Journal of Bioinformatics and Computational Biology
Vol. 16, No. 2 (2018) 1850005 (34 pages)

DOI: 10.1142/S0219720018500051

Principal component analysis in protein tertiary structure prediction

Óscar Álvarez[*,§], Juan Luis Fernández-Martínez[*,¶,§§], Celia Fernández-Brillet[*,‖], Ana Cernea[*,**], Zulima Fernández-Muñiz[*,††] and Andrzej Kloczkowski[†,‡,‡‡,§§]

[*]*Group of Inverse Problems, Optimization and Machine Learning*
Department of Mathematics, University of Oviedo
C. Federico García Lorca, 18, 33007 Oviedo, Spain

[†]*Battelle Center for Mathematical Medicine*
Nationwide Children's Hospital
Columbus, OH, USA

[‡]*Department of Pediatrics*
The Ohio State University
Columbus, OH, USA
[§]*UO217123@uniovi.es*
[¶]*jlfm@uniovi.es*
[‖]*celia.fernandez.98@gmail.com*
[**]*cerneadoina@uniovi.es*
[††]*zulima@uniovi.es*
[‡‡]*andrzej.kloczkowski@nationwidechildrens.org*

Received 7 February 2018
Accepted 7 February 2018
Published 22 March 2018

We discuss applicability of principal component analysis (PCA) for protein tertiary structure prediction from amino acid sequence. The algorithm presented in this paper belongs to the category of protein refinement models and involves establishing a low-dimensional space where the sampling (and optimization) is carried out via particle swarm optimizer (PSO). The reduced space is found via PCA performed for a set of low-energy protein models previously found using different optimization techniques. A high frequency term is added into this expansion by projecting the best decoy into the PCA basis set and calculating the residual model. This term is aimed at providing high frequency details in the energy optimization. The goal of this research is to analyze how the dimensionality reduction affects the prediction capability of the PSO procedure. For that purpose, different proteins from the Critical Assessment of Techniques for Protein Structure Prediction experiments were modeled. In all the cases, both the energy of the best decoy and the distance to the native structure have decreased. Our analysis also shows how the predicted backbone structure of native conformation and of alternative low energy states varies with respect to the PCA dimensionality. Generally speaking, the reconstruction can be successfully achieved with 10 principal components and the high frequency term. We also provide a computational analysis of protein energy landscape for the inverse problem of

[§§]Corresponding authors.

reconstructing structure from the reduced number of principal components, showing that the dimensionality reduction alleviates the ill-posed character of this high-dimensional energy optimization problem. The procedure explained in this paper is very fast and allows testing different PCA expansions. Our results show that PSO improves the energy of the best decoy used in the PCA when the adequate number of PCA terms is considered.

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 as 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 and that fold to a desired target protein structure. This importance is reflected by the fact that every two years since 1994, the performance of the state-of-the-art methods for protein structure prediction are assessed in the CASP (Critical Assessment of Techniques for Protein Structure Prediction) experiments. The results from recent CASP experiments show that prediction of protein structure from sequence has improved significantly but still remains challenging. The progress and challenges in protein tertiary structure prediction have been reviewed by Zhang.[1]

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 with 88% having been solved by crystallography, the majority of them, ∼93%, being proteins, 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.

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 homolog (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.[5–7]

Independently of the methodology that is used (template-based or template-free), protein tertiary structure prediction constitutes a very high dimensional optimization problem, whose dimensionality coincides with the total number of coordinates of atoms forming the protein (thousands of unknowns). Therefore, the tertiary structure protein prediction is hampered by the curse of the dimensionality, as these prediction methods are unable to explore the whole conformational space. The curse of dimensionality[8] 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). This result also describes the practical impossibility of sampling high dimensional spaces using random sampling methodologies. This problem is even more difficult in the case of optimization or inverse problems as the region of nonlinear equivalence has an elongated valley shape.[9] Therefore, there is a need to simplify the protein tertiary structure prediction problem by using model reduction techniques to alleviate its ill-posed character.

Protein refinement methods are a good alternative to approximate the native structure of a protein using template-based approximate models (see, for instance, Refs. 10–12). Some of these methods use molecular dynamics, coarse-grained models, and also spectral decomposition. In our earlier work,[13] 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. We represented the conformational space close to the native state by a set of decoys generated by the I-TASSER protein structure prediction server using template-free modeling, and we found that thermal motions for some substates overlap significantly with the deformations necessary to reach the native state. This suggested that structural refinement of poorly resolved protein models can be significantly enhanced by reduction of the conformational space to the motions imposed by the dominant normal modes.

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 different 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 sampling of the lowest-energy models in a low dimensional space close to the native conformation. This fact is very important because for most of the proteins the native structure is unknown and needs to be determined by

computational methods. However, a deeper understanding is required in order to successfully refine the protein structure, that is, the structure that corresponds to the lowest energy, with its corresponding uncertainty assessment (the median model and the interquartile range (IQR). The refined structure is affected by the number of PCA terms used to construct the reduced search space for energy optimization and sampling.[14–16]

Therefore, in this paper, we try to understand the effect of PCA dimensionality in the protein tertiary structure prediction problem. The structure plan of this paper is as follows: the protein tertiary structure prediction problem and the protein energy function landscape are analyzed theoretically in Sec. 2, showing the existence of equivalent protein configurations. Section 3 is devoted to the computational methods used in this paper, principal component analysis (PCA) and particle swarm optimization (PSO). Finally, in Sec. 4, the numerical results obtained for the MvR76 protein (CASP9 code T0545) are presented. Also, the analysis of other CASP target proteins is provided in Appendix A, confirming the results shown for T0545.

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 show the existing tradeoff between model reduction (number of PCA terms) and the backbone structure prediction. Therefore, determining the minimum number of PCA terms is a crucial step for achieving a successful tertiary structure protein refinement. Besides, the introduction of the high frequency term into the expansion is crucial for achieving good tertiary structure reconstruction. In addition, although the set of modeled proteins has been limited, the results show an improvement in results in all cases with respect to the best model found. These results could be further improved using advanced computational resources, given the intrinsic parallelism of the PSO algorithm.

2. The Protein Tertiary Structure Prediction Problem

Proteins are biological polymers composed of amino acids forming a linear polypeptide chain. Proteins can adopt a wide range of three-dimensional conformations due to possible rotation of the chain about two bonds of each amino acid. This rotation is responsible for differences in the protein structure. This fact makes the problem of determining the protein structure very challenging. Moreover, the knowledge of protein tertiary structure is of great importance for determining interactions within proteins and between proteins and their surroundings, as well as for annotating protein function. A consequence is the direct application of tertiary structure prediction to drug design.[17]

The most commonly used methods for determining this structure are X-ray crystallography and nuclear magnetic resonance (NMR).[18] However, due to the practical impossibility of determining the structure by experimental methods of all the proteins obtained by large-scale genome sequencing methods, accurate

computational determination of the tertiary structure of proteins has become a crucial problem. Nevertheless, one of the main challenges is to obtain an accurate model with a low uncertainty in the predictions.[19]

2.1. *The protein energy function landscape*

In the tertiary structure protein prediction problem, the model parameters are the protein 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 protein model that minimizes this energy function: $\mathbf{m}_p = \min_{\mathbf{m}\in\mathbf{M}} E(\mathbf{m})$.[20]

The main issue with this optimization problem is its high dimensionality. The optimization algorithm needs to tackle the high dimension of the model space with thousands of atoms, and also the complicated landscape of the energy function, with different isolated curvilinear valleys with almost null gradients.[21] This complicated topography can trigger the failure of local optimization methods that might get trapped in a different basin, or in one of these flat valleys, in a model far from the native conformation.

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 $M_{\text{tol}} = \{\mathbf{m} : E(\mathbf{m}) \leq E_{\text{tol}}\}$ whose energy is lower than a given energy cut-off E_{tol}. M_{tol} is the nonlinear equivalence region of energy E_{tol}. These models, in the neighborhood of $\mathbf{m}_p$, belong to the linear hyper-quadric[9,21]:

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

where $\text{HE}(\mathbf{m}_p)$ is the Hessian matrix calculated in $\mathbf{m}_p$. Nevertheless, the linear hyper-quadric can only describe locally throughout 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 and fail to converge close to the native structure. To avoid this problem, it is important to have at disposal very exploratory global optimization methods that are able to analyze the nonlinear equivalence region M_{tol} in a procedure that is known as "sampling while optimizing". Algorithms such as the binary genetic algorithms reported by Fernández-Álvarez *et al.*[22] and PSO by Fernández-Martínez *et al.*[23,24] are able to perform this task when used in their exploratory form. In our past work, we successfully applied PSO methodology (in combination with Extreme Learning Machines) to protein secondary structure prediction.[25,26] In this paper, we use an exploratory PSO family member (RR-PSO) to explore the Protein Energy Function Landscape in the PCA-reduced space. The acronym RR-PSO stands for the regressive-regressive finite differential schemes that are used to construct this algorithm of the PSO family.

3. Computational Methods

3.1. *Protein model reduction via PCA*

PCA is a 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 reducing the dimensionality while maintaining as much as possible the prior models' variability.[27,28] This procedure has been applied in different fields, but in protein tertiary structure, a preliminary application using the three largest PCs was carried out while optimizing via the Powel method.[29] However, in this paper, we perform the stochastic sampling of the nonlinear equivalence region using a member of the family of PSOs (RR-PSO).[30–33] These global optimization algorithms can perform a good posterior sampling of the nonlinear equivalence region when used in their exploratory version.

We study the protein structure prediction and how the number of PCA terms affects the final protein structure that is obtained. PCA is of great relevance in protein structure prediction as it allows the sampling of protein coordinates (parameters) while taking into account the correlation existing among the atom coordinates in the regions of lower energies (nonlinear equivalence region). PCA also alleviates the ill-posed character of the tertiary structure optimization problem as the solutions (protein structures) are found in a smaller dimensional space.

The optimization problem in the reduced PCA space consists in finding the coefficients

$$\mathbf{a}_k \in \mathbf{R}^d : \; E(\hat{\mathbf{m}}_k) = E(\boldsymbol{\mu} + \mathbf{U}_d \mathbf{a}_k) \leq E_{\text{tol}}, \tag{2}$$

where $\boldsymbol{\mu}, \mathbf{U}_d$ are provided by the model reduction technique that is used (PCA in this case).

The PCA dimensionality reduction is carried out as follows.[34,35]

(a) An ensemble of l decoys $\mathbf{m}_i \in \mathbf{R}^n$ is selected and arranged column wise into a matrix $\boldsymbol{X} = (\boldsymbol{m_1}, \boldsymbol{m_2}, \ldots, \boldsymbol{m_l}) \in \boldsymbol{M}(\boldsymbol{n}, \boldsymbol{l})$. The problem consists of finding a set of protein patterns $\boldsymbol{U_d} = (\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_d})$ that provides an accurate low dimensional representation of the original set with $d << l$. This is carried out by performing the diagonalization of the covariance matrix:

$$C_1 = (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T \in M(n, n), \tag{3}$$

where $\boldsymbol{\mu}$ is either the experimental mean of the decoys, the median, or any other decoy around which we want to perform the conformational search. The centered character of the experimental covariance C_1 is crucial to maintain consistency with the centroid model $\boldsymbol{\mu}$.[35] Matrix C_1 is symmetric and has a maximum rank of $l - 1$, because there is a maximum of $l - 1$ eigenvectors of C_1 that are required to expand the whole protein prior variability. Due to its symmetry, C_1 admits orthogonal diagonalization. Therefore, it can be expanded by the following spectral

decomposition:

$$C_1 = UD_1U^T = \sum_{k=1}^{l-1} \lambda_k \mathbf{u}_k \mathbf{u}_k^T, \tag{4}$$

where D_1 is the diagonal matrix containing $l-1$ nonnull eigenvalues λ_k, and $\mathbf{u}_k$ is the orthonormal basis set of $\mathbf{R}^n$, that is, the space of decoys that form the column space of the orthogonal matrix U. The vectors $\mathbf{u}_k$ are the protein decoys PCAs. Therefore, the column vectors forming $\mathbf{U}_d$ are the first d column vectors of U.

The sum of its eigenvalues coincides with the total variability of the decoys in the protein ensemble. The cumulative energy used to select the number of PCAs in the reduced basis of decoys is the cumulative sum of these eigenvalues ranked in the decreasing order. Ranking the nonnull eigenvalues $\{\lambda_j\}_{j=1,l-1}$ in the decreasing order, the cumulative energy percentage corresponding to the k first eigenvalues is $E(k) = \sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{l-1} \lambda_j \, 100$. Therefore, this procedure allows to select a number of PCA terms ($d << l-1 << n$) allowing to match most of the variability in the protein ensemble, which is $\sum_{j=1}^{l-1} \lambda_j$.

Numerically, it is simpler and faster to diagonalize the matrix, $C_2 = (\mathbf{X} - \boldsymbol{\mu})^T(\mathbf{X} - \boldsymbol{\mu}) \in M(l,l)$ and to obtain the first $l-1$ eigenvectors of C_1. Taking into account the connection with the singular value decomposition (SVD), we have:

$$\begin{aligned} &\mathbf{X} - \boldsymbol{\mu} = U\Sigma V^T, \\ &C_2 = (\mathbf{X} - \boldsymbol{\mu})^T(\mathbf{X} - \boldsymbol{\mu}) = V\Sigma^T\Sigma V^T, \end{aligned} \tag{5}$$

where U, V, Σ are the matrices obtained from the SVD of $\mathbf{X} - \boldsymbol{\mu}$. It is worthy to note that the matrix U is the same that appears in expression (4). Σ is the SVD of the centered decoy matrix $\mathbf{X} - \boldsymbol{\mu}$, and V is one of the orthogonal matrices that is issued from the SVD of $\mathbf{X} - \boldsymbol{\mu}$.

Therefore:

$$\begin{aligned} &B = (\mathbf{X} - \boldsymbol{\mu})V = U\Sigma, \\ &\mathbf{u}_k = \frac{B(:,k)}{\|B(:,k)\|_2}, \quad k = 1, \ldots, l-1, \end{aligned} \tag{6}$$

where $B(:,k)$ is the kth column vector of matrix B. The matrix $D_2 = \Sigma^T\Sigma$ is also diagonal and has the same nonnull eigenvalues as D_1.

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:

$$\mathbf{m}_{\text{BEST}} = \boldsymbol{\mu} + \sum_{i=1}^{d} m_i \mathbf{u}_i + \mathbf{u}_{d+1} \Leftrightarrow \mathbf{u}_{d+1} = \mathbf{m}_{\text{BEST}} - \boldsymbol{\mu} - \sum_{i=1}^{d} m_i \mathbf{u}_i, \tag{7}$$

where m_i are the coefficients of $\mathbf{m}_{\text{BEST}} - \boldsymbol{\mu}$ into the first d PCA basis terms. The inclusion of the high frequency term is vital to ensure a proper reconstruction of the protein model in Cartesian coordinates after the PCA sampling, as it provides

the high frequency details needed to decrease the energy cost function and achieve a final solution which is close to the native structure. The inclusion of this term allows a successful reconstruction of the backbone structure. Besides, an important feature is the use of Bioshell to compute the protein energies (forward problem of the tertiary structure protein inverse problem). In this approach, we assume that the energy function used to model the protein energy allows for the search of the native tertiary structure. The forward modeling (energy computation) is also an important field of research in protein prediction.

(b) Consequently, any protein model in the reduced basis set is represented as a unique linear combination of the main eigenvectors:

$$\hat{\mathbf{m}}_k = \boldsymbol{\mu} + \sum_{i=1}^{d+1} a_i \mathbf{u}_i = \boldsymbol{\mu} + \mathbf{Q}\mathbf{a}_k, \tag{8}$$

where $\mathbf{Q} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d, \mathbf{u}_{d+1}]$. The projection of any decoy $\hat{\mathbf{m}}_k$ is very fast, as matrix $\mathbf{Q}$ is orthogonal:

$$\mathbf{a}_k = \mathbf{Q}^T(\hat{\mathbf{m}}_k - \boldsymbol{\mu}). \tag{9}$$

This technique allows global optimization methods to perform efficiently the required sampling in the reduced search space, which is designed by calculating the maximum and the minimum values of the different components of the coefficients $\mathbf{a}_k$ of all the decoys used for the PCA expansion. This simple and fast procedure serves to establish the lower and upper bounds of the search space of the PCA coefficients for PSO to perform the search of new protein decoys (see, for instance, the graph shown in Fig. 2).

In this paper, we analyze how the number of PCA terms affects the final predicted protein configuration, and how the topography of the cost function in the reduced space changes accordingly.

3.2. *PSO*

For each backbone conformation, we have performed the optimization via PSO. This methodology is a stochastic and evolutionary optimization technique, which is inspired in an individual's social behavior (particles).[36] 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 $E(\hat{\mathbf{m}}_k) \leq E_{\text{tol}}$, where the mean model $\boldsymbol{\mu}$ and the matrix $\mathbf{V}$are provided by the PCA, as it has been explained before. 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 evaluate their energy.

The PSO algorithm is as follows:

(1) We define a prismatic space of admissible protein models, $\mathbf{M}$:

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

where l_j, u_j, are the lower and upper limits for the jth coordinate for each model, where n_{size} is the size of the swarm. 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 are required in order to carry out the sampling and to explore the solutions represented by the particle velocities. In the present case, the search space is found by projecting back all the decoys to the reduced PCA space and finding the lower and upper limits of each PCA coordinate.

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

   (a) The inertia term, a real constant, w that modifies the velocities.
   (b) The social term, the difference between the global best position found thus far in the entire swarm, $\mathbf{g}(tk)$ and the particle's current position, $\mathbf{a}_i(tk)$.
   (c) The cognitive term, the difference between the particle's best position found $\mathbf{l}_i(tk)$ and the particle's current position, $\mathbf{a}_i(tk)$.

The PSO algorithm is written as follows[36]:

$$\begin{aligned}
&\mathbf{v}_i(tk+1) = w\mathbf{v}_i(tk) + \phi_1(\mathbf{g}(tk) - \mathbf{a}_i(tk)) + \phi_2(\mathbf{l}_i(tk) - \mathbf{a}_i(tk)),\\
&\mathbf{a}_i(tk+1) = \mathbf{a}_i(tk) + \mathbf{v}_i(tk+1),\\
&\phi_1 = r_1 a_g, \quad \phi_2 = r_2 a_l, \quad r_1, r_2 \in U(0,1), \quad w, a_g, a_l \in \mathbf{R}.
\end{aligned} \tag{10}$$

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. $\overline{\phi} = (a_g + a_l)/2$ is the total mean acceleration, crucial in determining the algorithm's stability and convergence.[30–32]

In this paper, we used the RR-PSO algorithm obtained by adopting regressive discretization in acceleration and also in velocity in the PSO continuous model. The numerical analysis using different analytical benchmark functions has shown that RR-PSO is one of the most performing algorithms of the PSO family in terms of the balance between its exploration and exploitation capabilities. The RR-PSO algorithm for any time step Δt is:

$$\begin{aligned}
&\mathbf{v}_i(t+\Delta t) = \frac{\mathbf{v}_i(t) + \phi_1 \Delta t(\mathbf{g}(t) - \mathbf{x}_i(t)) + \phi_2 \Delta t(\mathbf{l}_i(t) - \mathbf{x}_i(t))}{1 + (1-w)\Delta t + \overline{\phi}\Delta t^2}, \quad i = 1, \ldots, N_p,\\
&\mathbf{x}_i(t+\Delta t) = \mathbf{x}_i(t) + \mathbf{v}_i(t+\Delta t)\Delta t.
\end{aligned} \tag{11}$$

Its version is obtained for $t = tk$ (iterations) and $\Delta t = 1$. The RR-PSO algorithm has regions of first- and second-order stochastic stability that are unbounded. It has been shown that the RR-PSO exploratory parameters sets are concentrated around the line $\overline{\phi} = 3(w - 3/2)$, mainly for inertia values $w > 2$. This line is independent of the cost function that is optimized and remains invariant when the number of optimization parameters increases. Furthermore, this line is located in a region of medium

attenuation and very high frequency for the swarm particle trajectories. This last property provides this algorithm with a good balance between exploration and exploitation, allowing a very efficient and explorative search around the oscillation center of each particle in the swarm.[33]

The protein energy calculations are performed in this case via the Bioshell computational package.[37–39] The PSO algorithm finishes by iterations since the procedure aims the sampling of the uncertainty space of the tertiary structure prediction problem. Other stopping criteria can be established to finish the PSO sampling, for instance, if the energy does not decrease in a given number of iterations, and/or if the root-mean-square deviation (RMSD) distance with respect to the model of lower energy found does not decrease.

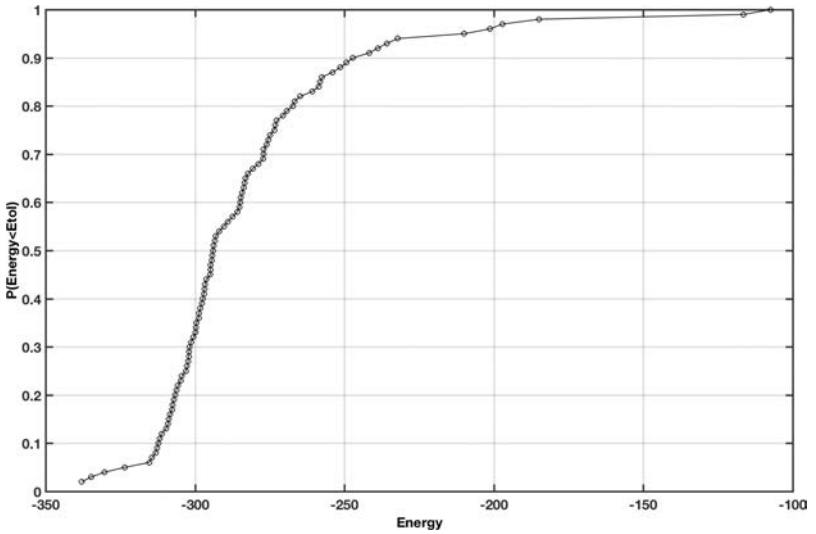
4. Numerical Results

In this section and in Appendix A, we show the application of this methodology to different proteins from the CASP experiment to show how this refinement method works in practice and how the number of PCAs used in the expansion affects to the final reconstruction and sampling.

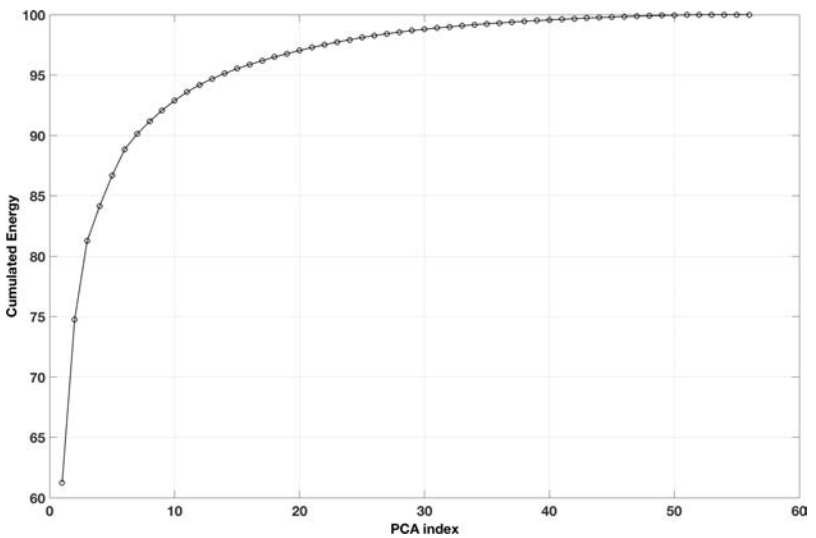
4.1. *MvR76 protein (CASP9 code T0545)*

In this section, we study how different PCA dimensions affect the prediction capabilities of the PSO algorithm when applied to the protein Uracil DNA glycosylase from *Methanosarcina acetivorans* (CASP9 code T0545) whose native structure is known and reported as the Northeast Structural Genomics Consortium Target.[40] This native structure has been obtained via NMR spectroscopy which makes it possible to obtain detailed and valuable information about the three-dimensional structure, dynamics, and function. Figure 1(a) represents the cumulative distribution function (cdf) of the energy values corresponding to the 185 decoys predicted by different research groups in the CASP9 competition. Each decoy comprises a total amount of 1271 atoms corresponding to 158 residues. It can be obseved that the maximum energy (worst decoy) is around -100 and the mimimum close to -350, and the median energy a little bit higher than -300. In order to generate an accurate PCA basis that accurately represents the tertiary protein structure, all the decoys (56) with an energy lower than -300 are selected. This energy cut-off corresponds to the percentile 30 of the cdf energy profile, that is, the 30% best models are selected to produce the reduced PCA basis set.

Optimizing protein tertiary structure models requires an efficient sampling of a search space formed by the decoys that have been selected. In this case, the procedure that is used consists in a combination of PCA and PSO, a sampling and global optimization algorithm which can be easily parallelized. RR-PSO is used in its sampling while optimizing modality, that is, promoting its exploratory behavior. Determining the prior number of PCA terms that might be needed for the expansion is a straightforward procedure. It consists of adding the eigenvalues of the covariance



(a)



(b)

Fig. 1. Protein T0545. (a) Energy cdf of the 185 decoys downloaded from the CASP9 experiment website. To generate the PCA reduced basis, we had considered 56 decoys with energy less than -300 that correspond to the percentile 30 of the energy cdf distribution. (b) Cumulated energy of the PCA decomposition. With the first PCA term, we expand 62% of the energy of the decoys database (prior total variability of the decoys). When a second PCA is added the cumulative energy is 75%, and around 40–50 PCAs we expand almost 100% of the variability.

matrix and calculating the ratio of the cumulated energy with respect to the total prior variance, as it has been explained before using the cumulative energy plot shown in Fig. 1(b). Generally speaking, as reported by Baker and co-workers,[18] the first PCA term can generally describe between 40% and 90% of the backbone conformational variation. In the present case, we were able to expand 60–75% of the decoys variability with one and two PCA terms, as it can be observed in Fig. 1(b). We achieved this by performing a preliminary selection of the 56 best decoys. This study suggests that we can efficiently sample and optimize a great number of conformational variation in tertiary protein structures by selecting the first PCA. Therefore, in order to study how the predicted protein structure varies with respect to the number of PCA dimensions, we select different PCA bases containing 3, 5, 7, 9, and 11 terms, corresponding to a cumulative energy of the covariance matrix ranging from 82% to 94%. Additionally, a last high frequency term is added in order to span the details of best model found (with the lowest energy). Nevertheless, it is always needed to check whether the prior number of PCAs is enough to explore the protein energy landscape and whether the number of PCA terms should be increased. This situation occurs when the minimum energy found gets stable and high along the PSO iterations. Figure 2 shows the search space used for search and optimization with 11 PCA terms plus the high frequency term. We also show the coordinates of the native structure in this basis set. The width of the first PCA coordinate interval is bigger and, afterwards, it starts getting narrower as the PCA index increases. Once the number of the PCA terms is fixed, we perform the PSO

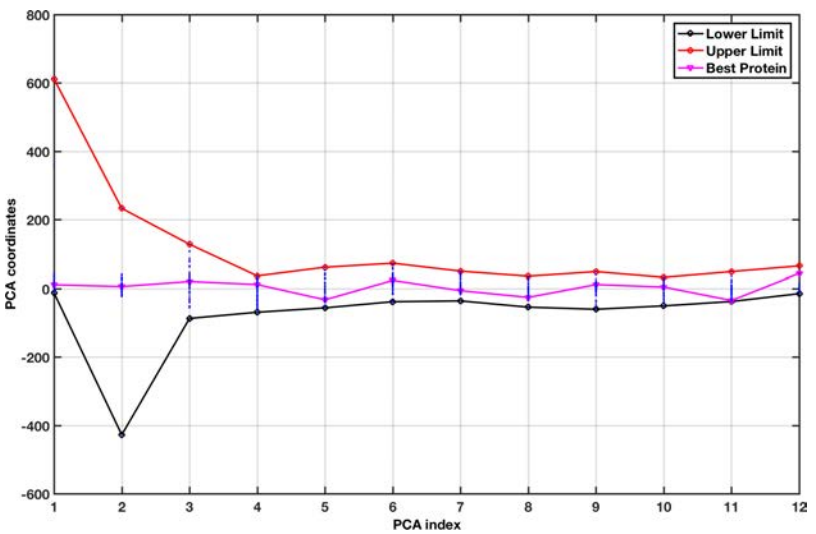
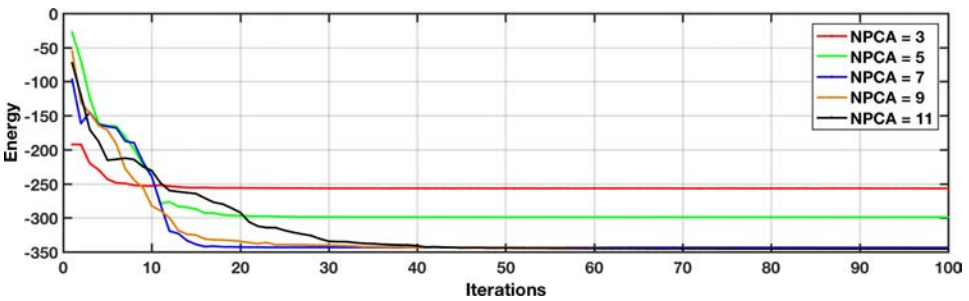


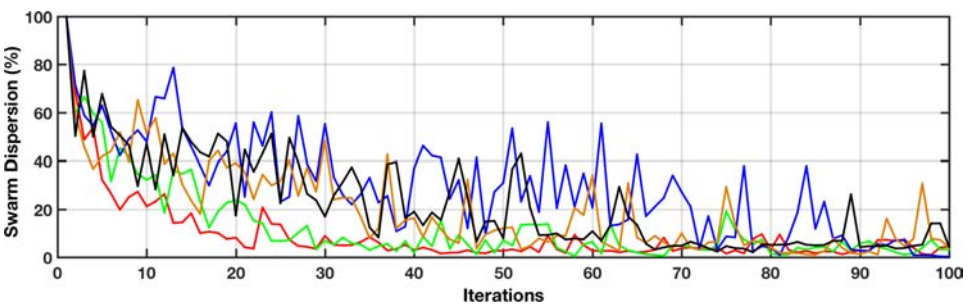
Fig. 2. Lower and upper limits of the search space for PSO of the PCA coefficients for 11 PCA terms. The high frequency term is added to span the details. These bounds are calculated projecting all the templates into the PCA basis set using expression[7] and finding their minimum and maximum for each coefficient. We also provide the projection of the best decoy into the PCA basis set.

search and optimization with a swarm of 40 particles during 100 iterations. The sampling and optimization are carried via RR-PSO, a member of the family of PSOs, whose exploration capabilities are very important and make it suitable to accomplish the approximate posterior sampling.[36] The width of the search space might be increased when a high percentage of particles hits the lower and upper bounds through iterations. Besides, it is possible to play with the time step Δt if needed. Values higher than 1 serve to increase the exploration and values lower than 1 to cool down the particle dynamics.

Figure 3 shows the convergence rate (Fig. 3(a)) and swarm dispersion (Fig. 3(b)) for different PCA expansions. For each case, the algorithm begins with a high energy between -10 and -200 and in approximately 20 iterations it reaches an energy plateau, corresponding to the optimum energy value. The optimum energy value of protein T0545 was found to be -256.8 with three PCA terms and -345.5 with 11 PCA terms, which is 1% lower than the energy of the best model found in the CASP experiment for this protein (-342.1). The energy of the native structure is -348.8. It is worth observing that the sampling occurs in a reduced space that it is one dimension higher than the number of PCA terms that have been adopted. Another point that is worth mentioning is the ability to explore the energy landscape of the PCA search space. The monitoring of the exploration is carried out by measuring the



(a)

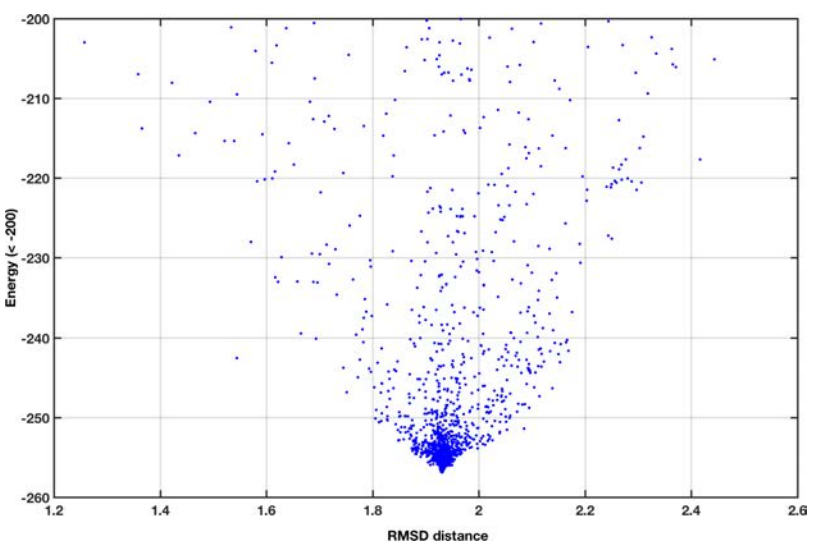


(b)

Fig. 3. T0545 protein: (a) convergence curve and (b) median dispersion curve (%).

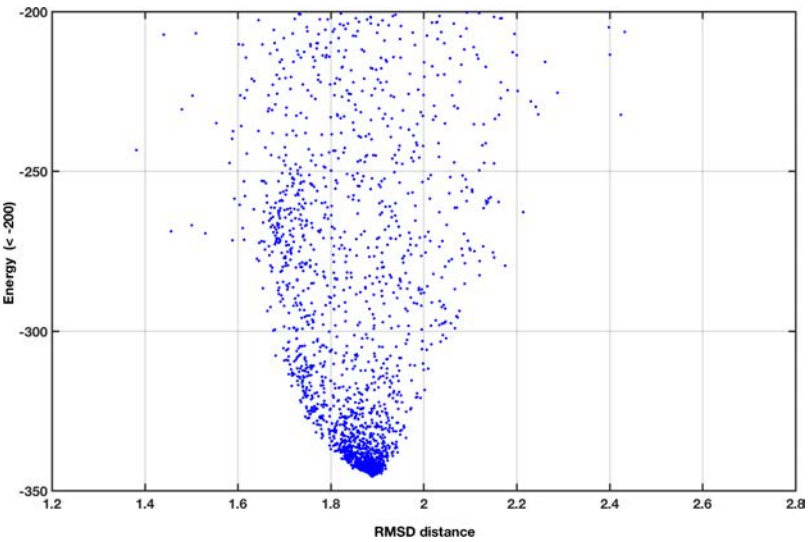
median distance for each particle and the center of gravity and, normalizing it with respect to the first iteration (considered to be a 100%). When the median dispersion falls below 3–5%, we can assume that the swarm has collapsed toward the global best, and we can either stop sampling or increase the exploration using time steps much greater than 1 to expand the swarm. When the swarm collapse happens, all the particles of the same iteration will be considered as a unique particle in the posterior sampling. Figure 3(b) shows smaller dispersion when a low number of PCA terms ($N_{\mathrm{PCA}} = 3$ and $N_{\mathrm{PCA}} = 5$) is used. This fact explains that when the number of PCA is very low, the algorithm is unable to optimize further the energy, and the swarm collapses toward the global best very early. This phenomenon of premature convergence is also related to the number of PCA terms adopted, as the PCA expansion does not allow to span the backbone structure of the protein model. In contrast, as the number of PCA terms increases, the dispersion also increases, showing that RR-PSO is able to sample the nonlinear equivalence region.

Figures 4(a) and 4(b) present the root mean square distance between the different models that have been sampled in the region of energy below −200. As observed, depending on the PCA terms utilized to construct the search space, the minimum energy achieved varies, as also shown in Fig. 3. Additionally, we can observe a symmetric behavior with a large quantity of models sampled within the region of 1.8 and 2.0 units with respect to the centroid. This illustrates the complexity of the energy landscape. Furthermore, Table 2 illustrates the RMSD of atomic positions of the optimized structure for each case compared with the best model submitted in the



(a)

Fig. 4. (a) RMS distance between different decoys sampled considering three PCA terms plus a high frequency term. (b) RMS distance between different decoys sampled with 11 PCA terms plus a high frequency term.



(b)

Fig. 4. (*Continued*)

CASP experiment. The RMSD is defined as follows:

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i}, \tag{12}$$

where δ_i is the distance between the atom I of the protein and the native structure that was used as reference, and N is the total number of atoms of the backbone structure.

PSO was found to succesfully reduce the RMSD of each structure except for T0580, where it is almost similar to the result found in the CASP experiment. Finally, Fig. 5 shows the best configurations obtained by RR-PSO for each number of PCA terms (N_{PCA}) compared with the best model submitted in the corresponding CASP experiment. In this sense, we can assess qualitatively the prediction capabilities of including a varying number of PCA terms. As observed, when three PCA terms are considered, the structure is not well defined compared with the native structure. In contrast, considering 11 PCA terms, the structure is better defined and gets closer to the native structure. In each case, the dimension of the optimization that is involved is the number of PCA terms plus one, due to the high frequency term. The PCAs come from the ensemble of decoys that are used to calculate the covariance matrix that is diagonalized, whereas the high frequency term is calculated from the best decoy, projecting it into the PCA basis set and calculating the residual. This result suggests as expected that a minimum amount of details (higher PCA terms) is needed to achieve a succesful tertiary reconstruction.

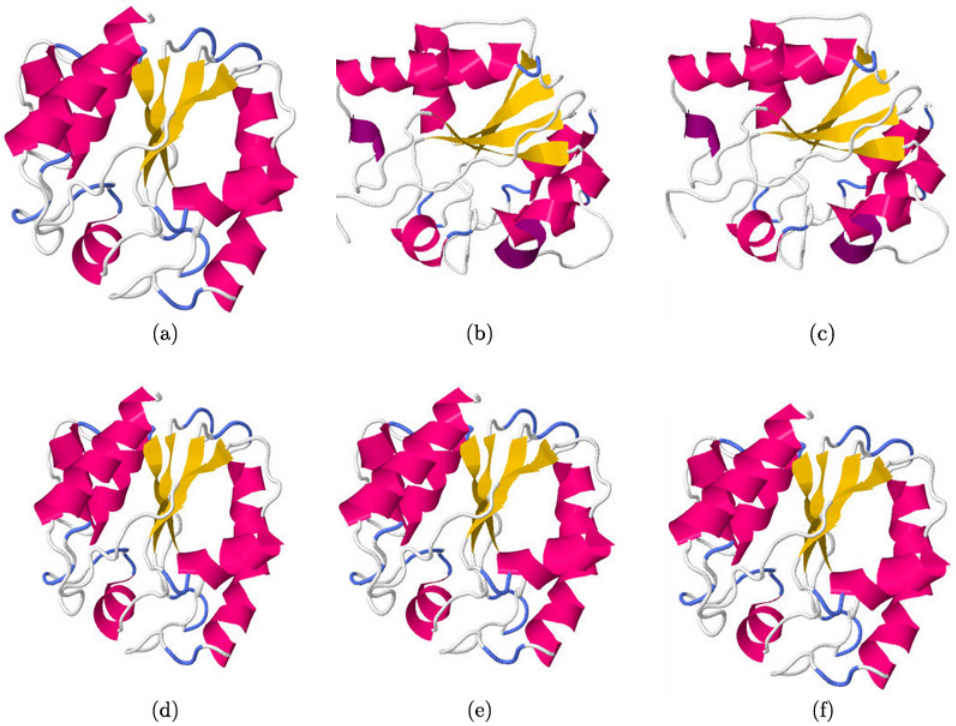
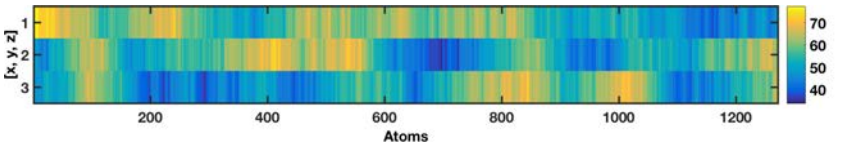
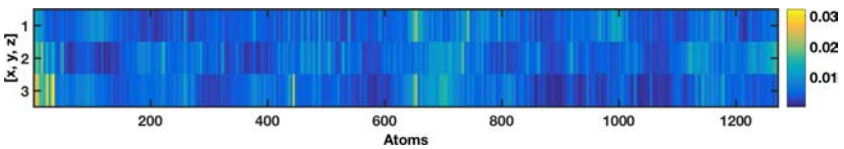


Fig. 5. (a) Best decoy structure compared with structures obtained with (b) 3 PCA terms, (c) 5 PCA terms, (d) 7 PCA terms, (e) 9 PCA terms, and (f) 11 PCA terms.

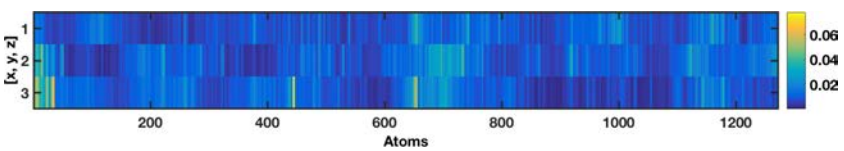
Figure 6 shows the results of the uncertainty analysis using the sampled protein decoys whose energy is below -200 for three PCA terms, whereas Fig. 7 shows the same graphics for 11 PCA terms. For each case, we show the median protein as a matrix with rows containing the coordinates x, y and z and the columns containing the atoms in the protein. In this graphic, we show for each coordinate the median of the coordinates of the decoys that have been sampled and fulfill the above-mentioned energy condition. The median protein obtained this way is to be compared with the native structure. Additionally, this figure also shows the IQR of each coordinate based on the sampled decoys and the IQR versus the median ratio. The methodology to produce the IQR plot is the same as for the median. These two last graphs are used to quantify the uncertainty and variations of the protein models around the median for different PCA expansions. This kind of representing the proteins is aimed at better visualizing the uncertainty corresponding to the tertiary structure prediction problem. These graphs show that the higher variations in the coordinates occur at the protein surface. Additionally, as the number of PCA terms decreases, the variations are smaller, that is, the ill-conditioned character of the tertiary protein structure prediciton problem is reduced. Nevertheless, the structures that are obtained are far from the native structure. In contrast, the more PCA terms, the



(a) Median protein

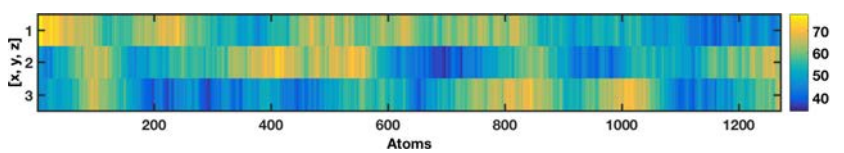


(b) IQR

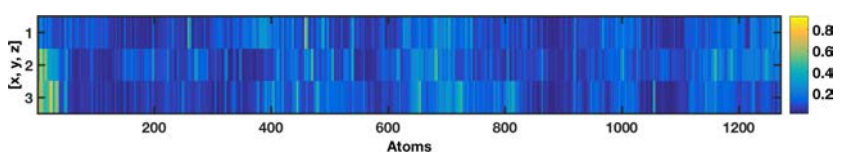


(c) IQR/Median (%)

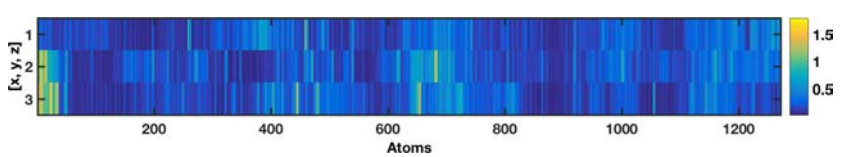
Fig. 6. Uncertainty analysis using the protein decoys that have been sampled in the energy lower than −200 for three PCAs. The graphs show the median protein configuration, the IQR, and the IQR versus the median ratio (%).



(a) Median protein



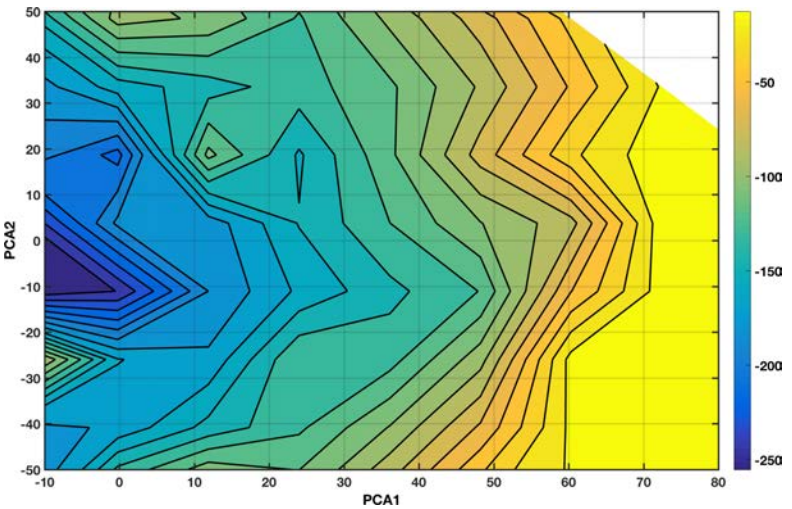
(b) IQR



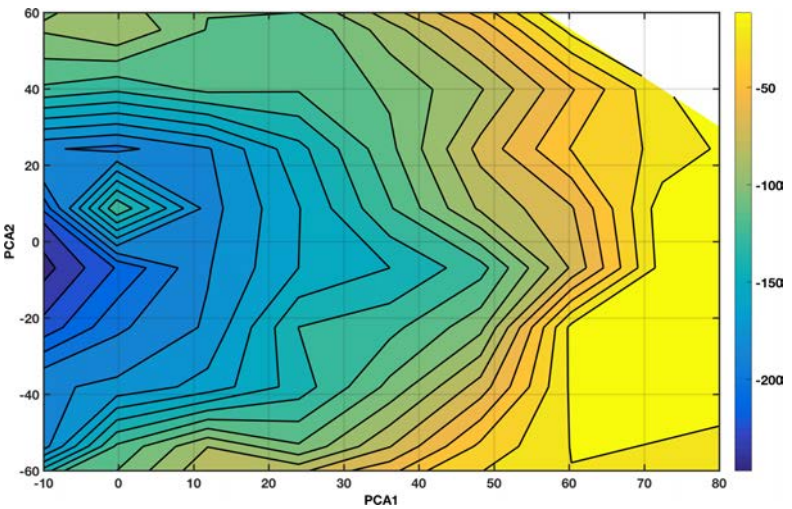
(c) IQR/Median (%)

Fig. 7. Uncertainty analysis using the protein decoys that have been sampled in the energy lower than −200 for 11 PCAs. The graphs show the median protein configuration, the IQR, and the IQR versus the median ratio (%).

more ill-conditioned the optimization problem is as it considers more spatial harmonics (PCAs) for the expansion. Therefore, a tradeoff exists between the ill-conditioning of the tertiary protein prediction problem in the reduced space and the realiability of the reconstruction as the number of PCAs increases. This result could be also expected taking into account the prior variablity curve (Fig. 1(b)) as a function of the number of PCAs, since it can be seen that as the number of PCA terms increases, the information of the initial matrix is greater, resulting in a greater

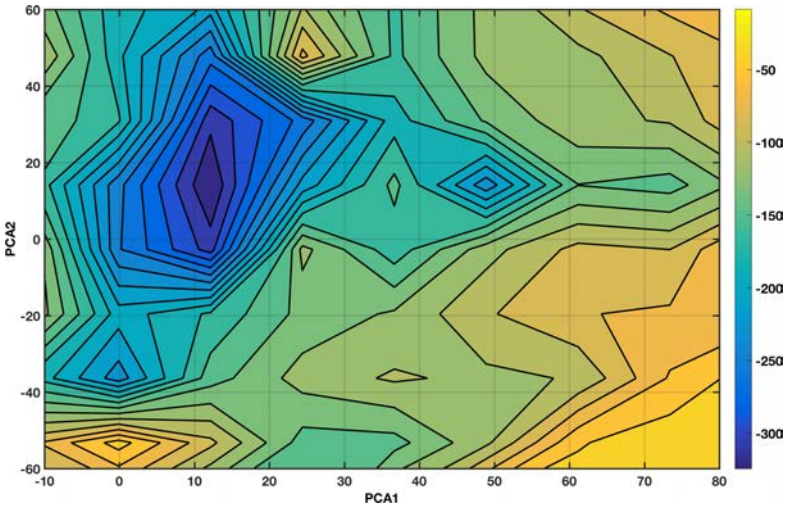


(a)

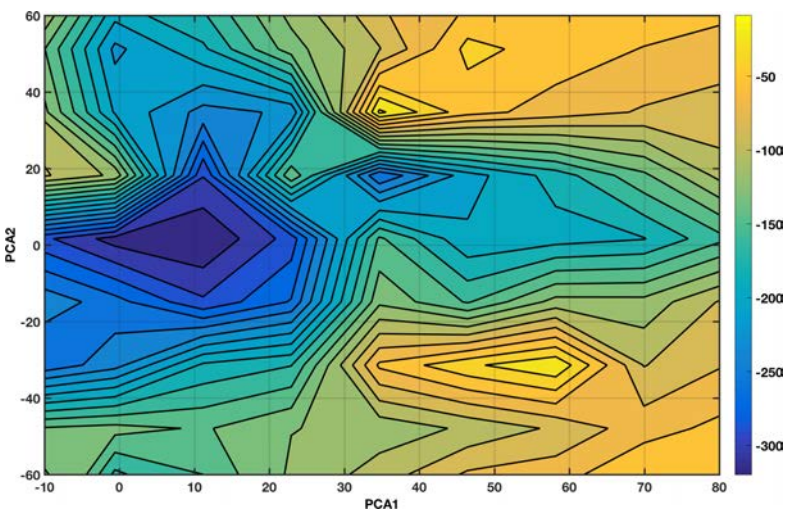


(b)

Fig. 8. T0545 protein energy landscape for four different PCA expansions: (a) 3 PCAs; (b) 5 PCAs; (c) 9 PCAs; and (d) 11 PCA.



(c)



(d)

Fig. 8. (*Continued*)

capability of minimizing the energy function. Therefore, as we decrease the dimensionality of the PCA search space, some crucial information required to get a good prediction is lost in the model simplification procedure, and the sampling algorithm accounts for fewer structural variations.

Figure 8 shows the topography of the cost function in the first two PCA coordinates of this protein, which has been interpolated from the PSO samples for the different PCA expansions treated in this paper. It can be observed that the

topography is more or less the same, with a central valley of low energies, whose orientation is North-South. The landscape becomes more complex when 11 PCAs are used, showing an East-West prolongation of the basin with its typical banana-shape.[20] This graph is useful to assess the mathematical complexity of the protein tertiary structure prediction problem by observing the complexity of the intricated valleys of the energy function in lower dimensions.

4.2. *Modeling of other proteins from CASP experiment*

Additional information is presented in order to support the theoretical benchmark described in the paper. We tested additional proteins utilizing PCA and PSO to prove its suitability for protein refinement purposes. Table 1 presents the summary of the computations carried out with different proteins detailing the energy and the number of PCA terms utilized to construct the search space. After this table, a more thorough desciption of each protein is introduced, including the algorithm

Table 1. Summary of the computational experiments performed in this paper, via PCA and PSO.

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**
			Proteins in Appendix A				
T0557	−278.9	−273.7	**−275.3**	**−275.2**	**−275.4**	**−277.2**	**−277.6**
T0580	−258.3	**−253.8**	−196.4	−250.8	−249.7	−249.5	−250.8
T0637	−384.5	−372.0	−46.7	−103.7	−369.2	−371.4	−372.4
T0643	−234.3	**−209.4**	−138.9	−209.2	**−209.5**	**−210.0**	**−210.0**

Note: Energy of the best decoy used in the PCA and lower energy found after PSO optimization. Bold faces indicate the cases in which the energy after optimization improved. The results concerning the proteins T0557 to T0643 are given in Appendix A.

Table 2. Summary of the computational experiments performed in this paper, via PCA and PSO.

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**
		Proteins in Appendix A				
T0557	1.617	1.696	**1.606**	**1.596**	**1.024**	**0.780**
T0580	**1.284**	1.716	1.331	1.303	1.304	1.291
T0637	4.961	12.610	7.468	4.966	**4.964**	**4.286**
T0643	3.882	20.670	19.800	**3.728**	**3.432**	**2.915**

Note: RMSD of atomic positions of the best decoy used in the PCA and the model with lower energy found after PSO optimization. The RMSD is calculated with respect to the native structure, which is known for these proteins. Bold faces indicate the cases in which the RMSD after optimization improved. The results concerning the proteins T0555 to T0643 are given in Appendix A.

performance, protein structures, and uncertainty analysis. Further details can be found in Appendix A. Other proteins could be equally modeled, but the aim of this section is not to show an exhaustive analysis of a large set of proteins, but to show that the methodology provides systematically good results. Nevertheless, some additional numerical results are shown in Tables A.1 and A.2 (Appendix A). This set of proteins has been randomly chosen, but the results show a systematic improvement with respect to the best model found.

5. Conclusions

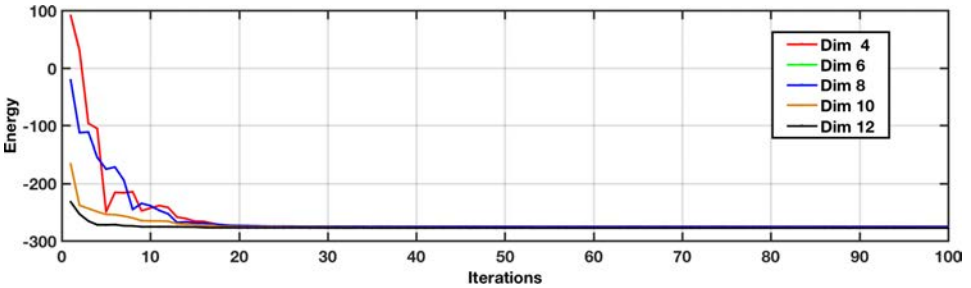
In this paper, we present a study of the PCA dimensionality and how this can affect the energy optimization and tertiary structure prediction of a protein from the CASP9 experiment (Uracil DNA glycosylase from *M. acetivorans*). The algorithm utilized succesfully establishes a low-dimensional space in order to apply the energy optimization procedure via a member of the family of PSOs. This model reduction has been performed in order to obtain four different search spaces (3, 5, 7, 9, and 11 dimensions in addition to a high frequency term) with views of performing the energy optimization later on. The optimizer was capable of modeling the protein sequence and sampling the selected decoys projected over the five 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 on the decoys used to construct the PCA is included, and a lower enegy and uncertainty is obtained in the predictions. The introduction of the high frequency term corresponding to the best model submitted is crucial to expand high frequency details of the tertiary structure and being able to lower the energy getting closer to the native structure. We believe that the results shown in this paper could be significantly improved using higher computational resources to improve the PSO sampling, which can be intrinsically parallelized. Also, in some cases, a higher number of PCAs would also help to improve the predictions.

Finally, this paper helps 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 appropiate model expansion by taking into account the existing trade-off between prior variability expansion and the energy optimization to find models close to the native structure.

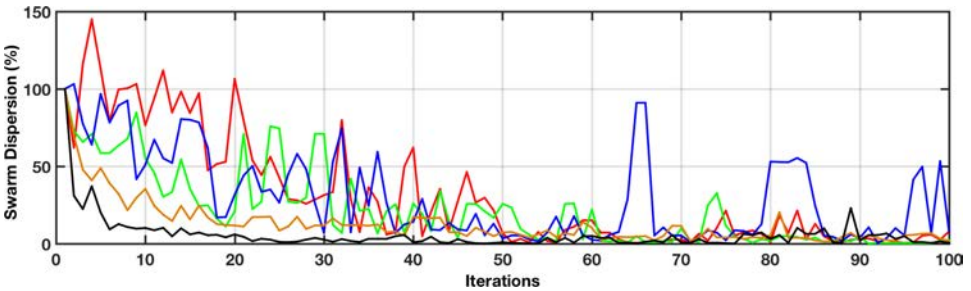
Appendix A

T0557— N-terminal domain of putative ATP-dependent DNA helicase RecG-related protein from Nitrosomonas European

The native structure of this protein has been obtained through NMR by Eletsky *et al.*[42] at the Northeast Structural Genomics Consortium. Figure A.1 represents the



(a)



(b)

Fig. A.1. T0557 protein. (a) Convergence curve for different numbers of PCAs. (b) Swarm dispersion (%).

convergence and dispersion curves for the T0557. As it can be observed, the behavior of the convergence curves is similar in order to obtain practically the same final energy regardless of the number of PCA terms that is employed. The algorithm reaches in all the cases the lower energy before 20 iterations. The main difference among these graphics is the lower energy at the first iterations for a higher number of PCA terms. This fact explains that when high frequency details are added to the expansion, it is easier to attend lower energy regions of the backbone structure. Besides, the dispersion is higher for lower dimensions. In this sense, analyzing solely the energy at which the algorithm collapses provides us with the tantalizing idea that the protein refinement barely improved the structure. Also, no variations are observed in the different structures predicted for different search space dimensions (Fig. A.2). However, imperceptible variations in atom coordinates could be observed in Fig. A.3, which shows the uncertainty analysis of the protein structure with 11 PCAs. Similar graphics can be produced for other PCA expansions. In conclusion, it is possible to observe how adding PCA terms to the search space expands the information contained in the decoys and, consequently, the PSO samples better describe the energy landscape of the protein, yielding a protein refinement with a lower uncertainty. In this case, the PCA-PSO procedure clearly improved the results concerning the best decoy. The native structure has energy of −278.3; the best decoy

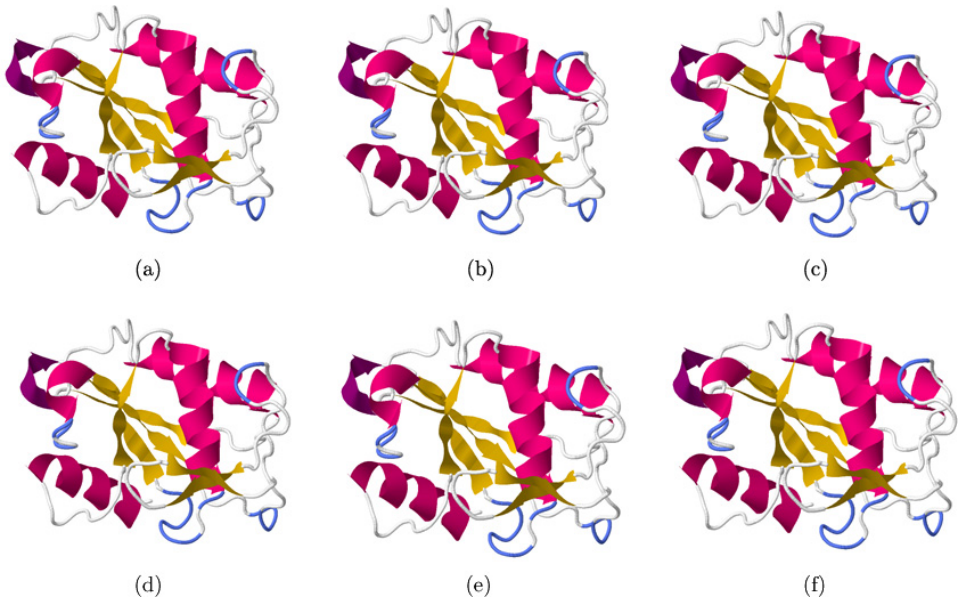


Fig. A.2. T0557 protein. (a) Native structure compared with structures obtained with (b) 3 PCA terms, (c) 5 PCA terms, (d) 7 PCA terms, (e) 9 PCA terms, and (f) 11 PCA terms.

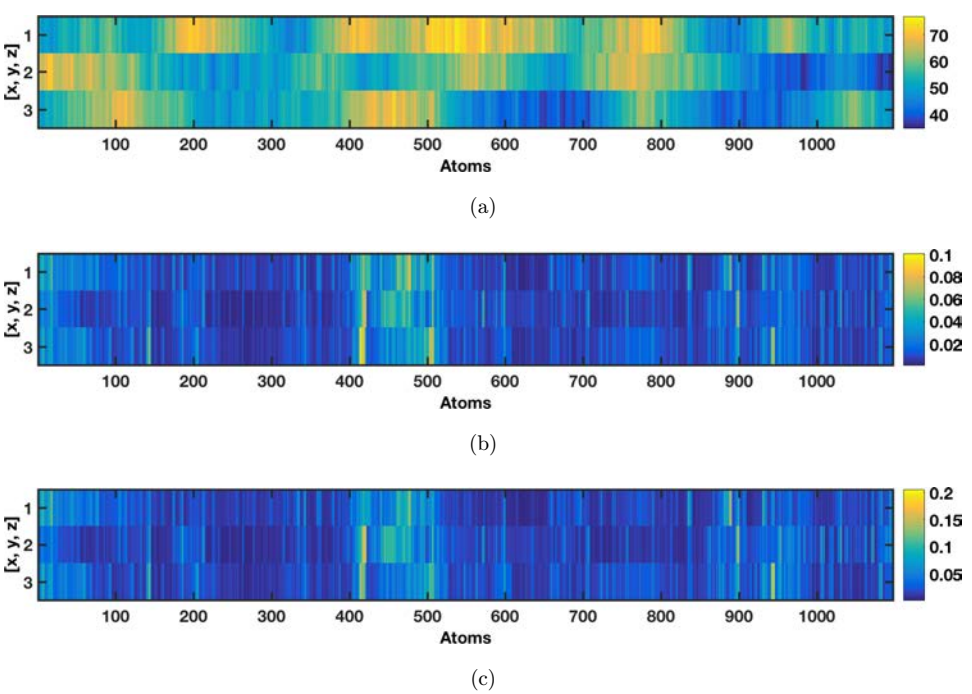
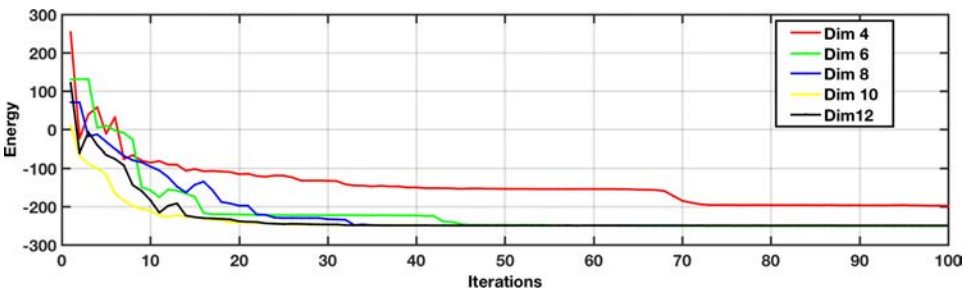


Fig. A.3. T0557 protein. Uncertainty analysis obtained for 11 PCAs. (a) Median protein. (b) Interquartile range (IQR). (c) IQR versus median ratio (%).

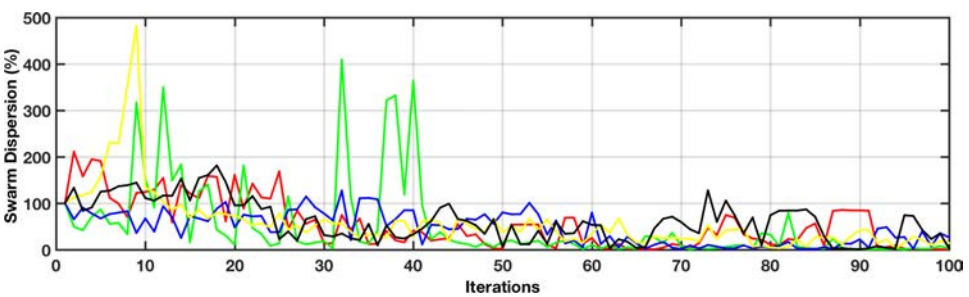
submitted −273.7; and the best model found with 11 PCAs −277.6 (Table 1). The RMSD distance (Table 2) between the best decoy and the native structure was decreased from 1.617 to 0.780 with 11 PCAs.

T0580 — The lactose-specific IIB component domain structure of the phosphoenolpyruvate: Carbohydrate phosphotransferase system (PTS) from Streptococcus pneumoniae

Protein T0580 corresponds to a different case, where the dimensionality of the search space plays a key role in the successful prediction of the protein structure. T0580 protein from CASP9 has been reported by Cuff *et al.*[43] and is considered a hypothetical structure. Figure A.4 shows the convergence and dispersion curves for different numbers of PCAs and how including a higher number of PCA terms allows bettering sample the energy landscape of the protein. With only four dimensions, part of the information is missing and the algorithm is not able to reconstruct a proper protein structure (Fig. A.5). Figure A.6 shows the uncertainty analysis with 11 PCA terms showing the recovered median structure and also the regions where the uncertainty in the reconstruction is bigger (red areas in the IQR). Although it is not shown, these graphics are similar for different PCA expansions. In this case, the PCA-PSO procedure did not improve the results concerning the best decoy. The



(a)



(b)

Fig. A.4. T0580 protein. (a) Convergence curve for different numbers of PCAs. (b) Swarm dispersion (%).

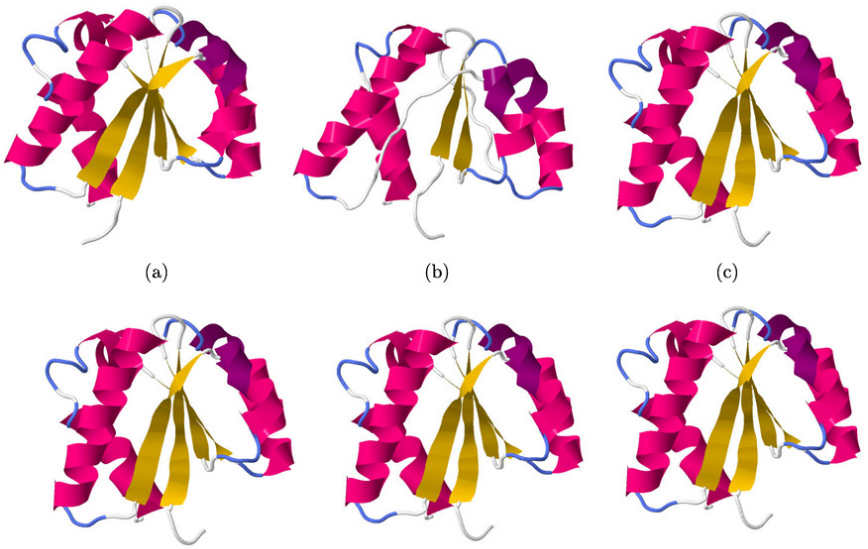


Fig. A.5. T0580 protein. (a) Native structure compared with structures obtained with (b) 3 PCA terms, (c) 5 PCA terms, (d) 7 PCA terms, (e) 9 PCA terms, and (f) 11 PCA terms.

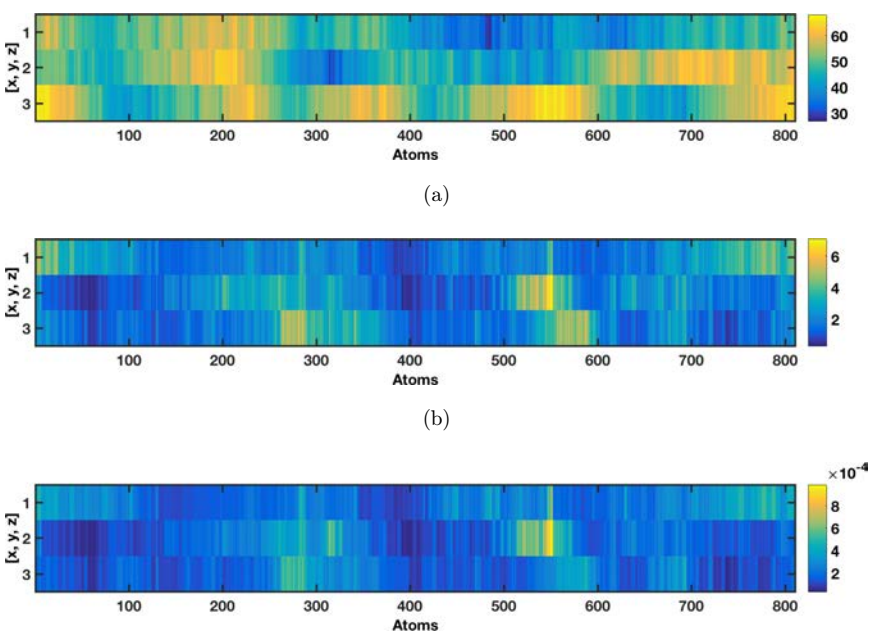
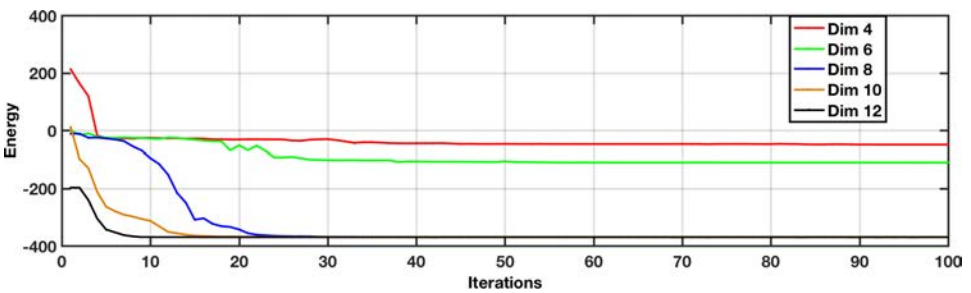


Fig. A.6. T0580 protein. Uncertainty analysis obtained for 11 PCAs. (a) Median protein. (b) IQR. (c) IQR versus median ratio (%).

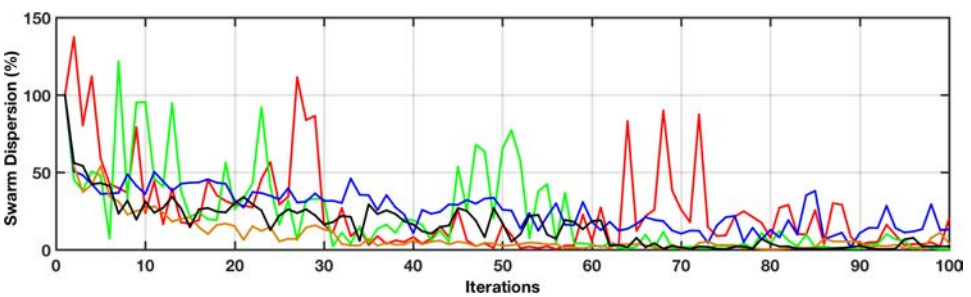
native structure has energy of −258.3; the best decoy −253.8; and the best model found with 11 PCAs has energy of −250.8 (Table 1). Interesting, the RMSD distance (Table 2) between the best decoy and the native structure (1.284) is very similar to the distance obtained with 11 PCAs (1.291). We believe that these results could be further improved augmenting the number of PCAs and the number of particles used in the PSO sampling.

T0639 — Crystal structure of functionally unknown protein from Neisseria meningitidis MC58

Protein T0639 from CASP9 competition is a protein from *Neisseria meningitidis* MC58, whose native structure was obtained by Zhang *et al.*[44] at the Midwest Center for Structural Genomics via X-ray diffraction. Similar to the case of T0637, this protein is susceptible to the search space dimensionality. In this sense, it possible to observe in Fig. A.7 that the more dimensions included the better the energy refinement. Also with only four PCAs it is impossible to attend the minimum energy, while for the other configurations less than 40 iterations are needed to attend it. Figure A.8 shows the reconstructed protein structures and Fig. A.9 shows the corresponding uncertainty analysis for 11 PCAs. It is possible to observe very low uncertainties. Nevertheless, for lower PCA dimensions, the uncertainty grows considerably (these graphs are not shown). In this sense, the algorithm is not capable of



(a)



(b)

Fig. A.7. T0637 protein. (a) Convergence curve for different numbers of PCAs. (b) Swarm dispersion (%).

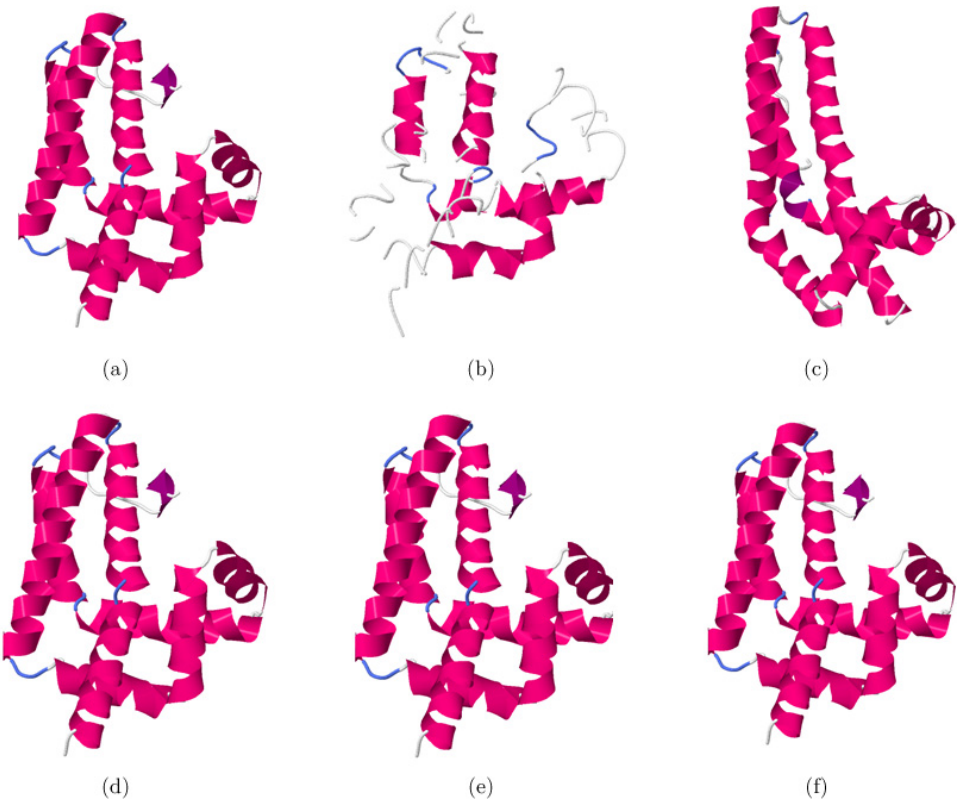
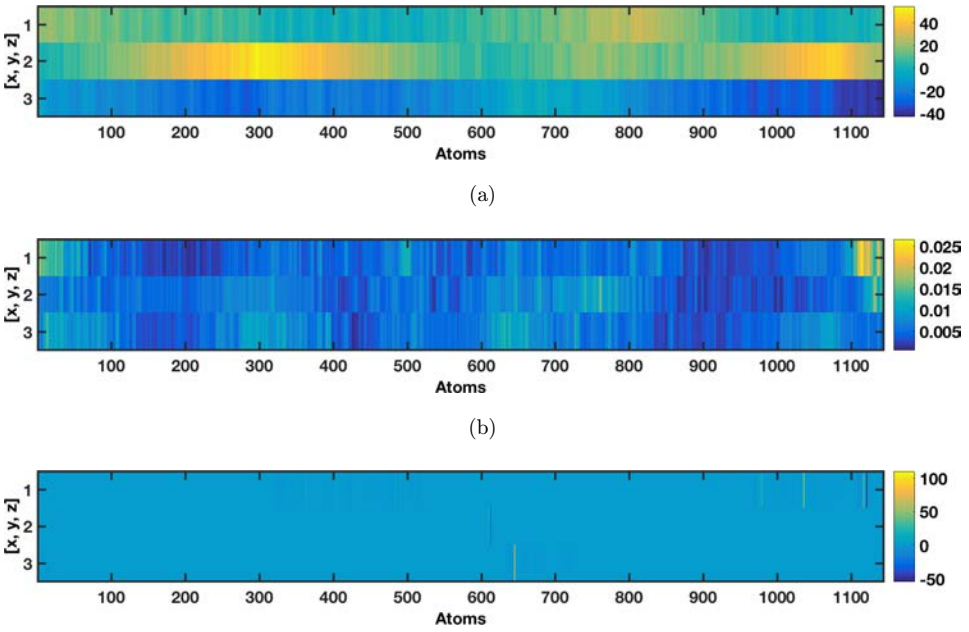


Fig. A.8. T0637 protein. (a) Native structure compared with structures obtained with (b) 3 PCA terms, (c) 5 PCA terms, (d) 7 PCA terms, (e) 9 PCA terms, and (f) 11 PCA terms.

sampling the entire conformational space of the protein structure and, moreover, it is not able to reconstruct it after carrying out the PSO sampling. In this case, the PCA-PSO procedure improved the results concerning the best decoy submitted. The native structure has energy of −380.6; the best decoy submitted −343.6; and the best model found with 9 PCAs −345.7 (Table 1). Therefore, the energy of the native structure is still far from the structure that has been found. The RMSD distance (Table 2) between the best decoy and the native structure (7.944) was decreased to 4.693 with 11 PCAs. We also think that these results could be further improved by increasing the number of PCAs used and the number of particles in the PSO sampling.

T0643 — Crystal structure of the N-terminal domain of DNA-binding protein SATB1 from Homo sapiens

Protein T0643 was also considered, which corresponds to the N-terminal domain of DNA-binding protein SATB1, whose native structure was obtained through X-ray



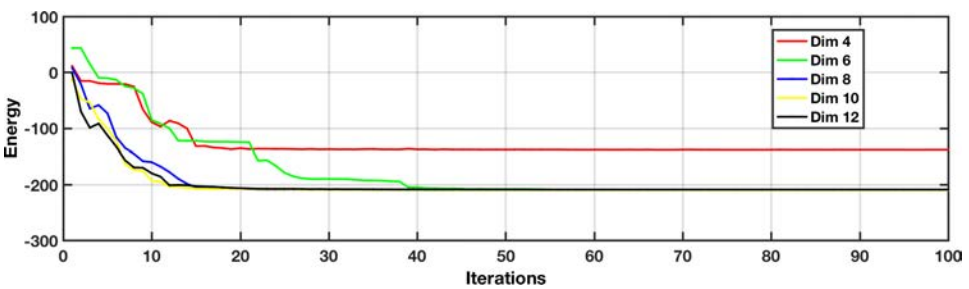
(a)

(b)

(c)

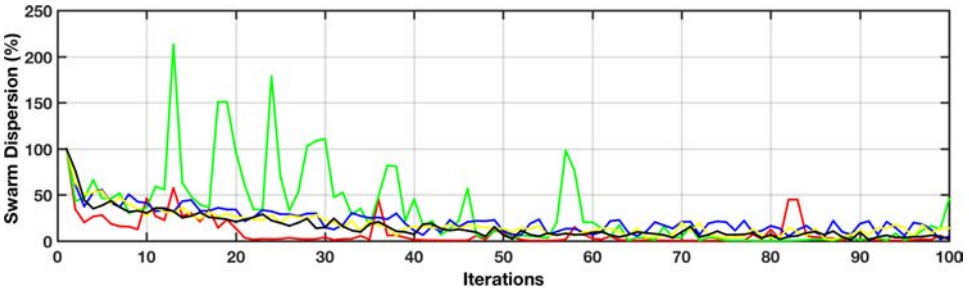
Fig. A.9. T0637 protein. Uncertainty analysis obtained for 11 PCAs. (a) Median protein. (b) IQR. (c) IQR versus median ratio (%).

diffraction by Forouhar *et al.*[45] Figure A.10 shows very high energies when only four dimensions are considered. At higher dimensions, the PSO algorithm is capable of sampling the conformational space of the protein structure, obtaining better energy predictions. Furthermore, this idea is also supported by Figs. A.11 and A.12, where at low dimensions, not enough information is obtained to reconstruct the proteins, yielding to structures that do not correspond to the reality yielding high uncertainties. At higher dimensions, PSO successfully samples the energy function



(a)

Fig. A.10. T0643 protein. (a) Convergence curve for different numbers of PCAs. (b) Swarm dispersion (%).



(b)

Fig. A.10. (*Continued*)

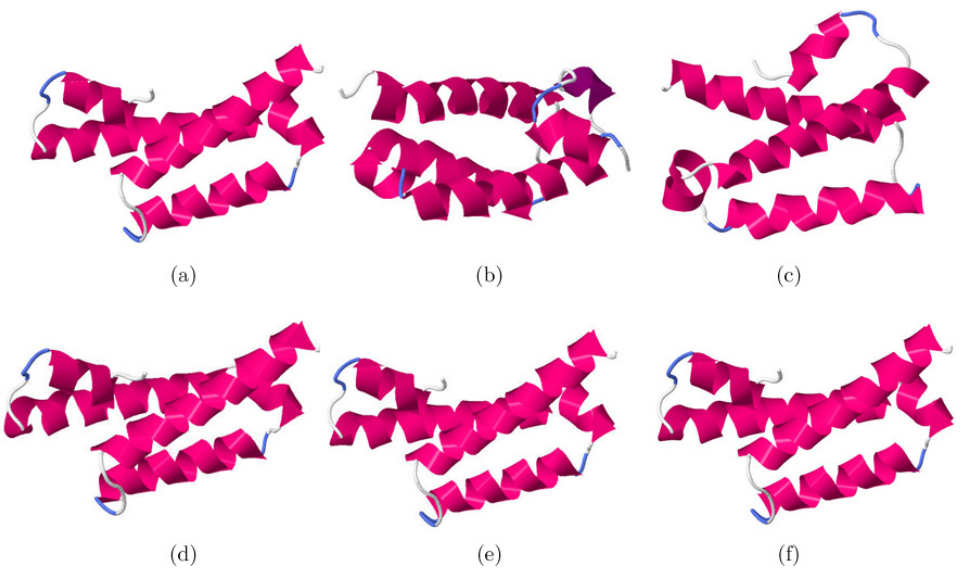


Fig. A.11. T0643 protein. (a) Native structure compared with structures obtained with (b) 3 PCA terms, (c) 5 PCA terms, (d) 7 PCA terms, (e) 9 PCA terms, and (f) 11 PCA terms.

landscape and the structures are similar to the best decoy in CASP9 competition with its corresponding low uncertainty.

In this case, the PCA-PSO provided results similar to the best decoy. The native structure has energy of −234.3; the best decoy −209.4 and the best model found with seven PCAs −209.5 (Table 1). The RMSD distance (Table 2) between the best decoy and the native structure (3.882) was clearly decreased with 11 PCAs (2.915). We believe that the results could be further improved by augmenting the number of PCAs and the number of particles in the PSO sampling using higher computational resources.

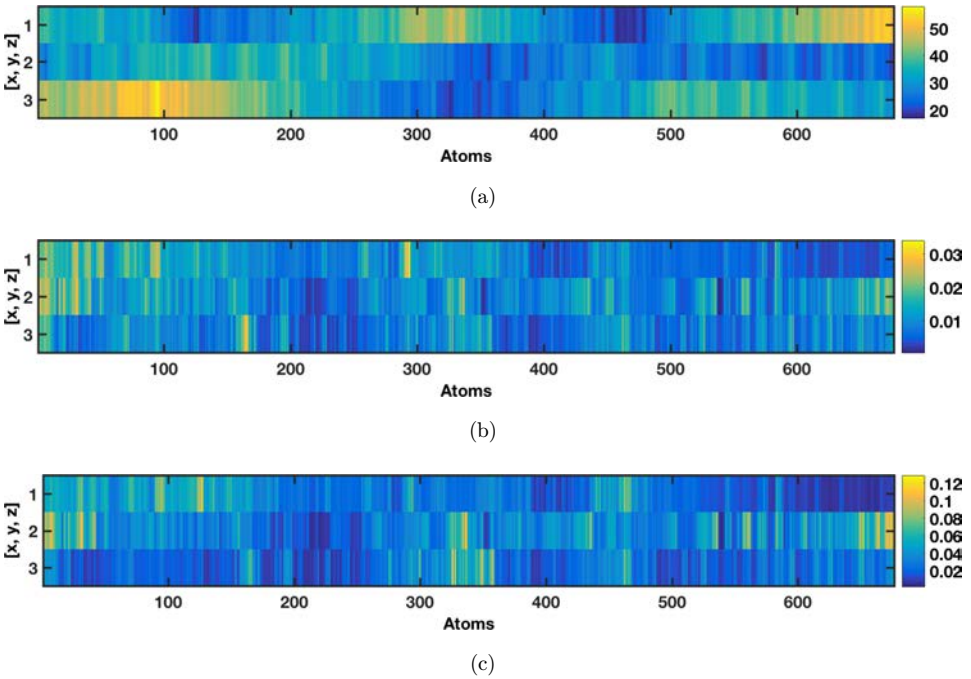


Fig. A.12. T0643 protein. Uncertainty analysis obtained for 11 PCAs. (a) Median protein. (b) IQR. (c) IQR versus median ratio (%).

Proteins shown in Tables A.1 and A.2

Tables A.1 and A.2 show some additional results for other proteins of the CASP competition. In all the cases, the energy and the RMSD have decreased with respect to the best decoy, although in some cases the energy is still far from the native structure. It might be recommended a higher exploration and adding more PCA terms to the expansion. For instance, we have predicted T0561 and T0639 increasing the number of PCAs to 18, using a swarm of 200 particles. The energy of T0561 decreased to −471.4 with RMSD with respect to the native of 5.814, whereas in the

Table A.1. Additional numerical results.

Protein CASP9 code	Native structure	Best decoy	3 PCA terms	5 PCA terms	7 PCA terms	9 PCA terms	11 PCA terms
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
T06358	−466.5	−462.8	−43.7	−324.1	−361.7	−463.1	−463.6
T0639	−380.6	−343.6	−102.3	−335.5	−345.4	−345.7	−345.4

Note: Summary of the computational experiments performed in this paper via PCA and PSO. Energy of the best decoy used in the PCA and lower energy found after PSO optimization. No graphical outputs are given for these proteins.

Table A.2. Additional numerical results.

Protein CASP9 code	Best decoy	3 PCA terms	5 PCA terms	7 PCA terms	9 PCA terms	11 PCA terms
T0555	8.566	14.411	8.568	8.566	8.522	8.516
T0561	5.898	14.156	5.941	5.899	5.895	5.892
T0635	2.450	12.520	9.238	6.388	2.225	2.222
T0639	7.944	13.390	10.310	8.967	6.068	4.693

Note: Summary of the computational experiments performed in this paper, via PCA and PSO. RMSD of the best decoy used in the PCA and lower energy found after PSO optimization. No graphical outputs are given for these proteins.

case of T0639 the energy decreased to −360.2 with RMSD of 4.586. Therefore, the results can be improved in most of the cases using advanced computational resources.

It should be understood that solving this optimization problem in higher dimensions and locating their global energy optimum is like finding a needle in a haystack (in words of Albert Tarantola[46]). Besides, in some cases, the energy function might not take into account all the physical phenomena involved in the protein energy landscape.

Acknowledgments

A.K. acknowledges financial support from NSF grand DBI1661391 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. Simons KT, Kooperberg C, Huang E, Baker D, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions, *J Mol Biol* **268**:209–225, 1997.
3. Xu D, Zhang Y, *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge-based force field, *Proteins Struct Funct Bioinform* **80**:1715–1735, 2012.
4. Li SC, Bu D, Xu J, Li M, Fragment-HMM: A new approach to protein structure prediction, *Protein Sci* **17**:1925–1934, 2008.
5. Bowie JU, Lüthy R, Eisenberg D, A method to identify protein sequences that fold into a known three-dimensional structure, *Science* **253**:164–170, 1991.
6. Jones DT, Taylor WR, Thornton JM, A new approach to protein fold recognition, *Nature* **358**:86–89, 1992.
7. Zhang Y, I-TASSER server for protein 3D structure prediction, *BMC Bioinformatics* **9**:40, 2008.
8. Bellman RE, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.
9. Fernández-Martínez JL, Model reduction and uncertainty analysis in inverse problems, *Leading Edge* **34**:1006–1016, 2015.
10. Stumpff-Kane A, Feig M, A correlation-based method for the enhancement of scoring functions on funnel-shaped energy landscapes, *Proteins* **63**:155–164, 2006.

11. Olson MA, Feig M, Brooks CL, Prediction of protein loop conformations using multiscale modeling methods with physical energy scoring functions, *J Comput Chem* **29**:820–831, 2008.
12. Mirjalili V, Feig M, Protein structure refinement through structure selection and averaging from molecular dynamics ensembles, *J Chem Theory Comput* **9**:1294–1303, 2013.
13. Gniewek P, Kolinski, A, Jernigan RL, Kloczkowski, A, Elastic network normal modes provide a basis for protein structure refinement, *J Chem Phys* **136**(19):195101, 2012.
14. Bradley P, Chivian D, Meiler J, Misura K, Rohl C, Schief WWW, Schueler-Furman O, Murphy P, Schonbrun J, Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation, *Proteins* **53**:457–468, 2003.
15. Bonneau R, Strauss CE, Rohl CA, 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.
16. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D, Automated prediction of CASP-5 structures using the Robetta server, *Proteins* **53**:524–533, 2003.
17. Stoker H, *Organic and Biological Chemistry*, Cengage Learning, 2015.
18. Ramelot T, Raman S, Kuzin A, Xiao R, Ma L, Acton T, Hunt J, Montelione G, Baker D, Kennedy M, Improving NMR protein structure quality by Rosetta refinement: A molecular replacement study, *Proteins* **75**:147–167, 2009.
19. Baker D, Sali A, Protein structure prediction and structural genomics, *Science* **294**:93–96, 2001.
20. Price S, From crystal structure prediction to polymorph prediction: Interpreting the crystal energy landscape, *Phys Chem Chem Phys* **2008**:1996–2009, 2008.
21. Fernández-Martínez J, Fernández-Muñiz M, Tompkins M, On the topography of the cost functional in linear and nonlinear inverse problems, *Geophysics* **77**:W1–W15, 2012.
22. Fernández-Álvarez JP, Fernández-Martínez JL, Menéndez-Pérez CO, Feasibility analysis of the use of binary genetic algorithms as importance samplers application to a geoelectrical VES inverse problem, *Math Geosci* **40**:375–408, 2008.
23. Fernández-Martínez JL, Fernández-Álvarez JP, García-Gonzalo ME, Menéndez-Pérez CO, Kuzma HA, Particle swarm optimization (PSO): A simple and powerful algorithm family for geophysical inversion, SEG Technical Program Expanded Abstracts, pp. 3568–3571, 2008.
24. Fernández-Martínez JL, García-Gonzalo E, Álvarez JPF, Kuzma HA, Menéndez-Pérez CO, A powerful algorithm to solve geophysical inverse problems. Application to a 1D-DC resistivity case, *J Appl Geophys* **71**:13–25, 2010.
25. Saraswathi S, Fernández-Martínez JL, Koliński A, Jernigan RL, Kloczkowski A, Fast learning optimized prediction methodology (FLOPRED) for protein secondary structure prediction, *J Mol Model* **18**(9):4275–4289, 2012.
26. Saraswathi S, Fernández-Martínez JL, Koliński A, Jernigan RL, Kloczkowski A, Distributions of amino acids suggest that certain residue types more effectively determine protein secondary structure, *J Mol Model* **19**(10):4337–4348, 2013.
27. Pearson K, On lines and planes of closest fit to systems of points in space, *Phylo Mag* **2**:559–572, 1901.
28. Jolliffe I, *Principal Component Analysis*, Springer, 2002.
29. Quian B, Ortiz A, Baker D, Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation, *Proc Natl Acad Sci USA* **101**:15346–15351, 2004.
30. Fernández-Martínez JL, García-Gonzalo E, The PSO family: Deduction, stochastic analysis and comparison, *Swarm Intell* **3**:245–273, 2009.

31. Fernandez-Martinez JL, Garcia-Gonzalo E, Stochastic stability analysis of the linear continuous and discrete PSO models, *IEEE Trans Evol Comput* **15**(3):405–423, 2011.
32. Fernández-Martínez JL, 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.
33. García-Gonzalo E, Fernández-Martínez JL, Convergence and stochastic stability analysis of particle swarm optimization variants with generic parameter distributions, *Appl Math Comput* **249**:286–302, 2014.
34. Fernández-Martínez JL, Mukerji T, García-Gonzalo E, Fernández-Muñiz Z, Uncertainty assessment for inverse problems in high dimensional spaces using particle swarm optimization and model reduction techniques, *Math Comput Model* **54**:2889–2899, 2011.
35. Fernández-Martínez JL, Mukerji T, García-Gonzalo E, Suman A, Reservoir characterization and inversion uncertainty via a family of particle swarm optimizers, *Geophysics* **77**(1):M1–M16, 2012.
36. Kennedy J, Eberhart R, A new optimizers using particle swarm theory, *Proc. Sixth Int. Symp. Micro Mach. Human Sci.* **1**:39–43, 1995.
37. Gront D, Kolinski A, Bioshell — A package of tools for structural biology prediction, *Bioinformatics* **22**:621–622, 2006.
38. Gront D, Kolinski A, Utility library for structural bioinformatics, *Bioinformatics* **24**:584–585, 2008.
39. Gniewek P, Kolinski A, Kloczkowski A, Gront D, BioShell — Threading: A versatile Monte Carlo package for protein threading, *BMC Bioinformatics* **22**:15–22, 2014.
40. Aramini J, Hamilton K, Ciccosanti C, Wang H, Lee H, Rost B, Acton T, Xiao R, Everett J, Montelione G, Solution NMR atructure of a putative Uracil DNA glycosylase from *Methanosarcina Acetivorans*. Northeast Structural Genomics Consortium Target, 2010.
41. Martínez JLF, Gonzalo EG, Muñiz ZF, Mariethoz G, Mukerji T, Posterior sampling using particle swarm optimizers and model reduction techniques, *Int J App Evol Comp* **1**:27–48, 2010.
42. Eletsky A, Mills JL, Lee H, Maglaqui M, Ciccosanti C, Hamilton K, Rost B, Acton TB, Xiao R, Everett JK, Montelione GT, Prestegard JH, Szyperski T, Solution NMR structure of the N-terminal domain of putative ATP-dependent DNA helicase RecG-related protein from *Nitrosomonas Europaea*. Northeast Structural Genomics Consortium Target NeR70A, 2010.
43. Cuff ME, Chhor G, Clancy S, Joachimiak A, The lactose-specific IIB component domain structure of the phosphoenolpyruvate:carbohydrate phosphotransferase system (PTS) from *Streptococcus Pneumoniae*, Midwest Center for Structural Genomics, 2010.
44. Zhang R, Tan K, Volkart L, Bargassa M, Joachimiak A, Crystal structure of functionally unknown protein from *Neisseria meningitidis* MC58, Midwest Center for Structural Genomics, 2010.
45. Forouhar F, Abashidze M, Seetharaman J, Kuzin A, Patel P, Xiao R, Ciccosanti C, Shastry R, Everett J, Nair R, Acton T, Rost B, Montelione G, Hunt J, Tong L, Crystal structure of the N-terminal domain of DNA-binding protein SATB1 from *Homo sapiens*. Northeast Structural Genomics Consortium, 2010.
46. Tarantola A, Popper, Bayes and the inverse problem, *Nat Phys* **2**:492–494, 2006.

Óscar Álvarez holds a Bachelor's Degree in Chemical Engineering by the University of Oviedo, being trained afterwards in computational modeling of physics systems at The University of Manchester where he pursued a Master of Science. His

research interest is to investigate the physicochemical behavior of Soft Matter systems. He applies computer simulation and modeling techniques to elucidate the physics that drive the equilibrium and dynamical properties of a system, and hence to predict and control its macroscopic response. Within the areas of utmost interest to him, he would remark proteomics, that is, predicting the tertiary structure of proteins, applying sampling techniques to elucidate protein mutations or to study new algorithms and methods to reduce the computational cost of modeling these biomolecules.

Juan Luis Fernández-Martínez received his Ph.D. in mining engineering from the University of Oviedo (Spain) in 1994 and was previously trained as a petroleum engineer in France (Ècole Nationale du Pètrole et des Moteurs, Paris, 1988) and England (Imperial College, Royal School of Mines, London, 1989). After years of working as a computing software engineer in France, he joined the Mathematics Department of Oviedo University in 1994 and has since held the position as a professor in applied mathematics, where he directs the Group of Inverse Problems, Optimization, and Machine Learning. During 2008–2010 he was a visiting and research professor at UC Berkeley-Lawrence Berkeley Laboratories and Stanford University. His areas of expertise include inverse problems, uncertainty analysis of very complex systems, feature selection and model reduction techniques, cooperative global optimization methods, with application in oil and gas, biometry, biomedicine, and finance.

Celia Fernández-Brillet studies Biomedical Engineering at Universidad Carlos III de Madrid (Spain). Passionate about biomedicine and technology, she wants to pursue a Master's Degree at Stanford University.

Ana Cernea has her M.Sc. and Ph.D. in Applied Mathematics. Her research interests include biometry, digital processing methods, and biomedicine.

Zulima Fernández-Muñiz is Mining Engineer with Ph.D. in Applied Mathematics. Her research interests include model reduction and uncertainty analysis in inverse problems.

Andrzej Kloczkowski is Principal Investigator in the Battelle Center for Mathematical Medicine in the Research Institute of the Nationwide Children's Hospital in Columbus, Ohio and Professor of Pediatrics in the Department of Pediatrics of The Ohio State University College of Medicine. His research focusses on various aspects of computational molecular biology and structural bioinformatics. Areas of interest include systems biology at multiple levels, prediction of protein structure, dynamics and function, protein packing, development of statistical potentials, prediction of binding sites, phosphorylation and other post-translational modification sites using machine learning methodologies, and application of these methods to various biomedical problems.