On the use of Principal Component Analysis and Particle Swarm Optimization in Protein Tertiary Structure Prediction

Óscar Álvarez[1], Juan Luis Fernández-Martínez[1], Celia Fernández-Brillet[1], Ana Cernea[1], Zulima Fernández-Muñiz[1], Andrzej Kloczkowski[2.3]

oscalmac@gmail.com, jlfm@uniovi.es, celia.fernandez.98@gmail.com, cerneadoina@uniovi.es, zulima@uniovi.es, Andrzej.Kloczkowski@nationwidechildrens.org

[1]Department of Mathematics, University of Oviedo
C. Calvo Sotelo S/N, 33007 Oviedo, Spain

[2]Battelle Center for Mathematical Medicine, Nationwide Children's Hospital, Columbus, OH

[3]Department of Pediatrics, The Ohio State University, Columbus, OH, USA.

Abstract. We discuss applicability of Principal Component Analysis and Particle Swarm Optimization in protein tertiary structure prediction. The proposed algorithm is based on establishing a low-dimensional space where the sampling (and optimization) is carried out via Particle Swarm Optimizer (PSO). The reduced space is found via Principal Component Analysis (PCA) performed for a set of previously found low- energy protein models. A high frequency term is added into this expansion by projecting the best decoy into the PCA basis set and calculating the residual model. Our results show that PSO improves the energy of the best decoy used in the PCA considering an adequate number of PCA terms.

Keywords: Principal Component Analysis, Particle Swarm Optimization, Tertiary Protein Structure, Conformational Sampling, Protein Structure Refinement

1 Introduction

The problem of protein tertiary structure prediction consists of determining the unique three- dimensional conformation of protein (corresponding to the lowest energy) from its amino acid sequence. Currently, this problem represents one of the biggest challenges for biomedicine and biotechnology since it is of utter relevance in areas such as drug design or design and synthesis of new enzymes with desired properties that have not yet been appeared naturally by evolution, that fold to a desired target protein structure [1,2].

Despite the constantly growing number of protein structures deposited in the Protein Data Bank (PDB), there is a rapidly increasing gap between the number of protein sequences obtained from large-scale genome and transcriptome sequencing and the number of PDB structures. Currently PDB contains over 130,000 macromolecular structures, while the UniProt Knowledge base contains around 50 million sequences (after recent redundancy reduction). Thus, less than 1% of protein sequences have the native structures in the PDB database. Therefore, accurate computational methods for protein tertiary structure prediction, which are much cheaper and faster than experimental techniques, are needed [1,2].

The main methodologies to generate protein tertiary structure models are divided into two categories: template-based and template-free modeling. Template-based homology modeling allows building a model of the target protein based on a template structure of a homologue (protein with known structure and high (at least 30%) sequence identity to the target protein), by simulating the process of evolution - i.e. introducing amino acid substitutions as well as insertions and deletions, while maintaining the same fold.

Template-free methods predict the protein tertiary structure from physical principles based on optimizing the energy function that describes the interaction between the protein residues to find the global minimum without using any template information. Some well-known programs in the literature use template-free modeling [2-4] mainly when no structural homologs exist in the PDB. Template-based modeling methods use the known structures (as templates) of the proteins that are analogous to the target protein to construct structural models [5].

Regardless the method utilized, the tertiary structure protein prediction is hampered by the curse of the dimensionality, since these prediction methods are unable to explore the whole conformational space. The curse of dimensionality describes how the ratio of the volume of the hyper sphere enclosed by the unit hypercube becomes irrelevant for higher dimensionality (more than 10 dimensions). Therefore, there is a need to simplify the protein tertiary structure prediction problem by using model reduction techniques to alleviate its ill-posed character [1].

Protein refinement methods are a good alternative to approximate the native structure of a protein using template-based approximate models. Some of these methods use molecular dynamics, coarse-grained models and also spectral decomposition. In our earlier work, [6] we applied Elastic Network Models to protein structure refinement. This mathematical model provides a reliable representation of the fluctuational dynamics of proteins and explains various conformational changes in protein structures. In this article, we use the tertiary structure information provided by other decoys to reduce the dimensionality of the protein tertiary structure prediction problem. We were able to accomplish this task by constraining the sampling within the subspace spanned by the largest principal components of a series of templates. These low-energy protein models (or templates) are previously found using dif-

ferent optimization techniques, or performing local optimization and using different initial and reference models, via template-free methods. In the present study we used as templates, models submitted by the different prediction groups during the CASP experiment.

This methodology allows the sampling of the lowest-energy models in a low dimensional space close to the native conformation. Due to the fact that, the native structure is unknown for most cases, the refined protein structure requires its uncertainty assessment in order to gain a deeper understanding of the protein and its alternate states [7]. The number of PCA terms (PCAs) used to construct the reduced search space for energy optimization and sampling affects the refined structure. Therefore, in this paper we try to understand the effect of PCA dimensionality in the protein tertiary structure prediction problem.

The main conclusions are that the dimensionality reduction alleviates the ill-posed character of this high-dimensional optimization problem, as well as the possibility to increase the uncertainty of the predicted backbone structure. Therefore, a tradeoff is required since, determining the minimum number of PCA terms is a crucial step for achieving a successful refinement.

2 Computational Methods

2.1 Protein Energy Function Landscape

In the tertiary structure protein prediction problem the model parameters are the proteins coordinates determined by n_a atoms, $\mathbf{m} = (m_1, m_2, \ldots, m_n) \in \mathbf{M} \subset \mathbf{R}^n$, with $n = 3n_a$, being $\mathbf{M}$ the set of admissible protein models elaborated taking into account their biological consistency. The tertiary structure of a given protein is defined by knowing the free-energy function, $E(\mathbf{m}): \mathbf{R}^n \rightarrow \mathbf{R}$ and finding the modal that minimizes that free energy function, $\mathbf{m}_p = \min_{\mathbf{m} \in \mathbf{M}} E(\mathbf{m})$ [8].

The main issue with this problem is its high dimensionality. That implies that the optimization algorithm utilized in this problem needs to tackle the high dimension of the model space consisting of thousands of atoms and also the landscape of the energy function.

Also, assuming that $\mathbf{m}_p$ is the global optimum for the energy function satisfying the condition, $\nabla E(\mathbf{m}_p) = \mathbf{0}$, there exist a set of models $M_{tol} = \{\mathbf{m} : E(\mathbf{m}) \leq E_{tol}\}$

whose energy is lower than a given energy cut-off E_{tol}. These models, in the neighborhood of $\mathbf{m}_p$, belongs to the linear hyperquadric [9] :

$$\frac{1}{2}\left(\mathbf{m}-\mathbf{m}_p\right)^T HE\left(\mathbf{m}_p\right)\left(\mathbf{m}-\mathbf{m}_p\right) \le E_{tol} - E\left(\mathbf{m}_p\right) \tag{1}$$

where $HE\left(\mathbf{m}_p\right)$ is the Hessian matrix calculated in $\mathbf{m}_p$. Nevertheless, the linear hyper quadric only describes locally in the neighborhood of $\mathbf{m}_p$ the global complexity of the energy landscape with one or more flat curvilinear elongated valleys with almost null gradients where the local optimization methods might get trapped.

2.2 Protein model reduction via Principal Component Analysis

Principal component analysis is mathematical model reduction technique that transforms a set of correlated variables into a smaller number of uncorrelated ones known as principal components. The resulting transformation has the advantage of being smaller and being more computationally advantageous while maintaining as much as possible the previous variability. This procedure has been applied in several fields but, in protein tertiary structure, it was carried out a preliminary application utilizing the three largest PCs while optimizing via the Powel method [3]. However, in this paper, we perform stochastic sampling in higher dimensions using a member of the family of Particle Swarm Optimizers (RR-PSO) [10, 11]. We study the protein structure prediction and how the number of PCA terms affects the final protein structure obtained via RR-PSO. This PCA is of great relevance in protein structure prediction as it aids us sampling the parameters when a correlation among exists, it also avoids the issue of a high dimensional problem and alleviates the ill-posed character of the tertiary structure optimization problem as the solutions are found in a smaller dimensional space: finding

$$\mathbf{a}_k \in \mathbb{R}^d : E\left(\hat{\mathbf{m}}_k\right) = E\left(\boldsymbol{\mu} + \mathbf{V}_d\,\mathbf{a}_k\right) \le E_{tol}, \tag{2}$$

where $\boldsymbol{\mu}, \mathbf{V}_d$ are provided by the model reduction technique that it is used.

The PCA dimensionality reduction is carried out as follows [12]:

An ensemble of *l* decoys $\mathbf{m}_i \in R^n$ is selected and arranged column wise into a matrix: $\mathbf{X} = \left(\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_l\right) \in M\left(n,l\right)$. The problem consists of finding a set of protein patterns $\mathbf{V}_d = \left(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d\right)$ that provides an accurate low dimensional representation of the original set with $d << l$. This carried out by diagonalizing the matrix X as follows:

$$C_{prior} = (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T \in M(n,n), \tag{3}$$

where $\boldsymbol{\mu}$ is either the experimental mean of the decoys, the median, or any other decoy around we desire to perform the search as a backbone structure.

Matrix C_{prior} has a maximum rank of $l-1$, therefore, as a maximum $l-1$ eigenvectors of C_{prior} are require to expand the whole prior variability. Thus, it is easier to diagonalize $C_{prior}^T \in M(l,l)$ and to obtain the $l-1$ first eigenvectors of C_{prior} as follows:

$$\begin{aligned} &\mathbf{X} - \boldsymbol{\mu} = V \Sigma U^T, \\ &C_{prior}^T = U \Sigma \Sigma^T U^T \Rightarrow B = V\Sigma = (\mathbf{X} - \boldsymbol{\mu})U, \\ &\mathbf{v}_k = \frac{B(:,k)}{\|B(:,k)\|_2}, k = 1,\ldots,l-1. \end{aligned} \tag{4}$$

The centered character of the experimental covariance C_{prior} is crucial to maintain consistency with the centroid model $\boldsymbol{\mu}$.

Ranking the eigenvalues of C_{prior}^T in decreasing order allows us to select a certain number of PCA terms ($d << l-1 << n$) to match most of the variability in the model ensemble. Additionally, a high frequency term is included within the PCA in order to consider the model with the lowest energy, and projecting it into the PCA basis as follows:

$$\mathbf{v}_{d+1} = \mathbf{m}_{BEST} - \boldsymbol{\mu} + \sum_{i=1}^{d} a_i \mathbf{v}_i. \tag{5}$$

Consequently, any protein model in the reduced base is represented as a unique linear combination of the eigen-modes:

$$\hat{\mathbf{m}}_k = \boldsymbol{\mu} + \sum_{i=1}^{d+1} a_i \mathbf{v}_i = \boldsymbol{\mu} + \mathbf{V}\mathbf{a}_k. \tag{6}$$

The projection of any decoy $\hat{\mathbf{m}}_k$ is very fast, since matrix $\mathbf{V}$ is orthogonal:

$$\mathbf{a}_k = \mathbf{V}^T (\hat{\mathbf{m}}_k - \boldsymbol{\mu}). \tag{7}$$

This technique allows global optimization methods to perform efficiently the required sampling in the reduced search space. The PCA procedure helps alleviating the ill-posed character of any highly dimensional problem and we look to study how the number of PCA terms affects the final predicted configuration.

2.3 The particle swarm optimizer

For each backbone conformation, we have performed the optimization via Particle Swarm Optimization (PSO). This methodology is a stochastic and evolutionary optimization technique, which is inspired in individual's social behavior (particles) [13-15]. The sampling problem consists of finding an appropriate sample of protein models $\hat{\mathbf{m}}_k = \boldsymbol{\mu} + \mathbf{V} \cdot \mathbf{a}_k$, such as $E(\hat{\mathbf{m}}_k) \le E_{tol}$. Although the search is carried out in the reduced search space (PCA), the sampled proteins must be reconstructed in the original atom space in order to correctly evaluate their energy. The PSO algorithm is as follows:

We define a prismatic space of admissible protein models, **M**:

$$l_j \le a_{ji} \le u_j, \quad 1 \le j \le n, \quad 1 \le i \le n_{size}, \tag{8}$$

where l_j, u_j, are the lower and upper limits for the *j*-th coordinate for each model. Each plausible model is a particle that is represented by a vector whose length is the number of PCA terms. Each model has its own position in the search space. The perturbations we produced in the PCA search space required in order to carry out he sampling and explore the solutions are represented by the particle velocities. In our case, the search space is designed by projecting back all the decoys to the reduced PCA space and finding the lower and upper limits that expand the variability in each PCA coordinate.

At each iterations, the algorithm updates the positions, $\mathbf{a}_i(k)$, and the velocities, $\mathbf{v}_i(k)$ of each particle swarm. The velocity of each particle, *i*, at each iteration, *k*, is a function of three major components:

The inertia term, a real constant, w that modifies the velocities.

The social term, the difference between the global best position found thus far in the entire swarm, $\mathbf{g}(k)$ and the particle's current position, $\mathbf{a}_i(k)$.

The cognitive term, the difference between the particle's best position found $\mathbf{l}_i(k)$ and the particle's current position, $\mathbf{a}_i(k)$.

Thus, the algorithm is written as follows: [15]

$$\begin{aligned} &\mathbf{v}_i(k+1) = \omega \mathbf{v}_i(k) + \phi_1(\mathbf{g}(k) - \mathbf{a}_i(k)) + \phi_2(\mathbf{l}_i^k - \mathbf{a}_i(k)) \\ &\mathbf{a}_i(k+1) = \mathbf{a}_i(k) + \mathbf{v}_i(k+1), \\ &\phi_1 = r_1 a_g, \quad \phi_2 = r_2 a_l, \quad r_1, r_2 \in U(0,1), \quad \omega, a_g, a_l \in \mathbf{R}. \end{aligned} \tag{9}$$

r_1, r_2 are vectors of random numbers uniformly distributed in (0,1) to weight the global and local acceleration constants, a_g, a_l. $\bar{\phi} = \dfrac{a_g + a_l}{2}$ is the total mean acceleration , crucial in determining the algorithm's stability and convergence [13].

Protein structure calculations are performed via the Bioshell computational package [16-18]. Additionally, Bioshell was considered and essential tool in our research as it was used to carry out tertiary structure in the different PCA basis dimensions, that is, it enabled us to eliminate the distortion of bond angles and lengths accompanying the displacement of protein coordinates when we sample moving along the PCA terms. Furthermore, Bioshell package help us maintaining the structure unchanged and, ultimately, obtaining a backbone structure closer to the determined structures via experiments. Finally, Bioshell also evaluates at each time step each protein conformation, calculating its residues and performing energy minimization to evaluate the energy conformation.

3 Results

In this section we look to study how different PCA dimension affect the prediction capabilities

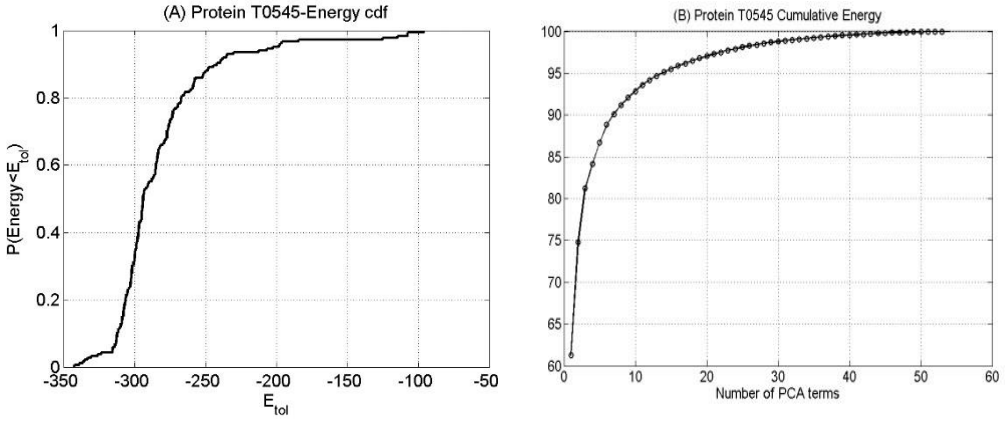

Fig. 1. Energy values of 185 different decoys for protein T0545 (Fig. 1A), is used to construct a reliable PCA base (Fig. 1B)

of the PSO algorithm when applied to different predictions found on the CASP database. We consider protein predictions whose native structures are known in order to assess how our prediction differs from the native structure. As it has been explained previously in the methodology section we utilize different decoys from proteins found in the CASP experiment, we randomly selected the protein T0545 to show the energy values of 185 different decoys and plotted in **Figure 1A**. If we select every single decoys that is in the 30^{th} energy percentile, that is, those with an energy less than this -300, we are capable of constructing a reliable PCA base (**see Figure 1B**). In this sense, it is possible to describe the vast majority of the backbone conformational variation, a fact that has also been reported

by Baker et al. (3) However, we were able to further tune the methodology in order to account for the highest energy details by adding an additional term, known as the high frequency term. This study suggests that we can efficiently sample and optimize a great number of conformational variations in tertiary protein structures by selecting the few first decoys.

The search space utilized is based on the PCA expansion. It is observed that, regardless the PCA coordinates we consider, the width of the first PCA coordinate interval is bigger and, afterwards, it starts getting narrower as the PCA index increases. Additionally, we consider another PCA with eleven terms plus the High Frequency term, in this case, a higher variability within the protein decoys is considered.

Once the PCAs are determined, we perform the PSO search and optimization by adopting a swarm of 40 particles and 100 iterations. To carry out the PSO sampling and optimization, we used the RR-PSO family member while its exploration capabilities were monitored in order to ensure that a good exploration of the PCA search space is performed. The monitoring is, then, carried out by measuring the median distance for each particles and the center of gravity and, normalizing it with respect to the first iteration, considered to be a 100%. When the median dispersion falls below 3%, we can assume that the swarm has collapsed towards the global best, and we can either stop sampling or increase the exploration utilizing steps much greater than 1. When the collapse happens, all the particles of the same iteration will be considered as a unique particle in the posterior sampling.

As shown in **Table 1**, the predictions utilizing three PCA terms are not of good quality with the majority of the predictions with energies far from the native structures. On the other hand, those predictions carried out with a higher dimensionality yield to lower energies. This point is due to the fact that the explorative character of the PSO is strongly correlated with the number of dimensions utilized in constructing the search space. That is, the more dimensions we use, the better the exploration of the protein structure conformational variations and, as a consequence, the final energy predicted.

Protein CASP9 Code	Native structure	Best decoy	3 PCA terms	5 PCA terms	7 PCA terms	9 PCA terms	11 PCA terms
T0545	-348.8	-342.1	-256.8	-299.0	**-343.5**	**-344.6**	**-345.5**
T0557	278.9	-273.7	**-275.3**	**-275.2**	**-275.4**	**-277.2**	**-277.6**
T0555	-389.4	-370.6	23.67	18.68	**-370.9**	**-370.9**	**-371.3**
T0561	-483.6	-448.6	13.28	-400.8	-447.7	**-449.4**	**-450.2**
T0580	-258.3	**-253.8**	-196.4	-250.8	-249.7	-249.5	-250.8
T0635	-466.5	-462.8	-43.7	-324.1	-361.7	**-463.1**	**-463.6**
T0637	-384.5	-372.0	-46.7	-103.7	-369.2	-371.4	**-372.4**
T0639	-380.6	-343.6	-102.3	-335.5	**-345.4**	**-345.7**	**-345.4**

T0643	-234.3	**-209.4**	-138.9	-209.2	**-209.5**	**-210.0**	**-210.0**

Table 1. Summary of the computational experiments performed in this paper, via Principal Component Analysis and Particle Swarm Optimization. Energy of the best decoy used in the PCA and lower energy found after PSO optimization. Bold faces indicate the cases where the energy after optimization improved.

The point remarked by the energy predictions is further confirmed when the Root Mean Squared (RMS) distance is scrutinized in Table 2. Predictions obtained with a PCA with low dimensions are structurally far from the native structures as shown by the RMS, whose values are extremely high. However, when we increase the dimensionality, it is possible to obtain better RMS closer to the native structure.

Protein CASP9 Code	Best decoy	3 PCA terms	5 PCA terms	7 PCA terms	9 PCA terms	11 PCA terms
T0545	1.942	9.231	**1.931**	**1.923**	**1.919**	**1.889**
T0555	8.566	14.411	8.568	**8.566**	**8.522**	**8.516**
T0557	1.617	1.696	**1.606**	**1.596**	**1.024**	**0.780**
T0561	5.898	14.156	5.941	5.899	**5.895**	**5.892**
T0580	**1.284**	1.716	1.331	1.303	1.304	1.291
T0635	2.450	12.520	9.238	6.388	**2.225**	**2.222**
T0637	4.961	12.610	7.468	4.966	**4.964**	**4.286**
T0639	7.944	13.390	10.310	8.967	**6.068**	**4.693**
T0643	3.882	20.670	19.800	**3.728**	**3.432**	**2.915**

Table 2. Summary of the computational experiments performed in this paper, via Principal Component Analysis and Particle Swarm Optimization. RMSD of the best decoy used in the PCA and lower energy found after PSO optimization. Bold faces indicate the cases where the RMSD after optimization improved.

It can be observed, when three PCA terms are considered, the structure is not well defined compared to the native structure, on the other hand, considering 11 PCA terms, the structure is better defined and closer to the native structure, as expected based on the previous analysis of the RMS and the energy function optimization results.

We computed the median coordinates of the sampled protein decoys that fullfil that the energy is below -200 for each PCA search space case. For each case, we presented the protein as a matrix with rows containing the coordinates x, y and z

and the columns containing the atoms in the protein. This way of representing the protein helps us visualizing better the uncertainty behind the coordinates. We observed, that larger variations ins the coordinates occurs in the protein borders. Additionally, as the number of PCA terms decreases, the variations are observed to be smaller, a possible confirmation that, as the terms gets reduced, the ill-conditioned character of the tertiary protein structure prediction problem is reduced. On the other hand, the more PCA terms, the more ill-conditioned the optimization problem is as it is considering more information. As it can be observed, there is a trade-off between the ill-conditioned character and the prediction capability of the model. This is due to the fact that, as we reduce the PCA Search Space, some crucial information required to get a good prediction is lost in the model reduction procedure when accounting for fewer structural variations.

4 Conclusions

In this study, we present an study of the Principal Component Analysis dimensionality and how this can affect the energy prediction and tertiary structure of proteins from the CASP9 competition. The algorithm utilized successfully establishes a low dimensional space in order to apply the energy optimization procedure via a member of the family of Particle Swarm Optimizers. This model reduction has been performed in order to obtain four different search spaces (3, 5, 7, 9 and 11 dimensions plus a high frequency term) to perform the energy optimization later on. The optimizer was capable o modelling the protein sequence and sample the selected decoys projected over the four different PCA Search spaces. Different energy optimum values were obtained depending on the dimensions of the PCA Search Space. It was concluded that as the number of PCA terms increases, it is possible to obtain a better refinement of both the protein energy and the backbone structure of the native protein and its alternative states. As the number of PCA increases, a greater level of information of the decoys utilized to construct the PCA is included and, a lower energy and uncertainty is obtained in the predictions.

Finally, this paper serves to explain how the model reduction technique serves to alleviate the ill-posed character of this high-dimensional optimization problem and how to choose an appropriate

5 Acknowledgements

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

6 References

1. Progress and challenges in protein structure prediction. Zhang, Y. 2008, Curr. Opin. Struc. Biol., Vol. 18, pp. 342-348.
2. De novo prediction of three-dimensional structures for major protein families R. Bonneau, C. E. Strauss, C. A. Rohl, D. Chivian, P. Bradley, L. Malmstrom, T. Robertson and D. Baker, 2002. *J. Mol. Biol.,* Vol. 322, pp. 65-78.
3. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation, P. Bradley, D. Chivian, J. Meiler, K. Misura, C. Rohl, W. W. W. Schief, O. Schueler-Furman, P. Murphy and J. Schonbrun, *Proteins,* 2003, Vol. 53, pp. 457-468.
4. Automated prediction of CASP-5 structures using the Robetta server , D. Chivian, D. E. Kim, L. Malmstrom, P. Bradley, T. Robertson, P. Murphy, C. E. Strauss, R. Bonneau, C. A. Rohl and D. Baker, 2003,*Proteins,* vol. 53, pp. 524-533.
5. The Extent of Cooperativity of Protein Motions Observed with Elastic Network Models Is Similar for Atomic and Coarser-Grained Models. Sen TZ, Feng Y, Garcia JV, Kloczkowski A, Jernigan RL., 2006, J Chem Theory Comput.,Vol. 2, 696-704.
6. Elastic network normal modes provide a basis for protein structure refinement, Gniewek P1, Kolinski A, Jernigan RL, Kloczkowski A., 2012, J Chem Phys. Vol.136, 195101.
7. Model reduction and uncertainty analysis in inverse problems. Fernández-Martínez, J.L. 2015, Leading Edge, Vol. 34, pp. 1006-1016.
8. From crystal structure prediction to polymorph prediction: interpreting the crystal energy landscape. Price, S.L. 2008, Phys. Chem. Chem. Phys., Vol. 2008, pp. 1996-2009.
9. On the topography of the cost functional in linear and nonlinear inverse problems. Fernández-Martínez, J.L. et al. 2012, Geophysics, Vol.77 pp. W1-W15.
10. Stochastic stability analysis of the linear continuous and discrete PSO models. Fernández-Martínez, J.L. and García-Gonzale, E. 2011, Trans. Evol. Comp., Vol. 15, pp. 405-423.
11. Stochastic stability and numerical analysis of two novel algorithms of the PSO family: PP-PSO and RR-PSO.
12. Fernández-Martínez, J.L. and García-Gonzalo, E. 2012, Int. J. Artif. Intell. Tools, Vol. 21, p. 1240011.
13. Jolliffe, I. Principal Component Analysis. s.l. : Springer, 2002.
14. A New Optimizers using Particle Swarm Theory. Kennedy, J. and Eberhart, R. 1995, Proc. Sixth Int. Symp. Micro Mach. Human Sci., Vol. 1, pp. 39-46.
15. The generalized PSO: a new door to PSO evolution. Fernández-Martínez, J.L. and García-Gonzalo, E. 2008, J. Artif. Evol. Appl., Vol. 2008 p. 861275.
16. The PSO family: deduction, stochastic analysis and comparison. Fernández-Martínez, J.L. and García-Gonzalo, E. 2009, Swarm Intell., Vol. 3, pp. 245-273.
17. Bioshell - A package of tools for structural biology prediction. Gront, D. and Kolinski, A. 2006, Bioinformatics, Vol. 22, pp. 621-622.
18. Utility library for structural bioinformatics. Gront, D. and Kolinski, A. 2008, Bioinformatics, Vol. 24, pp. 584- 585.
19. BioShell - Threading: A Versatile Monte Carlo Package for Protein Threading. Gniewek, P., Kolinski, A., Jernigan R.L. and Kloczkowski, A., 2014, BMC Bioinformatics, Vol. 22, p. Art. 22.
20. Solution NMR Structure of a putative Uracil DNA glycosylase from Methanosarcina acetivorans. Aramini, J.M. et al. 2010, Northeast Structural Genomics Consortium Target MvR76.

21. Solution NMR structure of the PBS linker polypeptide domain (fragment 254-400) of phycobilisome linker protein ApcE from Synechocystis sp. PCC 6803. Ramelot, T.A. et al., Northeast Structural Genomics Consortium Target SgR209C.
22. Solution NMR Structure of the N-terminal Domain of Putative ATP-dependent DNA Helicase RecG-related Protein from Nitrosomonas europaea, Eletsky, A. et al. 2010, Northeast Structural Genomics Consortium Target NeR70A
23. The Structural Basis for Recognition of J-Base containing DNA by a Novel DNA-Binding Domain in JBP1, Heidebrecht, T. et al. 2010, Northeast Structural Genomics Consortium and others.
24. The lactose-specific IIB component domain structure of the phosphoenolpy-ruvate:carbohydrate phosphotransferase system (PTS) from Streptococcus pneumoniae. Cuff, M.E. et al., 2010, Midwest Center for Structural Genomics Target TIGR4.
25. Structure of putative HAD superfamily (subfamily III A) hydrolase from Legionella pneumophila, Ramagopal.
26. U.A. et al. 2010, New York Structural Genomics Research Center Target 3N1U.
27. Crystal Structure of the Hypothetical Protein PA0856 from Pseudomonas Aeruginosa, Oke, M. et al. 2010, Joint Center for Structural Genomics NP_249547.1.
28. The Crystal Structure of Functionally Unknown Protein from Neisseria Meningitidis MC58, Zhang, R. et al.2010, Midwest Center for Structural Genomics Target 3NYM.
29. Crystal Structure of the N-Terminal Domain of DNA-Binding Protein SATB1 from Homo Sapiens, Forouhar, F. et al. 2010, Northeast Structural Genomics Consortium Target HR4435B.