Journal of Molecular Modeling (2019) 25:79
https://doi.org/10.1007/s00894-019-3956-0

ORIGINAL PAPER

Predicting protein tertiary structure and its uncertainty analysis via particle swarm sampling

Óscar Álvarez[1] · Juan Luis Fernández-Martínez[1] · Ana Cernea Corbeanu[1] · Zulima Fernández-Muñiz[1] · Andrzej Kloczkowski[2,3]

Received: 16 May 2018 / Accepted: 5 February 2019

Abstract

We discuss the relationship between the problem of protein tertiary structure prediction from the amino acid sequence and the uncertainty analysis. The algorithm presented in this paper belongs to the category of decoy-based modeling, where different known protein models are used to establish a low dimensional space via principal component analysis. The low dimensional space is utilized to perform an energy optimization via a family of very explorative particle swarm optimizers to find the global minimum. The aim of this procedure is to get a representative sample of the nonlinear equivalent region, that is, protein models that have their energy lower than a certain energy bound. The posterior analysis of this family provides very valuable information about the backbone structure of the native conformation and its possible alternate states. This methodology has the advantage of being simple and fast and can help refine the tertiary protein structure. We comprehensively illustrate the performance of our algorithm on one protein from the CASP-9 protein structure prediction experiment. We also provide a theoretical analysis of the energy landscape found in the tertiary structure protein inverse problem, explaining why model reduction techniques (principal component analysis in this case) serve to alleviate the ill-posed character of this high dimensional optimization problem. In addition, we expand the computational benchmark with a summary of other CASP-9 proteins in the Appendix.

Keywords Proteins · Tertiary structure prediction · PSO · Uncertainty analysis

Introduction

Tertiary protein structure prediction is computational elucidation of the three-dimensional structure of a protein for which an experimentally determined structure is unavailable from its amino acid sequence. Protein structure prediction is highly important in drug design, and in biotechnology, in the design of novel enzymes. The importance of this field is shown by the fact that every two years the performance of current methods is assessed in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction) and a worldwide community of researchers participate in this challenge.

The research has been focused mainly in two areas:

1. The first one focuses on understanding the mechanisms involved in protein structure and folding in order to find the correct energy function or model to be optimized [1]. Obviously all the energy functions are mathematical models that try to mimic the reality. This first part concerns what we call the forward problem: to compute the landscape of protein free energy. Obviously, the energy

✉ Juan Luis Fernández-Martínez
jlfm@uniovi.es

Óscar Álvarez
UO217123@uniovi.es

Ana Cernea Corbeanu
cerneadoina@uniovi.es

Zulima Fernández-Muñiz
zulima@uniovi.es

Andrzej Kloczkowski
Andrzej.Kloczkowski@nationwidechildrens.org

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

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

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

Published online: 27 February 2019

function choice will influence the result obtained in the optimization, since energy is the criterion that is used to classify the different templates that are obtained. Chemists, molecular biologists, and molecular physicists are usually involved in this research.

2. The second area concerns the energy optimization itself. Computer scientists, applied mathematicians, and physicists are mainly involved in this research that is not very different from classical and global optimization approaches, machine learning, and sampling (uncertainty analysis). Typically, the different optimization methods try to find the global minimum of the energy in a high dimensional space. Owing to the curse of dimensionality, the prediction methods are usually unable to explore the space of possible protein structures. These problems can be partially overcome by using some simplifications, assuming that the protein adopts a structure that is close to the experimentally determined structure of another homologous protein. The progress and challenges in protein structure prediction have been reviewed by Zhang et al. [2].

In spite of enormous efforts carried out by researchers and a growing number of protein structures experimentally solved and deposited in the Protein Data Bank (PDB), there is a huge constantly increasing gap between the number of protein sequences obtained from mass-scale genome sequencing and the number of PDB structures. Currently, after redundancy reduction, only around 1% of protein sequences have their native structures in the PDB database [3, 4]. Additionally, experimental solving of protein structures is costly, and time consuming — the main problem is obtaining the high quality crystals necessary for high resolution X-ray crystallography. Because of this, computational methods that lead to high accuracy predictions of protein structure from sequence become extremely important. Protein structure prediction methods can be divided into two categories: template-based and template-free modeling. Template-based modeling permits constructing a model of the target protein based on a template structure of a homolog, that is, a protein with known structure and high sequence identity. This is carried out by simulating the process of evolution; by introducing substitution of amino-acids while maintaining the same protein fold [5]. On the other hand, template-free modeling predicts protein structures from physics first principles by global minimization of the free energy of a protein [2, 3].

Regardless of the methodology utilized, protein tertiary structure represents a very high dimensional optimization problem, whose dimensionality coincides with the total number of atom coordinates of a protein. Thus, the problem is affected by the curse of dimensionality, as these prediction methods are unable to sample the entire 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 dimensions (more than 10). Therefore, there is a need to simplify the problem by using proper model reduction techniques that alleviate the ill-posed character as concluded in our earlier work, where we determined the reduced dimension range using principal component analysis [6].

Refinement methods are another alternative that offer a great opportunity to approximate the native structure of a given protein by using template-based models. Some of these methodologies utilize protein dynamics, coarse-graining, and spectral decomposition. In our previous research, we applied elastic network models to protein structure prediction. This model provided a reliable representation of the fluctuation in protein dynamics and explained several protein conformational changes [7, 8].

This paper is organized as follows: first the tertiary prediction problem is explained; second the energy function landscape of the tertiary protein prediction problem and the existence of equivalent protein configurations are analyzed theoretically; third, the parameter reduction using the principal component analysis (PCA) of different protein decoys is explained; fourth the particle swarm optimizers used in this paper are presented; fifth the numerical results for the MvR76 protein (CASP9 code T0545) are presented; and finally the conclusions are outlined. In addition, we expand the computational benchmark by the algorithm performance in other CASP-9 proteins randomly selected.

The tertiary structure prediction problem

Proteins are linear chains of amino acids linked by peptide bonds. Many conformations of the chain are possible owing to the rotation of the chain around each C_α atom. These conformational changes are responsible for differences in the three dimensional structure of proteins. The knowledge of protein tertiary structure is useful for determining protein-protein interactions, protein function and evolution, and drug design [3].

The importance of tertiary protein structure prediction is due to the massive amounts of protein sequence data that are produced by modern large-scale DNA sequencing efforts such as the Human Genome Project. X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the two most commonly used experimental methods employed to determine protein structure [9]. However, both methods are far too expensive and time consuming to be used to process thousands of genes encoding proteins produced by high-throughput genome sequencing. Computational methods are very interesting because they are fast and cheaply applied for protein structure prediction; nevertheless, the challenges reside in computing high resolution models of the tertiary

structure and reliable assessment of the structure prediction uncertainty [9, 10].

In this paper to perform the forward energy calculations we used the BioShell platform [11–14].

The following packages embedded in the BioShell platform were used in this research project:

- **Jbcl.data.dict-dictionaries**: this module holds various constants, variables common in protein files such as van der Waals forces, bonding energy, angular length and stiffness, etc.
- **jbcl.data.formats**: it enables us to handle the most popular formats, such as PDB, FASTA, DSSP, etc.
- **jbcl.data.types**: it classifies information typical in bioinformatics: proteins, residues, sequence profiles, etc.
- **jbcl.calc.structural**: this module calculates different protein structural properties, such as protein similarity, planar angles, bonding distances, etc.

In this sense, BioShell combined with the methodology presented in this paper, is crucial in order to predict protein structures while avoiding structural clashes. [13].

More specifically, since the protein energy landscape is very dependent on the forcefield utilized, we selected an all-atom distance dependent, pairwise energy function based on a distance-scaled, finite-ideal gas reference energy function (DFIRE), particularly a dipolar DFIRE, known as dDFIRE, because it accurately represents, over a wide range of proteins, the energy of the native structure and with high resolution hydrogen bonding, hydrophobic interactions, and structural properties [15].

Another important factor is the solvation model utilized, in this case, we used an implicit solvation model of water, embebbed in BioShell library, developed by Still and co-workers [16] known as generalized Born/surface area free-energy (GB/SA).

The energy function landscape in tertiary structure prediction

Let us refer to the model parameters by $\mathbf{m} = (m_1, m_2, \ldots, m_n)$ $\mathbf{M} \subset \mathbf{R}^n$ to the coordinates of a protein formed by n_{atoms}, with $n = 3n_{atoms}$. Here $\mathbf{M}$ is the set of admissible protein models formulated in terms of biological consistency. Later in this paper we will discuss how to characterize this set [17, 18].

The tertiary structure protein problem consists in knowing the free-energy function, $(\mathbf{m}) : \mathbf{R}^n \rightarrow \mathbf{R}$, finding the model $\mathbf{m}_p = \min_{\mathbf{m}\in\mathbf{M}} E(\mathbf{m})$ that provides the global optimum of the energy function. This is a challenging optimization problem because of the high dimension of the model space (thousands of atoms) and also because of the energy function landscape [17].

Let us suppose that $\mathbf{m}_p$ is the global optimum of the energy function, thus, $\nabla E(\mathbf{m}_p) = \mathbf{0}$. Considering a Taylor expansion of the energy function E around the model $\mathbf{m}_p$ we have:

$$E(\mathbf{m}) = E(\mathbf{m}_p) + \nabla E(\mathbf{m}_p)\cdot(\mathbf{m}-\mathbf{m}_p) + \frac{1}{2}(\mathbf{m}-\mathbf{m}_p)^T HE(\mathbf{m}_p)(\mathbf{m}-\mathbf{m}_p) + o\left(\|\mathbf{m}-\mathbf{m}_p\|_2^2\right) \quad (1)$$

where $HE(\mathbf{m}_p)$ stands for the Hessian of the energy function E (or curvature matrix) calculated in model $\mathbf{m}_p$, and $o\left(\|\mathbf{m}-\mathbf{m}_p\|_2^2\right)$ is a scalar function that vanishes faster than the squared distance between models $\mathbf{m}$ and $\mathbf{m}_p$. Neglecting the $o\left(\|\mathbf{m}-\mathbf{m}_p\|_2^2\right)$ term, that is, approaching$E(\mathbf{m})$ by its second order Taylor expansion:

$$E(\mathbf{m}) \simeq E_S(\mathbf{m}) = E(\mathbf{m}_p) + \nabla E(\mathbf{m}_p)\cdot(\mathbf{m}-\mathbf{m}_p) + \frac{1}{2}(\mathbf{m}-\mathbf{m}_p)^T HE(\mathbf{m}_p)(\mathbf{m}-\mathbf{m}_p), \quad (2)$$

and taking the facts that $\nabla E(\mathbf{m}_p) = \mathbf{0}$ and $HE(\mathbf{m}_p)$ has to be a positive definitive matrix if $\mathbf{m}_p$ is the global optimum, we have:

$$E_S^{tol}(\mathbf{m}) \leq E_{tol} \Rightarrow \frac{1}{2}(\mathbf{m}-\mathbf{m}_p)^T HE(\mathbf{m}_p)(\mathbf{m}-\mathbf{m}_p) \leq E_{tol} - E(\mathbf{m}_p) \quad (3)$$

that is, the proteins configurations with free energy less than E_{tol} belong locally to a hyper-quadric centered in $\mathbf{m}_p$ that have the Hessian, $HE(\mathbf{m}_p)$, as matrix. $E_S^{tol}(\mathbf{m})$ in Eq. (3) stands for the second order term of the energy landscape.

Because the Hessian is a symmetric matrix, then follows orthogonal decomposition, $HE(\mathbf{m}_p) = \mathbf{V}\mathbf{D}\mathbf{V}^T$, where $\mathbf{V}$ is an orthogonal matrix whose columns form an orthogonal base of the protein space, and the eigenvalues of $\boldsymbol{D}$ are positive real numbers (due to the definite positive character of $HE(\mathbf{m}_p)$). Now calling $\Delta\mathbf{m} = \mathbf{m} - \mathbf{m}_p$, the hyper-quadric that approximates locally the nonlinear equivalent region of value E_{tol} can be written as follows:

$$\Delta\mathbf{m}^T HE(\mathbf{m}_p)\Delta\mathbf{m} \leq 2(E_{tol} - E(\mathbf{m}_p)), \quad (4)$$

that is,

$$\Delta\mathbf{m}_V^T D \Delta\mathbf{m}_V \leq 2(E_{tol} - E(\mathbf{m}_p)), \quad (5)$$

when the model increments are referred to the $\mathbf{V}$ base, that is, $\Delta\mathbf{m}_V = \mathbf{V}^T\Delta\mathbf{m}$. In the case where $HE(\mathbf{m}_p)$ is full rank, the bounding hyper-quadric in [5] can be written:

$$\sum_{k=1}^{n} \lambda_k \Delta\mathbf{m}_{Vk}^2 = (E_{tol} - E(\mathbf{m}_p)), \quad (6)$$

which is an ellipsoid centered in $\mathbf{m}_p$, whose principal directions coincide with the eigenvectors $\mathbf{v}_k$ (columns of $\mathbf{V}$), and the length axes are $1/\sqrt{\lambda_k}$, λ_k being the eigenvalues of D. In the case where $HE(\mathbf{m}_p)$ is semi-definite positive, that is, $HE(\mathbf{m}_p)$ has a nontrivial null space associated with the null eigenvalues, the hyper-quadric [6] becomes an elliptical cylinder. As explained above, the bounding hyper-quadric [6] only locally delimitates — in the neighborhood of $\mathbf{m}_p$ — the nonlinear equivalent region, that is, the protein models fulfilling the energy condition $E^{tol}(\mathbf{m}) \leq E_{tol}$.

Now, considering the same type of analysis in a model $\mathbf{m}_n$ is located in the neighborhood of $\mathbf{m}_p$ and belonging to the nonlinear stability region, we have:

$$\nabla E(\mathbf{m}_p)\cdot(\mathbf{m}-\mathbf{m}_n) + \frac{1}{2}(\mathbf{m}-\mathbf{m}_n)^T HE(\mathbf{m}_n)(\mathbf{m}-\mathbf{m}_n) \leq tol - E(\mathbf{m}_n). \quad (7)$$

Several remarks are important now:

1. The center of the new hyper-quadric does not coincide with the model $\mathbf{m}_n$ but with the Gauss-Newton solution of the nonlinear optimization problem solved in $\mathbf{m}_n$. Local optimization methods wander around different models of the nonlinear equivalent region searching for the global optimum. Unfortunately, these methods might not converge, and they do not keep track of the good protein models that have been visited during the optimization process.
2. The matrix $HE(\mathbf{m}_n)$ might lose its semi-definite positive character and the hyper-quadric becomes a hyperboloid. In this case the function landscape shows sill points that indicate the presence of different basins of equivalent protein models.
3. The main orientations of the local hyper-quadric change with the model that is considered, since $HE(\mathbf{m}_n)$ coincides with the hyper-quadric matrix. Thus, the nonlinear region of equivalent protein models has to exhibit a croissant or banana-shaped structure. As explained above, the energy function landscape could be multimodal depending on the energy function that is adopted. Also, these basins are elongated with almost null gradients. Further details can be found in Fernández Martínez et al. [19–22] for the case of linear and nonlinear inverse problems.

In tertiary protein prediction, the native structure is expected to correspond to the global optimum of the energy function. Nevertheless, this does not have to be the compulsory case since the perfect energy function is unknown and it is only a model of reality. Tyka et al. [1] studied the alternate states of several protein families via a detailed analysis of their energy landscape. They have shown that most of the energy landscapes have steep funnels down to low-energy minima close to the experimentally determined structure, but in some cases the lowest-energy structures had significant local deviations from the experimental structure. They conclude that the structure prediction accuracy is limited mainly by the ability to sample close to the native structure. As we will show, this could be alleviated by the use of model reduction techniques, since all these models should share similar patterns. Further work was carried out by Sander and Schneider [4]. They developed a database that increases the number of known protein structures. The outcomes are of significant relevance in evaluating the structural significance of variation in protein structures, in elucidating patterns for structure prediction, and, additionally, in modeling three-dimensional detail by homology.

Also, the use of model reduction techniques implies that we are not looking for the real native structure but for a good approximation to it. Thus, the hypothesis used in this paper is that the native structure will be located in the nonlinear equivalent region $M_{tol} = \{\mathbf{m} : E(\mathbf{m}) \leq E_{tol}\}$, for a given energy value, E_{tol}, close to the global optimum of the energy function, and the energy landscape will be much simpler since we are looking for the solutions in a suitable linear variety of the protein model space. In that way, several disconnected basins of solutions might be connected in the reduced space. Details about how to establish this energy cut-off value and how to perform the sampling on the reduced base are explained in the following sections.

Gniewek et al. studied the problem of the effect of noise in protein force fields on protein structure. [12]. Also Fernández-Martínez et al. [23, 24] have studied the effect of noise and that of the regularization in linear and nonlinear problems proving that noise perturbs the location of the global optimum that is found and the regularization techniques do not impede the existence of other equivalent models. In the case of the tertiary protein prediction problem there is no observed data, but the modeling errors induced by the energy model could be interpreted as noise in data and the effect would be similar, that is, the optimization will provide only an approximation to the native structure of the protein.

The method that is discussed in this paper corresponds to the category of decoys-based optimization (and refinement) and includes the following steps:

1. Finding and selecting known decoys from existing protein databases that are all possible solutions of the target structure. During this stage some protein decoys might be discarded on the basis of similarity and energy considerations.
2. Once the decoys have been selected, they are superimposed onto their centroid, which is calculated utilizing the SPICKER method developed by Zhang and Skolnick [25] to account for rotation and translation of the protein set before PCA calculation. Translation and

rotation removal is important to correctly compute the PCA base, and this is carried out by Bioshell's internal functions. Generally speaking, that transforms the system from global to local coordinates.

A local coordinate system is based on three points in 3D space and is defined as follows:

$$v_x = \frac{v_3 - v_2}{|v_3 - v_2|} \quad v_y = \frac{v_3 + v_2}{|v_3 + v_2|} \quad v_z = \frac{v_3 \times v_2}{|v_3 \times v_2|}$$

where |v| denotes the length of a vector v. Practically, Y axis lies on the bisector of an angle defined by $\boldsymbol{v_1}$, $\boldsymbol{v_2}$ and $\boldsymbol{v_3}$, which are the coordinates of the first, second, and third atom, respectively. Obviously, Y lies in the plane defined by $\boldsymbol{v_1}$, $\boldsymbol{v_2}$, and $\boldsymbol{v_3}$. Axis X is perpendicular to Y and also lies in the plane defined by $\boldsymbol{v_1}$, $\boldsymbol{v_2}$, and $\boldsymbol{v_3}$. Axis Z is a vector product of X and Y. The rotation center is placed at $\boldsymbol{v_2}$, considered also to be the translation vector [13, 14, 26].

Dimensionality reduction using the spatial principal component base is computed in the set of selected protein templates. This parameter reduction enables us to perform sampling on the reduced model space using particle swarm optimization (PSO) in this case.

3. Sampling while optimizing using the RR-PSO algorithm. [20]. Their convergence property is related to the first and second order stability of the particle trajectories. In this paper we use a cloud version where the RR-PSO parameters are automatically tuned. Also, non-elitism and the discretization time step are important features to increase exploration and perform a good approximate sampling of the energy function landscape in the areas of interest (low energies).
4. Posterior analysis to estimate the posterior distribution of the protein model parameters from the samples gathered on the nonlinear equivalent region.

This methodology has been successfully applied in high dimensional and very challenging oil reservoir optimization problems and combined with extreme learning machines for proteins secondary structure prediction ([7, 12, 27]). Here we demonstrate its application to the tertiary structure prediction problem.

Parameter reduction using the PCA protein base

Principal component analysis is a well-known mathematical procedure that transforms a number of correlated variables into a smaller number of uncorrelated variables called principal components [28]. The resulting transformation is such that the first principal component accounts for as much of the variability and each succeeding component accounts for as much of the remaining variability as possible [29].

This procedure has been applied in several fields and it is known under different terminologies, such as Karhunen-Loeve expansion, proper orthogonal decomposition or empirical orthogonal bases. In the case of the tertiary protein prediction a preliminary application has been done by Qian et al. [29]. The sampling strategy only used the three largest PCs, using simplex, Powell method, and exhaustive grid sampling. The PCs were established through the backbone structures within a homologous family to define a small number of preferable sampling directions that helped to refine the proteins model quality. In this paper we perform stochastic sampling in higher dimensions using regressive-regressive particle swarm optimization (RR-PSO). The number of PCA terms is automatically determined using covariance matrix energy considerations. Additionally, we show how to automatically determine the search space in which the sampling/optimization procedure will take place. The principal components that are selected are those that expand most prior model variability that is expected in the homologous family.

Model reduction is needed in high dimensional problems due to the following reasons:

1. The model parameters (protein coordinates in this case) should not be sampled independently, since model correlations exist and are introduced by the energy function to achieve low energy scores. The model reduction that is used in this paper is based on analyzing these correlations for a set of decoys that are used to predict the native structure.
2. Owing to the curse of dimensionality, that is, the probability of sampling in the interior of an n-sphere that is inscribed in an n-dimensional hyper prism approaches zero for $n > 10$ [30, 31]. This result also would suggest that the correct reduced basis should not have more than ten principal modes.
3. Because model reduction alleviates the ill-posed character of the tertiary structure optimization problem, since the solutions are found in a much lower dimensional space: finding

$$\boldsymbol{a_k} \in \boldsymbol{R^d} : \boldsymbol{E}(\widehat{\boldsymbol{m}}_{\boldsymbol{k}}) = \boldsymbol{E}(\boldsymbol{\mu} + \boldsymbol{V_d a_k}) \leq \boldsymbol{E_{tol}} \tag{8}$$

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

The PCA dimensionality reduction works as follows:

a) First, an ensemble of l plausible decoys $\mathbf{m}_i\, R^n$ are selected and arranged column wise into the decoys experimental matrix: $\mathbf{X} = (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_l)$ belongs to $M(n, l)$. The problem consists of finding a set of protein patterns $\boldsymbol{V} = (\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_d})$ that provide an accurate low dimensional representation of the original set with $d << l$.

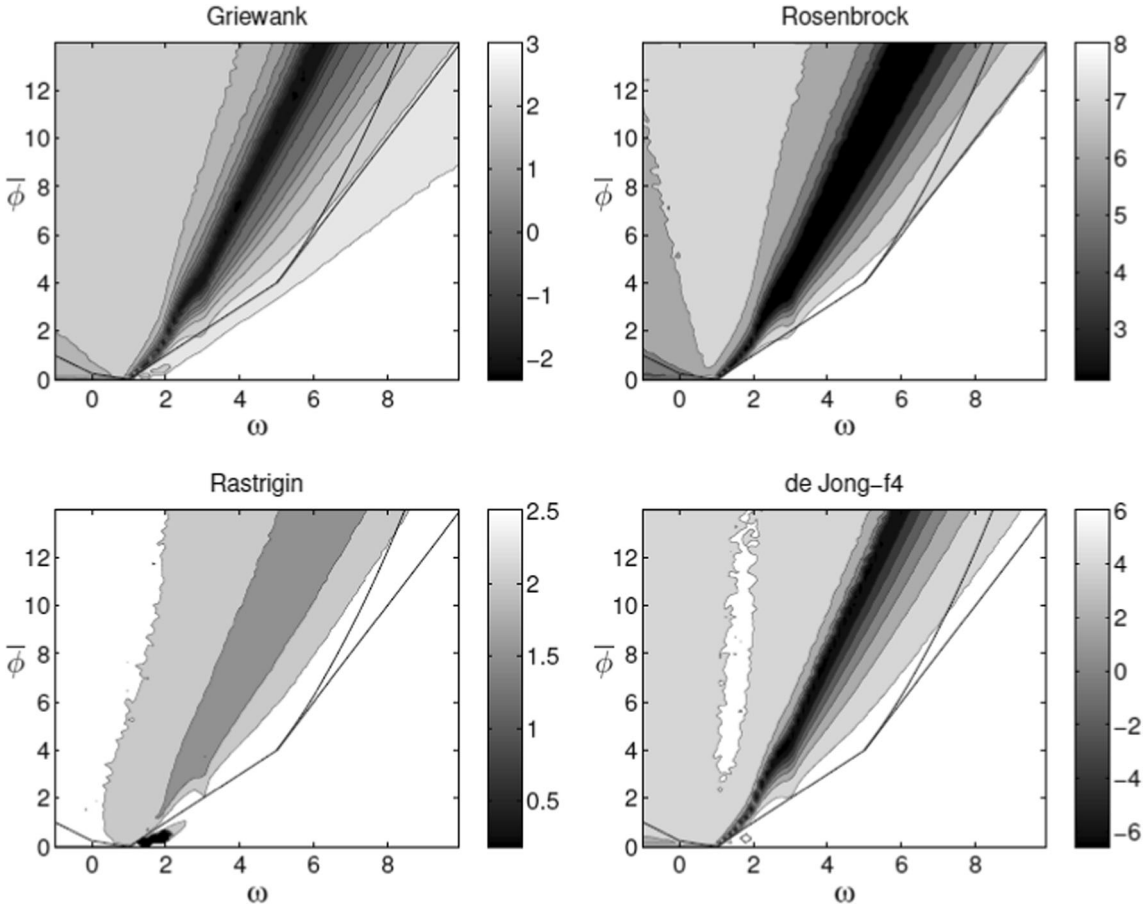

Fig. 1 RR-PSO: median misfit for different benchmark function in 30 dimensions having flat valleys and/or multimodality

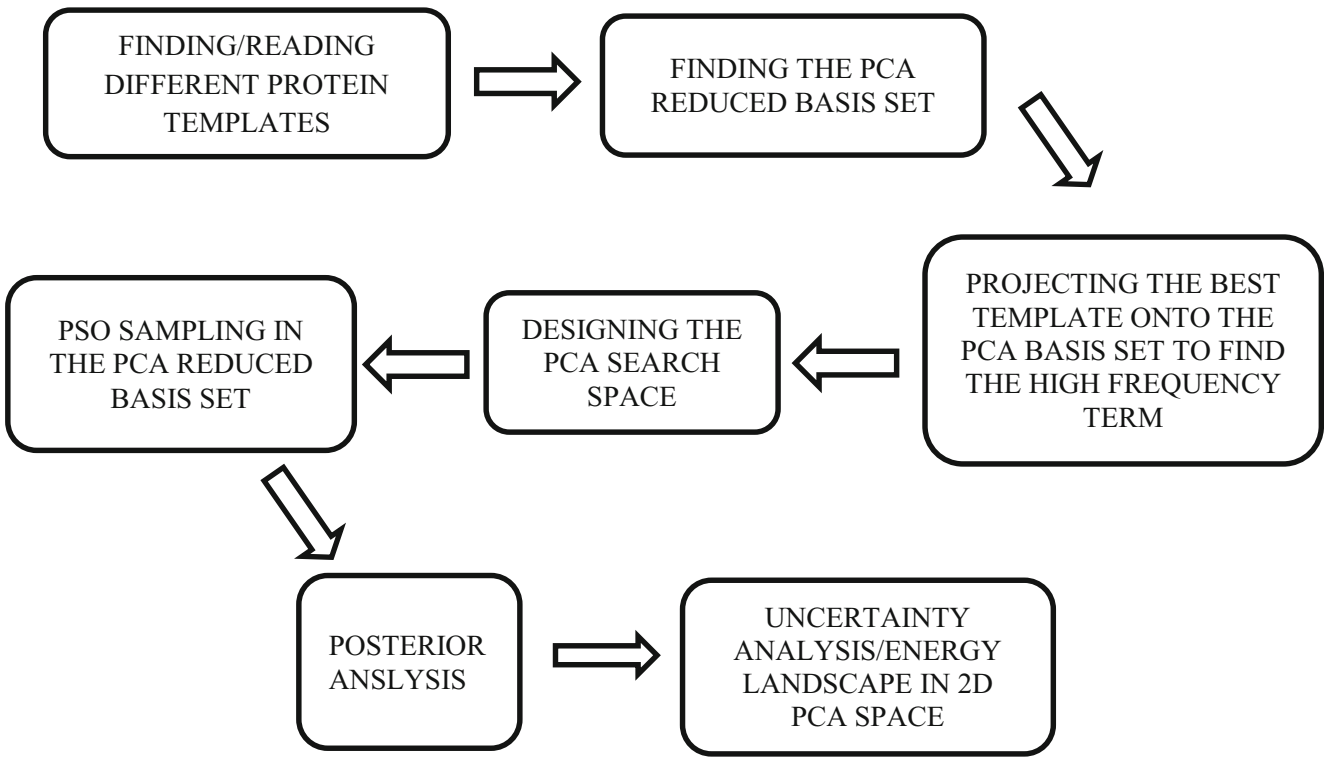

Fig. 2 Flow-chart of the methodology

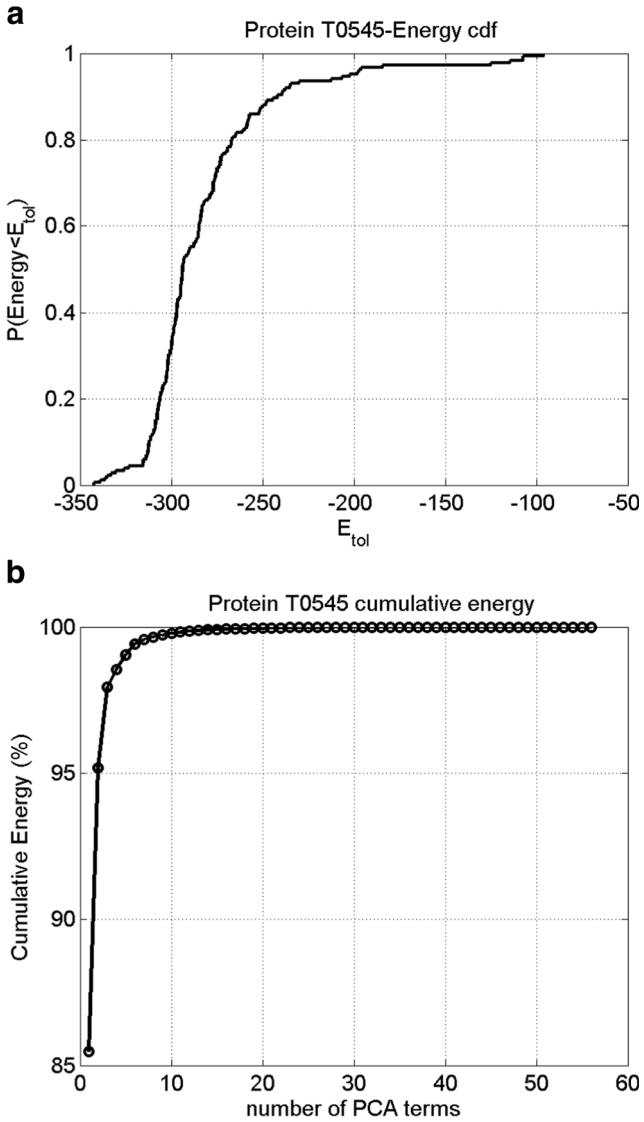

Fig. 3 Protein T0545. Step 1: Energy analysis of the decoys to produce the PCA basis and determination of the number of PCA basis terms based on energy considerations. a) Energy cdf of the 185 decoys downloaded from the CASP9 website. To generate the PCA reduced basis we considered 56 and 111 decoys with energy less than −280 and −300 that correspond to the 60th and 30th percetiles of the energy cdf distribution. b) Frobenius energy of the PCA decomposition. With the first PCA term we recover 86% of the cumulated energy from the decoys database. To perform the optimization we considered seven PCA terms and the high frequency term

PCA performs this task by diagonalizing the prior experimental centered covariance matrix:

$$C_{prior} = (\mathbf{X}-\mu)(\mathbf{X}-\mu)^T \in M(n,n), \tag{9}$$

where μ is either the experimental mean of the decoys, the medoid, or any other decoy around which we desire to perform the search of a backbone structure.

b) Matrix C_{prior} has a maximum rank of $l-1$, that is, at most $l-1$ eigenvectors of C_{prior} are needed to expand the whole prior variability. Therefore, 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:

$$\begin{aligned} &\mathbf{X}-\boldsymbol{\mu} = V\Sigma U^T, \\ &C_{prior}^T = U\Sigma\Sigma^T U^T \Rightarrow B = V\Sigma = (\mathbf{X}-\boldsymbol{\mu})U. \\ &\mathbf{v}_k = \frac{\mathbf{B}(:,k)}{\|\mathbf{B}(:,k)\|_2}, k = 1,\ldots,l-1. \end{aligned} \tag{10}$$

The centered character of the experimental covariance C_{prior} is crucial to maintaining consistency with the centroid model μ.

c) Ranking the eigenvalues of C_{prior}^T in decreasing order enables us to select a certain number of PCA terms ($d << l-1 << n$) to match most of the variability in the model ensemble. The automatic procedure is as follows: 1. calling $\lambda_k\,(k=1,\ldots,\,l-1)$ the non-null eigenvalues of C_{prior}, we define the cumulative energy of matrix C_{prior} as $E_{cumul}(j) = \sqrt{\sum_{k=1}^{j}\lambda_k / \sum_{k=1}^{l-1}\lambda_k}$, the number of PCA terms, d, is the minimum number of eigenvalues that we have to consider to fulfill the condition $E_{cumul}(d) \geq E_{por}$, where E_{por} is the percentage of the prior energy in the decoys that we want to keep during the search. Typically, we use $E_{por} = 99.5\%$ to conserve most of the high frequency content of the decoys in the reconstructed protein structures. The number of PCA terms depends on the prior decoys that are used, and on the experimental mean μ that is adopted. Finally, a high frequency term is added to the basis set considering the model with the lowest energy, and projecting it into the PCA basis as follows:

$$\mathbf{v}_{d+1} = \mathbf{m}_{BEST} - \boldsymbol{\mu} + \sum_{i=1}^{d} a_i \mathbf{v}_i. \tag{11}$$

Including the high frequency term is crucial for a successful protein model reconstruction in Cartesian coordinates after the PCA sampling. The combination of this high frequency term and the BioShell package forward calculations, included in the structural library (jbcl.calc.structural), makes the algorithm capable of reconstructing the protein without structural clashes. Otherwise, if this term is excluded, the algorithm yields to unrealistic and incomplete backbone structures [6].

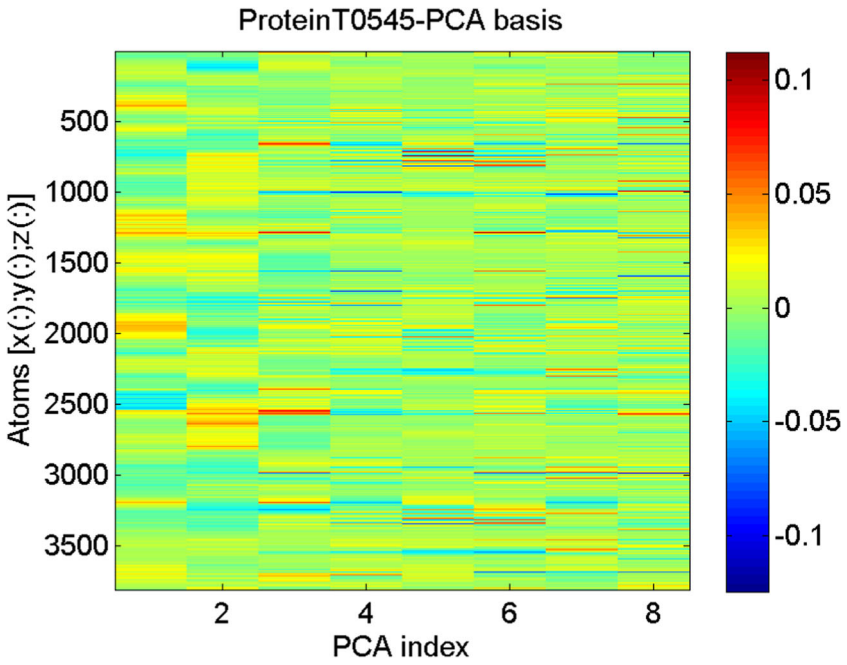

Fig. 4 Protein T0545. Step 2: PCA basis set construction. Unit basis vectors of the low dimensional subspace (dimension 8) of the original backbone structure where the PSO optimization takes place. The last term of the basis set gathers all the high frequency details needed to perform the tertiary structure protein refinement

d) Then, any protein model in the reduced base is represented as a unique linear combination of the eigenmodes:

$$\hat{\mathbf{m}}_k = \boldsymbol{\mu} + \sum_{i=1}^{d+1} a_i \mathbf{v}_i = \boldsymbol{\mu} + \mathbf{V}\mathbf{a}_k. \quad (12)$$

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

$$\mathbf{a}_k = \mathbf{V}^T(\hat{\mathbf{m}}_k - \boldsymbol{\mu}) \quad (13)$$

The model reduction allows global optimization methods to perform an efficient sampling in the reduced model protein space since the model parameters are not searched individually. The use of model reduction techniques serves to alleviate the ill-posed character of any highly underdetermined optimization problem and allows the optimization problem to be cast as a sampling problem.

The particle swarm optimizers

Particle swarm optimization is a stochastic evolutionary computation technique used in optimization, which is inspired in social behavior of individuals (called particles) in nature, such as bird flocking and fish schooling [32].

The sampling problem consists in finding enough representative samples of the model proteins $\hat{\mathbf{m}}_k = \boldsymbol{\mu} + \mathbf{V}\cdot\mathbf{a}_k$, such as $E(\hat{\mathbf{m}}_k) \leq E_{tol}$. Although the search is performed in the reduced PCA space, the sampled proteins have to be reconstructed in the original atom space to perform its fitness (energy) evaluation.

The algorithm consists of the following:

1) A prismatic space of admissible protein models $\mathbf{M}$, is defined

$$\mathbf{l}_j \leq \mathbf{a}_{ji} \leq \mathbf{u}_j, \quad 1 \leq j \leq n, \quad 1 \leq i \leq n_{size},$$

where l_j, u_j are the lower and upper limits for the jth coordinate for each geophysical model. In PSO terminology, each new plausible protein model will be called a particle, which is represented by a vector whose length is the number of PCA terms that are adopted. Each particle has its own position in the search space. The particle velocity represents the parameter perturbations in the PCA space needed for these particles to move around in the search space and explore solutions of the inverse problem. 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.

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

a The inertia term, which consists of the old velocity of the particle, $\mathbf{v}_i(k)$, weighted by a real constant, w, called inertia.

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

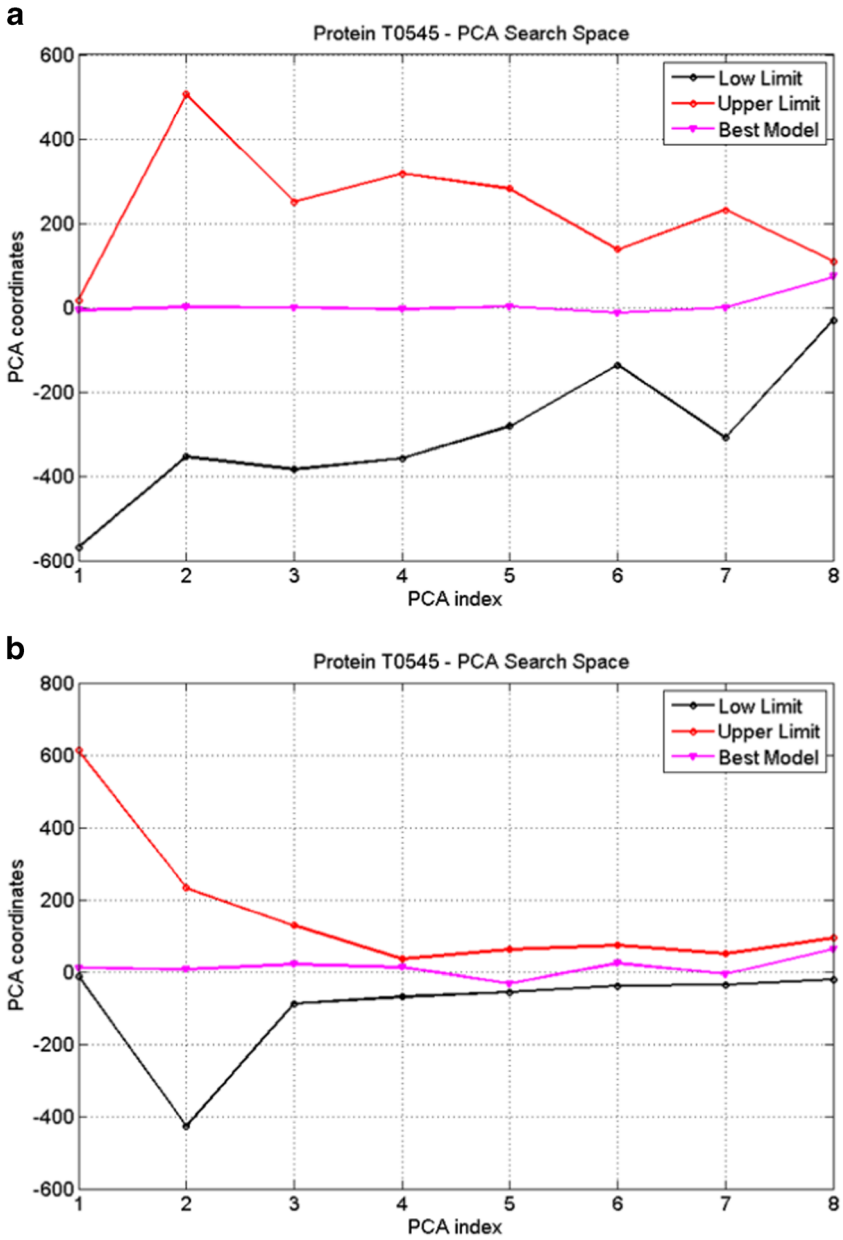

Fig. 5 Protein T0545. Step 3: Search space design. Low and upper limits of the search space and the best model from the CASP9 experiment projected onto this search space in (a) case 1 and (b) case 2

c The cognitive term, which is the difference between the particle's best position found so far (called $\mathbf{l}_i(k)$, the local best) and the particle's current position ($\mathbf{a}_i(k)$).

The PSO algorithm is as follows:

$$\begin{aligned} \mathbf{v}_i(k+1) &= \omega\mathbf{v}_i(k) + \phi_1(\mathbf{g}(k) - \mathbf{a}_i(k)) + \phi_2\left(\mathbf{l}_i^k - \mathbf{a}_i(k)\right) \\ \mathbf{a}_i(k+1) &= \mathbf{a}_i(k) + \mathbf{v}_i(k+1), \\ \phi_1 &= r_1 a_g, \;\; \phi_2 = r_2 a_l, \;\; r_1, r_2 \in U(0,1), \;\; \omega, a_g, a_l \in \mathbf{R}. \end{aligned} \tag{14}$$

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. $\overline{\phi} = \frac{a_g + a_l}{2}$ is the total mean acceleration and plays an important role on determining the algorithm's stability and convergence [33, 34].

The PSO algorithm can be physically interpreted as a particular discretization of a stochastic damped mass-spring system [19]. On the basis of this stochastic differential model, Fernández-Martínez and García-Gonzalo proposed a family of PSO members whose first and second order stability regions were analyzed [19, 20]. The stability regions of these algorithms can be defined in the space of $\omega - \overline{\phi}$, although the second order stability regions (controlling the exploration) also depend on the ratio of the local and global accelerations. This makes PSO a very singular algorithm with respect to other global optimization methods that are purely heuristic.

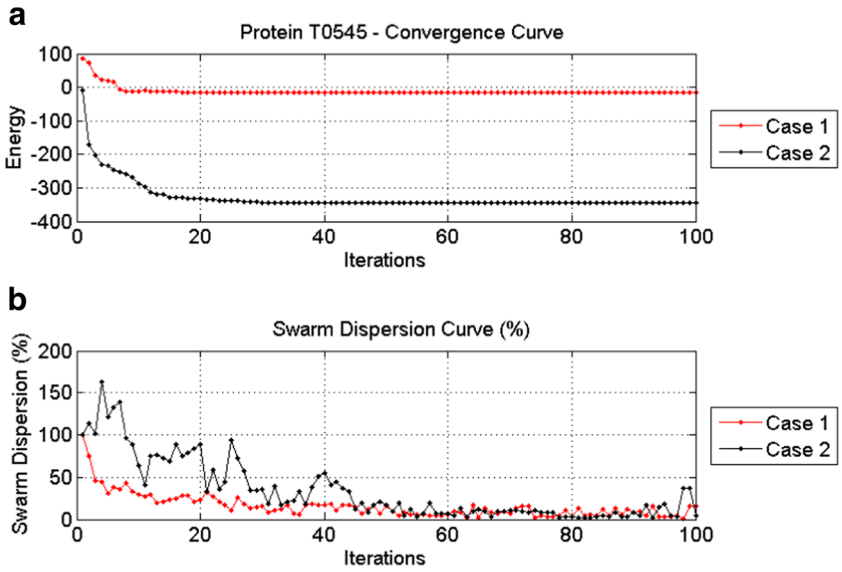

Fig. 6 T0545 protein. Step 4: PSO sampling. a) Convergence curve. b) Median dispersion curve (%)

Good PSO parameter sets are usually located in high explorative areas. In this paper we use the RR-PSO version due to its optimum balance between explorations. Also the good RR-PSO parameter sets can be analytically tuned.

Figure 1 shows, for the RR-PSO, the median misfit for different benchmark functions in 30 dimensions having flat valleys and/or multimodality. These graphics serve to understand where the performing RR-PSO parameters should be selected with respect to its respective first and second order stability regions. The good $(\omega, \overline{\phi})$ parameters sets for RR-PSO that provide the lowest misfits are concentrated around the line of equation $\overline{\phi} = 3\left(\omega - \frac{3}{2}\right)$ mainly for inertia values greater than 2. This line is the same for different flat valleys and multimodal functions and is invariant when the number of parameters increases. Therefore, this algorithm is very easy to implement.

Figure 2 shows a flowchart of the algortithm used to obtain numerical results presented in the next section.

Numerical results (CASP9 code T0545- MvR76 protein)

In this section we show the application of the PSO algorithm to the protein uracil DNA glycosylase from Methanosarcina acetivorans (CASP9 code T0545) whose native structure is known and reported by Aramini et al. at the Northeast Structural Genomics Consortium Target [35]. This native structure has been obtained through nuclear magnetic resonance spectroscopy (NMR) of proteins, which allows obtaining detailed information about the protein's structure, dynamics, and binding to DNA nucleotides.

Figure 3a shows the energy values of the 185 decoys that have been predicted by different teams. All the decoys have 1271 atoms corresponding to 185 residues. The lowest energy found was −342.15 and the highest −95.93. To generate the PCA reduced basis we considered two different cases. The purpose of this is to demonstrate the algorithm performance when low quality proteins are considered, and, furthermore, how the algorithm works with proteins with large differences in energy topographies. For case 1, we selected 111 decoys with an energy lower than −280 that corresponds to the 60th percentile of energy cumulative distribution function (cdf) values. For case 2, we utilized 56 decoys with an energy less than −300 that correspond to the 30th percetile of the energy cdf distribution. The selection of the decoys is important to drive the PSO search with the correct back-bone structure. Figure 3b shows the cumulative energy of the PCA decomposition. It is possible to observe that with the first PCA term we expand 86% of the energy of the decoys database (56 best decoys), and with seven terms we achieve 99.5% of the total energy. Therefore, to perform the optimization we considered seven PCA terms and the last high frequency term which is very important to ensure the good reconstruction of the protein in cartesian coordinates after the optimization. Otherwise, we would miss important details in the search space that would lead to structural clashes during the protein reconstruction.

Figure 4a and b show the unit basis vectors of the low dimensional subspace (dimension 8) of the original backbone structure where the PSO optimization takes place for both cases. The last term of the basis set gathers all the high frequency details needed to perform the tertiary structure protein refinement.

Figure 5 shows the search space used to perform the PSO search. The search space was determined projecting the 56 selected decoys into the PCA basis set and finding the minimum and maximum coordinates. The width of the first PCA coordinate interval is observed to be larger, and this interval narrows with the PCA index.

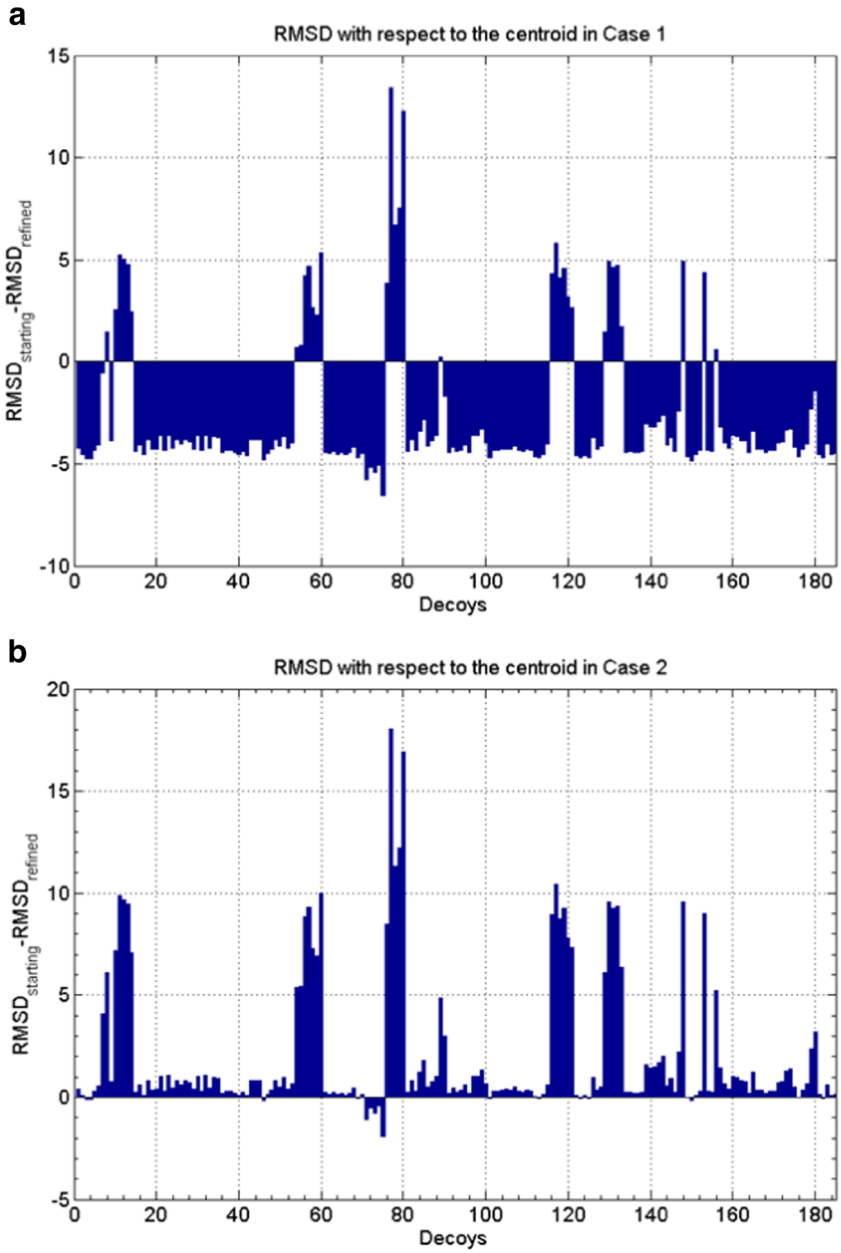

Fig. 7 Protein T054. Step 5: Posterior analysis. RMSD improvement for a) case 1 and b) case 2. The improvement of RMSD is measured by computing the difference between the RMSD the of starting models and the RMSD of the refined model, so that positive values indicate an improvement of backbone conformations

To perform the PSO sampling we adopted a swarm composed of 40 particles and 100 iterations. We used the RR-PSO family member, whose exploration capabilities of the algorithm were monitored in order to assure a good exploration of the PCA search space. For that purpose, we defined for any iteration the median distance between all the particles of the swarm and the center of gravity. This distance was then normalized by the dispersion in the first iteration (random sampling), that is considered to be 100%. When the median dispersion is lower than 3% the swarm has collapsed toward the global best, and either we can stop the sampling, or we can increase the exploration using time 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.

Figure 6 shows, for protein T0545, the convergence rate and the dispersion for the cases studied. In case 1, the algorithm starts with an initial energy around 85 and, after 30 iterations, it reaches a plateau around −16. The minimum energy achived was −16.9, which is 91.5% higher than the best protein model found in the CASP9 experiment. In case 2, the algorithm begins with an energy of −9 and in iteration 12 reaches the region of energy lower than −300 where the decoys where selected. The minimum energy reached was −344.6, which is 0.7% lower than the energy of the best protein model in the CASP9 experiment. From iteration 27 till the end, RR-PSO samples the nonlinear equivalence region (Fig. 5A). The dispersion remains greater than 5% till iteration 53, and only

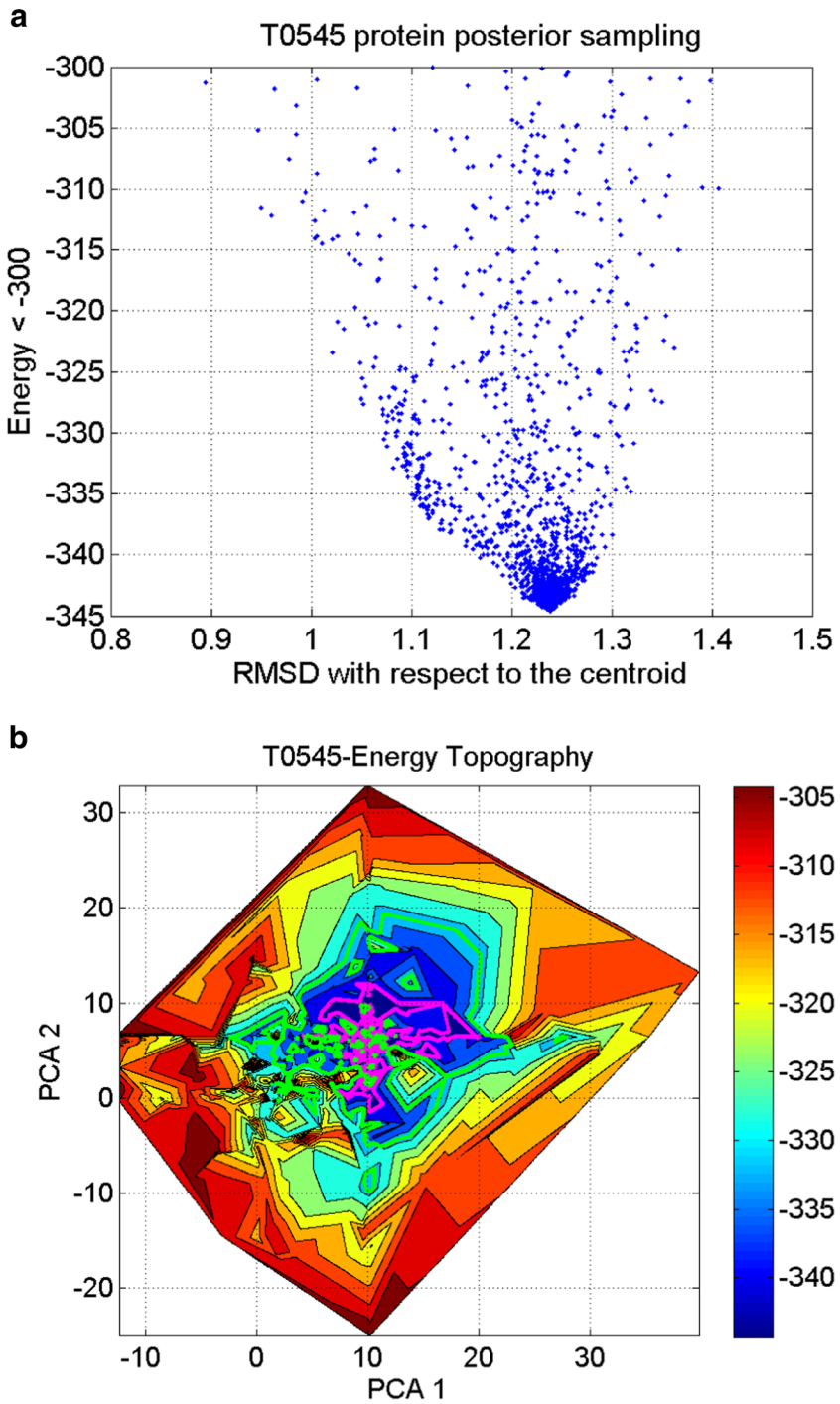

Fig. 8 T0545 protein. Step 5: Posterior analysis. a) Energy plot with respect to the root mean squared deviation (RMSD) for the sampled decoys with energy lower than −100 in case 1 and −300 in case 2. The minimum region is achieved for distances between 1.1 and 1.32. c) Energy zoom in the region of minimum misfit. The optimum is marked with an asterisk

in nine iterations (55, 63, 78, 81, 82, 83, 84, 89, 93) is smaller than 3%.

The refinement in the backbone structure achieved via RR-PSO is shown in Fig. 7. We show the improvement of the root mean squared distance (RMSD) as the RMSD of the starting models minus the RMSD of the refined model. In this sense, positive values would indicate an improvement in the backbone structure.

Figure 7a shows the RMSD improvement for case 1, where decoys with higher energies and different topologies have been included. In this sense, the algorithm is not capable of improving the prediction for the majority of the decoys

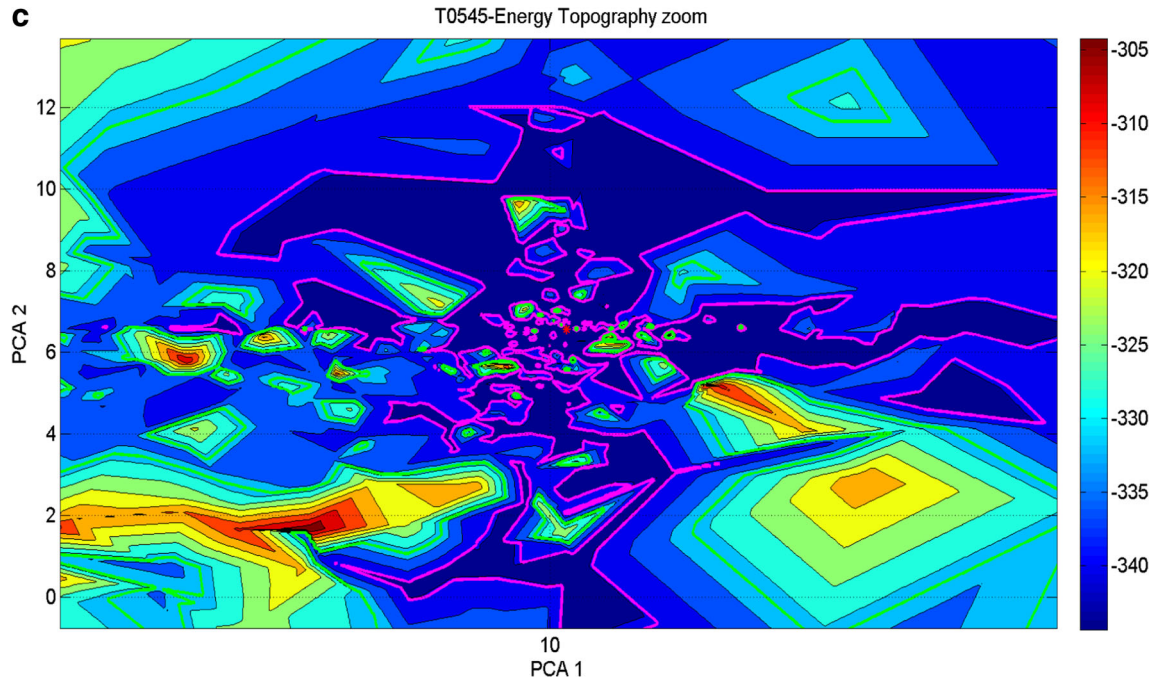

Fig. 8 (continued)

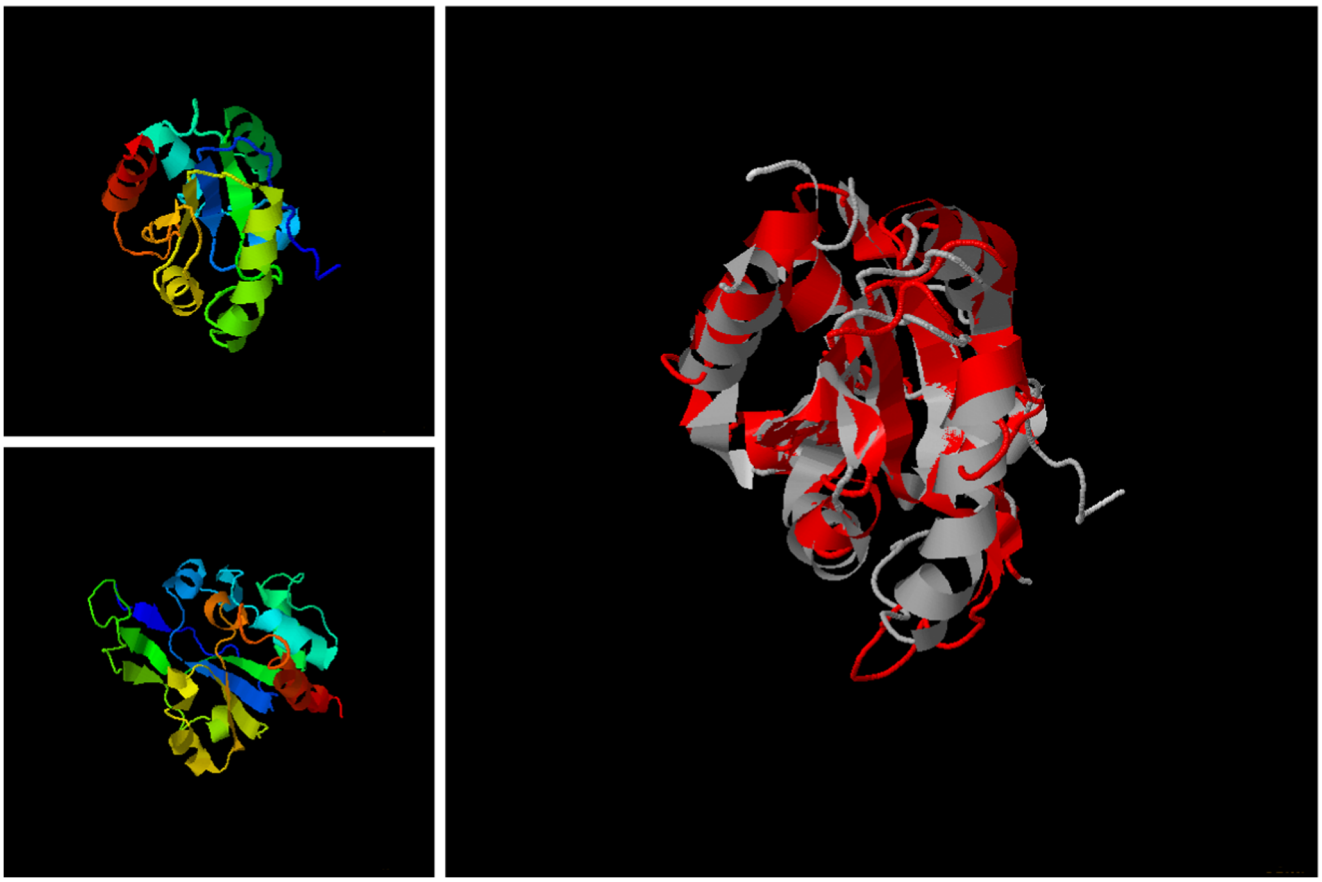

Fig. 9 T0545 protein stucture. Step 5: Posterior analysis. (left top) T0545 global optimal structure obtained via particle swarm optimization. (left bottom) T0545 protein native structure obtained through NMR and reported by Aramini et al. at the Northeast Structural Genomics Consortium Target. (right) Superimposed structure of predicted over the native structure

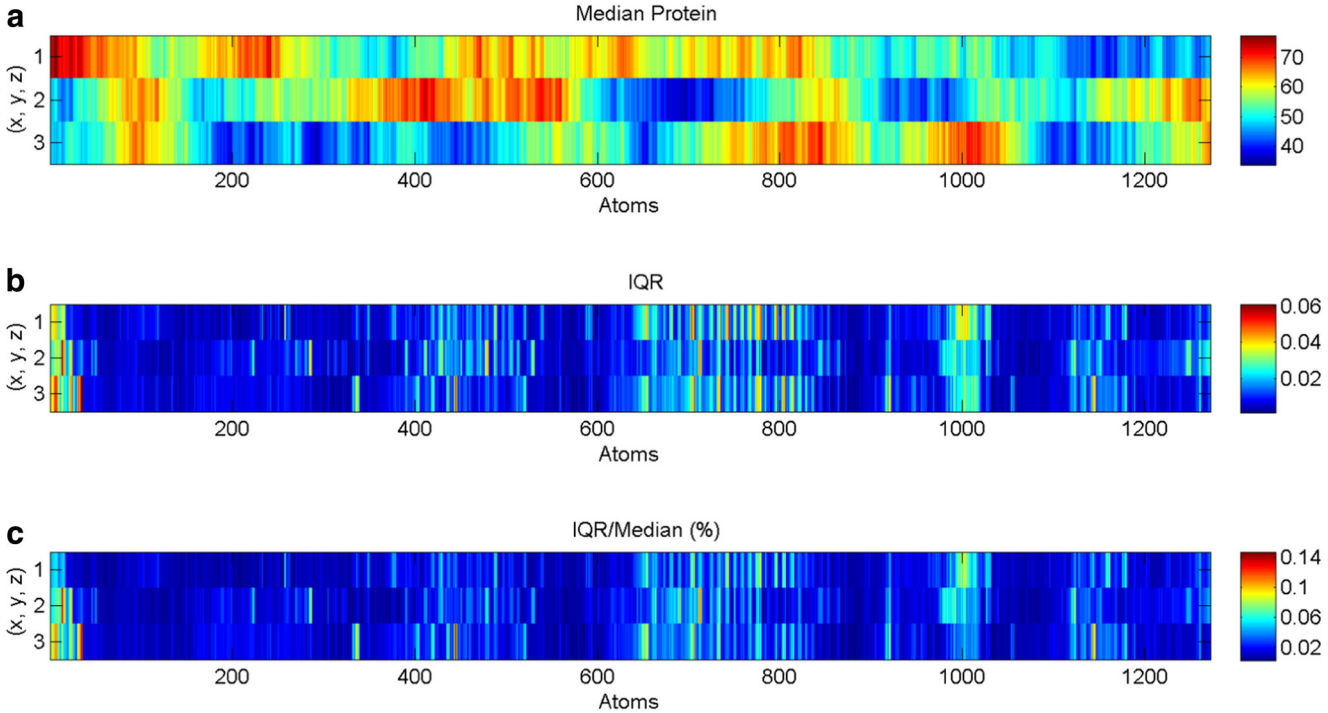

Fig. 10 T0545 protein stucture. Step 5: Posterior analysis. Uncertainty analysis in the region of energy lower than −300. a) Median protein of the decoys sampled in the region of energy lower than −330. b) Median protein plus the interquartile range of the coordinates of these decoys. c) Median protein minus the interquartile range of the coordinates of these decoys

analyzed. On the other hand, Fig. 7b shows the refinement achieved when a more strict selection is carried out (case 2). In this case, how the RR-PSO is capable of further improving the structures is observed. However, if high performance computational facilities are utilized, the algorithm may be capable of refining all the structures in both cases, that is, utilizing higher swarms and carrying out more iterations.

In order to give further nuance about the algorithm explorative character, we present, in Fig. 8, the protein conformational space sampling for the best case (case 2). Figure 8a shows the root mean squared distance between the models that have been sampled in the region of energy lower than −300. The minimum is achieved for an RMS distance with respect to the centroid of the prior decoys of 1.24 distance units. An almost symmetrical behavior with respect to this point between 0.9 and 1.4 can be observed. A high number of decoys have been sampled in the M_{tol} energy region of $E_{tol} = -335$ with RMSDs between 1.1 and 1.3. This is an illustration of the complex landscape of the energy function. To further clarify this fact, Fig. 8b shows the interporlated energy function in the 2D-PCA space. We also show the equivalent regions for $E_{tol} = -330$ (green isoline) and $E_{tol} = -340$ (magenta isoline) showing a complex topography with isolated basins for $E_{tol} = -330$. We also show a zoom of the topography landscape in the neighborhood of the minimum that have been found (Fig. 8c). It can be observed the complex topography of the cost function.

Finally we show the best configuration (lower energy) that has been obtained by RR-PSO (Fig. 9b) compared to the protein native structure obtained through NMR (Fig. 9a).

Figure 10 shows the results of the posterior analysis in the region of energy lower than −300. Figure 10a shows the median coordinates of the sampled protein decoys that fulfill this energy condition. In this case we show the protein as a matrix with rows *x, y*, and *z* and the columns are the number of the atoms of the protein. This graphic allows better visualization of the protein coordinates uncertainty. Figure 10b and c show the interquartile range (IQR) of the coordinates of these models and the IQR vs median ratio. The biggest variations in the coordinates (Fig. 10b and c) occur in the right border and in the middle of the protein. The maximum IQR/median ratio is 0.14%, that is, the distance between all the equivalent configurations is not very big. This is a confirmation of the ill-conditioned character of the tertiary structure prediction optimization problem.

We performed computations for nine additional proteins from CASP9 experiments. Detailed results are shown in Appendix A.

Table 1 presents the summary of computations carried out for ten different proteins detailing the energy of the best model in the CASP9 experiment and the energy achieved with PSO. Following this table, we present computed convergence curves, protein structures, and the corresponding uncertainties, as they are the key parameters to understand the protein backbone structure and the

Table 1 Details of the computational experiments performed with the methodology presented in this paper, via principal component analysis and particle swarm optimization, with the corresponding p value information to test the statistically significance between the energy obtained by PSO and initial energy

Protein CASP9 code	Number of residues	Number of PCA terms	Percentile of decoys	Number of iterations	Swarm size	Energy of best decoy	Energy obtained through PSO	p- value
T0545	166	9	20	100	40	−342.1	−344.6	$2.84\ 10^{-32}$
T0551	82	9	10	100	70	−161.9	−162.3	$1.44\ 10^{-34}$
T0555	155	9	5	100	80	−369.9	−371.4	$8.80\ 10^{-32}$
T0557	153	9	5	100	80	−273.7	−277.2	$6.03\ 10^{-32}$
T0561	170	9	10	100	60	−448.6	−450.3	$1.87\ 10^{-31}$
T0580	108	9	10	100	60	−253.6	−249.5	$6.50\ 10^{-34}$
T0635	191	9	5	100	70	−464.4	−465.5	$1.28\ 10^{-31}$
T0637	240	9	10	100	70	−369.2	−372.0	$9.48\ 10^{-34}$
T0639	128	9	7	100	70	−343.6	−343.7	$1.02\ 10^{-35}$
T0643	82	9	7	100	70	−209.4	−210.0	$3.98\ 10^{-31}$

alternative states. We also report the p value infomation to analyze the statistically significance (Wilcoxon signed rank test) between the energy obtained by PSO and the initial energy of the optimization procedure. Table 2 shows RMSDs and TM-scores of predicted structures.

Conclusions

In this paper, an algorithm that corresponds to the category of decoys-based modeling is presented. The algoritm successfully establishes a low dimensional space based on a priori energy considerations in order to apply an energy optimization procedure throughout a family of particle swarm optimizers. This optimizer is capable of modeling the protein sequence and sample the decoys of the predictions used to find a global optimum that satisfies the energy target.

Table 2 RMSDs and TM-scores of predicted structures via principal component analysis and particle swarm optimization

Protein CASP9 code	Number of residues	RMSD	TM-score
T0545	166	1.923	0.7597
T0551	82	4.256	0.4485
T0555	155	8.566	0.2665
T0557	153	1.596	0.7624
T0561	170	5.899	0.5500
T0580	108	1.303	0.7851
T0635	191	6.388	0.5234
T0637	240	4.966	0.4550
T0639	128	8.967	0.2135
T0643	82	3.728	0.6892

The nonlinear equivalence region corresponding to proteins that have energy lower than a certain energy bound was also sampled. This equivalence region was used to understand the backbone structure of the native structure and the alternative states of the protein according to the protein predictions obtained from the CASP9 experiment. Using this method concludes its key advantages of fastness and explorative character that greatly aid the refinement of the tertiary protein structure. Finally, this paper explains how the model reduction technique serves to alleviate the ill-posed character of this high-dimensional optimization problem.

Acknowledgments A. K. acknowledges financial support from NSF grant DBI 1661391, NIH grants R01 GM127701 and R01 GM127701-01S1, and from Bridge funding from The Research Institute at Nationwide Children's Hospital. We also acknowledge Ms. Celia Fernández-Brillet for her help in revising this paper.

Appendix. Supporting Information

In this section, we aim to expand the paper benchmark by presenting results for an additional set of nine proteins from CASP9. We tested our methodology in order to prove its suitability for protein refinement purposes.

T0551 – X-ray crystal structure of protein SP_0782 (7-79) from Streptococcus pneumoniae. Northeast Structural Genomics Consortium Target SpR104

We present the numerical results of the application of PSO in order to obtain the tertiary structure of protein SP_0782 (7-79) from Streptococcus pneumoniae., whose native structure has been obtained through X-ray by Kuzin et al. [36].

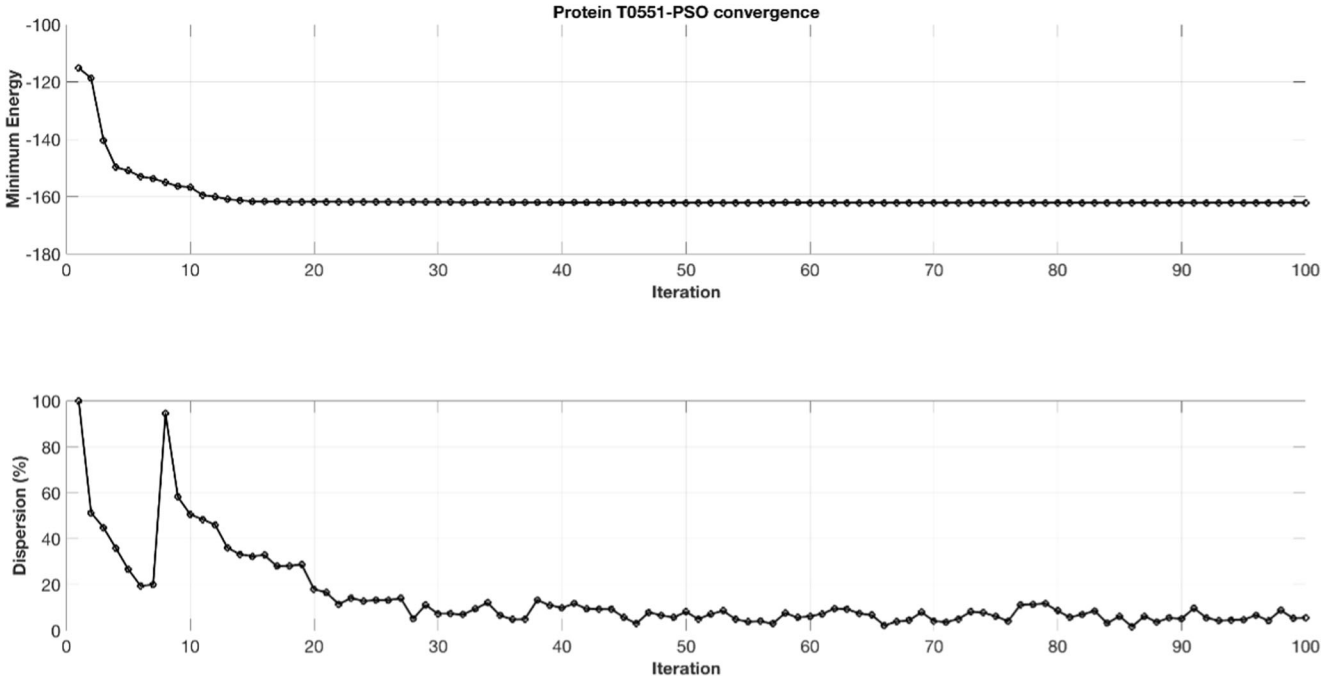

Fig. 11 T0551 protein. a) Convergence curve. b) Median dispersion curve (%)

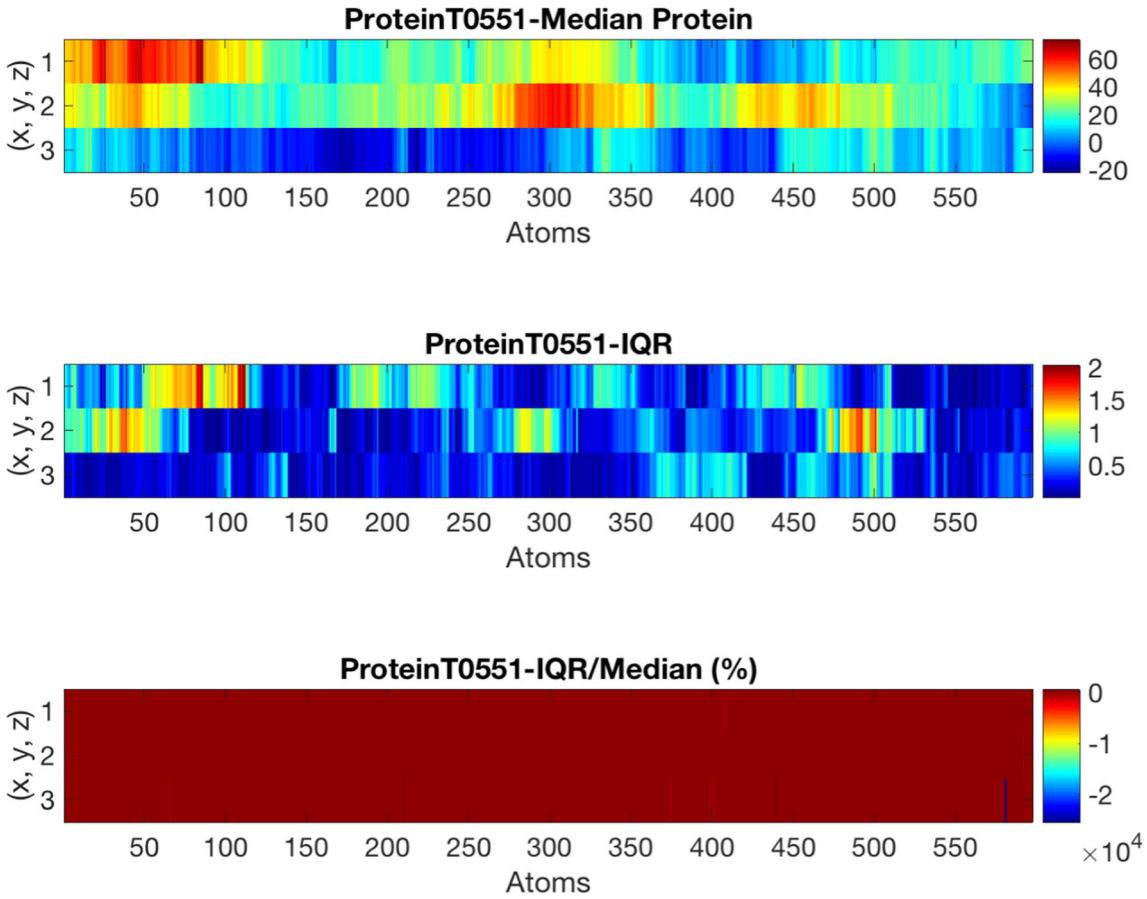

Fig. 12 T0551 posterior sampling in the region of energy lower than –200. a) Median protein of the decoys sampled in the region of energy corresponding to the 10th percentile. b) Median protein plus the interquartile range of the coordinates of these decoys. c) Median protein minus the interquartile range of the coordinates of these decoys

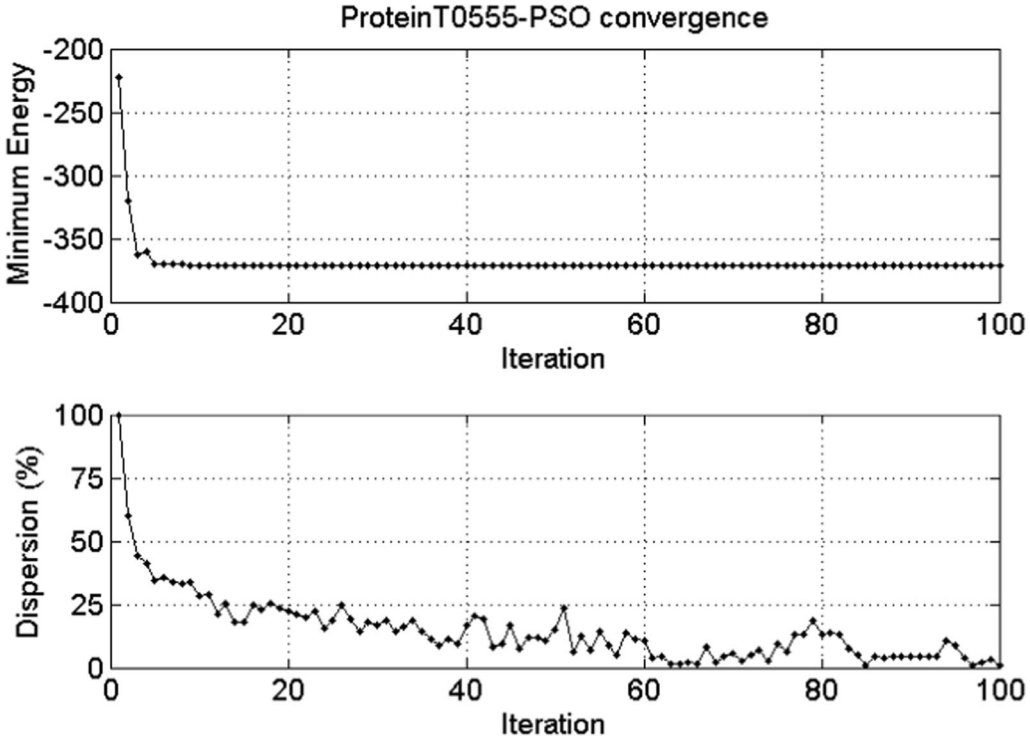

Fig. 13 T0555 protein. a) Convergence curve. b) Median dispersion curve (%)

Owing to the complexities experienced in performing the optimization of the T0551 structure, a swarm composed of 70 particles was applied. Additionally, the tenth percentile of the best templates was chosen. By

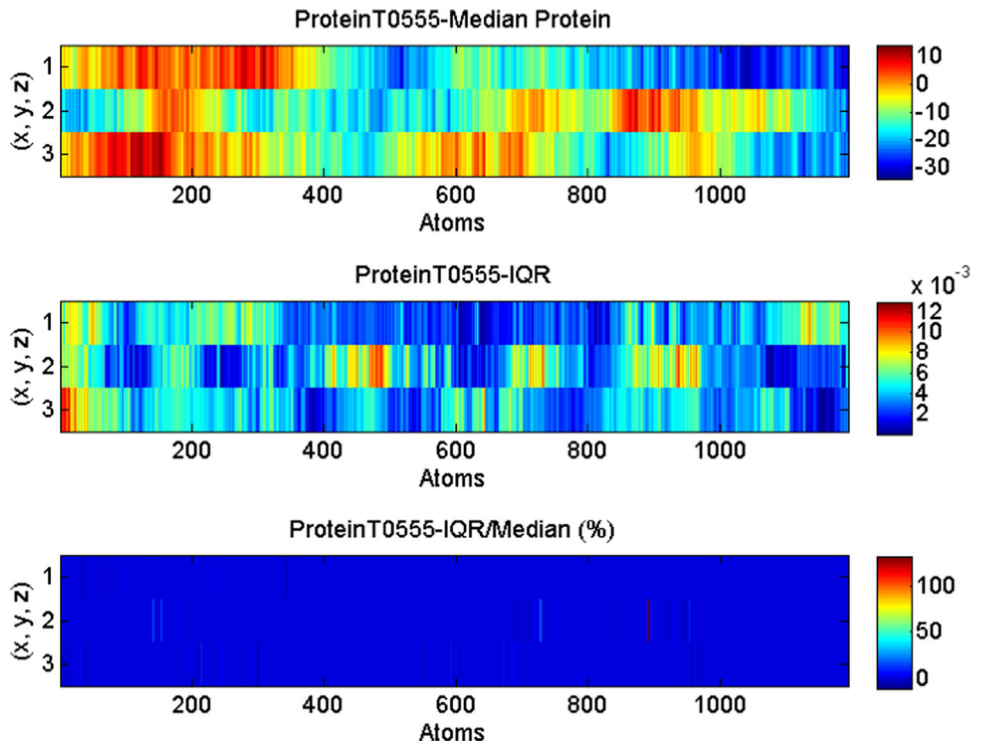

Fig. 14 T0555 posterior sampling in the region of energy lower than –200. a) Median protein of the decoys sampled in the region of energy corresponding to the 5th percentile. b) Median protein plus the interquartile range of the coordinates of these decoys. c) Median protein minus the interquartile range of the coordinates of these decoys

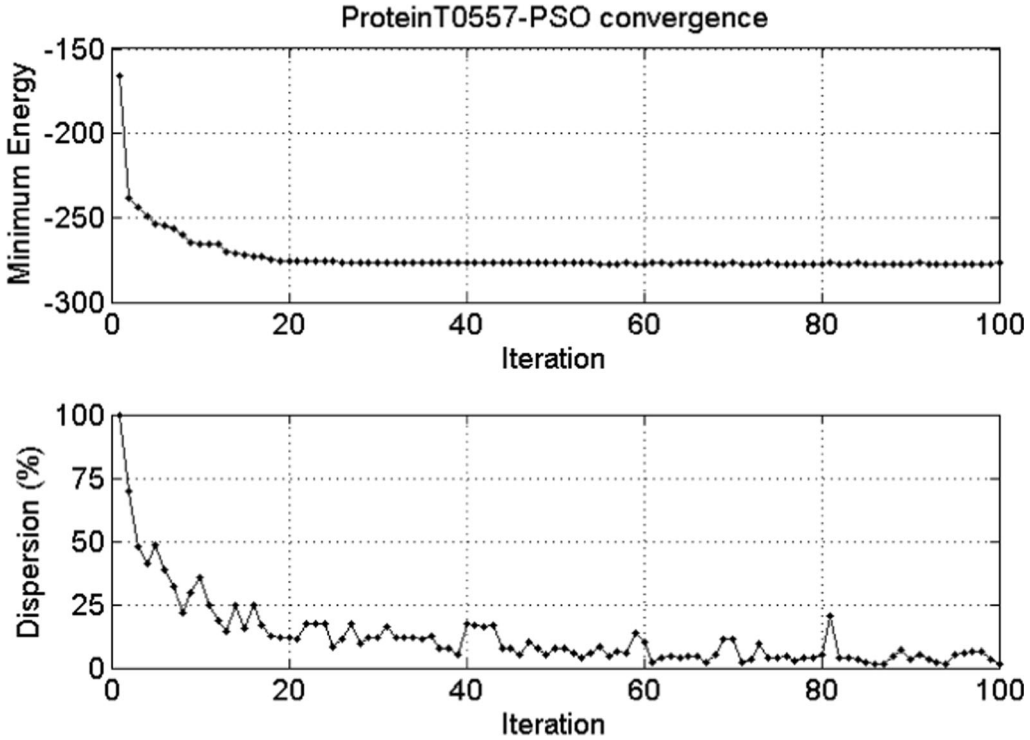

Fig. 15 T0557 protein. a) Convergence curve. b) Median dispersion curve (%)

taking into account these considerations, we ensure a good convergence and protein refinement, while also carrying out a wide sampling over a Search Space constructed with good a priori models.

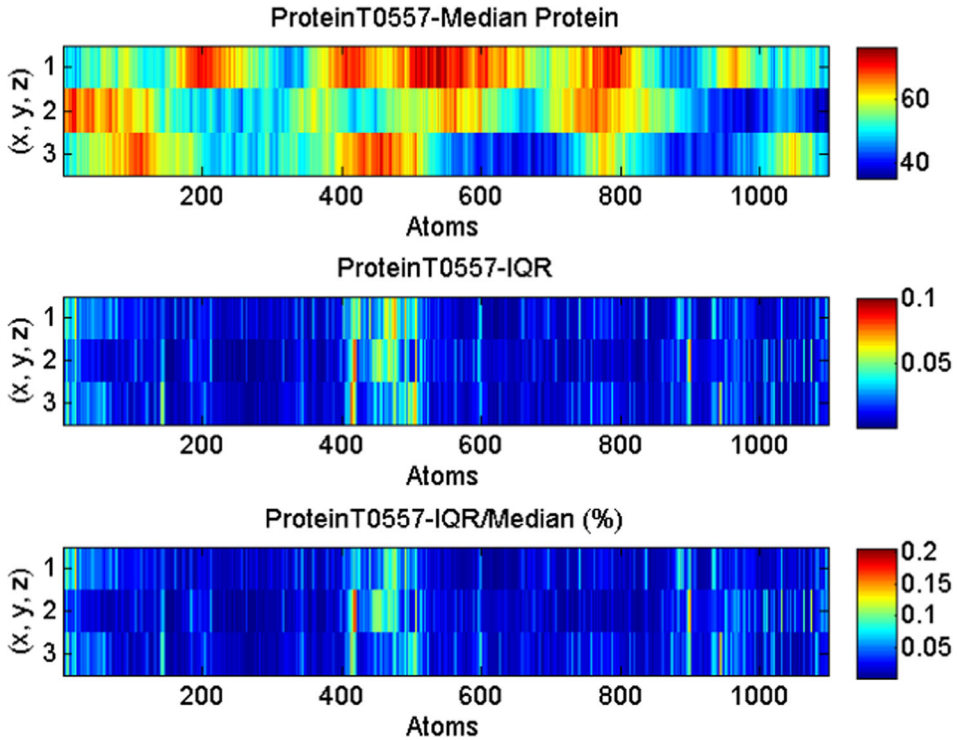

Fig. 16 T0557 posterior sampling in the region of energy lower than –150. a) Median protein of the decoys sampled in the region of energy corresponding to the 5th percentile. b) Median protein plus the interquartile range of the coordinates of these decoys. c) Median protein minus the interquartile range of the coordinates of these decoys

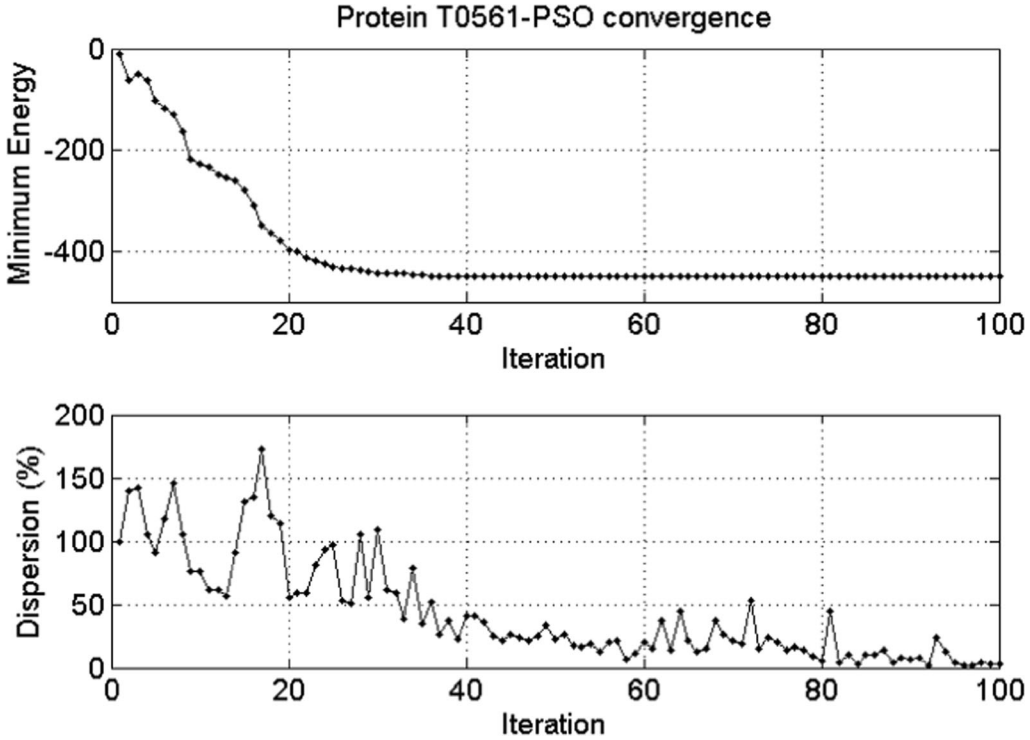

Fig. 17 T0561 protein. a) Convergence curve. b) Median dispersion curve (%)

As observed in Fig. 11, the energy converges in the first 20 iterations and achieves energy of –162.3. Because the majority of the models fluctuate around this energy, we obtain a protein with a low uncertainty as shown in Fig. 12.

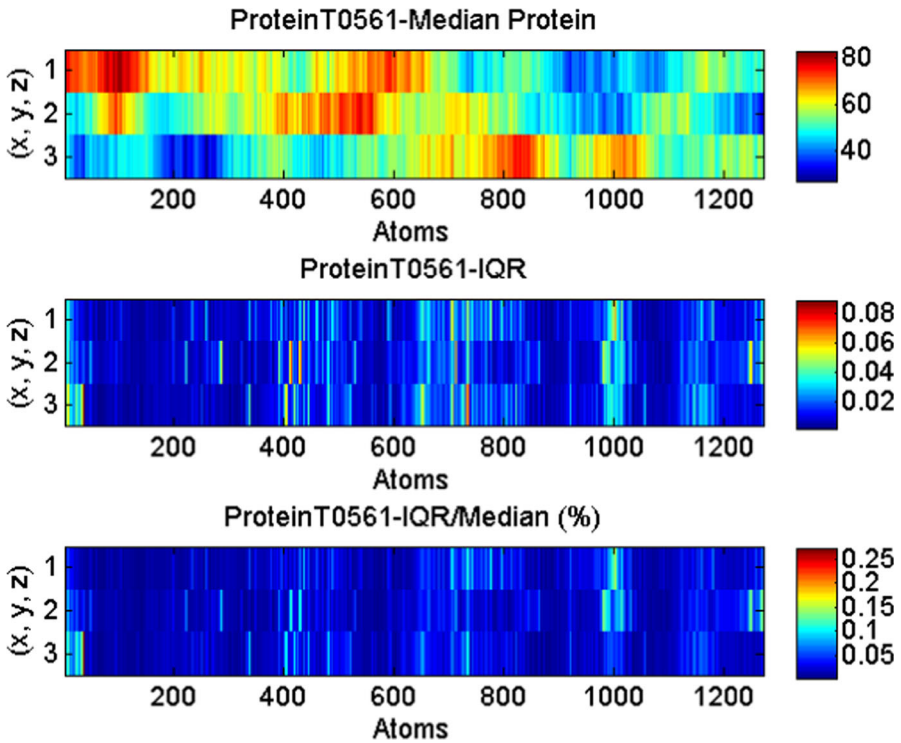

Fig. 18 T0561 posterior sampling in the region of energy lower than 0. a) Median protein of the decoys sampled in the region of energy corresponding to the 10th percentile. b) Median protein plus the interquartile range of the coordinates of these decoys. c) Median protein minus the interquartile range of the coordinates of these decoys

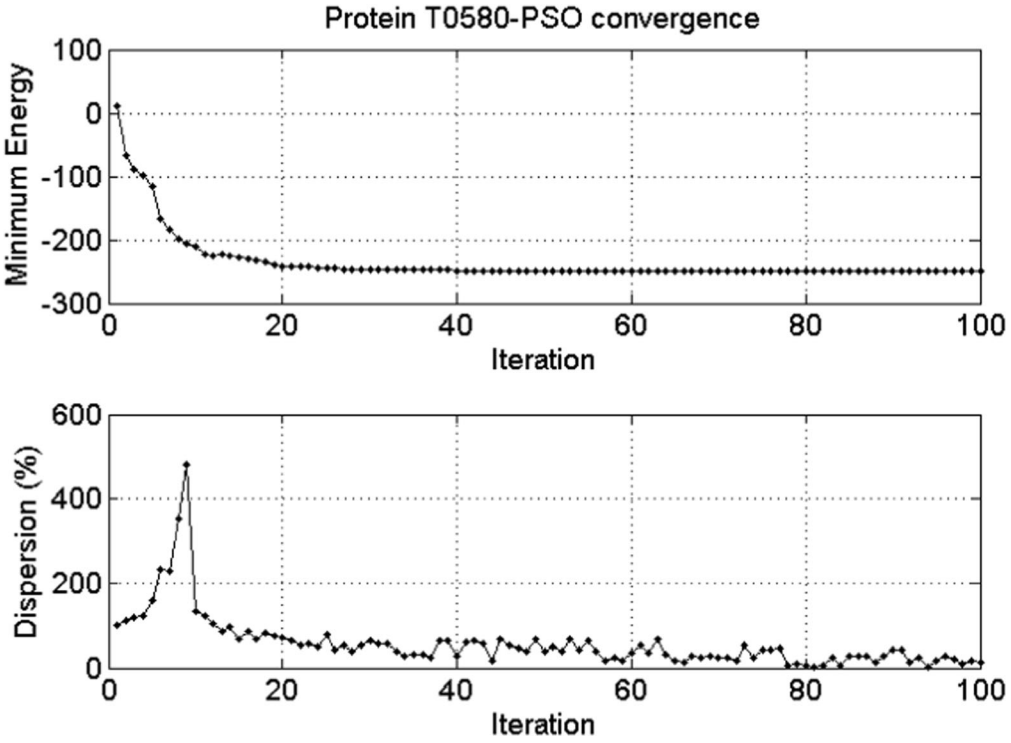

Fig. 19 T0580 protein. a) Convergence curve. b) Median dispersion curve (%)

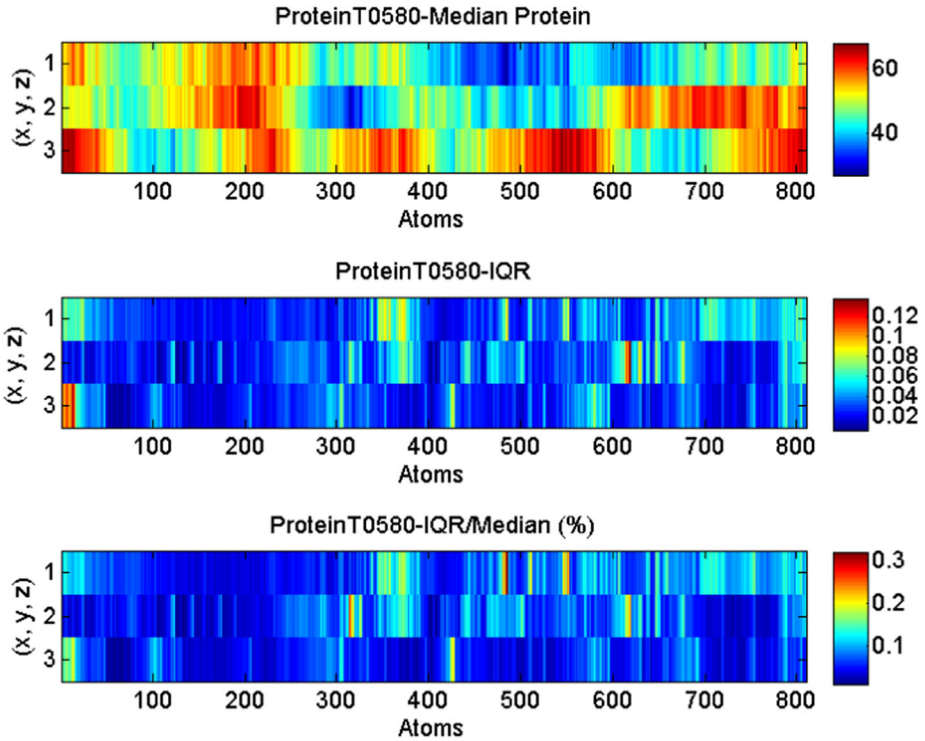

Fig. 20 T0580 posterior sampling in the region of energy lower than 0. a) Median protein of the decoys sampled in the region of energy corresponding to the 10th percentile. b) Median protein plus the interquartile range of the coordinates of these decoys. c) Median protein minus the interquartile range of the coordinates of these decoys

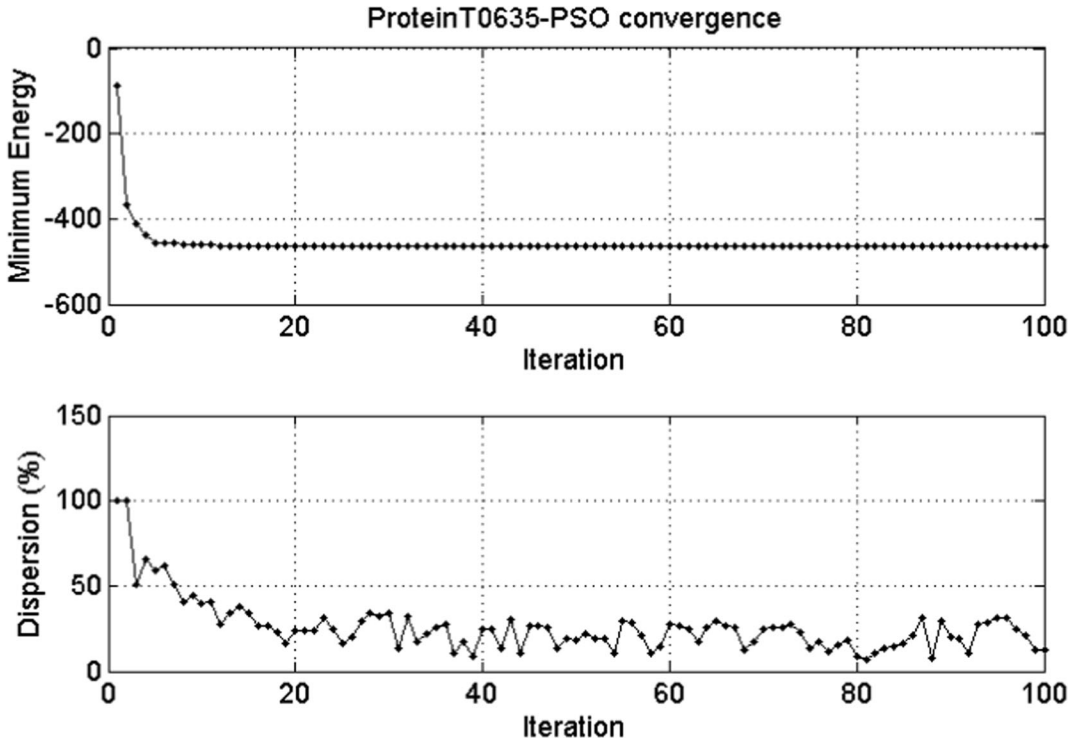

Fig. 21 T0635 protein. a) Convergence curve. b) Median dispersion curve (%)

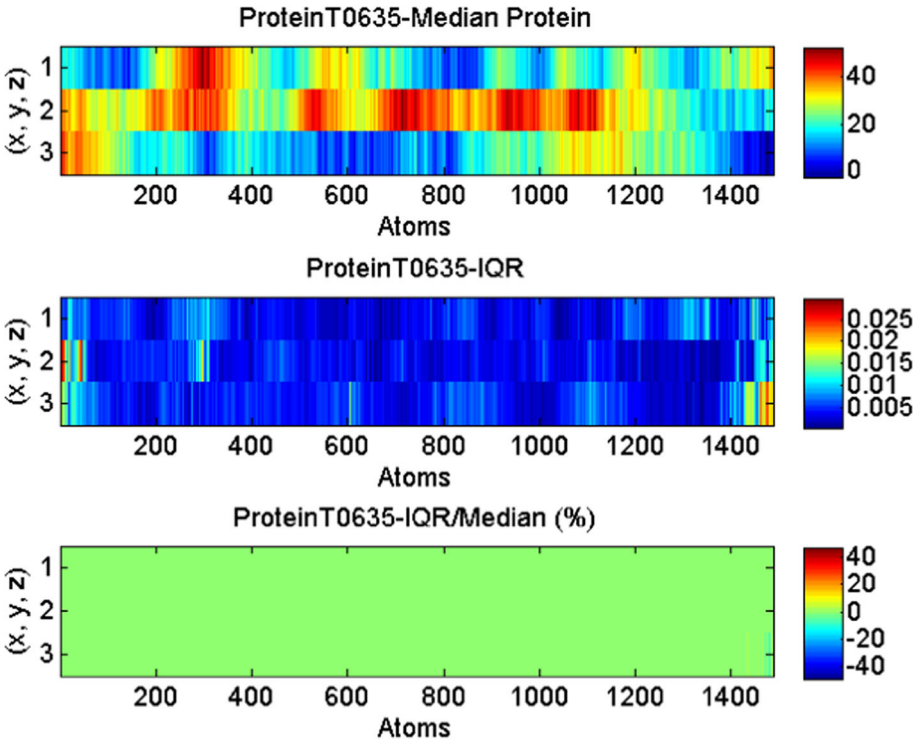

Fig. 22 T0635 posterior sampling in the region of energy lower than 0. a) Median protein of the decoys sampled. b) Median protein plus the interquartile range of the coordinates of these decoys. c) Median protein minus the interquartile range of the coordinates of these decoys

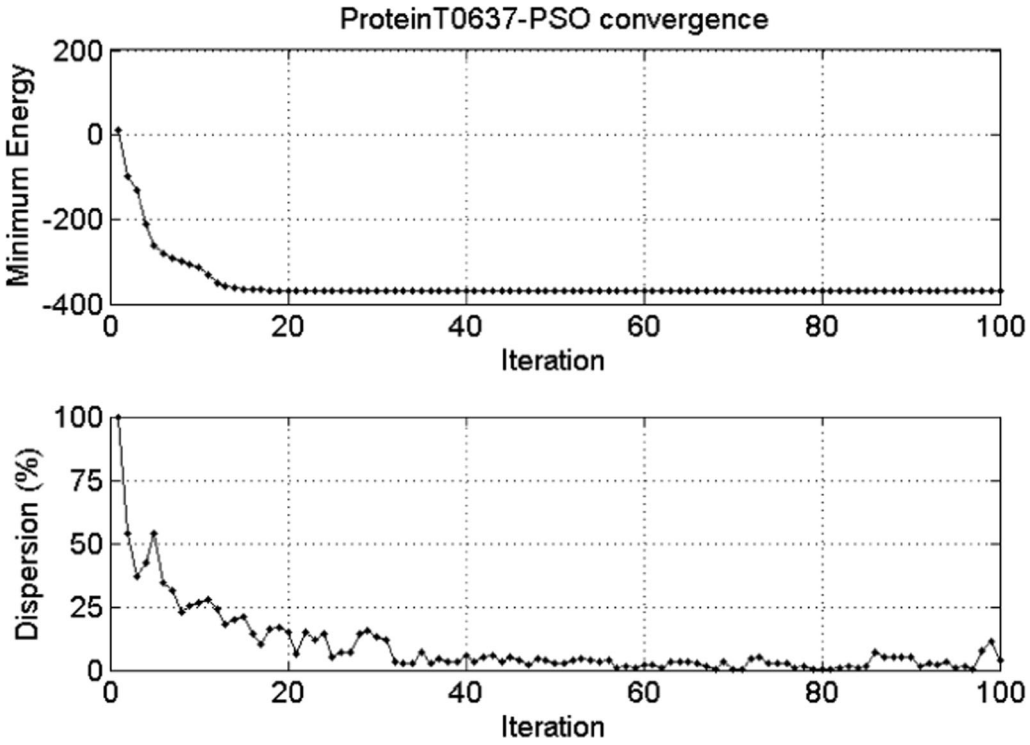

Fig. 23 T0637 protein. a) Convergence curve. b) Median dispersion curve (%)

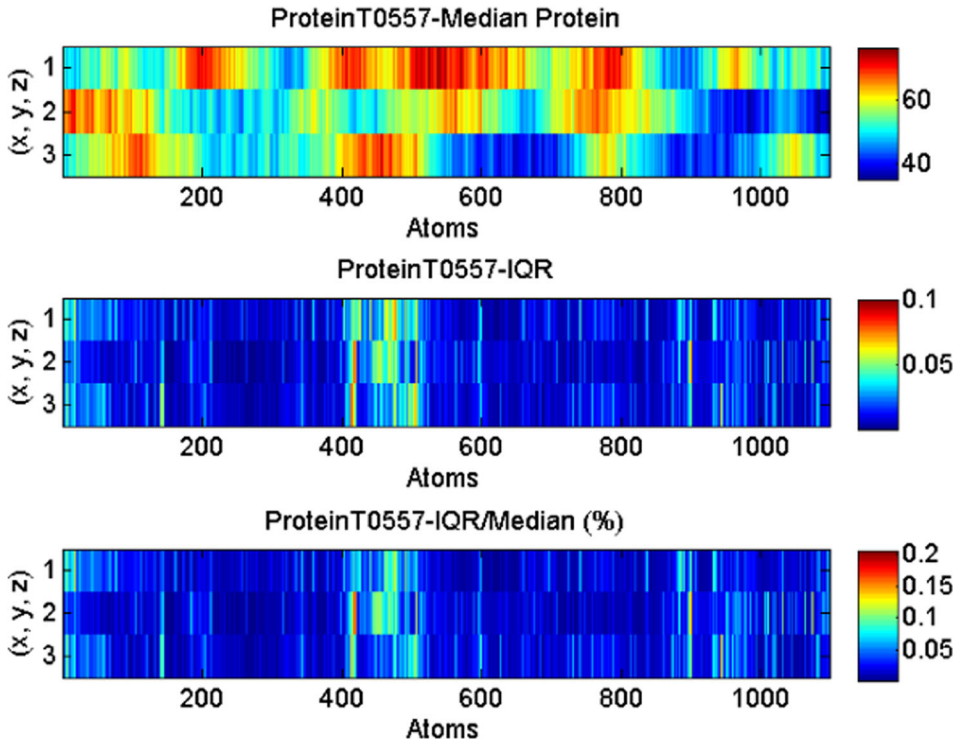

Fig. 24 T0637 posterior sampling in the region of energy lower than 0. a) Median protein of the decoys sampled. b) Median protein plus the interquartile range of the coordinates of these decoys. c) Median protein minus the interquartile range of the coordinates of these decoys

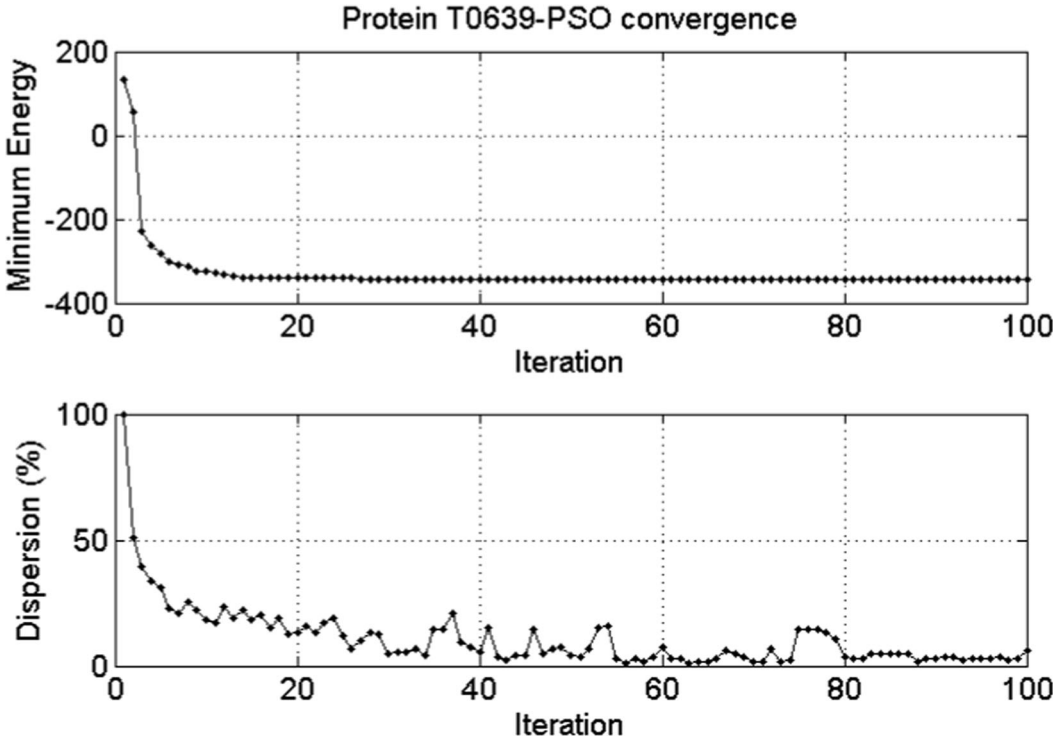

Fig. 25 T0639 protein. a) Convergence curve. b) Median dispersion curve (%)

T0555 – PBS linker polypeptide domain of phycobilisome linker protein

We present the numerical results of the application of PSO in order to obtain the tertiary structure of the PBS linker polypeptide domain of phycobilisome linker protein, whose native structure has been obtained through RMN by Ramelot et al. [37].

Owing to the complexities experienced in performing the optimization of the T0555 structure, a swarm composed of 80 particles was applied. Additionally, the fifth percentile of the best templates was chosen. By taking into account these considerations, we ensure a good convergence and protein refinement, while also carrying out a wide sampling over a search space constructed with good a priori models.

As observed in Fig. 13, the energy converges fast in the first 5 iterations and achieves energy of –371.4. Because the majority of the models fluctuate around this energy, we obtain a protein with a low uncertainty as shown in Fig. 14.

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

CASP9 T0557 protein performance under the algorithm presented in this paper is shown graphically in Figs. 15 and 16. The native structure of this protein has been obtained via NMR by Eletsky et al. [38] at the Northeast Structural Genomics Consortium.

We require the utilization of a swarm composed of 80 particles for protein T0555. Additionally, the fifth percentile of the best templates was chosen in order to ensure a good convergence and a good protein refinement.

T0561 – The structural basis for recognition of J-base containing DNA by a novel DNA-binding domain in JBP1

We performed PSO methodology to CASP9 protein T0561, whose native structure has been obtained through X-ray diffraction by Heidebrecht et al. [39]. To accomplish a proper convergence, we utilized a swarm size composed of 60 particles and the tenth percentile of the best protein decoys (Figs. 17 and 18).

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

CASP9 protein T0580 obtained by Cuff, M.E. [40] through X-Ray Diffraction has been optimized by PSO. In this sense a swarm size of 60 particles and the tenth percentile of the best decoys have been selected (Figs. 19 and 20).

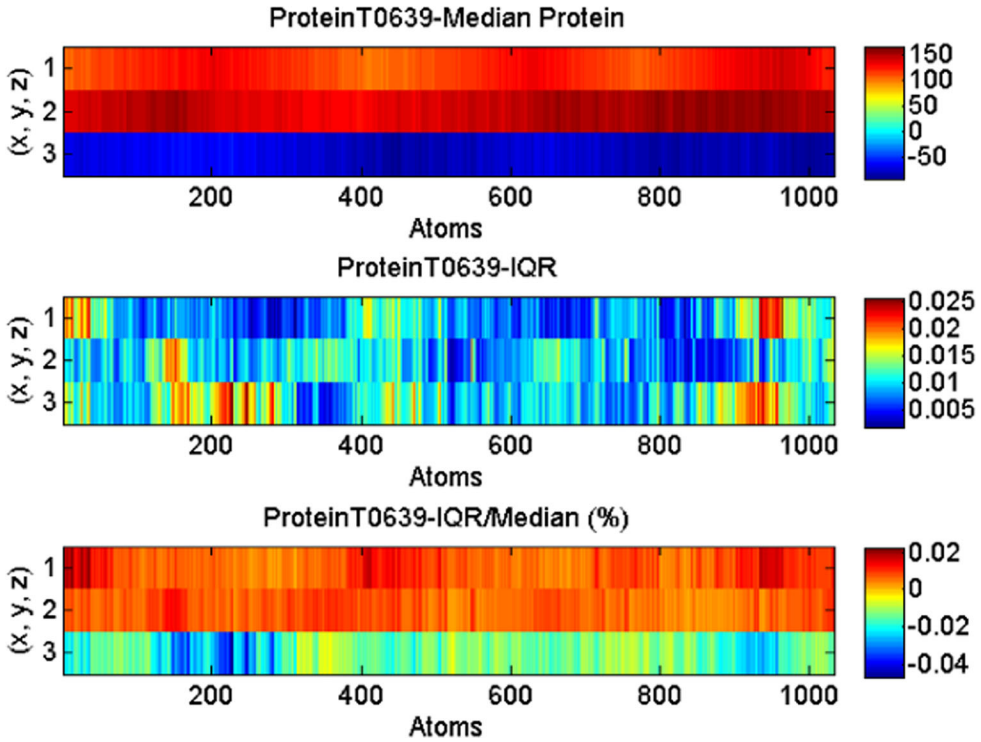

Fig. 26 T0639 posterior sampling in the region of energy lower than –300. a) Median protein of the decoys sampled. b) Median protein plus the interquartile range of the coordinates of these decoys. c) Median protein minus the interquartile range of the coordinates of these decoys

T0635 – The putative HAD superfamily (subfamily III A) hydrolase from Legionella pneumophila

We present the numerical results of the application of PSO in order to obtain the tertiary structure of the purative HAD superfamily (subfamily III A) hydrolase from Legionella pneumophila, whose native structure has been obtained through X-ray diffraction by Ramagopal et al. [41].

Since this protein has a very difficult topology to perform the algorithm, a swarm composed of 70 particles was applied. Additionally, the fifth percentile of the best templates was chosen. By taking into account these considerations, we ensure a good convergence and a good protein refinement by carrying out a wide sampling and good a priori models, as performed for previous proteins.

As observed in Fig. 21, the energy converges fast in the first five iterations and achieves an energy of –465.5. Because the majority of the models fluctuate around this energy, we obtain a protein with a low uncertainty as shown in Fig. 22, where only a small variation is observed in the extremes of the protein, while negligible variations are observed in the central atoms.

T0637 – Crystal structure of the hypothetical protein PA0856 from Pseudomonas aeruginosa.

Additionally, PSO capabilities have been tested for a hypothetical protein listed in the CASP9 experiment. It has been reported by Oke et al. [42] in the Scottish Structural Proteomics Facility. The PSO was performed utilizing a swarm of 70 particles and the tenth percentile of the best protein decoys submitted at the experiment. In this sense, the energy obtained was –372.0 with a very low variability within the models. Consequently, the uncertainty of the protein is low and only minor variations are observed in the extremes, the most sensitive part (Figs. 23 and 24).

T0639 – Crystal structure of functionally unknown protein from Neisseria meningitidis MC58.

Protein T0639 from the CASP9 experiment, a protein from Neisseria meningitidis MC58, whose native structure was obtained by Zhang. et al. [43] from the Midwest Center for Structural Genomics via X-ray diffraction. Similar to the case of T0637, this protein rapidly achieves its minimum at –342.7. The RMSD improvement confirms that the protein is successfully refined through PCA and RR-PSO optimization (Figs. 25 and 26).

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

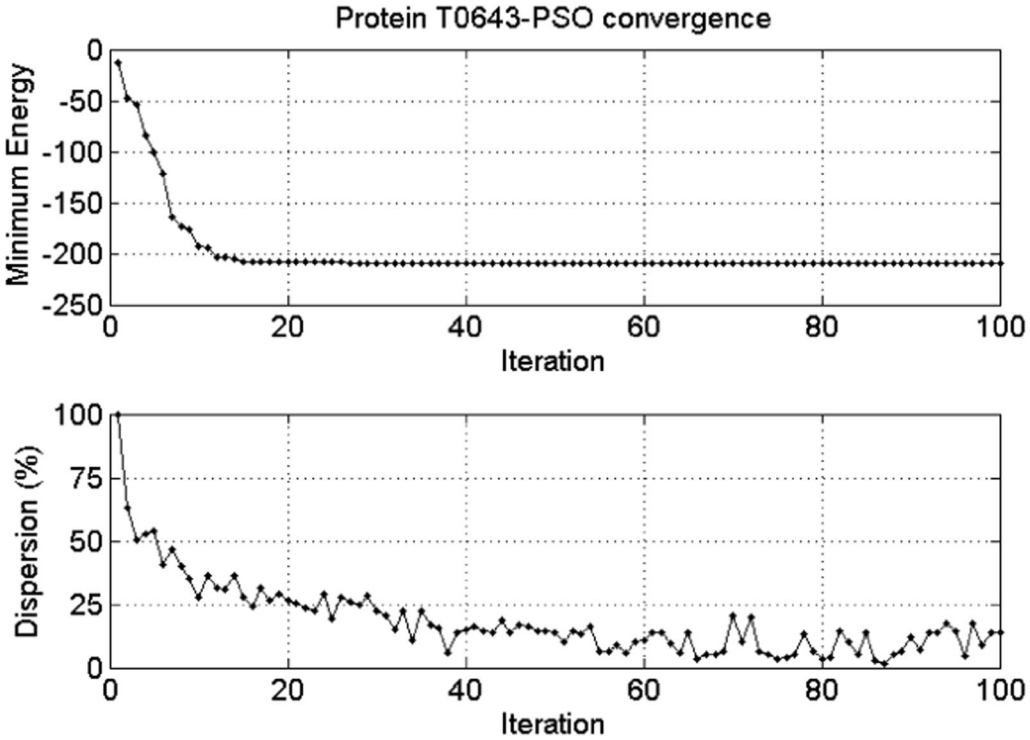

Fig. 27 T0639 protein. a) Convergence curve. b) Median dispersion curve (%)

SATB1, whose native structure was obtained through X-ray diffraction by Forouhar et al. [44]. The algorithm performance is presented in Figs. 27 and 28. Figure 27 shows how PSO was capable of optimizing the energy of the protein while carrying out the sampling in the region below –200. The sampling in the region below –200 quantifies the structure uncertainty as shown in Fig. 28.

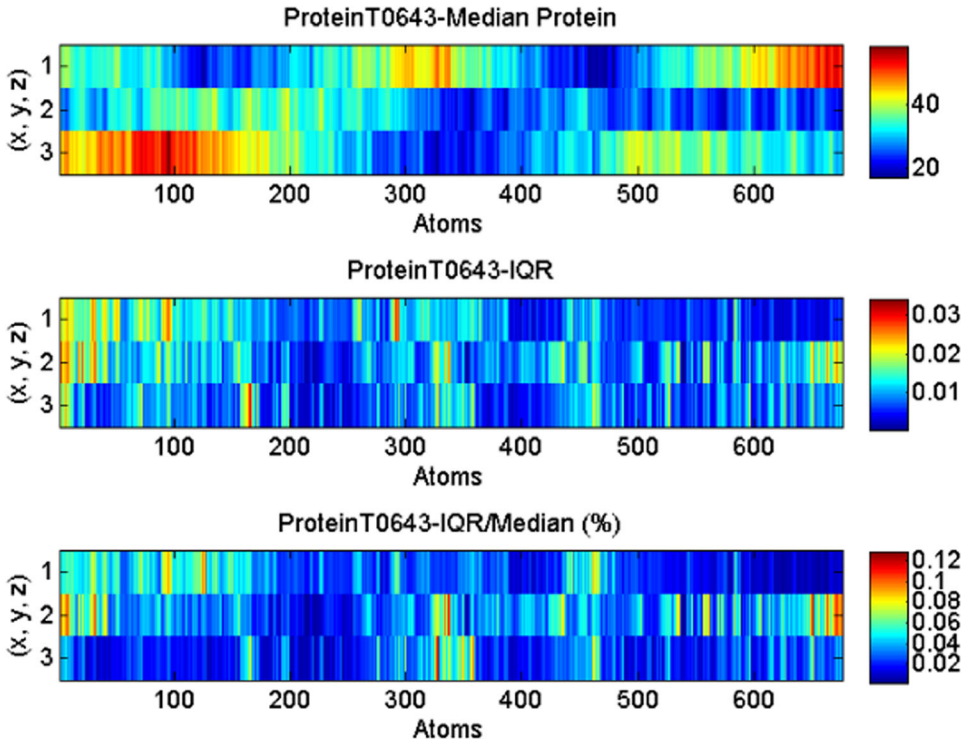

Fig. 28 T0643 posterior sampling in the region of energy lower than –150. a) Median protein of the decoys sampled. b) Median protein plus the interquartile range of the coordinates of these decoys. c) Median protein minus the interquartile range of the coordinates of these decoys

References

1. Tyka MD et al (2011) Alternate states of proteins revealed by detailed energy landscape mapping. J Mol Biol 405:607–618
2. Zhang Y (2008) Progress and challenges in protein structure prediction. Curr Opin Struc Biol 18:342–348
3. Stoker HS (2015) Organic and biological chemistry. Cengage Learning, Boston
4. Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68
5. Jowie BU et al. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170
6. Alvarez-Machancoses O et al (2018) Principal component analysis in protein tertiary structure. J Boinf Comp Biol 16:1850005
7. Sarawasthi S, Fernández-Martínez JL et al. (2012) Fast learning optimized prediction methodology (FLOPRED) for protein secondary structure prediction. J Mol Model 18:4275–4289
8. Araswathi S, Fernández Martínez JL et al. (2013) An aminoacid perspective to secondary structure prediction. J Mol Model 19: 4337–4348
9. Baker D, Sali A (2001) Protein stucture prediction and structural genomics. Science 294:93–96
10. Ramelot TA et al. (2009) Improving NMR protein structure quality by Rosetta refinement: a molecular replacement study. Proteins 75: 147–167
11. Gniewek P et al. (2014) BioShell - threading: a versatile Monte Carlo package for protein threading. BMC Bioinform 22:22
12. Gniewek P et al. (2012) How noise in force fields can affect the structural refinement of protein models. Proteins: Stuct Funct Bionf 80:335–341
13. Gront D, Kolinski A (2006) Bioshell - A package of tools for structural biology prediction. Bioinformatics 22:621–622
14. Gront D, Kolinski A (2008) Utility library for structural bioinformatics. Bioinformatics 24:584–585
15. Yang Y, Zhou Y (2008) Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 72: 793–803
16. Qiu D et al. (1997) The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. J Phys Chem A 101:3005–3014
17. Price SL (2008) From crystal structure prediction to polymorph prediction: interpreting the crystal energy landscape. Phys Chem Chem Phys 2008:1996–2009
18. Goldenberg DP, Creighton TE (2004) Energetics of protein structure and folding. Biopolymers 24:167–182
19. Fernández-Martínez JL, García-Gonzale E (2011) Stochastic stability analysis of the linear continuous and discrete PSO models. Trans Evol Comp 15:405–423
20. Fernández-Martínez JL, García-Gonzalo E (2012) 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
21. Fernández-Martínez JL et al. (2013) From Bayes to Tarantola: New insights to understand uncertainty in inverse problems. J Appl Geophys 98:62–72
22. Fernández-Martínez JL et al. (2012) On the topography of the cost functional in linear and nonlinear inverse problems. Geophysics W1-W15:77
23. Fernández-Martínez JL et al. (2014) The effect of the noise and Tikhonov's regularization in inverse problems. Part I: the linear case. J Appl Geophys 108:176–185
24. Fernández-Martínez JL (2014) The effect of the noise and Tikhonov's regularization in inverse problems. Part II: the nonlinear case. J Appl Geophys 108:186–193
25. Zhang Y, Skolnick J (2004) SPICKER: a clustering approach to identify near-native protein folds. J Comp Chem 25:865–871
26. Pearson K (1901) On lines and planes of closest fit to systems of points in space. Phylo Mag 2:559–572
27. Fernández-Martínez JL et al. (2012) Reservoir characterization and inversion uncertainty via a family of particle swarm optimizers. Geophysics 77–1:1–16
28. Jolliffe I (2002) Principal component analysis. Springer, New York
29. Quian B et al. (2004) Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. Proc NatL Acad Sci USA 101:15346–15351
30. Tarantola A (2005) Inverse problem theory and methods for model parameter estimation. SIAM, Philadelphia
31. Fernández-Martínez JL (2015) Model reduction and uncertainty analysis in inverse problems. Leading Edge 34:1006–1016
32. Kennedy J, Eberhart R (1995) A new optimizers using particle swarm theory. Proc Sixth Int Symp Micro Mach Human Sci
33. Fernández-Martínez JL, García-Gonzalo E (2008) The generalized PSO: a new door to PSO evolution. J Artif Evol Appl: 861275
34. Fernández-Martínez JL, García-Gonzalo E (2009) The PSO family: deduction, stochastic analysis and comparison. Swarm Intell 3:245–273
35. Aramini JM et al. (2010) Solution NMR structure of a putative uracil DNA glycosylase from Methanosarcina acetivorans. Northeast structural genomics consortium target MvR76
36. Fernández-Martínez JL et al. (2012) 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
37. Fernández-Martínez JL, García Gonzalo E (2011) Stochastic stability analysis of the linear continuous and discrete PSO models. IEEE Trans Evol Comput 15:405–423