Protein Tertiary Structure Prediction via SVD and PSO Sampling

Óscar Álvarez[1], Juan Luis Fernández-Martínez[1(✉)], Ana Cernea[1], Zulima Fernández-Muñiz[1], and Andrzej Kloczkowski[2]

[1] Group of Inverse Problems, Optimization and Machine Learning, Department of Mathematics, University of Oviedo, C/ Federico García Lorca, 18, 33007 Oviedo, Spain
{UO217123,jlfm,cerneadoina,zulima}@uniovi.es
[2] Batelle Center for Mathematical Medicine, Nationwide Children's Hospital, Department of Pediatrics, The Ohio State University, Columbus, OH, USA
Andrzej.Kloczkowski@nationwidechildrens.org

Abstract. We discuss the use of the Singular Value Decomposition as a model reduction technique in Protein Tertiary Structure prediction, alongside to the uncertainty analysis associated to the tertiary protein predictions via Particle Swarm Optimization (PSO). The algorithm presented in this paper corresponds to the category of the decoy-based modelling, since it first finds a good protein model located in the low energy region of the protein energy landscape, that is used to establish a three-dimensional space where the free-energy optimization and search is performed via an exploratory version of PSO. The ultimate goal of this algorithm is to get a representative sample of the protein backbone structure and the alternate states in an energy region equivalent or lower than the one corresponding to the protein model that is used to establish the expansion (model reduction), obtaining as result other protein structures that are closer to the native structure and a measure of the uncertainty in the protein tertiary protein reconstruction. The strength of this methodology is that it is simple and fast, and serves to alleviate the ill-posed character of the protein structure prediction problem, which is very highly dimensional, improving the results when it is performed in a good protein model of the low energy region. To prove this fact numerically we present the results of the application of the SVD-PSO algorithm to a set of proteins of the CASP competition whose native's structures are known.

AQ1

Keywords: Particle Swarm Optimization · Protein refinement
Singular Value Decomposition · Model reduction
Protein tertiary structure prediction

1 Introduction

In computational biology, there is a wide range of problems that can be formulated as a sampling problem over a search space with multiple dimensions. Protein tertiary structure prediction and refinement is solved as the optimization (minimization) of the energy function of the protein. The protein structure prediction problem it is considered as one of the foremost challenges in computational biology [1].

I. Rojas and F. Ortuño (Eds.): IWBBIO 2018, LNBI 10813, pp. 1–10, 2018.
https://doi.org/10.1007/978-3-319-78723-7_18

Proteins are biopolymers that are composed of a set of peptide-bonded amino-acids. The fact that, many spatial conformations of proteins are possible due to the rotation of the chain on each C_α atom, imply that a wide range of structural differences exist. These conformational differences are crucial to fully understand protein interactions, functions and evolution. Large efforts are made in protein structure prediction since the experimental methods used to study their structure are very costly. The computational prediction of protein structures implies the understanding of the mechanisms involved in protein structure and folding, in order to construct good physical models of the protein energy function and accurately mimic the reality, and also the development of mathematical approaches to handle this problem [1, 2]. These methods are based on the optimization of the protein energy function that depends on the protein atoms' coordinates. The forward model is crucial, because if the energy function is able of fully describing the energetics of the protein folds, the minimum energy will correspond to the native structure, but other plausible configurations might also coexist. The fact that these algorithms are not able of sampling the entire protein conformational search space implies that some modelling simplifications are needed. The use of Principal Component Analysis performed in a set of templates to reduce the dimension and performing Protein Tertiary Structure Prediction via Particle Swarm Optimization has been presented [3], showing that the accuracy of the structure prediction will depend on how the reduced PCA basis set is constructed. Particularly, the quality of the a priori templates, the number of PCAs terms, and also on the introduction of a high frequency term which is able to span high frequency details of the protein structure, play key roles in the algorithm performance.

In this paper, we assume an accurate energy function, focusing on the method developed to sample the conformational space via the SVD-PSO algorithm. The model reduction is different from PCA since only one good template is needed to achieve the model reduction. Also, independently of the number of atoms in the protein the sampling is performed in a three-dimensional space. This drastic dimensionality reduction serves to sample other templates closer to the native structure and whose tertiary structure is compatible with the three SVD basis terms.

2 Protein Tertiary Structure Modelling

Our aim is to model protein tertiary structures using SVD as model reduction technique and PSO as global optimizer and sampler. Therefore, the algorithm presented belongs to the category of template-based modelling [4]. Proteins are modelled by their free-energy function, $E(\mathbf{m}):\mathbb{R}^n \rightarrow \mathbb{R}$, by finding the protein model that achieves the minimum energy value, $\mathbf{m}_p:E(\mathbf{m}_p) = \min E(\mathbf{m})$. In this case the model parameters ($\mathbf{m}$) are the protein coordinates and its dimension is three times the number of atoms of the protein. Therefore, the prediction of the best protein structure involves the optimization of the energy function in a high dimensional space with an intricate energy function landscape [5]. These two issues have to be carefully considered as they may cause the failure of the optimization problem if the algorithm gets trapped in one flat valley corresponding

to a local minimum which could be located far from the native backbone structure. If we assume that $\mathbf{m}_p$ is the global optimum, it satisfies the condition, $\nabla E(\mathbf{m}_p) = \mathbf{0}$.

Consequently, there is a set of models $M_{TOL} = \{\mathbf{m}: E(\mathbf{m}) < E_{TOL}\}$, whose energy is lower than a specific cut-off value E_{TOL}. This set in the neighbourhood of $\mathbf{m}_p$:, can be approximated by the linear hyper-quadric [5, 6]:

$$\frac{1}{2}(\mathbf{m} - \mathbf{m}_p)^T HE(\mathbf{m}_p)(\mathbf{m} - \mathbf{m}_p) \leq E_{TOL} - E(\mathbf{m}_p) \tag{1}$$

where $HE(\mathbf{m}_p)$ is the Hessian matrix evaluated at $\mathbf{m}_p$. To avoid, the global optimization method to be trapped in flat, curvilinear elongated and intricate valleys, we require high explorative global optimization methods to explore the non-linear equivalence region M_{TOL}. Algorithms such as the binary genetic and Particle Swarm Optimization (PSO) are capable of performing this task [7]. In this paper, we use an explorative member of the PSO family, denoted as RR-PSO to sample the free-energy function in a reduced space. The main difference with respect to others heuristic approaches is that RR-PSO parameters are tuned based in stochastic stability analysis results [15].

3 Protein Tertiary Structure Refinement Algorithm

3.1 The SVD-PSO Algorithm

Protein prediction, as other real problems from science, has a large number of parameters. As pointed, the relatively high number of atoms and its associated coordinates determine the value of the free-energy function. This feature, alongside the accuracy required to make good predictions, make these problems highly undetermined and ill-posed. Consequently, good "a priori" information is required to make good predictions using global optimization methods. The high numbers of atoms precludes the use of highly explorative optimization algorithms (RR-PSO). In this paper, we show how to construct a reduced search space utilising SVD. Constructing a reduced search space via SVD helps us regularizing the inverse problem and finds the atom coordinates that minimise the protein free-energy function [8].

The utilization of SVD allows the optimization of the free-energy function to be performed in a very low dimensional search space and can be written as follows: finding

$$\mathbf{a}_k \in \mathbb{R}^d: \mathrm{E}(\hat{\mathbf{m}}_\mathbf{k}) = E(\boldsymbol{\mu} + \mathbf{V}_\mathbf{d}\mathbf{a}_\mathbf{k}) \leq E_{\mathrm{TOL}}, \tag{1}$$

where $\boldsymbol{\mu}$ is the mean protein (it could be null) and $\mathbf{V}_\mathbf{d}$ contains as columns the basis set of vectors provided by the SVD.

Focusing on the SVD model reduction, the idea consists in writing the protein in a matrix format $\hat{\mathbf{m}}_\mathbf{k} \in \mathrm{M}(3, \mathrm{n}_{\mathrm{atoms}})$, storing in each column the [x, y, z] coordinates of each atom of the protein structure. Then, it can be factorized, as follows via the SVD:

$$\hat{\mathbf{m}}_\mathbf{k} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\mathbf{T} = \sum_{k=1}^{3} \alpha_k \mathbf{u}_\mathbf{k}\mathbf{v}_\mathbf{k}^\mathbf{T} \tag{2}$$

where $\mathbf{U}, \mathbf{V}$ are orthogonal matrices whose column vectors are respectively $\mathbf{u_k}$, $\mathbf{v_k^T}$, and Σ is the SVD of $\hat{\mathbf{m}}_\mathbf{k}$, that has 3 non-null singular values $(\alpha_1, \alpha_2, \alpha_3)$. The previous expression is known as the spectral decomposition of a matrix and, in this case, it implies that the protein tertiary structure prediction problem can be performed over the reduced basis $\mathbf{u_k v_k^T}$ without any loss of energy (information). In this reduced basis set the protein $\hat{\mathbf{m}}_\mathbf{k}$ has only these 3 coordinates.

Once the reduced base is defined, any other protein decoy will be spanned as a unique linear combination as $\hat{\mathbf{m}}_\mathbf{new} = \sum_{k=1}^{3} \beta_k \mathbf{u_k v_k^T}$, and the coordinates $(\beta_1, \beta_2, \beta_3)$ are found via PSO optimization. The SVD allows a drastic dimensionality reduction from $3n_{atoms}$ to 3 dimensions provided by the spectral basis set $\left\{\mathbf{u_1 v_1^T}, \mathbf{u_2 v_2^T}, \mathbf{u_3 v_3^T}\right\}$. Then, the PSO sampling (while optimzing) is performed efficiently in a reduced search space as the protein atoms coordinates are not sampled independently. Consequently, the ill-determination of the problem is reduced [9]. This procedure works fairly well due to the deterministic nature of the protein energy function landscape, and should be considered as a protein refinement method, with the advantage that the PSO sampling allows to assess the uncertainty of the protein structure reconstruction in the SVD basis set. The aim of this paper is not demonstrating superiority with respect to existing methods, but to provide a new algorithm for tertiary protein prediction refinement.

3.2 Minimisation of the Free-Energy Function

Most of the advances in reducing computational costs and efficiency are based on amino-acid sequences homology [3, 10–13] However, other algorithms are capable of storing the ongoing protein structure information during the sampling [14]. In this sense, PSO has been confirmed as a major improvement on sampling a specific protein backbone structure and evaluating its alternate states by Fernández-Martínez et al. [3].

Hence, we perform the minimization of the energy function for each through an explorative member of the family of Particle Swarm Optimizers (RR-PSO) [15]. RR-PSO is a stochastic and evolutionary optimization algorithm, which was motivated by individual's (particle) social behaviour [16]. The task consists of sampling an appropriate protein model that satisfies the condition, $E\left(\hat{\mathbf{m}}_\mathbf{k}\right) \leq E_{\mathrm{TOL}}$. The sampled model must be reconstructed in the original atom space in order to evaluate the energy, atom coordinates and forces. These forward calculations are performed through the Bioshell package developed by Gont et al. [17–19].

The PSO algorithm starts by defining a prismatic space of admissible protein models:

$$l_j \leq \mathrm{a_{ji}} \leq u_j, 1 \leq j \leq n, 1 \leq i \leq n_{size}$$

where l_j, u_j are the lower and upper limits for the j-th coordinate for each model and n_{size} is the size of the swarm. In this particular case, the sampling is performed in the three-dimensional SVD reduced base. In the algorithm, each particle (model) has its own position in the reduced search space. The particle velocity corresponds to the applied atom coordinates perturbations required in order the particle to explore the search space.

4 Numerical Results

4.1 2L3F PDB Code Protein

We applied the model reduction technique utilizing a SVD to the protein Uracil DNA glycolase from Methanosarcina acetivorans whose native structure is known and reported by the Northeast Structural Genomics Consortium Target [20]. This native structure has been obtained via Nuclear Magnetic Resonance (NMR) which helps obtaining valuable information about the 3D protein structure, dynamics, nucleic acids and its derived complexes.

The assessment of the algorithm performance over a reduced search space is carried out by evaluating two different decoys corresponding to the best decoy and the 10th percentile decoy listed in the CASP9 competition. Each decoy comprises 1271 atoms corresponding to 158 residues. When these two decoys are projected over the reduced search space, the energy of each basis term comprises the three decoy eigenvalues; consequently, the protein sampling would be carried out with a lower ill-posed character while maintaining the prediction accuracy. Information about the algorithm performance over the reduced Search Space is given in Fig. 1.

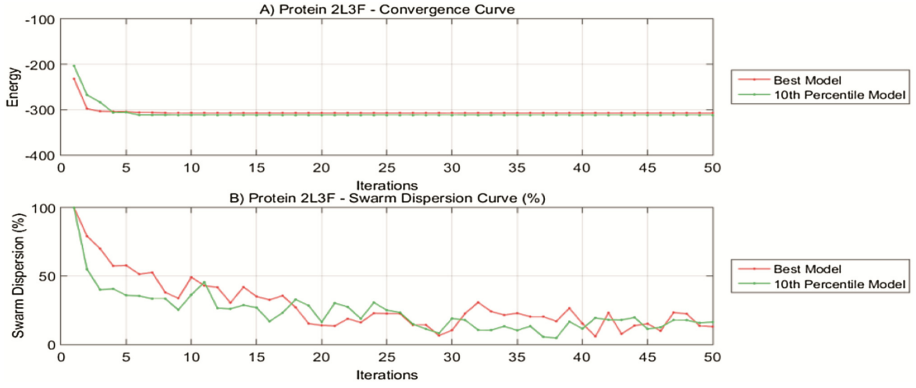

Fig. 1. Protein 2L3F. (A) Convergence curve. (B) Median dispersion curve (%).

As observed in Fig. 1A, the algorithm starts with an energy value which is very close to the optimum. Additionally, the protein refinement algorithm is strongly influenced by the "a priori" model utilized, that is, better initial models yield to better refinements. In Fig. 1B, we show the algorithm performance by plotting the median distance for each particle with respect to the centre of gravity normalized with respect to the first iteration (considered to be 100% dispersion). The qualitative assessment of the protein refinement is shown in Fig. 2, where the best configuration found for each case is presented and compared to the best prediction in CASP9 competition. In this sense, good predictions, similar to the native structure were obtained.

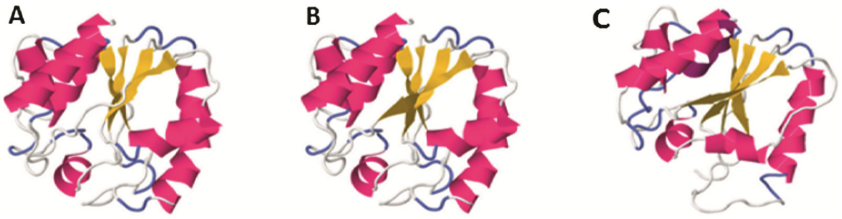

Fig. 2. Protein 2L3F backbone structures corresponding to the (A) best model, (B) result of best model refinement, (C) result of 10^{th} percentile decoy refinement.

We quantitatively analyse the refined structures via the Root Mean Squared Distance (RMSD) with respect to the native. Table 1 summarizes the results obtained with different expanded initial models. It can be observed how the algorithm is capable of improving almost the entire decoy set from CASP9 competition. The major drawback is that a good "a priori" model, situated within the valley where the optimum value exists, is required as a starting point as observed by the poor improvement in the energy function. However, despite the energy function is seldom improved, the RMSD suffers improvements; due to the fact that, RR-PSO samples within the valley where the energy function does not vary substantially, however, it is capable of finding a new backbone conformation with a lower RMSD.

Table 1. Summary of the computational experiments performed in this paper, via Singular Value Decomposition and Particle Swarm Optimization. The table shows the results obtained with different initial models to perform the SVD expansion (initial energy).

Protein PDB code	Model	Initial energy	Best fit energy	"Initial" RMSD	Best fit RMSD
2L3F	Best model	−342.1	−341.5	1.9424	1.8884
	10th percentile	−311.8	−312.5	2.0179	2.0178
2L06	Best model	−369.9	−371.4	5.9876	5.9570
	10th percentile	−322.3	−323.6	6.6480	4.6003
2KYY	Best model	−273.7	−277.1	1.6171	1.6051
	10th percentile	−247.4	−248.3	3.6767	3.6508
2L02	Best model	−448.6	−450.1	7.2553	7.1511
	10th percentile	−373.3	−376.0	14.5460	9.8897
3NBM	Best model	−253.6	−249.6	0.9829	0.9055
	10th percentile	−233.5	−233.9	1.4245	1.3309
3N1U	Best model	−464.4	−465.8	0.6949	0.6945
	10th percentile	−438.1	−439.9	0.8601	0.8945
2X3O	Best model	−369.2	−369.2	8.2840	3.2162
	10th percentile	−334.9	−335.3	11.3852	8.070
3NYM	Best model	−343.6	−343.0	8.9442	6.1692
	10th percentile	−299.3	−301.8	10.8898	6.3731
3NZL	Best model	−209.4	−210.0	3.8829	3.8128
	10th percentile	−177.1	−177.8	4.1682	4.1648

Figure 3 shows the median coordinates of the sampled protein models over the energy region below −200. We show the protein as a matrix with rows containing the coordinates x, y and z and the columns containing the atoms. This representation helps us visualizing better the uncertainty behind the coordinates in the form of coordinate variation and interquartile range.

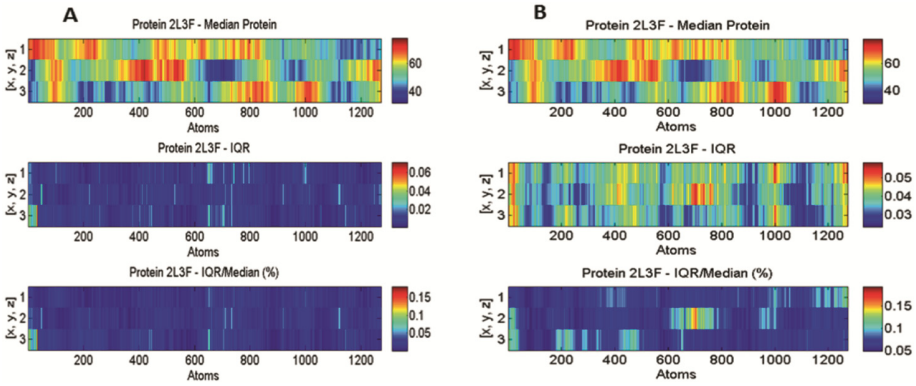

Fig. 3. Median protein and the protein uncertainty for predictions based on the sampling performed over the (A) best model in the energy region below −320 PCAs; (B) 10th percentile decoy and within the energy region below −300.

This graph is used to quantify different conformations of the protein structure. In this case, the higher variations occur in the border coordinates, those atoms corresponding to the protein ends. Besides, the uncertainty is bigger for case (B) concerning the 10th percentile decoy. Therefore, the protein structure seems to be better constraint when the sampling is performed using the best decoy found.

Figure 4 shows the topography of the energy in the first two PCA coordinates; those two reduce search space coordinates that store the majority of the information. As

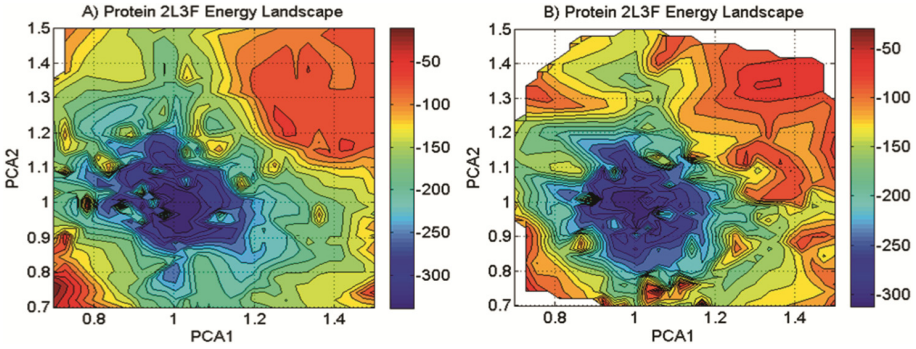

Fig. 4. Protein 2L3F energy landscape for samplings performed over a search space obtained via SVD performed on: (A) the best model, (B) 10th percentile best decoy. It could be observed that both maps have a similar structure.

observed, the topography is similar, with a central valley of low energies, whose orientation is North-South. These graphics serve to assess the mathematical complexity of the protein tertiary structure prediction problem, by observing the intricate valleys of the energy function in lower dimensions. In this case we have used PCA as a visualization tool to produce this plot, because the projection onto a 2D dimensional space has to be done using different sampled templates with different energy.

4.2 Prediction of Other Proteins via SVD and PSO

We present additional information to expand this research benchmark. We have tested additional protein (21–26) utilising SVD model reduction and the RR-PSO algorithm to prove this technique as protein tertiary structure refinement technique. Table 1 summarizes the results obtained, where we show the energy and RMSD of the initial model used to perform the expansion, and the same descriptors obtained after optimisation in the SVD reduced basis set. The case of 2L06, 2X30 and 3NBM are special, since a drastic improvement with respect each decoy has not been achieved as in the other cases. We propose to use this method as a final refinement step, after a good protein model has been found via other existing methodologies. Besides, the No-free lunch theorem in search and optimization [20] states that no algorithm is superior to the rest when it is used over the whole set of problems. Therefore, research is always needed to provide new mathematical-based, elegant and simple algorithms.

5 Conclusions

In this paper, we describe a model reduction technique applied to a decoy-based modelling algorithm. The application of SVD is capable of successfully establishing a three-dimensional Search Space in order to perform the sampling of protein structures via RR-PSO. The SVD model reduction technique is able of preserving the complete information of a given protein backbone structure, consequently, it has been proven to further refine and lower the energy when the optimization is carried out. In this sense, it has been shown that a better refinement is achieved compared to other model reduction techniques such as Principal Component Analysis. The main difference with respect to PCA is that the SVD model reduction is performed in one protein template, while PCA needs different templates to diagonalize their experimental covariance matrix and finding the reduced basis set. Besides, independently of the number of atoms, the sampling is always performed in the SVD spectral basis set, which is three dimensional. Additionally, the SVD model reduction combined with PSO, allows us to sample the equivalent nonlinear region, that helps us understand the protein backbone structure and its alternate states. The SVD model reduction serves to alleviate the ill-posed character of this highly-dimensional optimization problem without losing information when the protein that is used to calculate the basis set is expressed in this reduced search space. Therefore, the SVD-PSO methodology should be used as a protein structure refinement method.

Acknowledgements. A. K. acknowledges financial support from NSF grant DBI 1661391 and from The Research Institute at Nationwide Children's Hospital.

References

1. Zhang, Y.: Progress and challenges in protein structure prediction. Curr. Opin. Struct. Biol. **18**, 342–348 (2008)
2. Bonneau, R., Strauss, C.E., Rohl, C.A., Chivian, D., Bradley, P., Malmstrom, L., Robertson, T., Baker, D.: De novo prediction of three-dimensional structures for major protein families. J. Mol. Biol. **322**, 65–78 (2002)
3. Álvarez-Machancoses, O., Fernández-Martínez, J.L., Fernández-Brillet C., Cernea A., Fernández-Muñiz, Z., Kloczkowski, A.: Principal component analysis in protein tertiary structure prediction, J. Bioinf. Comput. Biol. (2018). Accepted for publication
4. Fiser, A.: Template-based protein structure modeling. Methods Mol. Biol. **673**, 73–94 (2010)
5. Fernández-Martínez, J.L., Fernández-Muñiz, M.Z., Tompkins, M.J.: On the topography of the cost functional in linear and nonlinear inverse problems. Geophysics **77**, W1–W7 (2012)
6. Fernández-Martínez, J.L.: Model reduction and uncertainty analysis in inverse problems. Lead. Edge **34**, 1006–1016 (2015)
7. Fernández-Martínez, J.L., Fernández-Álvarez, J.P., García-Gonzalo, M.E., Ménendez-Pérez, C.O., Kuzma, H.A.: Particle swarm optimization (PSO): a simple and powerful algorithm family for geophysical inversion. In: SEG Technical Program Expanded Abstracts, pp. 3568–3571 (2008)
8. Fernández-Martínez, J.L., Tompkins, M., Fernández-Muñiz, Z., Mukerji, T.: Inverse problems and model reduction techniques. In: Borgelt, C., et al. (eds.) Combining Soft Computing and Statistical Methods in Data Analysis. Advances in Intelligent and Soft Computing, vol. 77, pp. 255–262. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14746-3_32
9. Fernández-Muñiz, Z., Fernández-Martínez, J.L., Srinivasan, S., Mukerji, T.: Comparative analysis of the solution of linear continuous inverse problems using different basis expansion. J. Appl. Geophys. **113**, 95–102 (2015)
10. Quian, B., Ortiz, A., Baker, D.: Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. Proc. Nat. Acad. Sci. **101**, 15346–15351 (2004)
11. Leach, A.R.: Molecular Modelling—Principle and Applications. Prentice Hall, Upper Saddle River (1991)
12. Jones, D.T., Thornton, J.M.: Potential energy functions for threading. Curr. Opin. Struct. Biol. **6**, 210–216 (1996)
13. Frantz, D.D., Freeman, D.L., Doll, J.D.: Reducing quasi-ergodic behavior in Monte Carlo Simulations by J-walking: applications to atomic clusters. J. Chem. Phys. **93**, 2769–2784 (1990)
14. Brunette, T.J., Brock, O.: Improving protein prediction with model-based search. Bioinformatics **21**, 66–74 (2005)
15. Fernández-Martínez, J.L., García-Gonzalo, E.: Stochastic stability and numerical analysis of two novel algorithms of the PSO family: PP-PSO and RR-PSO. Int. J. Artif. Intell. Tools **21**, 1240011 (2012)
16. Kennedy, J., Eberhart, R.: A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium Micro Machine Human Science (1995)

17. Gont, D., Kolinski, A.: Bioshell - a package of tools for structural biology prediction. Bioinformatics **22**, 621–622 (2006)
18. Gont, D., Kolinski, A.: Utility library for structural bioinformatics. Bioinformatics **24**, 584–585 (2008)
19. Gniewek, P., Kolinski, A., Kloczkowski, A., Gront, D.: Bioshell - threading: a versatile Monte Carlo package for protein threading. BMC Bioinf. **22**, 22 (2014)
20. Wolper, D.H., Mcready, W.G.: No free lunch theorems for optimization. IEEE Trans. Evol. Comput. **1**, 67–82 (1997)