PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

# MINOES: A new approach to select a representative ensemble of structures in NMR studies of (partially) unfolded states. Application to Δ25-PYP

**Mickaël Krzeminski, Gloria Fuentes, Rolf Boelens, and Alexandre M.J.J. Bonvin***

Bijvoet Center for Biomolecular Research, Science Faculty, Utrecht University, 3584 CH Utrecht, The Netherlands

ABSTRACT

**In nature, some proteins partially unfold under specific environmental conditions. These unfolded states typically consist of a large ensemble of conformations; their proper description is therefore a challenging problem. NMR spectroscopy is particularly well suited for this task: information on conformational preferences can be derived, for example, from chemical shifts or residual dipolar couplings. This information, which is measured as a time- and ensemble-average, can be used to model these states by generating large ensembles of conformations. The challenge is then to select a minimum representative set of conformations out of a large ensemble to represent the unfolded state. We have developed for this purpose an algorithm called MINOES (MINimum Optimal Ensemble Selection), which is based on an iterative procedure based on a driven expansion/contraction selection process. MINOES aims at selecting an optimal and minimal ensemble of conformations that, on average, maximizes the agreement between back-calculated and experimental (NMR) data, without any *a-priori* assumption about the required ensemble size. This approach is demonstrated by modeling the partially unfolded state of a deletion mutant of the Photoactive Yellow Protein, Δ25-PYP, which has been previously characterized by NMR (Bernard *et al.*, Structure 2005;13:953–962).**

Proteins 2009; 74:895–904.


**Key words: partially unfolded states; photoactive yellow protein; ensemble selection; NMR.**

## INTRODUCTION

High-resolution NMR spectroscopy is a powerful method that has become a routine for the structural characterization of biomolecules. In particular, the ability to study dynamical properties makes the technique complementary to X-ray crystallography. It normally relies on a dense network of distance restraints derived from nuclear Overhauser effects between nearby hydrogen atoms in the protein[1,2] to calculate the three-dimensional (3D) structure of a protein in solution.

The preferred representation of an NMR protein structure in solution is an ensemble, in the order of 20 models, which explore regions of conformational space that satisfy at the same time some physical parameters and experimentally derived restraints. The most used procedures to select these structures from a large set of calculated ones are based on energy (selection of a subensemble consisting of the lowest energy structures) and/or restraints violation criteria (selection of structures with no distance and/or dihedral angle violation above a particular value or threshold). Both approaches involve user-defined, arbitrary criteria for the cutoffs used and the number of structures selected; this together with possible differences in force field parameters and in the treatment of the experimental restraints used in the process lead to the "structure selection problem" in the structural NMR field. This problem is even more acute when it comes to describing highly flexible molecules such as short peptides or poorly defined (partially) unfolded states of proteins.

Chemical shifts have been long recognized as a potentially important structural information source because of their dependency and high sensitivity on multiple electronic and geometric factors and the high accuracy of their measurement.[1,3] A large variety of conformational effects, including backbone torsion angles, side-chain orientations, hydrogen bonding, and the type and conformation of neighboring residues, influence the chemical shift behavior in proteins.[4,5] All these effects make the chemical shifts good reporters of secondary structure elements as implemented in

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Netherlands Organization for Scientific Research (NWO); Grant numbers: 805.47.121 (van Molecuul tot Cell grant), 700.56.442 (VICI grant); Grant sponsor: European Community (FP6 STREP "UPMAN"); Grant number: LSHG-CT-2005-512052.

Gloria Fuentes's current address is Structural Computational Biology, Centro Nacional de Investigaciones Oncológicas (C.N.I.O.), Melchor Fernández Almagro, 3, E-28029 Madrid, Spain.
*Correspondence to: Alexandre M.J.J. Bonvin, Bijvoet Center for Biomolecular Research, Science Faculty, Utrecht University, 3584 CH, Utrecht, The Netherlands. E-mail: a.m.j.j.bonvin@uu.nl

Received 21 February 2008; Revised 2 June 2008; Accepted 25 June 2008
Published online 14 August 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22197

several softwares (e.g., CSI,[6] Talos,[7] Pecan[8]), as well as of structural changes observed upon different conditions (e.g., ligand binding[9]). However, the multiple dependencies mentioned above make both the interpretation and accurate prediction of chemical shifts exceedingly difficult, particularly in large systems, and their use, so far, has been limited. Fortunately, significant computational progresses in chemical shift prediction have been made[10–14] opening new possibilities for including them on a regular basis in structural studies.

In this article, we investigate the use of chemical shifts as information source for the selection of a representative ensemble of NMR structures. For this, we chose a specific class of proteins containing a photoreceptor that partially unfolds upon illumination. We concentrate on the photoactive yellow protein (PYP), which has been extensively studied in our laboratory in the past.[15,16] The transition to the excited, partially unfolded state, which corresponds actually to the signaling state of these proteins, has been described to follow a protein quake model.[17,18] The transient and unstable nature of such a state makes the acquisition of NMR data difficult. As a result, only sparse data can be typically obtained. The transient nature of the intermediates practically exclude the possibility to measure and/or use residual dipolar couplings that have become popular in the characterization of unfolded states.[19–22] Hence, characterizing the partially unfolded state of such proteins becomes rather complicated and a challenge to the current structural methods. A protein that partially unfolds can usually be split into two parts: a core, which maintains to a large extend a conformation similar to the native state, and a partially unfolded moiety.

In a previous work,[23] we have demonstrated that it is possible to model *in silico* partially unfolded states using native-like interresidue restraints for those residues that do not show appreciable chemical shift changes upon partially unfolding. Here, assuming that a partially unfolded state corresponds to an intermediate between a fully folded and fully unfolded state, we introduce a structure calculation protocol based on a gradual unfolding of the partially unfolded moiety by progressively decreasing the weight of the corresponding native-like restraints in the structural calculation protocol. In addition, a radius of gyration ($R_g$)[24] restraint derived from NMR diffusion measurements is introduced for better characterization of the partially unfolded state. In that way, a large ensemble of conformations can be generated covering the conformational space from the native to the (partially) unfolded state. The challenge is then to select an ensemble of structures that best describes the experimental observables. This selection problem has been addressed in the past for native structures among others by Smith *et al.*[25]: in that work, the number of structures to be selected was first determined based on the restraint energy profile and was used in a second step to select a representative ensemble out of all generated structures. In the case of partially unfolded proteins, the classical approach that selects the lowest energy structures and/or those having the minimum number of violations, is no longer suitable; because of the scarcity of the experimental data, those parts of the protein for which no or only little experimental data are available will strongly depend on the forcefield and calculation protocol used. The selection problem in the case of unfolded systems has been addressed previously by Forman-Kay's group who developed an algorithm called ENSEMBLe.[26,27] It makes use of the available experimental data to define some kind of energy function in which averaged back-calculated data and experimental data are compared; the latest version of the algorithm tries to select in a Monte Carlo process a user-defined number of structures that best represent the experimental data.

Here, we propose an innovative selection method which extracts an optimal subset out of a large ensemble by maximizing the agreement between observables back-calculated from the generated models and the experimental data without making any assumption about the ensemble size. Although the approach is generic and could make use of a variety of experimental data, we demonstrate it using protons $H_\alpha$ chemical shifts for the selection process. We first validate our selection method using synthetic data for lysozyme and then apply it to model the partially unfolded state of a deletion mutant of the Photoactive Yellow Protein (Δ25-PYP),[28] for which experimental NMR data are available[15]; this allows us to compare the resulting structures with the NOE-based one and validate our proposed selection method.

## THEORY

To evaluate how a selected model or ensemble thereof will represent the experimental data, one can calculate the "distance" between the available experimental data and back-calculated data obtained from these models. This distance can be expressed in the form of a $\chi^2$ function like in the following formula:

$$\chi^2 = \sum_{i=1}^{D} \omega_i \left(x_i - y_i\right)^2 \Big/ \sigma_i^2$$

with $D$, the number of available data; $x_i$, the data value predicted from models; $y_i$, the experimental data value; $\omega_i$, weight put on data $i$; $\sigma_i = \frac{s_i}{c_i} = \frac{\text{sensitiveness}}{\text{confidence}}$ = Experimental error associated with data $i$.

This scoring function has the advantage that it allows the combination of various kinds of data by use of the sensitiveness factor (which allows to correct for different data ranges) and takes experimental errors into account with the confidence factor. In addition, each data point

or set of points can be weighted separately by adjusting the value of $\omega_i$.

In our selection algorithm, we make use of $\chi^2$ to determine from a pool of generated structures, the subensemble that best fits the experimental data, without a-priori knowledge of the optimum ensemble size. This is a combinatorial problem that would lead to an explosion in computational time requirements if all possibilities were to be tested for a large ensemble. To avoid this problem, we use a recursive approach based on a driven expansion/contraction process of the system. The selection algorithm consists of the following steps:

For $N$ trials:

1. Randomly select one model from the initial pool of generated structures and define a combination ratio $C_R = 0.9$
2. Expansion process
   a. Generate new ensembles by:
      i. adding each of the nonselected structures in turn
      ii. removing each of the selected structures in turn
      iii. exchanging each of the selected structure by each of the nonselected ones
   b. From step 2a, select the ensemble with the best score (lowest $\chi^2$) and compare it with the previously defined one:
      i. If $\chi^2_{new} < \chi^2_{old}$, accept the new ensemble and start again from step 2.
      ii. If $\chi^2_{new} \geq \chi^2_{old}$, temporary accept the new ensemble and start again from step 2. If the score is not improved after repeating this selection process five times, proceed to step 4.
3. Contraction process
   a. From the ensemble reached in the previous step, generate new ensembles by:
      i. removing each of the selected structures in turn
      ii. exchanging each of the selected structure by each of the nonselected ones
   b. Keep the ensemble with the best score (lowest $\chi^2$)
   c. Repeat step 3 until the number of structures of the ensemble becomes equal to the predefined $C_R$ times the number of structures in the initial ensemble obtained in step 2
4. Increment the number of iterations and start from step 2 with the new ensemble defined from step 3. If the number of iterations reaches 20 without any new best score, decrease $C_R$ by 0.1, reinitialize the number of iterations and start from step 2. If $C_R$ reaches 0.1, stop the process.

In this process, we make use of the tabu search method.[29] This heuristic mechanism allows visiting more possible ensembles by avoiding returning to a previous state: if a modification results in an ensemble that has already been met, it is not kept, but instead the second best one is selected, and so on. In total, two tabu search tables are created, one for the expansion process and one for the contraction process.

This selection method has been coded in C in the MINOES program, standing for *MIN*imum *O*ptimal *E*nsemble *S*election.

## METHODS

### Generation of a synthetic partially unfolded state for the lysozyme

Structure ensembles were generated with CNS[30] using protocols derived from ARIA[31] as implemented in the RECOORD scripts.[32] For each system, we first performed 13 independent structure calculation runs (without water refinement), each of them producing 2000 structures. All runs started from the same extended conformation using a different initial random seed. Native-like $C_\alpha$-$C_\beta$ distance restraints were imposed on the α-domain of the protein (core region) as described previously.[23] In addition, considering that partial unfolding starts from the native state, native-like restraints were also applied to the unfolded part (β-domain, residues $G^4$ to $S^{36}$ and $T^{89}$ to $L^{129}$), with decreasing force constants (from 50 to 0 kcal $mol^{-1}$ $Å^{-2}$) for the different runs (Table I). This protocol allows the non-native region of the protein to slowly relax from the fully folded state to a fully or partially unfolded one. Distance restraints are only defined between residues that were at least three positions apart in the sequence. The definition of the synthetic native-like restraints was based on a previously described stochastic dynamic simulation of the hen egg-white lysozyme (PDB entry: 1AKI) aimed at unfolding the β-domain (for details see Fuentes *et al.*). In short, a

**Table I**
Weighted Force Constants (kcal $mol^{-1}$ $Å^{-2}$) for $C_\alpha$-$C_\beta$ Distance Restraints (Core/Noncore and Noncore/Noncore) Used Through the Simulated Annealing Protocol for the Various Runs to Generate Partially Unfolded Structures

| Run | Initial weight | Final weight |
|---|---|---|
| 0 | 10 | 50 |
| 1 | 9 | 45 |
| 2 | 8 | 40 |
| 3 | 7 | 35 |
| 4 | 6 | 30 |
| 5 | 5 | 25 |
| 6 | 4 | 20 |
| 7 | 3 | 15 |
| 8 | 2 | 10 |
| 9 | 1 | 5 |
| 10 | 0.2 | 1 |
| 11 | 0.1 | 0.5 |
| 12 | 0 | 0 |

In all runs, force constant for core-core atom interactions was set to 10/50. Each run yielded 2000 structures.

native-like distance restraint is defined only if the corresponding distance is shorter than 7.5 Å in at least half of structures of the native state NMR ensemble. In this case, the restraint is set to the average of all distances below 7.5 Å, with a lower distance bound of 1.8 Å and an upper one equal to the average plus one standard deviation (core/core distance) or to the average plus the standard deviation plus 2.0 Å (core/noncore or noncore/noncore).

Additionally, the radius of gyration $R_g$ was included as a restraint in such a way that the target values imposed for the entire ensemble of structures calculated follow a log-normal probability density function. Hence, for $r \geq 0$, we have the following formula:

$$F(r + R_g^{\text{eff}}) = e^{-\frac{1}{2}\left(\frac{\ln(r)-\mu}{\sigma}\right)^2} \Big/ \sqrt{2\pi}\sigma r$$

where $R_g^{\text{eff}}$ is the experimental value of the radius of gyration. We arbitrary set the values of $\mu$ and $\sigma$ to 0.2 and 0.4, respectively. By taking values from a log normal distribution, we allow for more extension than compaction of the structure. For each individual structure calculation, a different value of $R_g^{\text{eff}}$ is used, taken from the log-normal distribution. $R_g^{\text{eff}}$ was set to 14.3 Å.

For each run, the 500 lowest energy structures obtained after simulated annealing were submitted to refinement in explicit solvent. From these refined models, only those with less than 5 distance restraint violations below 0.5 Å were kept for further analysis; their $H_\alpha$ chemical shifts were calculated using ShiftX.[12,14]

To validate our method, we selected out of all generated models some random structures and created reference chemical shifts based on the average of these on the randomly selected structures. Then, using the entire ensemble (6465 models), we tried to retrieve with MINOES the reference ensemble. For validation, the selection was repeated 10 times from various reference ensembles consisting of 10, 15, and 20 structures, respectively.

### Experimental data for Δ25-PYP and generation of models *in silico*

The $^1H^{15}N$-HSQC spectra of both the native (dark state, pG) and partially unfolded (light state, pB intermediate) states of Δ25-PYP have been previously reported.[15] The chemical shift perturbation calculated from these two spectra allowed us to define the core (61 residues out of 100, comprising by the segments: $G^{29}$ to $N^{43}$, $K^{55}$ to $F^{62}$, $G^{82}$ to $F^{96}$, and $T^{103}$ to $V^{125}$) and the unfolded part of the protein. This information was used to define native-like restraints from the NMR pG ensemble (PDB entry: 1XFN).

The same structure calculation protocol as described for the lysozyme above was followed. The radius of gyration was set to 14.0 Å, based on NMR diffusion experiments (Nico van Nuland, Universidad de Granada, Spain; personal communication).

### Analysis of models

The selected ensembles were compared with the reference one or with each other by calculating the average pairwise positional root mean square deviations (RMSD). For this, the structures were fitted on the backbone atoms of the defined core regions and the RMSDs were calculated on the entire backbone of the protein, using ProFit (Martin, A.C.R., http://www.bioinf.org.uk/software/profit/).

Secondary structure analysis was performed with PROCHECK,[33,34] which makes use of the Kabsch and Sander secondary structure definitions.[35] A consensus secondary structure for each residue was adopted when at least half of the structures of an NMR ensemble had the same classification.

## RESULTS

### Validation of the method with synthetic data

We first validated our algorithm by performing the selection of a subensemble that best fits the reference synthetic lysozyme data. For this, we tested whether we could recover the selected reference ensemble from the combination of all generated models. These consist of 6465 structures with less than 5 consistent violations (of the defined native-like restraints) below 0.5 Å out of the 6500 (13 × 500) generated ones. In all cases, our algorithm successfully retrieved the reference ensemble of 10, 15, or 20 structures from the pool of 6465 models with a correlation coefficient between back-calculated and reference $H_\alpha$ chemical shifts of 1.0 and a score obviously equal to 0.

This optimal selection could not have been obtained by selecting the same number of structures with the lowest individual scores (Table II); indeed, the individual members of the reference ensemble have scores spread over the entire range of scores and only their average results in a score of 0.

### Application to the partially unfolded state of Δ25-PYP (pB)

The PYP is found in the bacterium called *Ectothiorhodospira halophila*. It is localized in the cell machinery that promotes the swimming away of the organism when exposed to intense blue light.[36–38] This function has been found to rely on a chromophore embedded in the core of the protein, which absorbs light at 446 nm and triggers a reorganization of the protein leading to partial unfolding. The latter state corresponds to an active/signaling intermediate that triggers a cascade of events leading to the escape of the bacterium. The partially unfolded

**Table II**
Results of the MINOES Selection Processes

| Perfect ensemble size (*n*) | MINOES score | Lowest score of al individual structures | Average score of the *n* lowest individual score structures |
|---|---|---|---|
| 10 | 0 ± 0 | 9.89 ± 0.53 | 1.60 ± 0.42 |
| 15 | 0 ± 0 | 9.23 ± 0.53 | 1.33 ± 0.23 |
| 20 | 0 ± 0 | 9.25 ± 0.44 | 1.25 ± 0.25 |

Indicated values have been evaluated over the 10 trials we performed. In this table are specified the lowest individual score and the score of the *n* lowest individual score structures, *n* being the number of structure in the perfect ensemble.

state of a deletion mutant of this protein, Δ25-PYP, in which the 25 first N-terminal residues have been deleted, shows a longer lifetime. This behavior makes it a better candidate for structural studies by NMR and as consequence, the structure of both the native and partially unfolded states could be determined using classical NOE restraints.[15]

Following the protocol described earlier (see Methods), we generated an ensemble of 6494 water-refined structures with less than 5 violations within the core. The native-like restraints for the core and noncore regions were defined based on the ground state structure of Δ25-PYP (pG). Running MINOES, we selected, based on Hα CS, a set of 14 structures (Ntrial was set to 20) [Fig. 1(a)]. This final ensemble has an average internal energy of $-18179 \pm 675$ kcal mol$^{-1}$, a score of 3.95 and a correlation coefficient between back-calculated and experimental $H_\alpha$ chemical shifts of 0.94 (0.94 and 0.91 for the core and the partially unfolded moieties, respectively). The average pairwise backbone RMSD between the structures is 2.4 ± 0.4 Å and the average radius of gyration is 13.8 ± 0.4 Å. For comparison, the ensemble of 20 lowest energy structures has an average energy of $-20646 \pm 412$ kcal mol$^{-1}$ and a $R_g$ of 15.9 ± 0.5 Å. This ensemble has a correlation coefficient of only 0.74 (0.78 and 0.28 for the core and the partially unfolded moieties, respectively) and a score of 13.60. The average pairwise backbone RMSD in this case is 6.2 ± 3.7 Å. The agreement is even worse than when considering the full ensemble of 6494 structures ($\chi^2 = 10.00$, $R = 0.83$).

These results clearly indicate that our algorithm, making use of $H_\alpha$ proton chemical shifts in this case, performs well in selecting a representative ensemble for the partially unfolded state of Δ25-PYP. This is evidenced both by the improved correlation between back-calculated and experimental chemical shifts (which is not surprising since the selection was based on them), but also by the better agreement between the average radius of gyration of the selected ensemble and the experimental value (14.0 Å) obtained from NMR diffusion measurements. The latter provides an independent measure of the quality of the selected ensemble since it was not used in the selection procedure. The selected ensemble of structures samples $R_g$ values between 13.1 and 14.9 Å (the $R_g$ of all generated structures range between 12.9 and 17.7 Å) because of the way the radius of gyration restraints are defined following a log-normal distribution; still, our algorithm, based on $H_\alpha$ chemical shift criteria, selects an ensemble that matches the experimental values while the energy-based selection results in a looser ensemble. Note that the full ensemble also fits reasonably well the chemical shift ($R = 0.83$), although the quality of the fit is poorer for the noncore region ($R = 0.57$). This ensemble, however, covers a rather large conformational space (see supplementary figure), which makes it difficult to work with or even deposit in a structure database like the PDB.

We compared the MINOES ensemble with the NOE-based deposited ensemble (PDB entry: 1XFQ), which was solved previously in our laboratory.[15] The MINOES ensemble clearly better reproduces the chemical shift data (Table III and Fig. 2). We also analyzed the violations from the NOE restraints involving backbone amide protons; the latter were not used in the structure calculation, instead, native-like restraints were defined based on chemical shift differences between the native and unfolded states (see Methods). Out of 252 NOE restraints involving amide protons, the MINOES ensemble only shows 12 consistent violations larger than 0.5 Å, 4 of which exceed 1.0 Å and none above 2.0 Å. The MINOES ensemble is more compact than the NOE-based ensemble (see Fig. 1), with a higher secondary structure content, especially in the α-helices region (α1, α2, and α3) (see Fig. 3). This is also reflected in the average pairwise RMSD with each ensemble (Table III). The average pairwise RMSD between the two ensembles calculated only on the secondary structure elements of the NMR-based structure is 1.56 ± 0.27 Å, which indicates a rather good agreement between the core regions.

We also compared the MINOES ensemble with the crystal structure of the ground state of PYP (PDB entry: 1NWZ)[39]: it turns out that the selected ensemble is quite close to the ground state crystal structure, both in overall 3D structure [Fig. 1(b)] and in secondary structure content (see Fig. 3). The main difference comes from the side chain of $Cys^{69}$ bearing the chromophore that is flipped out in the pB state. The protonated $Glu^{46}$, which is hydrogen bonded to the chromophore and gives its hydrogen upon excitation, remains inside. Interestingly, most of the secondary structures are reasonably well preserved with some changes in helical regions, for instance from α-helical into hydrogen-bonded turn conformations, e.g., α1, α2 and part of α4 (see Fig. 2).

These results, showing a more compact pB state than what was reported previously, are in line with a recent SAXS study[40]: the average radius of gyration of our ensemble is lower than that of the deposited model and is

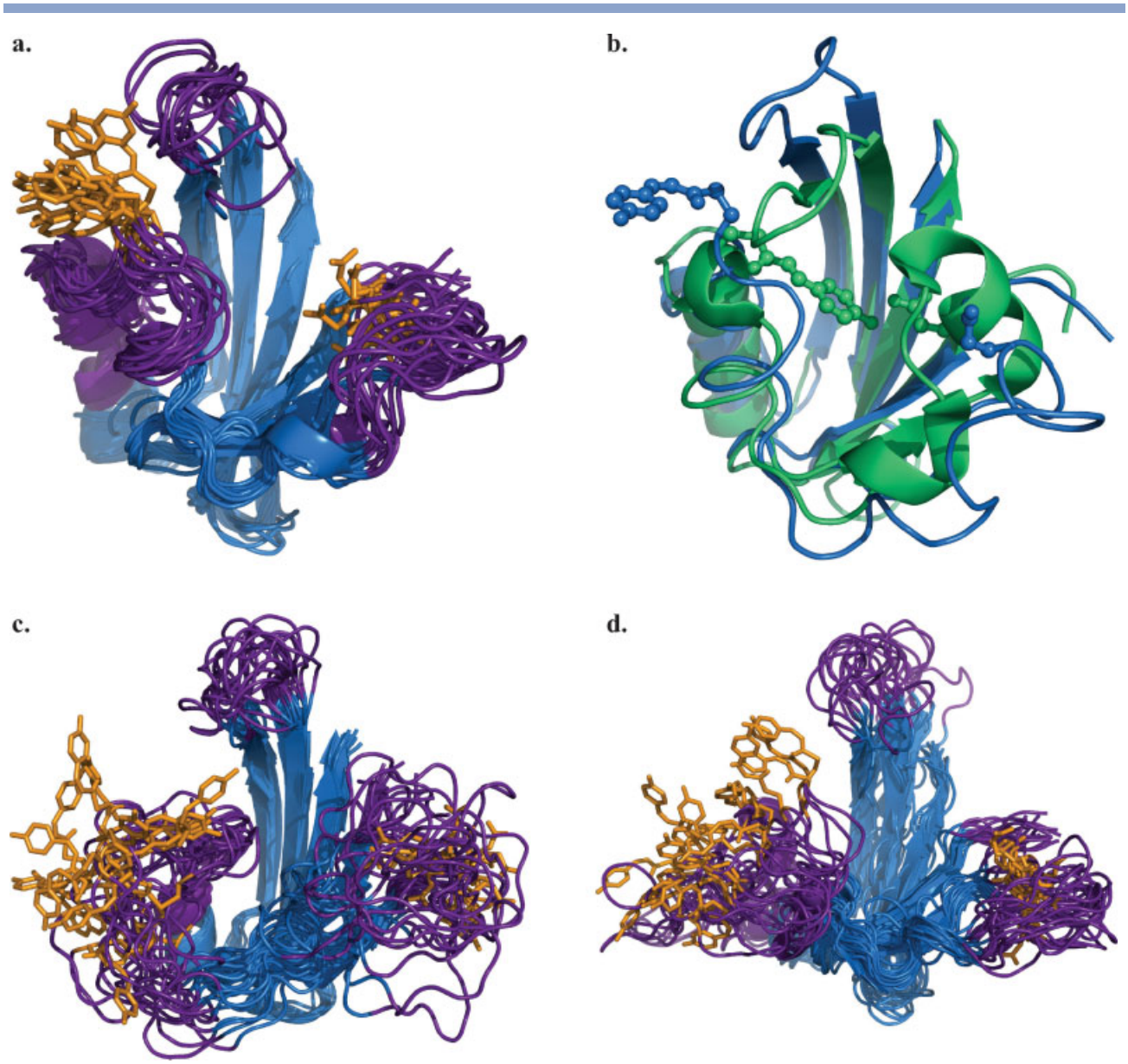


**Figure 1**

Cartoon representations of the pB state of Δ25-PYP for the MINOES ensemble selected based on $H_\alpha$ chemical shifts (**a**), deposited (**c**), and lowest energy (**d**) ensembles. The chromophore attached to $Cys^{69}$ and $Glu^{46}$, which serves as hydrogen donor upon excitation, are shown in sticks. The native-like (ground-state) core defined in this work is shown in dark gray. In (**b**) is shown a comparison between the X-ray ground state (dark color) at high resolution and one of our models (light color). The residues 69 and 46 are enlightened with a stick and sphere representation.

more consistent with the estimated radius of gyration from SAXS, which indicates only ∼5% increase of $R_g$ upon excitation. The increase of $R_g$ for the MINOES pB ensemble compared to pG is ∼9%, whereas the deposited NOE-based ensemble (PDB entry: 1XFQ) shows ∼17% increase. Furthermore, the predicted SAXS data for our MINOES ensemble (back-calculated with the CRYSOL program[41]) displays the experimentally observed bimodal profile (data not shown). The observed relative difference of 4% between SAXS results and our model corresponds to an absolute difference of 0.5 Å. This difference is not really significant considering the experimental errors and model assumptions associated with the interpretation of the experimental NMR diffusion experiments and the fact that the SAXS study was performed on a slightly shorter PYP construct (Δ23).

**Table III**
Comparison of Excited State (pB) Ensembles Obtained by Different Approaches: NOE-Based Ensemble (Deposited Model), the 20 Lowest Energy Structures Obtained Using Native-Like Restraints and the Ensemble Selected by the MINOES Algorithm

| | Number of structures | Average $R_g$ (Å)[a] | RMSD (Å)[b] | *R* Between Hα Calculated vs. Experimental[c] | | |
|---|---|---|---|---|---|---|
| | | | | Full protein | Core | Noncore |
| Deposited model (1XFQ) | 20 | 14.7 ± 0.3 | 4.3 ± 0.8 | 0.85 | 0.86 | 0.63 |
| Lowest energy structures | 20 | 15.9 ± 0.5 | 6.2 ± 3.7 | 0.74 | 0.78 | 0.34 |
| MINOES ensemble | 14 | 13.8 ± 0.4 | 2.4 ± 0.4 | 0.94 | 0.94 | 0.92 |
| All generated structures | 6494 | 13.8 ± 0.4 | 4.3 ± 3.0 | 0.83 | 0.85 | 0.57 |

[a]The experimental $R_g$ value estimated from NMR diffusion experiments (Nico van Nuland, personal communication is 14.0 Å).
[b]Average pairwise backbone RMSD (see Material and Methods).
[c]Correlation coefficient between experimental and back-calculated Hz chemical shifts. The Hα chemical shifts were back-calculated with SHIFTX.[14]

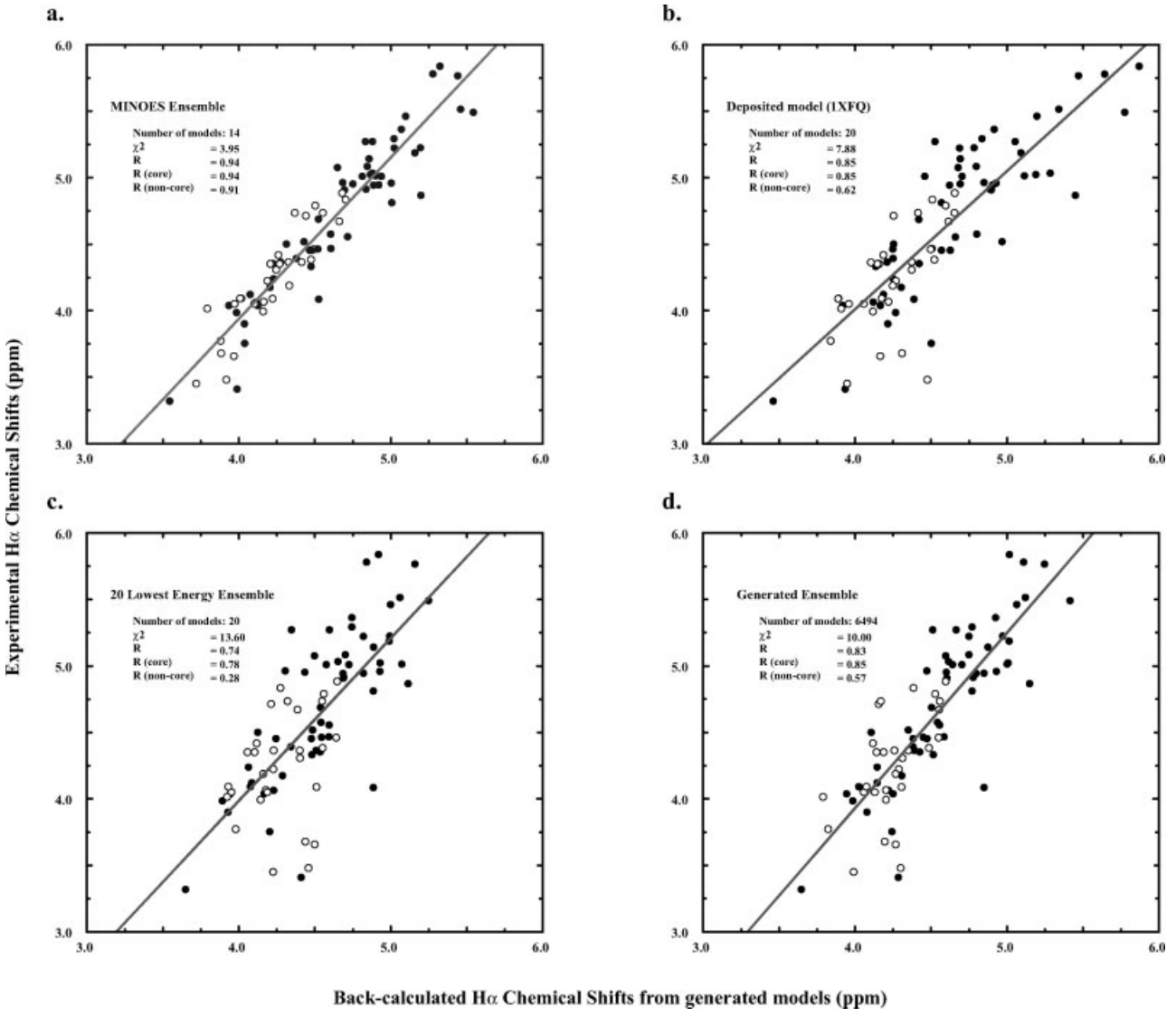


**Figure 2**
Comparison of average data of each ensemble with experimental data: (**a**) MINOES ensemble, (**b**) deposited model, (**c**) 20 lowest energy structures, and (**d**) full ensemble generated with RECOORD.[32] Closed and open circles correspond to the core and the noncore regions of the protein, respectively. The trend line is drawn in gray. The number of structures of each ensemble, the score ($\chi^2$), and the correlation coefficients *R* are indicated on the left side.

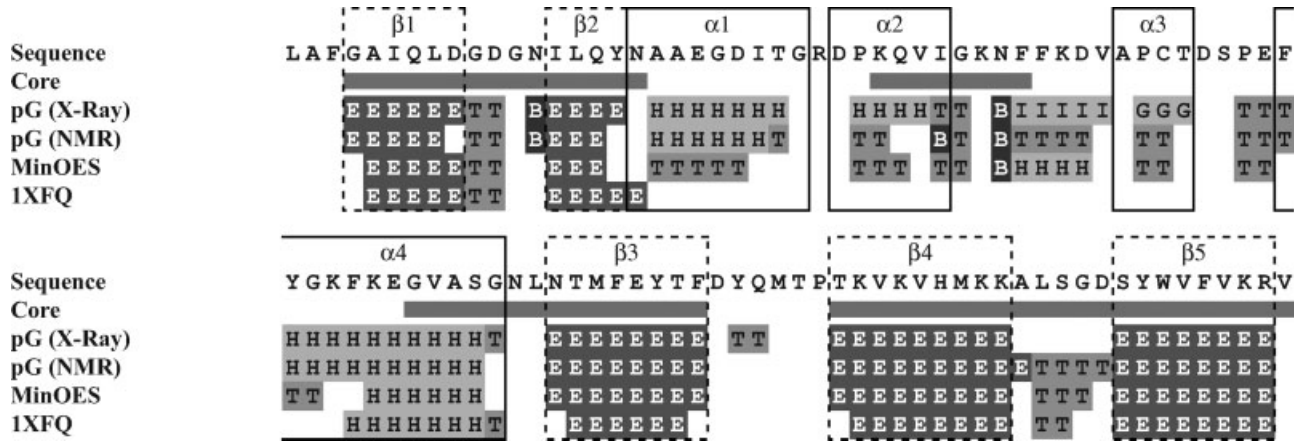


**Figure 3**

Secondary structure of Δ25-PYP calculated with PROCHECK[33,34] for (**a**) the crystal structure of the ground state (pG) (1NWZ), (**b**) the solution NMR structure of the ground state (pG-1XFN), (**c**) the partially unfolded state (pB) ensemble selected by MINOES, and (**d**) the NOE-based partially unfolded state ensemble (pB-1XFQ). H, G, and I indicate alpha helices, 3/10-helices and π-helices, respectively (all with the same gray shading). T correspond to hydrogen bonded turns and B and E are residues in isolated β-bridges and extended strands, respectively. Bends are not indicated for the sake of clarity. The core defined in this work is indicated by filled gray boxes. The secondary structure elements defined in the crystal structure are specified.

## DISCUSSION

We have shown here that, in the case of partially unfolded states, selecting an ensemble of structures based on the agreement between back-calculated data from theoretical models and experimental data results in a better representation of the experimental data than simple energy and violation criteria. The algorithm we have developed for this purpose finds in an efficient expansion/contraction process the optimum ensemble without any *a priori* assumption about ensemble size.

The difficulty in such a selection method relies on the almost impossible task of considering all possible subensembles, because this is a combinatorial problem. For instance, assuming that we generated only 200 structures and want to determine which ensemble best fits the experimental data, the number of possible combinations would be:

$$\sum_{k=0}^{n} C_n^k = 2^{200} \approx 1.61 \times 10^{60}$$

If we consider that a regular computer performs roughly $10^{12}$ calculations per second, this means we would need around $5 \times 10^{40}$ years to achieve all required calculations, in other words about $3 \times 10^{27}$ times the age of the Universe!

The complexity class of the problem addressed here is NP-Hard.[42] Each subensemble can be interpreted as a unique state having a score indicating the agreement with the experimental data. A simple way to represent these states is to consider all binary numbers made of as many (0,1) as the size of the initial ensemble. For example, 01010 is the subensemble composed with the structures 2 and 4 of an initial ensemble constituted of 5 structures. Hence, the core of our algorithm tries to exchange "0" and "1", in a stepwise manner. The main characteristics of such a process are that the final state is highly dependent on the starting one and that the more structures there are in the initial pool, the bigger will be the selected ensemble, as there is always a possibility to add a structure that will make the correlation improve on average. To overcome this problem, a contraction step has been included in the selection procedure. In this way, we create the possibility to start from some smaller states that would not be reachable otherwise. As the outcome depends potentially on the initial state, the entire selection is repeated several times (10 in our case) starting from different structures and the best ensemble of all runs is kept.

The most important parameter in MINOES is the fraction of structures ($C_R$) that are kept after each expansion step. Indeed, the size of the best ensemble (unknown parameter) could be much smaller than the number of structures in the current selected subensemble; in such a case, the program would never be able to reach the best ensemble. To overcome this problem, the $C_R$ is progressively decreased from 0.9 (the ensemble size shrinks by 10%) to 0.1 (the ensemble size shrinks by 90%).

The number of iterations is another important parameter. Indeed, the more iterations are performed, the more ensembles will be met. This, together with the tabu search approach, increases the chance of finding the optimal ensemble.

### Timing of the algorithm

For each run, the program tests the possibility of removing, exchanging, and/or adding structures, one by one, and finally keeps the combination that leads to the largest decrease in score. The computing time needed by the algorithm to find an optimal ensemble depends on several parameters, such as the number of iterations, the size of the initial pool of structures, the number of available experimental data (chemical shifts, RDC...), etc. Benchmarking has indicated that the calculation time is proportional to the total number of structures in the selection pool and to the square of the number of data points. As an indication, in the case of Δ25PYP (6494 structures in the initial pool and 100 data points), it took a couple of hours per selection process on a 3.0 GHz Xeon Linux PC.

### Δ25-PYP pB structure

Our model of *Δ25-PYP pB* obtained with MINOES differs somewhat from the previously reported NMR structure[15] in two aspects. First, it shows a better defined α4 helix; this was already reported by Fuentes *et al.*[23] and is supported by the secondary chemical shifts that show a clear helical propensity for that helix (see Fig. 7 of Fuentes *et al.*[23]). Second, the partially unfolded part is now better defined since the selection based on the agreement with Hα chemical shifts, while no information was present for that region in the structure of Bernard *et al.*[15] The main purpose of this work was to validate our new ensemble selection procedure and we used for this *Δ25-PYP pB*. Having demonstrated the validity of our approach, we should be able in the future to generate an even better model of the pB state of PYP by making use of the available NOE distance restraints in combination with our chemical-shift based selection.

## CONCLUSIONS

We have described an innovative and efficient algorithm called MINOES that allows to select a minimum representative set of conformations out of a large ensemble of structures and this without any a-priori knowledge about the optimal ensemble size. MINOES aims at selecting a minimum optimal ensemble of conformations that, on average, maximizes the agreement between back-calculated and experimental (NMR) data. Although the method was developed and demonstrated for the selection of a representative partially unfolded state ensemble based on chemical shift information, it is by far not limited to such a problem and/or type of data; the only condition is that the data can be back-calculated from the generated models. MINOES is coded in generic manner and can accept in principle any kind of data. The current implementation assumes a linear averaging, but this can be easily modified. In the case of (partially) unfolded states, RDC and/or paramagnetic relaxation enhancement data could provide other very useful sources of information[20] that could be optimally used in our selection procedure.

### Availability

The source code for MINOES is available for free from the authors upon request.

## REFERENCES


1. Wüthrich K. NMR of proteins and nucleic acids. New York: Wiley; 1986.
2. Neuhaus D, Williamson MP. The nuclear overhauser effect in structural and conformational analysis. New York: Wiley; 2000.
3. Wishart DS, Sykes BD. Chemical shifts as a tool for structure determination. Methods Enzymol 1994:363.
4. Le H, Oldfield E. Correlation between 15N NMR chemical shifts in proteins and secondary structure. J Biomol NMR 1994;4:341–348.
5. Spitzfaden C, Braun W, Wider G, Widmer H, Wuthrich K. Determination of the NMR solution structure of the cyclophilin A-cyclosporin A complex. J Biomol NMR 1994;4:463–482.
6. Wishart DS, Sykes BD. The 13C chemical-shift index: a simple method for the identification of protein secondary structure using $^{13}$C chemical-shift data. J Biomol NMR 1994;4:171–180.
7. Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 1999;13:289–302.
8. Eghbalnia HR, Wang L, Bahrami A, Assadi A, Markley JL. Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements. J Biomol NMR 2005;32:71–81.
9. Zuiderweg ER. Mapping protein-protein interactions in solution by NMR spectroscopy. Biochemistry 2002;41:1–7.
10. Case DA. The use of chemical shifts and their anisotropies in biomolecular structure determination. Curr Opin Struct Biol 1998;8:624–630.
11. Meiler J. PROSHIFT: Protein chemical shift prediction using artificial neural networks. J Biomol NMR 2003;26:25–37.
12. Neal S, Nip AM, Zhang HY, Wishart DS. Rapid and accurate calculation of protein H-1. C-13 and N-15 chemical shifts. J Biomol NMR 2003;26:215–240.
13. Williamson MP, Asakura T. Protein chemical shifts. Methods Mol Biol 1997;60:53–69.
14. Wishart DS, Case DA. Use of chemical shifts in macromolecular structure determination. Methods Enzymol 2001;338:3–34.
15. Bernard C, Houben K, Derix NM, Marks D, van der Horst MA, Hellingwerf KJ, Boelens R, Kaptein R, van Nuland NA. The solution structure of a transient photoreceptor intermediate: delta25 photoactive yellow protein. Structure 2005;13:953–962.
16. Dux P, Rubinstenn G, Vuister GW, Boelens R, Mulder FA, Hard K, Hoff WD, Kroon AR, Crielaard W, Hellingwerf KJ, Kaptein R. Solution structure and backbone dynamics of the photoactive yellow protein. Biochemistry 1998;37:12689–12699.
17. Ansari A, Berendzen J, Bowne SF, Frauenfelder H, Iben IE, Sauke TB, Shyamsunder E, Young RD. Protein states and proteinquakes. Proc Natl Acad Sci USA 1985;82:5000–5004.
18. Itoh K, Sasai M. Dynamical transition and proteinquake in photoactive yellow protein. Proc Natl Acad Sci USA 2004;101:14736–14741.
19. Bernado P, Bertoncini CW, Griesinger C, Zweckstetter M, Blackledge M. Defining long-range order and local disorder in native alpha-synuclein using residual dipolar couplings. J Am Chem Soc 2005;127:17968–17969.

20. Bernado P, Blanchard L, Timmins P, Marion D, Ruigrok RW, Blackledge M. A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering. Proc Natl Acad Sci USA 2005;102:17002–17007.
21. Kristjansdottir S, Lindorff-Larsen K, Fieber W, Dobson CM, Vendruscolo M, Poulsen FM. Formation of native and non-native interactions in ensembles of denatured ACBP molecules from paramagnetic relaxation enhancement studies. J Mol Biol 2005;347:1053–1062.
22. Mukrasch MD, Markwick P, Biernat J, Bergen M, Bernado P, Griesinger C, Mandelkow E, Zweckstetter M, Blackledge M. Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. J Am Chem Soc 2007;129:5235–5243.
23. Fuentes G, Nederveen AJ, Kaptein R, Boelens R, Bonvin A. Describing partially unfolded states of proteins from sparse NMR data. J Biomol NMR 2005;33:175–186.
24. Kuszewski J, Gronenborn AM, Clore GM. Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration. J Am Chem Soc 1999;121:2337–2338.
25. Smith JA, GomezPaloma L, Case DA, Chazin WJ. Molecular dynamics docking driven by NMR-derived restraints to determine the structure of the calicheamicin gamma(I)(1) oligosaccharide domain complexed to duplex DNA. Magn Reson Chem 1996;34:S147–S155.
26. Choy WY, Forman-Kay JD. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. J Mol Biol 2001; 308:1011–1032.
27. Marsh JA, Neale C, Jack FE, Choy WY, Lee AY, Crowhurst KA, Forman-Kay JD. Improved structural characterizations of the drkN SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure. J Mol Biol 2007;367:1494–1510.
28. van der Horst MA, van Stokkum IH, Crielaard W, Hellingwerf KJ. The role of the N-terminal domain of photoactive yellow protein in the transient partial unfolding during signalling state formation. FEBS Lett 2001;497:26–30.
29. Glover F, Laguna F. Tabu search. Boston: Kluwer Academic Publishers; 1997. xix, 382 pp.
30. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang J-S, Kuszewski J, Nilges N, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography and NMR system (CNS): a new software system for macromolecular structure determination. Acta Crystallogr D Biol 1998;54:905–921.
31. Linge JP, O'Donoghue SI, Nilges M. Automated assignment of ambiguous nuclear overhauser effects with ARIA. Methods Enzymol 2001;339:71–90.
32. Nederveen AJ, Doreleijers JF, Vranken WF, Miller Z, Spronk CAEM, Nabuurs SB, Güntert P, Livny M, Markley JL, Nilges M, Ulrich EL, Kaptein R, Bonvin AMJJ. RECOORD: a REcalculated COORdinates database of 500+ proteins from the PDB using restraint data from the Bio Mag Res Bank. Proteins: Struct Funct Bioinformatics 2005;59:662–672.
33. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993:283–291.
34. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 1996;8:477–486.
35. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
36. Armitage JP. Behavioural responses of bacteria to light and oxygen. Arch Microbiol 1997;168:249–261.
37. Armitage JP, Hellingwerf KJ. Light-induced behavioral responses ('phototaxis') in prokaryotes. Photosynth Res 2003;76:145–155.
38. Hellingwerf KJ. The molecular basis of sensing and responding to light in microorganisms. Antonie Van Leeuwenhoek 2002;81:51–59.
39. Getzoff ED, Gutwin KN, Genick UK. Anticipatory active-site motions and chromophore distortion prime photoreceptor PYP for light activation. Nat Struct Biol 2003;10:663–668.
40. Kamikubo H, Shimizu N, Harigai M, Yamazaki Y, Imamoto Y, Kataoka M. Characterization of the solution structure of the M intermediate of photoactive yellow protein using high-angle solution X-ray scattering. Biophys J 2007;92:3633–3642.
41. Svergun D, Barberato C, Koch M. CRYSOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr 1995;28:768–773.
42. Garey MR, Johnson DS. Computers and intractability: a guide to the theory of NP-completeness. New York: W.H. Freeman & Company; 1979. 340 p.