volume	Header (bytes)	Raw data size
aaaa	1760	220x220x220 floats (4 bytes)
bbbb	1680	210x210x210 floats (4 bytes)
cccc	1360	170x170x170 floats (4 bytes)
dddd	1280	160x160x160 floats (4 bytes)

Contact us

All email correspondence associated to this Challenge will be made from the following email address: 3dembenchmark@cnb.csic.es

Bibliography

[Chandonia2006] Chandonia, J.-M. & Brenner, S. E. The impact of structural genomics: expectations and outcomes. Science, 2006, 311, 347-351

[Grigorieff2000] Grigorieff, N. Resolution measurement in structures derived from single particles Acta Crystallographica section D, 2000, 56, 1270-1277

[Henderson2004] Henderson, R. Realizing the potential of electron cryo-microscopy Quarterly Review of Biophysics, 2004, 37, 3-13

[Overington2006] Overington, J. P.; Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? Nat Rev Drug Discov, 2006, 5, 993-996

[Sali2003] Sali, A.; Glaeser, R.; Earnest, T. & Baumeister, W. From words to literature in structural proteomics Nature, 2003, 422, 216-225

[Scheres2010] Scheres, S. H. W.; Núñez-Ramírez, R.; Sorzano, C. O. S.; Carazo, J. M. & Marabini, R. Image processing for electron microscopy single-particle analysis using XMIPP Nature Protocols, 2008, 3, 977-990

[Sorzano2012] C.O.S. Sorzano, J. M. de la Rosa Trevín, J. Otón, J. J. Vega, J. Cuenca, A. Zaldívar-Peraza, J. Gómez-Blanco, J. Vargas, A. Quintana, R. Marabini, J. M. Carazo. Semiautomatic, high-throughput, high-resolution protocol for three-dimensional reconstruction of Single Particles in Electron Microscopy. Nanoimaging: Methods and Protocols. Methods in Molecular Biology, 950: 171-193. Eds. Alioscka Sousa, Michael Kruhlak. Humana Press. (2012)

[Wolf2010] Wolf, M.; Garcea, R. L.; Grigorieff, N. & Harrison, S. C. Subunit interactions in bovine papillomavirus. Proc. Natl. Acad. Sci. USA, 2010, 107, 6298-6303

[Young2009] Young, D. C. Computational drug design John Wiley & Sons, 2009

[Zhang2011] Zhang, X. & Zhou, Z. H. Limiting factors in atomic resolution cryo electron microscopy: no simple tricks. J Struct Biol, 2011, 175, 253-263