Quadratic Programming Approach to Fit Protein Complexes into Electron Density Maps

Roman Pogodin, Alexander Katrutsa, Sergei Grudinin

Quadratic Programming Approach to Fit Protein Complexes into Electron Density Maps

Roman Pogodin
Skolkovo Institute of Science and Technology, Nobel St., 3, Moscow, 143026, Russia
Moscow Institute of Physics and Technology, Institutskiy Lane 9, Dolgoprudny, Moscow, 141700, Russia
pogodin@phystech.edu

Alexander Katrutsa
Moscow Institute of Physics and Technology, Institutskiy Lane 9, Dolgoprudny, Moscow, 141700, Russia
Skolkovo Institute of Science and Technology, Nobel St., 3, Moscow, 143026, Russia
aleksandr.katrutsa@phystech.edu

Sergei Grudinin
University of Grenoble Alpes, LJK, F-38000 Grenoble, France
CNRS, LJK, F-38000 Grenoble, France
Inria, F-38000 Grenoble, France
sergei.grudinin@inria.fr

Abstract

The paper investigates the problem of fitting protein complexes into electron density maps. They are represented by high-resolution cryoEM density maps converted into overlapping matrices and partly show a structure of a complex. The general purpose is to define positions of all proteins inside it. This problem is known to be NP-hard, since it lays in the field of combinatorial optimization over a set of discrete states of the complex. We introduce quadratic programming approaches to the problem. To find an approximate solution, we convert a density map into an overlapping matrix, which is generally indefinite. Since the matrix is indefinite, the optimization problem for the corresponding quadratic form is non-convex. To treat non-convexity of the optimization problem, we use different convex relaxations to find which set of proteins minimizes the quadratic form best.

Keywords: cryoEM, electron microscopy fitting, quadratic programming, protein structure prediction

1. Introduction

The problem of proteins fitting into cryoEM density maps of protein complexes remains important for biophysical studies of cell processes. Some examples of its importance can be found in [1], which presents EMDataBank, an online database for electron microscopy.

Two approaches to the problem are noticeable. The first one [2] uses a genetic algorithm that discovers and then recombines good solutions which fit the density map. This parallel approach increases efficiency, but the accuracy decreases with the number of components inside the complex and the map's resolution. The second solution [3] uses a cryoEM map directly and uses a set of predefined possible positions. It divides a set of fitting variables into uncoupled subsets, solves combinatorial optimization problems independently and finally gather all the solutions into the global minimum. This approach reduces the size of the problem from exponential in the number of all components to exponential in the number of components of the largest subset. However, it is sensitive to the accuracy of the component models and clustering into the subsets. Methods investigated in the current paper use a set of predefined positions like in the last paper. However, the minimization problem over a binary set is relaxed into a problem over a continuous set. The problem is then solved with continuous optimization methods, which are more efficient that discrete ones. The solution is then rounded to a binary one.

For continuous optimization methods, quadratic programming approaches and a stochastic algorithm are used in this paper. General overview of quadratic programming, main ideas and results of convex optimization are presented in [4].For our purposes, the first idea is convex relaxation. The paper [5] is a fulfilling re-

view of a basic methods semidefinite (SDP). The second idea is sequential quadratic programming which finds a local minimum of a problem [6]. The stochastic algorithm is called Simulated Annealing which also finds a local minimum of the problem [7].

Efficiency of the method is tested in two principle ways. Firstly, it is testing against artificial cryoEM maps, which are based on a known structure of a protein. This method allows one to test efficiency at different map resolutions and was used in [2,3,8–11]. The second way is to use experimental density maps of protein complexes which structure we know, e.g. in [2,3,9,10].

2. Problem statement

A protein complex consists of m proteins and has N computed spatial positions for each protein, which are different for different proteins. To predict the structure of the protein, we look for positions of a given set of proteins which fit into the density map best. Introduce correlation parameters for proteins' density maps and state the prediction problem formally.

Definition 1 *Overlapping of two proteins' positions is overlapping of corresponding electron density maps.*

To measure overlapping between two density maps, we use the cross-correlation function [12], CCF:

$$\mathrm{CCF} = \sum_i \rho_i^1 \rho_i^2, \tag{1}$$

where ρ_i^1 and ρ_i^2 are densities of i-th element of two maps. Laplacian-filtered CCF (LAP) allows one to compare matching of maps' edges rather than the whole volumes [12]. It can be achieved with modifying both maps with Laplacian filter before computing the CCF. Motivation of its usage and the filter kernel are described in [13].

The ideas of CCF and LAP allows us to introduce four overlapping scores. Firstly, CCF itself can be computed. Then overlapping shows how incompatible are the positions of two proteins. For example, if they are too close or even has two atoms in a same position, the CCF will be big. This approach is denoted as CCF, and the goal is to minimize it.

Secondly, both maps can be filtered with the Laplacian filter to find the best match of their contours. This score is called the Contact score, and the goal is to maximize it.

The last two approaches imply applying the Laplacian filter to only one map, hence these scores shows how the contour of one map fits the volume of the other. These scores are called Skin-Core and Core-Skin scores and must be maximized to find the best match.

All scores except CCF are computed and then multiplied by -1 to write the optimization problem as the minimization one for all four cases. These scores allows us to introduce a matrix where each element considers overlapping between two positions of proteins.

Definition 2 *Let $\mathbf{Q} \in \mathbb{R}^{n\times n}$ be an overlapping matrix that corresponds to a density map, where $n = m \cdot N$.*

Each matrix element q_{ab} shows overlapping between i-th and j-th components of the complex, which are in k-th and l-th positions respectively, so $a = (i-1)\cdot N + k$ and $b = (j-1)\cdot N + l$. Overlapping between two positions of a single protein is set to zero.

Besides relative positions of proteins, their fitting to the density map should be considered. It means that it is more important to fit a protein's map to the complex' map contour rather than its volume, because we need to achieve the original position within the complex.

Definition 3 *Relevance of a protein's position is a measure of quality of it's fit into a density map's contour.*

Relevance can be measured with the LAP, since it can be used for contour matching.

With that, a relevance vector, which describes each possible position, can be introduced.

Definition 4 *Let $\mathbf{b} \in \mathbb{R}^n$ be a component relevance vector. Each element b_a shows relevance of i-th component to k-th position in the complex, where $a = (i-1)\cdot N + k$.*

Further, the problem's variable shows taken positions for each protein and is defined as following.

Definition 5 *Let $\mathbf{x} \in \{0,1\}^n$ be a binary vector that represents the proteins positions in the complex:*

$$x_i^k = \begin{cases} 1, & \textit{if } i\textit{-th protein is in the } k\textit{-th position,} \\ 0, & \textit{otherwise,} \end{cases}$$

where $a = (i-1)\cdot N + k$.

The vector $\mathbf{x}$ is divided into m subvectors for the proteins, each with length of N for possible positions.

Since the best set of positions should have minimum overlapping between proteins and maximum relevance between each protein and the map, formulate the problem as a constrained binary quadratic optimization problem. It considers the overlapping in the quadratic term and the relevance in the linear term:

$$\begin{aligned} \mathbf{x}^* &= \underset{\mathbf{x}\in\{0,1\}^n}{\arg\min} \left(\mathbf{x}^\mathsf{T}\mathbf{Q}\mathbf{x} - \mathbf{b}^\mathsf{T}\mathbf{x}\right), \\ \text{s.t. } & \mathbf{A}\mathbf{x} = \mathbf{1}_m, \end{aligned} \tag{2}$$

where $\mathbf{1}_m$ is a vector of ones and $\mathbf{A} \in \mathbb{R}^{m\times Nm}$ is a matrix ensures that each protein within the complex takes a

single position. Hence, it has the following structure:

$$\mathbf{A} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ \underbrace{0 \cdots 0}_{N} & & & \underbrace{0 \cdots 0}_{N} & & & \cdots & \underbrace{1 \cdots 1}_{N} & & \end{bmatrix}. \quad (3)$$

Since matrix $\mathbf{Q}$ is indefinite in general case, the problem (2) is non-convex. Moreover, the problem (2) is NP-hard. Because of that, relax integer constraints of the problem into continuous variables. The problem then can be solved with continuous optimization methods. The last approach will be discussed later, but reformulation in a continuous form can be written now as

$$\begin{aligned} \mathbf{y}^* = &\arg\min_{\mathbf{y}\in[0,1]^n} \left(\mathbf{y}^\top \mathbf{Q}\mathbf{y} - \mathbf{b}^\top \mathbf{y}\right), \\ \text{s.t. } &\mathbf{A}\mathbf{y} = \mathbf{1}_m. \end{aligned} \quad (4)$$

To define which proteins takes which place, return to the binary vector. To do that, replace the biggest element in each of N position subvectors by 1 and others by 0:

$$i = 1, \ldots, m: \; x_a = \begin{cases} 1, & \text{if } k = \arg\max_{k=1,\ldots,N} y_a, \\ 0, & \text{otherwise} \end{cases}$$

for $a = (i-1)\cdot N + k$.

2.1. Convex relaxations

Since (4) is a non-convex problem over a convex set, it can not be solved directly with guarantees of global minimum. However, an approximate solution can be found with relaxing the problem into a convex one. In the next sections spectrum shift and semidefinite relaxations of the problem are introduced. These approaches help in finding an approximate solution, using solutions of the relaxed problems.

Spectrum shift relaxation (Shift). We shift the spectrum of matrix $\mathbf{Q}$ to achieve positive-semidefiniteness and hence make the problem convex. The corresponding transformation is

$$\hat{\mathbf{Q}} = \mathbf{Q} - \lambda_{\min}\mathbf{I}, \quad (5)$$

where $\lambda_{\min}$ is the smallest eigenvalue of $\mathbf{Q}$ and $\mathbf{I}$ is an identity matrix of the size of $\mathbf{Q}$. Then the problem (4) can be rewritten with the new matrix $\hat{\mathbf{Q}}$ as

$$\begin{aligned} \mathbf{y}^* = &\arg\min_{\mathbf{y}\in[0,1]^n} \left(\mathbf{y}^\top \hat{\mathbf{Q}}\mathbf{y} - \mathbf{b}^\top \mathbf{y}\right), \\ \text{s.t. } &\mathbf{A}\mathbf{y} = \mathbf{1}_m. \end{aligned} \quad (6)$$

The problem (6) can now be easily solved as a convex on. However, this method does not guarantee the global minimum of the initial problem (2).

Semidefinite relaxation (SDP). To introduce the semidefinite relaxation, rewrite the problem (4) as

$$\begin{aligned} \mathbf{y}^* = &\arg\min_{\mathbf{y}\in[0,1]^n} \left(\mathrm{Tr}\left(\mathbf{Q}\mathbf{Y}\right) - \mathbf{b}^\top \mathbf{y}\right), \\ \text{s.t. } &\mathbf{A}\mathbf{y} = \mathbf{1}_m, \\ &\mathbf{Y} = \mathbf{y}\mathbf{y}^\top. \end{aligned} \quad (7)$$

To get a lower bond of the solution, relax the last constraint from equalities to inequalities, so now it is positive semidefinite:

$$\mathbf{Y} - \mathbf{y}\mathbf{y}^\top \succeq 0.$$

But the initial binary program implies

$$\mathrm{diag}\left(\mathbf{x}\mathbf{x}^\top\right) = \mathbf{x}.$$

Hence, we use an additional constraint

$$\mathrm{diag}\left(\mathbf{Y}\right) = \mathbf{y}$$

to bound the problem. The relaxed problem is now convex and can be written as

$$\begin{aligned} \mathbf{y}^* = &\arg\min_{\mathbf{y}\in[0,1]^n} \left(\mathrm{Tr}\left(\mathbf{Q}\mathbf{Y}\right) - \mathbf{b}^\top \mathbf{y}\right), \\ \text{s.t. } &\mathbf{A}\mathbf{y} = \mathbf{1}_m, \\ &\mathbf{Y} - \mathbf{y}\mathbf{y}^\top \succeq 0, \\ &\mathrm{diag}\left(\mathbf{Y}\right) = \mathbf{y}. \end{aligned} \quad (8)$$

Sequential quadratic programming (SQP). The basic ideas of sequential quadratic programming are described in Chapter 18 of the book [6]. This approach finds a local minimum for a non-convex problem and implies solving a quadratic subproblem at each iteration. The subproblem is a convex second-order approximation of the Lagrangian function of the (4), i.e. it involves a positive-semidefinite approximation of the Hessian. We use the implementation of an SQP algorithm from MATLAB Optimization Toolbox [14]. The initial point for the algorithm is obtained with solving the linear part of the problem (4):

$$\begin{aligned} \mathbf{y}^* = &\arg\min_{\mathbf{y}\in[0,1]^n} \left(-\mathbf{b}^\top \mathbf{y}\right), \\ \text{s.t. } &\mathbf{A}\mathbf{y} = \mathbf{1}_m. \end{aligned} \quad (9)$$

In this work, the solution of the problem (9) can be treated as an approximation that only fits the density map, but does not consider overlapping between proteins.

Simulated annealing (SA). The simulated annealing method [7] is a probabilistic global optimization method, implemented in MATLAB Global Optimization Toolbox [14]. This approach simulates a physical

process of heating and then slow lowering the temperature of a material to decrease defects. At each iteration, it generates a new point near the current one, with a uniformly random direction and step length equals the current temperature. If the new point is better than the current one, the algorithm accepts it. If not, it accepts the point with probability

$$\mathbb{P}(\text{accept } x_{k+1}) = \left(1 + \exp\left(\frac{\Delta}{T_k}\right)\right)^{-1}, \qquad (10)$$

where $\Delta = f(x_{k+1}) - f(x_k)$ for an objective function f, current and new points x_k and x_{k+1} respectively and the current temperature T_k, which changes as

$$T_{k+1} = 0.95 \cdot T_k.$$

For this work, the method can not be directly implemented for the problem (4) because it has linear constrains, but the method is designed for unconstrained and bound-constrained problems. However, constraints can be implemented as a penalty function to the objective one, so the problem is converted as

$$\mathbf{y}^* = \underset{\mathbf{y}\in[0,1]^n}{\arg\min} \left(\mathbf{y}^\intercal \mathbf{Q}\mathbf{y} - \mathbf{b}^\intercal \mathbf{y} + w\|\mathbf{A}\mathbf{y} - \mathbf{1}_m\|_1\right), \qquad (11)$$

where $w \in \mathbb{R}$ is a penalty weight and $\|g\|_1$ denotes the l_1-norm of a vector g. The initial point for the algorithm is obtained by solving (9). Moreover, since the algorithm has two parameters, the initial temperature T_0 and the penalty weight w, the method is denoted as $\mathrm{SA}(T_0, w)$.

2.2. Scoring functions to measure quality of fit

As the simplest scoring function which implies knowledge of the real structure of the protein, one can use root-mean-square deviation (RMSD) [15]. It measures the distance δ_i between pairs of i-th atoms of a protein in two positions, one in predicted position and one in the native position. Both atoms in a pair take the same place in a corresponding protein. With M pairs of such atoms, it can be written as

$$\mathrm{RMSD} = \sqrt{\frac{1}{M}\sum_{i=1}^{M} \delta_i^2}. \qquad (12)$$

This approach helps to measure quality of other criteria on test data, but can not be used with direct determination of an unknown structure.

Another approach is quality criteria that use only information about the density map. For future work, we propose two scoring functions recommended in the paper [12]. Both of them use electron density maps of an initial structure and one from the solution. A way to produce a probe density map from the discrete solution $\mathbf{x}^*$ is described in [12]. It includes following steps.

1. Get the atomic structure from fitted proteins using the discrete solution $\mathbf{x}^*$.
2. Impose a 3D grid with voxel size of 1 Å.
3. For every non-hydrogen atom increase the density value of the nearest voxel by the atomic number of the atom.
4. Apply the Gaussian Fourier filter to blur the map. The recommended in [12] sigma is $0.187\times$ resolution. The Gaussian kernel size is $2 \cdot \lceil 2\sigma \rceil + 1$, where $\lceil x \rceil$ is the smallest integer greater than or equal to x.
5. Resample the grid using Fourier method to match the sampling of the target map.

When the probe map is obtained, compare the original (target) map and the probe one with the scoring functions describing above.

For 10Å resolution or less the authors propose the Laplacian-filtered cross-correlation function. The cross-correlation function itself is described above (1). For the LAP, modify both target and probe maps with Laplacian filter before computing the CCF. For other cases, the mutual information score (MI) is proposed. The scoring function is

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \qquad (13)$$

Here, X and Y correspond to the density values in the probe and target maps. Functions $p(x)$ and $p(y)$ are the percentage of values in maps equal to x and y. The aligned maps are maps where elements with equal coordinates represent one point in space. For aligned target and probe maps, $p(x,y)$ is percentage of elements with value x in the probe map and y in the target one. Because of wide range of values and noise, X and Y have limited number of values, e.g. 20 in [12].

3. Dataset

All methods were tested on simulated maps, since it allows one to compare the methods for different protein complexes that are tested in the same conditions, i.e. the same resolution of the map. The maps were generated as described below from 7 protein complexes. Their PDB entries are *1e6v* [17], *1gte* [16], *1tyq* [18], *1z5s* [19], *2p4n* [20], *4a6j* [21], *4bij* [22]. The map resolution is 10 Å, a voxel size is 1 Å. Map's generation is similar to creating a map form a solution and includes following steps.

1. Impose a 3D grid with voxel size of 1 Å.
2. For every non-hydrogen atom increase the density value of the nearest voxel by the atomic number of the atom.

3. Apply the Gaussian Fourier filter to blur the map. The recommended in [12] sigma is $0.187\times$resolution. The Gaussian kernel size is $2 \cdot \lceil 2\sigma \rceil + 1$.

4. Computational experiment

Quality of fit is characterized by achieved RMSD of each protein in a complex (2). The solution is treated as correct if each protein has $\text{RMSD} \leq 10$ Å. Every results is characterized by correct answers ratio

$$\beta = \frac{N_c}{N},$$

where N_c is a number of correctly determined protein's positions and N is a number of proteins within the complex. Results for simulated annealing method are presented for following parameters: initial temperature is 100, penalty weight is 1. Results for other parameters were the same for all simulations (data not shown), which is connected with the quality of the linear solution described below. Moreover, only RMSD is used as a scoring function, since the purpose of the current work is to test the proposed optimization approach, and RMSD completely represents how precise an obtained solution is.

Comparison of quality of fit and the objective function's value. Results presented in Table (1) for the complex *1gte* show that the highest β corresponds to the lowest approximate objective value. Moreover, the SQP and SA methods find a continuous solution, which gives almost the same objective value as the binary solution.

Comparison of the methods with the linear approximation. Table (2) with results for *1e6v*, *1gte*, *1z5s* shows that the highest β was obtained by simulated annealing and sequential programming approaches. However, the initial point for both methods was obtained from the solution of the linear problem (9), and β for SQP, SA and the linear problem is almost the same. Hence, correct solutions for these datasets can be achieved using only the linear approach.

Overall performance. Despite for some complexes mentioned above the linear approach leads to the correct solution (and to the correct solution with SQP and SA), in general it is not true. Moreover, SQP and SA sometimes make the linear solution even worse, which can be observed on the following Figure 1, that shows average β for all complexes and map scores. Therefore, in these cases fitting to the given map is more important than arranging proteins' positions among each other.

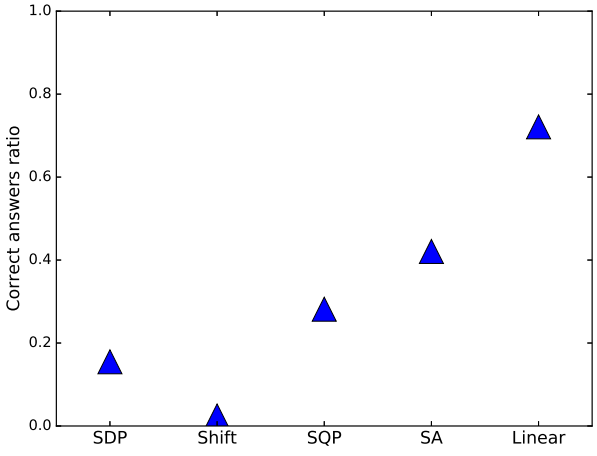

Figure 1: Averaged for all complexes and all map scores β

5. Conclusion

The paper investigates quadratic programming approach for the problem of determining proteins' position inside a complex by its EM density map. The mathematical optimization problem is formulated using overlapping of proteins in computed positions between each other, which forms an overlapping matrix, and with the given math, which gives a relevance vector.

We tested semidefinite relaxation, spectrum shift relaxation, sequential quadratic programming as quadratic programming approaches and simulated annealing and linear approximation for comparison with quadratic methods. The datasets were formed from simulated density maps. The best performance was shown by SQP, SA and linear approximation methods, which shows the importance of fitting a protein to the given map rather than looking for the best positions with relate to other proteins in the complex.

6. Acknowledgements

This work is supported by RFBR, grant 16-37-00485.

References

[1] C.L. Lawson, M.L. Baker, C. Best, C. Bi, M. Dougherty, P. Feng, G. Van Ginkel, B. Devkota, I. Lagerstedt, S.J. Ludtke, R.H. Newman, T.J. Oldfield, I. Rees, G. Sahni, R. Sala, S. Velankar, J. Warren, J.D. Westbrook, K. Henrick, G.J. Kleywegt, H.M. Berman, and W. Chiu. Emdatabank.org: Unified data resource for cryoem. *Nucleic Acids Research*, 39(SUPPL. 1):D456–D464, 2011.

Table 1: Objective function's optimal values for *1gte*, Contact, CCF, Skin-Core and Core-Skin scores.

	Contact			CCF		
Method	Continuous	Binary	β	Continuous	Binary	β
SDP	-62711135	23648296	0.50	-2261626	563438	0.25
Shift	1990252	1483316	0.00	78968	43276	0.00
SQP	-3707363	-3707363	1.00	1086	1086	0.00
SA$(100, 1)$	-3707354	-3707363	1.00	205462	205462	1.00
Linear	-80	-3707363	1.00	-80	205462	1.00
	Skin-Core			Core-Skin		
Method	Continuous	Binary	β	Continuous	Binary	β
SDP	-26429514	9343938	0.50	-26196482	2055248	0.25
Shift	801317	499195	0.00	789572	210241	0.00
SQP	-2382769	-2382769	1.00	-2352385	-2352385	1.00
SA$(100, 1)$	-2382764	-2382769	1.00	-2352380	-2352385	1.00
Linear	-80	-2382769	1.00	-80	-2352385	1.00

Table 2: Results for three complexes. Each column represents a method, each value is a correct answers ratio β.

1e6v					
Scoring	SDP	Shift	SQP	SA	Linear
Contact	0.33	0.17	0.67	0.67	0.67
CCF	0.33	0.00	0.33	0.67	0.67
Skin-Core	0.50	0.00	0.67	0.67	0.67
Core-Skin	0.67	0.17	0.67	0.67	0.67
1gte					
Scoring	SDP	Shift	SQP	SA	Linear
Contact	0.50	0.00	1.00	1.00	1.00
CCF	0.25	0.00	0.00	1.00	1.00
Skin-Core	0.50	0.00	1.00	1.00	1.00
Core-Skin	0.25	0.00	1.00	1.00	1.00
1z5s					
Scoring	SDP	Shift	SQP	SA	Linear
Contact	0.25	0.00	0.50	1.00	1.00
CCF	0.25	0.00	0.25	1.00	1.00
Skin-Core	0.25	0.00	0.75	1.00	1.00
Core-Skin	0.25	0.00	0.75	1.00	1.00

[2] A.P. Pandurangan, D. Vasishtan, F. Alber, and M. Topf. γ-tempy: Simultaneous fitting of components in 3d-em maps of their assembly using a genetic algorithm. *Structure*, 2015.

[3] K. Lasker, M. Topf, A.b Sali, and H.J. Wolfson. Inferential optimization for simultaneous fitting of multiple components into a cryoem map of their assembly. *Journal of Molecular Biology*, 2009.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] Alexandre d'Aspremont and Stephen Boyd. Relaxations and randomized methods for nonconvex qcqps. *EE392o Class Notes, Stanford University*, 2003.

[6] Jorge Nocedal and Stephen J Wright. Numerical optimization, second edition. *Numerical optimization*, pages 497–528, 2006.

[7] L. Ingber. Adaptive simulated annealing (asa): Lessons learned. *Control and Cybernetics*, 25(1):32–54, 1996.

[8] F. DiMaio, M.D. Tyka, M.L. Baker, W. Chiu, and D. Baker. Refinement of protein structures into low-resolution density maps using rosetta. *Journal of Molecular Biology*, 392(1):181–190, 2009.

[9] A.P. Pandurangan and M. Topf. Finding rigid bodies in protein structures: Application to flexible fitting into cryoem maps. *Journal of Structural Biology*, 177(2):520–531, 2012.

[10] M. Topf, K. Lasker, B. Webb, H. Wolfson, W. Chiu, and A. Sali. Protein structure fitting and refinement guided by cryo-em density. *Structure*, 16(2):295–307, 2008.

[11] S. Zhang, D. Vasishtan, M. Xu, M. Topf, and F. Alber. A fast mathematical programming procedure for simultaneous fitting of assembly components into cryoem density maps. *Bioinformatics*, 26(12):i261–i268, 2010.

[12] D. Vasishtan and M. Topf. Scoring functions for cryoem density fitting. *Journal of Structural Biology*, 174(2):333–343, 2011.

[13] P. Chacón and W. Wriggers. Multi-resolution contour-based fitting of macromolecular structures. *Journal of Molecular Biology*, 317(3):375–384, 2002.

[14] The MathWorks, Inc., Natick, Massachusetts, United States. *MATLAB and Optimization Toolbox, Global Optimization Toolbox Release 2015b*.

[15] V.N. Maiorov and G.M. Crippen. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *Journal of Molecular Biology*, 235(2):625–634, 1994.

[16] Dobritzsch, D., Ricagno, S., Schneider, G., Schnackerz, K.D. and Lindqvist, Y. (2002). Crystal Structure of the Productive Ternary Complex of Dihydropyrimidine Dehydrogenase with Nadph and 5-Iodouracil. Implications for Mechanism of Inhibition and Electron Transfer. J.Biol.Chem., 277, 13155.

[17] Grabarse, W., Mahlert, F., Shima, S., Thauer, R.K. and Ermler, U. (2000). Comparison of Three Methyl-Coenzyme M Reductases from Phylogenetically Distant Organisms: Unusual Amino Acid Modification, Conservation and Adaptation. J.Mol.Biol., 303, 329.

[18] Nolen, B.J., Littlefield, R.S. and Pollard, T. D. (2004). Crystal structures of actin-related protein 2/3 complex with bound ATP or ADP. Proc.Natl.Acad.Sci.Usa, 101, 15627–15632.

[19] Reverter, D. and Lima, C. D. (2005). Insights into E3 ligase activity revealed by a SUMO-RanGAP1-Ubc9-Nup358 complex. Nature, 435, 687–692.

[20] Sindelar, C. V., and Downing, K. H. (2007). The beginning of kinesin's force-generating cycle visualized at 9-A resolution. Journal of Cell Biology, 177(3), 377–385.

[21] Gayathri, P., Fujii, T., Moller-Jensen, J., van den Ent, F., Namba, K., and Lowe, J. (2012). A Bipolar Spindle of Antiparallel ParM Filaments Drives Bacterial Plasmid Segregation. Science, 338(6112), 1334–1337.

[22] Daudén, M. I., Martiń-Benito, J., Sánchez-Ferrero, J. C., Pulido-Cid, M., Valpuesta, J. M., and Carrascosa, J. L. (2013). Large terminase conformational change induced by connector binding in bacteriophage T7. Journal of Biological Chemistry, 288(23), 16998–17007.