BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 19 2008, pages 2184–2192
doi:10.1093/bioinformatics/btn396

Structural bioinformatics

Mixture models for protein structure ensembles

Michael Hirsch[1] and Michael Habeck[1,2,*]

[1]Department of Empirical Inference, Max-Planck-Institute for Biological Cybernetics, Spemannstrasse 38 and [2]Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, Spemannstrasse 35, 72076 Tübingen, Germany

Received on May 6, 2008; revised on July 17, 2008; accepted on July 26, 2008

Advance Access publication July 28, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Protein structure ensembles provide important insight into the dynamics and function of a protein and contain information that is not captured with a single static structure. However, it is not clear a priori to what extent the variability within an ensemble is caused by internal structural changes. Additional variability results from overall translations and rotations of the molecule. And most experimental data do not provide information to relate the structures to a common reference frame. To report meaningful values of intrinsic dynamics, structural precision, conformational entropy, etc., it is therefore important to disentangle local from global conformational heterogeneity.

Results: We consider the task of disentangling local from global heterogeneity as an inference problem. We use probabilistic methods to infer from the protein ensemble missing information on reference frames and stable conformational sub-states. To this end, we model a protein ensemble as a mixture of Gaussian probability distributions of either entire conformations or structural segments. We learn these models from a protein ensemble using the expectation–maximization algorithm. Our first model can be used to find multiple conformers in a structure ensemble. The second model partitions the protein chain into locally stable structural segments or core elements and less structured regions typically found in loops. Both models are simple to implement and contain only a single free parameter: the number of conformers or structural segments. Our models can be used to analyse experimental ensembles, molecular dynamics trajectories and conformational change in proteins.

Availability: The Python source code for protein ensemble analysis is available from the authors upon request.

Contact: michael.habeck@tuebingen.mpg.de

1 INTRODUCTION

In many situations it is instructive to represent a protein structure as an ensemble of different conformational states rather than a single structure. Such a representation arises naturally in molecular dynamics studies where protein ensembles reflect the internal flexibility of the molecule and can, for example, be used to estimate the conformational entropy. Solution structures determined from nuclear magnetic resonance (NMR) data are also represented as ensembles (Wüthrich, 1986). But here the conformational variability mainly reflects the quality and completeness of the data: NMR or other experimental data are often not sufficient in their own right to determine the complete structure, the remaining uncertainty has to be represented in some way. Although they are time and ensemble averages, NMR measurements are most commonly modelled with a single structure. At such an approximate level of description, it is problematic to distinguish variability due to limitations of the data from true dynamics. Therefore, a standard NMR ensemble will reflect both uncertainty due to data imperfections as well as 'systematic' errors resulting from simplifications in the modelling of the data. Structure ensembles are the agreed-upon way to capture this uncertainty (Havel and Wüthrich, 1985; Markley *et al.*, 1998; Rieping *et al.*, 2005; Snyder *et al.*, 2005; Spronk *et al.*, 2003; Sutcliffe, 1993; Wüthrich, 1986). Also predicted structures often come as ensembles. In homology modelling, for example, it is common practice to calculate several structural models instead of just a single one. Again, the intention is to provide a sort of local reliability measure or error bar in very much the same way as NMR structure ensembles do. Recently, it has been proposed that also crystal structures should be presented as ensembles, particularly if they are based on low-resolution data (Furnham *et al.*, 2006).

Irrespective whether it represents true dynamics or positional imprecision, a structure ensemble needs to be analysed in the light of a model in order to quantify the intrinsic variability one is usually interested in. Without making modelling assumptions it is not clear how to separate local from global variability. Indeed, in a molecular dynamics simulation, the molecule will start to tumble, such that the conformational variability is a mixture of local and global movements. Similarly in NMR structure determination, data such as scalar couplings or nuclear Overhauser effects do not provide information to relate the ensemble members to a common reference frame. An exception are orientational data such as residual dipolar couplings (RDCs). These relate the structures to an external reference frame provided by the alignment tensor. However, the implied superposition will mix global with local dynamics, because RDCs are time- and ensemble-averaged.

The ensemble members need to be superimposed before the intrinsic variability can be quantified in a meaningful way. But which parts should be optimally superimposed is not clear. Naively, one would consider the entire polypeptide chain and find the optimal superposition by least-squares fitting. However, one will then risk to distribute the enhanced variability in loop regions over the entire protein chain. Also, the ensemble may comprise subgroups of structures that fluctuate around structurally distinct, but well-defined conformers. Another difficulty arises when conformational change

*To whom correspondence should be addressed.

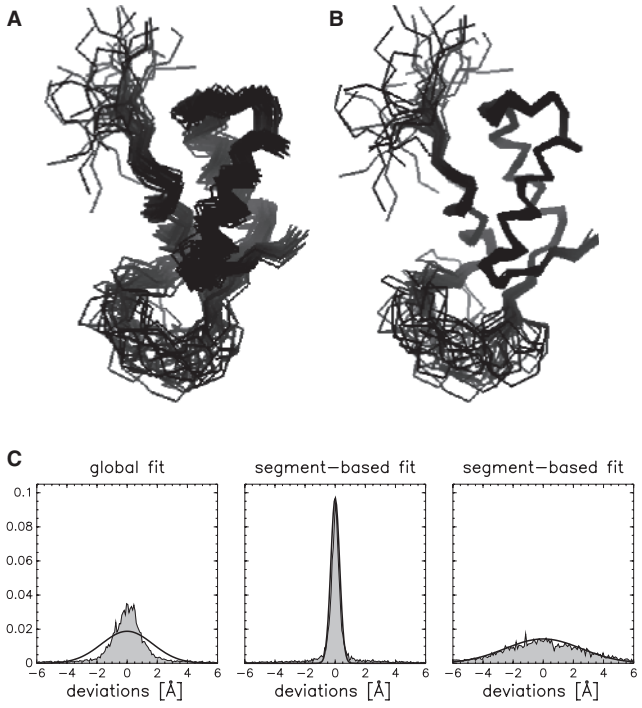

Fig. 1. Problems with global superposition of protein structure ensembles. (**A**) protein structure ensemble 1FOX superimposed by least-squares fitting over the entire chain. (**B**) superposition based on segments comprising residues 7–21 and residues 32–76 only. (**C**) comparison of empirical histograms of positional deviations when using a global superposition (left) and a segment-based superposition (middle and right). The middle and right panels show the differences within and outside the segments used for superposition, respectively. Black curves indicate fitted Gaussian distributions.

involves a hinge motion of structurally rigid domains (Gerstein *et al.*, 1994), which can also be observed in experimental ensembles (Snyder and Montelione, 2005). A *global* superposition of all atoms will be misleading in this case (Arnold and Ornstein, 1997).

Figure 1 illustrates some of these issues for an NMR ensemble of the C-terminal domain of 50S ribosomal protein L11 (PDB accession code 1FOX): panel A shows a structure ensemble based on a superposition of all residues. Panel B shows the same ensemble superimposed onto the segments spanning residues 7–21 and 32–76, which make up 75% of the entire chain. The average root mean square deviation (RMSD) in Cα positions of the ensemble members to the average structure is 2.9 Å for the global superposition and 3.3 Å for the segment-based superposition. However, the RMSD of the well-fitting segments alone is 1.2 Å for the global and 0.4 Å for the segment-based superposition. These observations indicate that there is the need to balance the number of positions taken into account for determining the optimal superposition against the tightness of the fit in the well-matching regions. The methods outlined in this article perform this balancing in an automated, model-driven way.

1.1 A probabilistic model for structure ensembles

The raw ensemble is not useful unless we analyse it in the light of a model. The model makes assumptions that might be partially idealistic or unrealistic. Therefore, we might have to consider several models and compare them against each other. A technical problem is to find the best fitting model. We use Bayesian inference techniques to address these problems. We consider the protein ensemble as a statistical sample drawn from an unknown underlying conformational distribution. We introduce two probabilistic models to describe this underlying distribution: the first model assumes that the ensemble can be grouped into sub-ensembles that fluctuate around distinct conformational states. The second model identifies rigid structural parts and sub-elements that are stable intrinsically but flexible in their relative orientation. Both models are formulated as finite Gaussian mixture models (Titterington *et al.*, 1985).

To motivate the use of mixture models for protein ensemble analysis, let us come back to Figure 1C showing the empirical distributions of positional deviations for a global and a segment-based superposition (i.e. the histograms of coordinate differences between the ensemble members and the average structure). These histograms are compared to the theoretical curves based on a fit of a Gaussian distribution. It becomes evident that in case of a global superposition the distribution of positional differences is only poorly captured by a single Gaussian. If the superposition is based on segments, a Gaussian is a much better model for describing positional deviations: the middle panel of Figure 1C shows the distribution within the segments used for superposition. The distribution of differences is very narrow and can be fit with a Gaussian. But also the atoms that are outside the segments can be modelled using a Gaussian distribution, now with a broader width. Therefore, the deviations of all atoms can be parameterized using a mixture of two Gaussians: the first component is narrow and carries 75% of the total probability mass, whereas the second component is fairly broad and accounts for deviations in the remaining 25% of all positions.

1.2 Generative model

A protein ensemble consists of M members; the atom positions of the m-th member are represented by a $N \times 3$ matrix X_m where, N is the number of atoms; the three-dimensional vector X_{mn} denotes the position of the n-th atom in the m-th structure. Both models assume that atom positions can be related to one of K unknown conformations encoded in the $N \times 3$ matrices Y_k. The fundamental generative model is

$$X_{mn} \approx R_{mk} Y_{kn} + t_{mk}. \tag{1}$$

That is, the coordinates of the n-th atom in the m-th structure have been generated by globally transforming the corresponding position in the k-th conformer/segment. The global transformation relating the m-th structure to the k-th conformer/segment is described by the rotation matrix R_{mk} and the translation vector t_{mk}. This transformation is the same for all atoms within an ensemble member assigned to the k-th conformer/segment. The rotation matrices R_{mk} and translation vectors t_{mk} need to be introduced to optimally superimpose the m-th ensemble member onto the k-th conformer/segment. Such a superposition is necessary because a common reference frame of all ensemble members is usually not known. Further, we will assume that the above relation is not exactly valid but within some error. We model this error using an isotropic Gaussian distribution with no correlations or atom dependence. The spread about the k-th conformer/segment is quantified by the SD σ_k.

In the first model, which we will refer to as the conformer model, Y_k are the stable conformational sub-states in the ensemble. Each ensemble member X_m can be attributed to one of the K conformers. The fraction of ensemble members that are populating the k-th sub-ensemble is w_k. In the second model, which we will refer to as the segment model, Y_k encode the structural segments. Each atom with position X_{mn} can be attributed to one of K segments for all M ensemble members. The fraction of atoms that belong to the k-th segment is w_k.

Both models are formulated as Gaussian mixture models. The probability of the entire ensemble under the conformer model is:

$$p(X|Y,R,t,w,\sigma) = \prod_{m=1}^{M}\sum_{k=1}^{K} w_k \prod_{n=1}^{N} \frac{e^{-\Delta^2_{mnk}/2\sigma_k^2}}{(2\pi\sigma_k^2)^{3/2}}; \tag{2}$$

under the segment model this probability is:

$$p(X|Y,R,t,w,\sigma) = \prod_{n=1}^{N}\sum_{k=1}^{K} w_k \prod_{m=1}^{M} \frac{e^{-\Delta^2_{mnk}/2\sigma_k^2}}{(2\pi\sigma_k^2)^{3/2}}; \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean norm and $\Delta_{mnk} = \|X_{mn} - R_{mk}Y_{kn} - t_{mk}\|$ is the distance between the observed and the idealized position of the n-th atom in the m-th ensemble member assuming that it has been generated from the k-th conformer under model (1). Here, Y,R,t,w,σ are parameter sets, i.e. $Y = \{Y_1,\ldots,Y_K\}$, $w = \{w_1,\ldots,w_K\}$, etc. In both models, the proportions w_k are non-negative and must sum to one. Note, that the only formal difference between model (2) and (3) is that the product over all N atoms and the product over all M ensemble members are interchanged. In the case of a single conformer/segment, i.e. $K = 1$, both models are identical and constitute the limiting case of standard superposition.

We use the expectation–maximization (EM) algorithm (Dempster *et al.*, 1977) to infer the conformer and the segment model from a protein ensemble. The only parameter that needs to be set is the number of conformers in case of the first model and the number of segments in case of the second model. We apply both models to NMR ensembles and experimental protein structures exhibiting conformational change.

2 METHODS

2.1 Expectation–maximization

We have to estimate all unknown parameters Y,R,t,w,σ from the protein ensemble. That is, given an ensemble of M structures there are $3NK + 6MK + 2K - 1$ parameters that need to be determined. This can be accomplished efficiently by using the EM algorithm (Dempster *et al.*, 1977). The optimal parameters maximize the probability of observing the protein ensemble under the assumed model. That is, we have to maximize the log-likelihood $\log p(X|Y,R,t,w,\sigma)$ as a function of Y,R,t,w,σ. This leads to a non-linear optimization problem (Bishop, 1995). The idea of the EM algorithm is to simplify the maximization of $\log p(X|Y,R,t,w,\sigma)$ by the introduction of auxiliary variables. These auxiliary variables are z_{mk} and z_{nk} for model (2) and (3), respectively, and are treated as 'missing data' in the EM framework. The binary variables z_{mk} indicate if the m-th structure is assigned to the k-th conformer, in which case $z_{mk} = 1$. The constraint $\sum_k z_{mk} = 1$ ensures that the structure must be assigned to one of the K conformers. Analogously, for the segment model $z_{nk} = 1$ indicates that the n-th atom is part of the k-th structural segment. Here, we have the constraint $\sum_k z_{nk} = 1$ meaning that an atom must belong to one of the K segments.

The EM algorithm works by cycling through an expectation step (E-step) used to estimate the assignment variables (either z_{mk} or z_{nk} depending on the model) and a maximization step (M-step) for estimating the model parameters Y,R,t,w,σ. The E-step estimates the auxiliary variables by calculating their expectation value for given model parameters. For the conformer model we have:

$$z_{mk} = \frac{w_k \sigma_k^{-3N} e^{-\Delta^2_{mk}/2\sigma_k^2}}{\sum_k w_k \sigma_k^{-3N} e^{-\Delta^2_{mk}/2\sigma_k^2}} \tag{4}$$

where $\Delta^2_{mk} = \sum_n \Delta^2_{mnk}$, and for the segment model:

$$z_{nk} = \frac{w_k \sigma_k^{-3M} e^{-\Delta^2_{nk}/2\sigma_k^2}}{\sum_k w_k \sigma_k^{-3M} e^{-\Delta^2_{nk}/2\sigma_k^2}} \tag{5}$$

where $\Delta^2_{nk} = \sum_m \Delta^2_{mnk}$; in both steps the distances Δ_{mnk} are evaluated for given Y,R and t. Note that the E-step relaxes the condition that the auxiliary parameters are binary valued: In the EM framework, z_{mk} and z_{nk} can take any real value between zero and one. But the normalization constraints for the rows of the assignment matrix Z are maintained.

The computation of the assignment variables, z_{mk} for the conformer model and z_{nk} for the segment model, mainly determines the complexity of the EM algorithms. Since in both cases the calculation of the distances Δ_{mnk} is required [cf. Equations (4) and (5)], both models have the same complexity $O(MNK)$, which was also confirmed empirically. Given Z, it is straightforward to update the parameters of primary interest, i.e. Y,R,t,w,σ. The M-step of the conformer and segment model will be explained in the next two subsections. The EM algorithm is initialized randomly and proceeds until a convergence criterion based on the log-likelihood is reached.

2.2 M-step of the conformer model

The joint likelihood of observing the given ensemble and a specific assignment z_{mk} is:

$$p(X,Z|Y,R,t,w,\sigma) = \prod_k \left[\frac{w_k}{(2\pi\sigma_k^2)^{3N/2}}\right]^{n_k} e^{-\sum_m z_{mk}\Delta^2_{mk}/2\sigma_k^2}; \tag{6}$$

where $n_k = \sum_m z_{mk}$. In the M-step, we maximize this probability for given Z as a function of w,Y,R,t,σ to obtain:

$$w_k = n_k / \sum_k n_k = n_k / M. \tag{7}$$

For every component k, the parameters Y_k, R_{mk}, t_{mk} are obtained by minimizing the loss:

$$L(Y_k, R_{\cdot k}, t_{\cdot k}) = \sum_{m,n} z_{mk} \|X_{mn} - R_{mk}Y_{kn} - t_{mk}\|^2. \tag{8}$$

Minimization with respect to t_{mk} gives:

$$t_{mk} = \frac{1}{N}\sum_n (X_{mn} - R_{mk}Y_{kn}). \tag{9}$$

If we insert this estimate into the loss (8) and omit constant terms, we obtain a loss for the remaining parameters Y_k and R_{mk}:

$$L(Y_k, R_{\cdot k}) = n_k \mathrm{tr}[Y_k^T H Y_k] - 2\mathrm{tr}[Y_k^T H \sum_m z_{mk} X_m R_{mk}] \tag{10}$$

where the $N \times N$ matrix H is the centering matrix with elements $H_{ij} = \delta_{ij} - 1/N$. To derive this loss function, we made use of the orthogonality constraint on the rotation matrices R_{mk}.

The problem with loss function (10) is that the parameters Y_k and R_{mk} become intertwined. Therefore, we optimize iteratively: a minimization with respect to Y_k for a given R_{mk} is followed by an optimization of the R_{mk} for the new Y_k. Minimization of the loss function (10) with respect to Y_k gives:

$$Y_k = \frac{1}{n_k}\sum_m z_{mk} H X_m R_{mk}. \tag{11}$$

The conformers calculated by this formula will always be centred, because $\sum_i H_{ij} = 0$ for all j. Therefore, if we center all ensemble members X_m right

from the start, the translation vectors t_{mk} will always be zero and can be neglected. For given Y_k, the function that we need to maximize to obtain R_{mk} is

$$\mathrm{tr}[Y_k^T H X_m R_{mk}] \tag{12}$$

subject to an orthogonality constraint. That is, R_{mk} is the rotation matrix that is nearest to $(X_m^T H Y_k)$. This matrix nearness problem can be solved by calculating the polar decomposition of $X_m^T H Y_k$ (Higham, 1989). This is in fact identical to applying the Kabsch algorithm (Kabsch, 1976) to every centred pair of X_m and Y_k.

The remaining parameters σ are estimated as:

$$\begin{aligned}\sigma_k^2 &= \frac{1}{3Nn_k}\sum_{m,n} z_{mk}\|X_{mn}-R_{mk}Y_{kn}-t_{mk}\|^2 \\ &= \frac{1}{3Nn_k}\sum_{m} z_{mk}\Delta_{mk}^2.\end{aligned} \tag{13}$$

The above equations show that the conformer model can be viewed as a soft clustering of the ensemble members. The variables z_{mk} are membership probabilities for each structure m and each cluster k.

2.3 M-step of the segment model

For given assignments z_{nk} (5), the likelihood of the ensemble under the second model is:

$$p(X,Z|Y,R,t,w,\sigma)=\prod_k\left[\frac{w_k}{(2\pi\sigma_k^2)^{3M/2}}\right]^{n_k} e^{-\sum_n z_{nk}\Delta_{nk}^2/2\sigma_k^2}; \tag{14}$$

where now $n_k=\sum_n z_{nk}$. We maximize this probability for given Z as a function of w,Y,R,t,σ to obtain:

$$w_k = n_k/\sum_k n_k = n_k/N. \tag{15}$$

For every segment k, the parameters Y_k, R_{mk}, t_{mk} are obtained by minimizing the loss:

$$L(Y_k,R_{.k},t_{.k})=\sum_{mn} z_{nk}\|X_{mn}-R_{mk}Y_{kn}-t_{mk}\|^2. \tag{16}$$

In contrast to (8), this is a *weighted* fit between X_m and Y_k: each atom position n contributes with weight z_{nk} to the goodness of fit. The weighting of positions with z_{nk} guarantees that for determining the transformation to the k-th structure only the segment positions are taken into account. That is, the structures are placed in a local reference frame attached to the segment. minimization with respect to t_{mk} gives:

$$t_{mk}=\frac{1}{N}\sum_n z_{nk}(X_{mn}-R_{mk}Y_{kn})=\overline{X}_{mk}-R_{mk}\overline{Y}_k \tag{17}$$

where, $\overline{X}_{mk}=\frac{1}{N}\sum_n z_{nk}X_{mn}$ and $\overline{Y}_k=\frac{1}{N}\sum_n z_{nk}Y_{kn}$. If we insert this estimate into the loss (16) and consider terms depending on the rotations only, we observe that for each rotation R_{mk}, we have to maximize:

$$\mathrm{tr}\left[R_{mk}\sum_n z_{nk}(Y_{kn}-\overline{Y}_k)(X_{mn}-\overline{X}_{mk})^T\right]\equiv\mathrm{tr}[R_{mk}S_{mk}]. \tag{18}$$

Again, this is a matrix nearness problem which we solve by computing the polar decomposition of S_{mk}. Minimization of the loss function (16) with respect to the segments Y_k for fixed rotations and translations yields:

$$Y_{kn}=\frac{1}{M}\sum_m R_{mk}^T(X_{mn}-t_{mk}). \tag{19}$$

That is, we estimate Y_k, R_{mk}, t_{mk} iteratively by cycling through Equations (18), (17) and (19). After convergence, the error parameters σ are obtained from:

$$\begin{aligned}\sigma_k^2 &= \frac{1}{3Mn_k}\sum_{mn} z_{nk}\|X_{mn}-R_{mk}Y_{kn}-t_{mk}\|^2 \\ &= \frac{1}{3Mn_k}\sum_{n} z_{nk}\Delta_{nk}^2.\end{aligned} \tag{20}$$

Let us summarize the differences in the estimation of Y,R,t in both models: in the conformer model (2), the conformers Y_k are *weighted* averages of the ensemble members; the optimal translation and rotation are obtained by an *unweighted* least-squares fit of each member onto each conformer. In the segment model (3), the segments Y_k are *unweighted* averages of the ensemble members; the optimal translation and rotation are obtained by a *weighted* least-squares fit of each member onto each segment where atoms are weighted according to how much they contribute to the segment.

2.4 Initial conditions

The EM algorithm maximizes the likelihood only locally. Therefore, one needs to run the algorithm multiple times from randomly chosen starting conditions. In the conformer model, we choose K ensemble members randomly as initial conformers and then start with the EM iterations. In the segment model, the initialization is less straightforward. One expects structural segments to be contiguous, which is not enforced by our segment mixture model (3). To ensure that neighbouring atoms belong to the same class, we use the following initialization procedure: first we randomly choose K weights $w_k=u_k/\sum_k u_k$ where, u_k are uniformly distributed random numbers between zero and one. We then sample the lengths of the K segments from a multinomial distribution using the weights as probabilities. We assign all atoms within these sampled segments by setting $z_{nk}=1$ and start the EM iterations with the M-step. Another improvement is likely to be achieved by using secondary structure assignments for the initialization of the segment model. However, we have not tested this so far since in our applications the convergence of learning the segment model was sufficiently fast.

3 RESULTS

In the following, we will discuss applications where a mixture model analysis aids in the interpretation of protein structure ensembles. We will first investigate the behaviour of the conformer and the segment model using experimental structure ensembles. We then show for the case of adenylate kinase (ADK) that the segment model can be used to characterize conformational changes in proteins. Finally, we illustrate how to choose K and discuss aspects of assessing precision and accuracy of experimental structures. All results are based on computations performed on Cα coordinates, but other choices for the coordinate sets should perform similarly.

3.1 Conformer model

Let us first address the problem of grouping an ensemble of structures into a predefined number of conformational sub-states. Many algorithms for solving this problem have already been developed. Here, we only mention the NMRCLUST program developed for NMR ensembles (Kelley *et al.*, 1996). NMRCLUST combines a hierarchical clustering method with a heuristic to choose automatically the optimal number of clusters.

Using the conformer model (2), we analysed the NMR ensemble 4HIR comprising 32 structures. The 4HIR ensemble also serves as a test case for NMRCLUST. We find the same sub-groups of structurally similar ensemble members as NMRCLUST. Figure 2A shows the superimposed ensemble obtained with five conformers. Also shown are the conformer structures (Fig. 2B), which differ mainly in one of the loops.

Figure 3 provides additional details on the performance of the algorithm. Figure 3A shows the development of the log-likelihood function during EM. We notice that the log-likelihood function increases monotonically and converges rapidly to its final value. However, for unfavourable starting conditions the algorithm may not be able to find the global maximum of the likelihood function. Therefore, one needs to run the EM iterations several times. In all

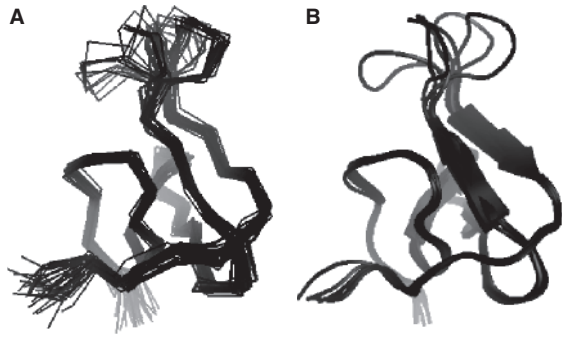

Fig. 2. Conformer analysis of the NMR ensemble 4HIR. (**A**) Structure ensemble comprising 32 structures that were analysed using the conformer model with 5 different conformers. (**B**) Ribbon plots of the conformers estimated with the EM algorithm.

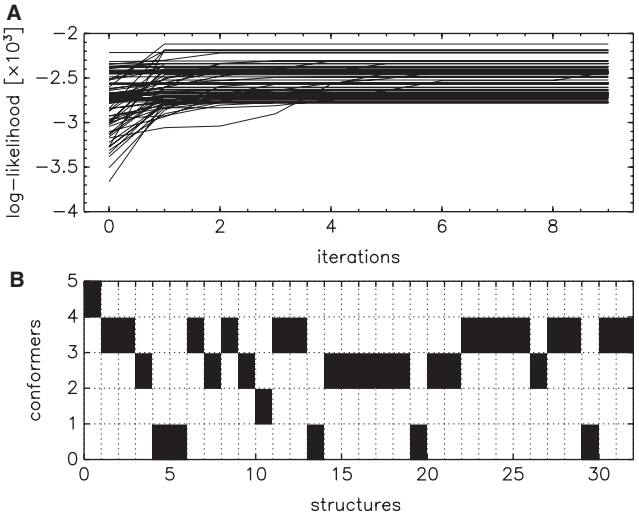

Fig. 3. Algorithmic details of the conformer analysis of 4HIR. (**A**) Development of the log-likelihood function in the EM. (**B**) Final configuration of the assignment variables z_{mk} where m is the index of the ensemble members. Each grid cell corresponds to a particular z_{mk} value. The magnitude of the z_{mk} values is coded in grey levels where black corresponds to $z_{mk}=1$ and white to $z_{mk}=0$.

applications we looked at, we found that a single EM run converges within 20 iterations and that among 50 restarts the optimal solution is usually found. In case of the 4HIR ensemble, five among the 50 fitted models were equivalent to the optimal model. Figure 3B shows the final configuration of the assignment variables z_{mk}. The membership probabilities are coded as grey levels: black meaning $z_{mk}=1$ and white $z_{mk}=0$. The lack of grey grid cells in Figure 3 is indicative of a very sharply defined clustering.

The EM updates are reasonably fast and simple to implement. On a 2.8 GHz pentium 4 processor, 100 EM iterations take 10 CPU seconds for the 4HIR ensemble. The code is implemented in Python and not optimized for speed.

3.2 Segment model

The structures of globular proteins typically consist of rigid or well-defined structural elements and less ordered parts often found in loop or linker regions. Moreover, the rigid parts may be flexible in their relative orientation and thus require separate superposition (Snyder and Montelione, 2005). The segment model (3) can be used to analyse ensembles showing this behaviour. It estimates a partitioning of the protein chain into rigid and non-rigid domains in an automated unsupervised manner. We obtain such a segmentation of the chain by assigning atoms to the segments with largest z_{nk} value. Each region has its own associated rotation and translation. Therefore, there is no risk of loosing information by fitting segments that cannot be superimposed meaningfully. If only a single rigid domain is present, the ensemble can be described with a segment mixture model using two components ($K=2$). The low-variance component then corresponds to the structural core, and the corresponding assignments z_{nk} indicate the partitioning of the structure. If several rigid elements with flexible relative orientations are present, one has to choose $K\geq 2$. Again, the core regions correspond to components with low variance σ_k.

We analysed an NMR ensemble of calcium-free calmodulin (PDB code 1CFC) to demonstrate an analysis based on the segment model. As reported in Kuboniwa *et al.* (1995), 1CFC is composed of an N- and a C-terminal domain which can be superimposed internally but not globally. To model this ensemble we used a segment mixture with two components. Figure 4 shows the resulting structure ensembles superimposed onto the N-terminal segment and onto the C-terminal segment. The estimated spread of the segments is $\sigma_1=0.29$ Å for the N-terminal domain, and $\sigma_2=0.45$ Å for the C-terminal domain. This is consistent with the observation that the C-terminal domain of 1CFC is less well-defined by the NMR data than the N-terminal domain (Kuboniwa *et al.*, 1995).

It has been shown that dynamics and conformational change in proteins is often highly correlated and that functionally important changes occur as concerted motion. It may seem that the segment model neglects these correlations. The *internal* covariance matrix of the k-th segment is indeed $\sigma_k^2 I$ where I is the identity matrix. That is, there are no correlations between atom positions nor atom-specific positional variances. However, the definition of structural correlations depends on the reference frame. Usually, one chooses the origin of the coordinate system at the mean structure. In case of multiple segments, such a unique canonical reference frame does no longer exist (Arnold and Ornstein, 1997).

Let us make this more explicit for the calmodulin example. If one adopts the local coordinate system of the N-terminal segment (i.e. the coordinate system in which the intra-segment correlations of the N-terminal segment are minimal), the atoms of the second segment undergo large, concerted movements. Likewise one can choose the C-terminal segment as reference frame, in which case the largest correlations will occur in the N-terminal domain. Figure 4C and D show the correlation matrices in both reference frames. If one adopts a reference frame attached to a segment, the coordinates of the other segment become highly correlated, whereas the intra-segment correlations are not as strong and mainly concentrate on the diagonal.

The block structure of the correlation matrices directly reflects the segmentation as encoded in the z_{n1} and z_{n2} values. The segment model captures large inter-atomic correlations not explicitly in terms of a full covariance matrix but through the assignment variables

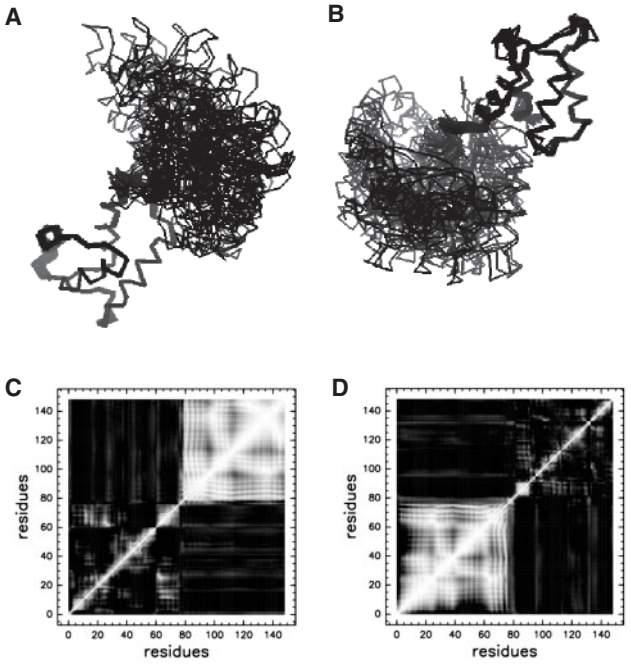

Fig. 4. Segment analysis of the NMR ensemble 1CFC. (**A**) Structure ensemble comprising 25 structures that were analysed using the segment model with two segments superimposed onto the N-terminal segment. (**B**) Superimposition using the C-terminal segment as reference. (**C**) Correlation matrix in the reference frame of the first segment, (**D**) Correlation matrix in the reference frame of the second segment.

z_{nk} and by the sharing rotational and translational degrees of freedom. Atoms that are assigned to the same segment (i.e. they have high z_{nk} values for the same k) are modelled as a rigid body that has some intrinsic flexibility σ_k which is approximated as being constant over the whole segment. This is in spirit similar to translation/libration/screw (TLS) models used in X-ray crystallography (Painter and Merritt, 2006).

3.3 Analysis of conformational change

The segment model can be used to analyse proteins that undergo conformational changes. A well-studied example of a protein manifesting large conformational rearrangements when switching between active and inactive state is ADK. ADK is a nucleoside monophosphate kinase that catalyses the transfer of a phosphoryl group from ATP to AMP. ADK is composed of three domains, the CORE, the LID and the NMP binding domain (Vonrhein *et al.*, 1995). The CORE domain is the main domain, the LID domain binds ATP and the NMP domain binds AMP. In the unligated conformation, the LID and the NMP domain are 'open'. Upon nucleotide binding, large motions of the LID and the NMP domains result in a 'closed' conformation.

Figure 5 shows results for a segment model analysis based on a crystal structure of the closed conformation (PDB code 1AKE) and of the open conformation 4AKE. A three component segment model was fitted to an 'ensemble' containing only these two structures. The resulting segmentation of the structure is shown in Figure 5. The EM algorithm reproduces the domain composition of ADK. The values of the assignment variables z_{nk} are almost binary valued and result

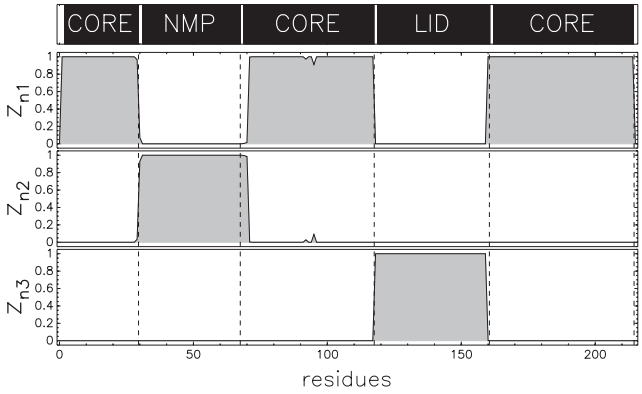

Fig. 5. Estimated domain composition of ADK when applying the segment model to an ensemble consisting of the closed conformation 1AKE and the open conformation 4AKE. The top row and the dashed vertical lines indicate the domain composition of ADK according to Whitford *et al.* (2008). The bottom rows show the estimated assignment z_{nk}. The x-axis is the atom/residue index (one Cα atom per residue), the y-axis indicates the value of the assignment variable z_{nk}.

in a segmentation that is very similar to the domain structure of ADK as defined in Whitford *et al.* (2008).

3.4 Choice of K

The single free parameter of both the conformer and the segment mixture model is K, the number of conformers or segments, respectively. A fully Bayesian approach employing posterior sampling would allow us to select this parameter by model comparison techniques (MacKay, 2003). However, in the optimization approach pursued here we need to resort to heuristics such as cross-validation, which is a well-established technique for estimating hyperparameters (Stone, 1974). We used 10-fold cross-validation to choose K. The idea of cross-validation is to avoid over-fitting by splitting the data into a training set used for learning the probabilistic model and a test set for evaluating the generalization capabilities of the model. The data are partitioned randomly into 10 sets each of which is used as a test set, while the remaining make up the training data. In case of the conformer model, we treat ensemble members as data points, i.e. the training is done on $9M/10$ randomly chosen structures and the testing on the remaining 10%. In case of the segment model, atoms are treated as data points, meaning that for training we leave out the coordinates of $N/10$ randomly chosen atoms in *all* structures of the ensemble and use them only for testing.

Figure 6 shows 10-fold cross-validation curves for the conformer and the segment model. Shown are the average negative log-likelihood values for the training and the test set. As expected, the average negative log-likelihood of the training set attains smaller and smaller values with increasing K indicating the risk of over-fitting. However, the negative log-likelihood of the test set exhibits a broad minimum or reaches a plateau. The cross-validated choice for K is the smallest K (least complex model) for which the minimum or plateau is attained. Figure 6A shows an application of the conformer model to the 4HIR ensemble. Here, the optimal number of conformers is somewhere around $K=5$. This is in accord with the results from Kelley *et al.* (1996) who developed an alternative strategy to determine K. Figure 6B shows an application of the

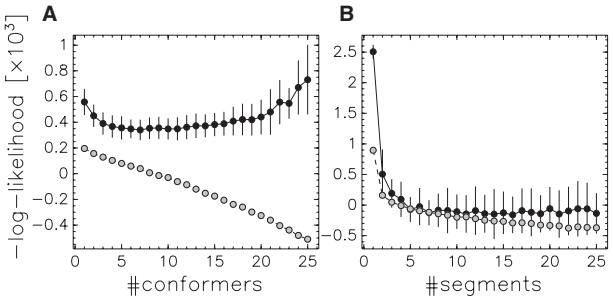

Fig. 6. Cross-validated choice of K. A 10-fold cross-validation analysis was carried out for the conformer and the segment model. Grey dots indicate the average negative log-likelihood values of the training set normalized by the number of data. Black dots are the averaged negative log-likelihood values of the 10 subsets used for testing. For a given number of conformers, an error bar was calculated as the SD in negative log-likelihood over the 10 test sets. (**A**) Cross-validation analysis of the conformer model applied to 4HIR. (**B**) Cross-validation curve of a segment model applied to 1CFC.

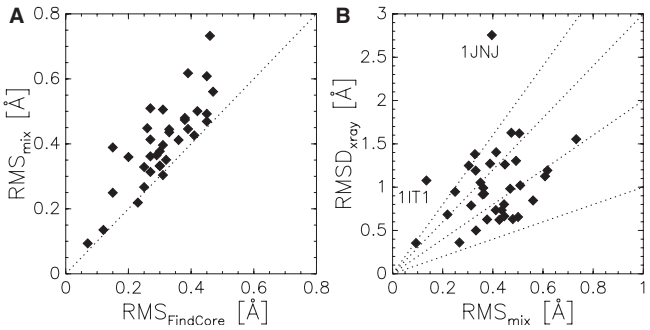

Fig. 7. (**A**) Joint RMSD calculated by FindCore versus the RMS of the low-variance component. (**B**) Precision of the ensemble as evaluated by the RMS (22) versus weighted RMSD to the crystal structure. Dotted lines with slope 1–4. Two outliers are labelled with their PDB codes.

segment model to the 1CFC ensemble. The largest improvement in cross-validated log-likelihood is obtained for $K=2$, which reflects the fact that 1CFC has an internal repeat structure consisting of two structurally similar domains. The variability in the calmodulin ensemble mainly involves a movement of these two rigid domains relative to each other through a 'kinking' of the central helix. This movement can be described approximately by one degree of freedom. For a larger number of segments, the segment model introduces segments mainly in the central helix, i.e. in the linker region between the two domains. For $K>4$ the improvements in cross-validated log-likelihood become insignificant when taking errors into account.

These examples show that it is difficult to automize model selection by cross-validation. The log-likelihood curves often do not exhibit a clear minimum but rather converge to a plateau. Even if a minimum is observed, its depth is small in comparison to the magnitude of the errors. Therefore, even the cross-validated choice of K will be subjective to a certain degree, and probably the safest is to choose the model with the least number of parameters after the elbow region, right at the beginning of the plateau.

3.5 Assessing structural precision and accuracy

One of the goals in NMR structure calculation is to produce ensembles that indicate which coordinates are reliable and which are not, and to derive from this ensemble a measure of precision that correlates with the accuracy of the structure. This latter aspect involves a superposition problem because NMR structure calculations typically do not produce superimposed structures, and depending on the superposition, atoms will change their position to a lesser or greater extent. Therefore, the well-defined positions used for superposition are often hand picked. As a consequence, the resulting measure of ensemble precision is subjective and will depend on the choices made in the superposition. There is no generally accepted definition of structural precision. We argue here that the segment model could constitute a step towards a more objective definition of ensemble precision.

A probabilistic model of a protein structure ensemble encodes a notion of closeness and similarity between structures. Structures that are more likely under the ensemble model are also structurally closer to the ensemble members. The standard measure for comparing protein structures quantitatively is the RMSD of atomic coordinates, most often Cα positions. The corresponding measure to quantify the precision of structure ensembles is the RMS fluctuation over the ensemble. We calculate this precision measure for the probabilistic ensemble models in terms of an average coordinate fluctuation. For the k-th conformer or segment, we have:

$$\mathrm{RMS}_k = \sqrt{\frac{1}{N}\sum_n \langle \|X_n - Y_{kn}\rangle\|^2\rangle_k} = \sqrt{3}\sigma_k \qquad (21)$$

where $\langle\cdot\rangle_k$ denotes an average over the k-th component of the mixture model. The RMS values of several components are combined by averaging the component-wise variances weighted with their mixture proportion w_k:

$$\mathrm{RMS} = \sqrt{3\sum_k w_k \sigma_k^2}. \qquad (22)$$

If one includes only the low-variance components of the segment model in the above sum, this definition is equivalent to the 'joint RMSD', a precision measure calculated by the FindCore method of Snyder and Montelione (2005).

In a recent publication, Andrec *et al.* (2007) analysed protein structure ensembles using FindCore. We trained a two-component segment model for each of the 35 NMR ensembles that constitute the 'Reduced test set' in Andrec *et al.* (2007). For each mixture, we identify the core segment as the component that has the smallest contribution $w_k\sigma_k^2$ to the overall RMS [Equation (22)]. In a few cases where the ensemble is tight, the non-core component may consist of a small set of extremely well-matching positions which are then mistaken to be the core. Therefore, whenever w_k drops below 0.2 we choose the other component to be the core segment. For the core segment, we evaluate the RMS, called $\mathrm{RMS}_{\mathrm{mix}}$ in the following, according to Equation (21). Figure 7 shows that $\mathrm{RMS}_{\mathrm{mix}}$ correlates well with $\mathrm{RMS}_{\mathrm{FindCore}}$, the joint RMSD of the FindCore method [denoted $\mathrm{RMS}_{\mathrm{env}}$ in Andrec *et al.* (2007)]. Particularly, we find that $\mathrm{RMS}_{\mathrm{mix}} \geq \mathrm{RMS}_{\mathrm{FindCore}}$. That is, there is a trend that the RMS values of the segment model are larger than those from FindCore. A comparison of the weights of the core segment with the number of core atoms produced by FindCore shows that FindCore tends to define smaller cores than the segment model.

Another issue in NMR structure determination is to assess the accuracy of a solution structure. Most often the crystal structure, if available, is taken as a reference (Spronk *et al.*, 2003). To compare the crystal structure with the NMR ensemble, we first estimate the optimal rotation and translation of the crystal structure using the assignment matrix and segment coordinates trained on the NMR ensemble. We then update the σ_k values and define as a measure of accuracy $\mathrm{RMSD}_{\mathrm{xray}} = \sqrt{3}\sigma_{\mathrm{core}}$ where σ_{core} is the deviation from the core segment used in the definition of $\mathrm{RMS}_{\mathrm{mix}}$.

The Figure 7B shows that there is a moderate correlation of 0.3 between the precision of the ensemble as evaluated by $\mathrm{RMS}_{\mathrm{mix}}$ and the accuracy $\mathrm{RMSD}_{\mathrm{xray}}$. Thus the precision of an ensemble evaluated by $\mathrm{RMS}_{\mathrm{mix}}$ gives some indication on the accuracy of the solution structure: for 16 structure pairs (46%) $\mathrm{RMSD}_{\mathrm{xray}} \leq 2\mathrm{RMS}_{\mathrm{mix}}$, for 25 pairs (71%) $\mathrm{RMSD}_{\mathrm{xray}} \leq 3\mathrm{RMS}_{\mathrm{mix}}$ and for 33 pairs (94%) $\mathrm{RMSD}_{\mathrm{xray}} \leq 4\mathrm{RMS}_{\mathrm{mix}}$. There are two outliers, 1JNJ and 1IT1, for which $\mathrm{RMSD}_{\mathrm{xray}} > 4\mathrm{RMS}_{\mathrm{mix}}$. One of these ensembles, 1JNJ, is also noted to be problematic in Andrec *et al.* (2007). Visual inspection indicates that the main reason for the discrepancy seems to be a loop that shows little conformational variability in the NMR ensemble, but a systematic difference from the crystal structure. This region may have been over-restrained in the structure calculation. When we ignore this outlier the correlation increases to 0.4. This is an improvement over the precision measure $\mathrm{RMS}_{\mathrm{FindCore}}$ found by the FindCore method: there the correlation is 0.1 with and without 1JNJ. Overall, these findings indicate that the segment model might provide a useful definition of structural precision that is also indicative of the accuracy of the structure. A more thorough investigation based on the full set of 148 structure pairs in Andrec *et al.* (2007) should provide more insights into these issues.

4 DISCUSSION

We introduced two probabilistic models for describing and analysing protein structure ensembles. Both models employ Gaussian mixtures to model the conformational distribution underlying a protein structure ensemble. Each of the two mixture models is a weighted superposition of unimodal Gaussian components with a global isotropic spread σ_k. The components of the first mixture model are stable conformers around which protein conformations are more densely distributed. This model can be used to cluster protein structure ensembles. The components of the second mixture model are structural elements into which the protein chain can be segmented. These segments correspond to well-defined and less well-defined regions in experimental structure ensembles, or to moveable rigid parts in protein structures undergoing conformational change. Each segment is superimposed independently, thereby accounting for the fact that segments are rigid internally but flexible in their relative orientation. The segment model identifies these parts in an automated unsupervised fashion.

Many algorithms for clustering protein structure ensembles have been developed. Often these algorithms rely on heuristics and require several parameters to be set. The conformer model has the advantage that it has a sound probabilistic foundation and interpretation. The only parameter that needs to be specified is the number of conformers; it can be determined using cross-validation.

Algorithms for segmenting protein structures into rigid domains that are moveable relative to each other have been developed mainly in the context of studying conformational changes in protein structures. In the context of assessing the precision of NMR structures, the most recent method is FindCore by Snyder and Montelione (2005) who build on the method of Kelley *et al.* (1997). In essence, FindCore is a two-step procedure: in the first step, structurally well-defined 'core' atoms are identified using a distance-based order parameter. In the second step, the core atoms are partitioned into segments that require separate superposition ('RMSD-stable domains'). The segment model serves the same purpose but combines these two steps into a single one. For $K = 2$, we obtain a classification of the atoms into core (belonging to the low-variance component) and non-core atoms (high-variance component). For $K \geq 2$ with several low-variance components, the ensemble encompasses several well-defined domains that require separate superposition. Snyder and Montelione also defined a new measure of structural precision, the 'joint RMSD'. A correlated measure is provided by the segment model. The segment mixture model is a simple and principled alternative to the FindCore algorithm. It has the advantage of being a physically motivated probabilistic model with easily interpretable parameters.

Recently, Theobald and Wuttke (2006a) introduced a Bayesian framework for protein ensemble analysis that also utilizes an EM algorithm to learn the model parameters. This approach has been implemented in the THESEUS software (Theobald and Wuttke, 2006b). At the heart of THESEUS is a generative model similar to ours [Equation (1)] except for the fact that only a single rotation and translation is used (i.e. $K = 1$). Another important difference to the models presented here is how coordinate deviations are modelled. THESEUS uses a unimodal multivariate Gaussian distribution that also takes into account correlations between the positions of different atoms. To some extent, the segment model also accounts for differences in the precision of atom positions (i.e. the diagonal entries of the covariance matrix): given the assignments z_{nk}, atoms have the internal precision $\sum_k z_{nk}/\sigma_k^2$ which can vary from atom to atom. However, inter-atomic correlations are not modelled explicitly and only captured in the sharing of rotational and translational degrees of freedom (cf. Section 3.2). Although correlations are not explicitly modelled, the superposition produced by the segment model are very similar to those obtained by THESEUS. In case of the 1FOX ensemble shown in Figure 1, the superimposed ensembles look almost identical (data not shown). However, in cases where the ensemble contains multiple rigid domains that are flexible relative to each other, THESEUS runs into problems, because all coordinates are related to the same global reference frame. For the calmodulin ensemble 1CFC, for example, THESEUS finds only the superposition obtained in the reference frame of the N-terminal segment (data not shown). As a consequence, the resulting ensemble is very similar to the ensemble shown in Figure 4A but the internal similarity of the C-terminal segment is completely missed.

It should be possible to combine aspects of the model underlying THESEUS and of the mixture model approach. One problem could be the explosion of the number of parameters to be estimated. If the number of parameters exceeds the number of data points, the role of prior distributions becomes more important. Here, we have only used uninformative 'flat' prior distributions which fail to encode obvious properties such as correlations between the segment assignments z_{nk} of neighbouring residues. The 1FOX example also shows that identifiability of the parameters may be a problem when combining the segment model with THESEUS: both methods give

a very similar superposition and therefore map to similar likelihood values. Again, a remedy is to use more informative prior distributions over the model parameters. An extension of the conformer model would result in a mixture of multivariate Gaussians with different covariance matrices. One simple requirement for the prior of an extended conformer model could be that the conformers should be farer apart than the scale of fluctuations encoded in the covariance matrix. An extension of the segment model is less obvious and will be the subject of future research.

In the current implementation, we use an optimization framework to learn the mixture models from a protein ensemble. In ongoing work, we are developing a Gibbs sampling procedure that estimates the model parameters in a fully probabilistic way. Such an approach will have all benefits of a fully Bayesian treatment in providing not only parameter estimates but also estimates of their uncertainty. It will then be possible to compare the likelihood of modelling an ensemble with a conformer or a mixture model. Moreover, by using an infinite Gaussian mixture model (Rasmussen, 2000) it is possible to alleviate the necessity of choosing the number of components by personal judgement or cross-validation. How many conformers or segments should be used to describe the ensemble, will then be determined automatically by the sampling procedure.

Conflict of Interest: none declared.

REFERENCES

Andrec,M. *et al*. (2007) A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. *Proteins Struct. Funct. Bioinform.*, **69**, 449–465.

Arnold,G.E. and Ornstein,R.L. (1997) Molecular dynamics study of time-correlated protein domain motions and molecular flexibility: cytochrome P450BM-3. *Biophys. J.*, **73**, 1147–1159.

Bishop,C. (1995) *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. B*, **39**, 1–38.

Furnham,N. *et al*. (2006) Is one solution good enough? *Nat. Struct. Biol.*, **13**, 184–185.

Gerstein,M. *et al*. (1994) Structural mechanisms for domain movements in proteins. *Biochemistry*, **33**, 6739–6749.

Havel,T.F. and Wüthrich,K. (1985) An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformations in solution. *J. Mol. Biol.*, **182**, 281–294.

Higham,N.J. (1989) Matrix nearness problems and applications. In Gover,M.J.C. and Barnett,S. (eds) *Applications of Matrix Theory*. Oxford University Press, Oxford, UK, pp. 1–27.

Kabsch,W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, **A32**, 922–923.

Kelley,L.A. *et al*. (1996) An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Eng.*, **9**, 1063–1065.

Kelley,L.A. *et al*. (1997) An automated approach for defining core atoms and domains in an ensemble of NMR-derived protein structures. *Protein Eng.*, **10**, 737–741.

Kuboniwa,H. *et al*. (1995) Solution structure of calcium-free calmodulin. *Nat. Struct. Biol.*, **2**, 768–776.

MacKay,D.J.C. (2003) *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, Cambridge, UK.

Markley,J.L. *et al*. (1998) Recommendations for the presentation of NMR structures of proteins and nucleic acids. *J. Mol. Biol.*, **280**, 933–952.

Painter,J. and Merritt, E.A. (2006) Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr. D Biol. Crystallogr.*, **62**, 439–450.

Rasmussen,C.E. (2000) The infinite gaussian mixture model. In Solla,S.A.T.K. and Müller,K.R. (eds) *NIPS 12*. MIT Press, Cambridge, MA, pp. 554–560.

Rieping,W. *et al*. (2005) Inferential structure determination. *Science*, **309**, 303–306.

Snyder,D.A. and Montelione,G.T. (2005) Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. *Proteins Struct. Funct. Bioinform.* **59**, 673–686.

Snyder,D.A. *et al*. (2005) Assessing precision and accuracy of protein structures derived from NMR data. *Proteins Struct. Funct. Bioinform.* **59**, 655–661.

Spronk,A.E.M. *et al*. (2003) The precision of NMR structure ensembles revisited. *J. Biomol. NMR*, **25**, 225–234.

Stone,M. (1974) Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. B*, **36**, 111–147.

Sutcliffe,M.J. (1993) Representing an ensemble of NMR-derived protein structures by a single structure. *Protein Sci.*, **2**, 936–944.

Theobald,D.L. and Wuttke,D.S. (2006a) Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian procrustes problem. *Proc. Natl Acad. Sci. USA*, **103**, 18521–18527.

Theobald,D.L. and Wuttke,D.S. (2006b) THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics*, **22**, 2171–2172.

Titterington,D.M. *et al*. (1985) *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Vonrhein,C. *et al*. (1995) Movie of the structural changes during a catalytic cycle of nucleoside monophosphate kinases. *Structure*, **3**, 483–490.

Whitford,P.C. *et al*. (2008) Conformational transitions in adenylate kinase. Allosteric communication reduces misligation. *J. Biol. Chem.*, **283**, 2042–2048.

Wüthrich,K. (1986) *NMR of Proteins and Nucleic Acids*. John Wiley, New York.