PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2016, **18**, 5832

Bayesian inference of protein ensembles from SAXS data

L. D. Antonov,*[a] S. Olsson,[bc] W. Boomsma[d] and T. Hamelryck*[a]

Received 17th August 2015,
Accepted 28th October 2015

DOI: 10.1039/c5cp04886a

www.rsc.org/pccp

The inherent flexibility of intrinsically disordered proteins (IDPs) and multi-domain proteins with intrinsically disordered regions (IDRs) presents challenges to structural analysis. These macromolecules need to be represented by an ensemble of conformations, rather than a single structure. Small-angle X-ray scattering (SAXS) experiments capture ensemble-averaged data for the set of conformations. We present a Bayesian approach to ensemble inference from SAXS data, called Bayesian ensemble SAXS (BE-SAXS). We address two issues with existing methods: the use of a finite ensemble of structures to represent the underlying distribution, and the selection of that ensemble as a subset of an initial pool of structures. This is achieved through the formulation of a Bayesian posterior of the conformational space. BE-SAXS modifies a structural prior distribution in accordance with the experimental data. It uses multi-step expectation maximization, with alternating rounds of Markov-chain Monte Carlo simulation and empirical Bayes optimization. We demonstrate the method by employing it to obtain a conformational ensemble of the antitoxin PaaA2 and comparing the results to a published ensemble.

Introduction

Recent years have witnessed increased recognition of the ubiquity and importance of intrinsically disordered proteins (IDPs) and multi-domain proteins with disordered intra-domain linker regions (IDRs).[1–5] Long unstructured regions can be found in more than half of eukaryotic proteins and at least 25% are completely disordered.[6] It is becoming evident that structural plasticity plays an important role in the function of biological macromolecules, *e.g.* in areas such as transcription regulation, cell signaling, and the function of chaperones.[1,7,8] Misfolding and aggregation of IDPs are associated with many human diseases, such as Alzheimer's and Parkinson's.[9,10] These flexible proteins comprise dynamic systems that explore a conformational space that cannot be adequately described by a single state, but requires an ensemble of conformations.

Small-angle X-ray scattering (SAXS) and nuclear magnetic resonance (NMR), as solution structure methods, are well-suited to characterize structural ensembles. SAXS, in particular, is a powerful technique, yielding averaged, low-resolution structural information across multiple spatial orders of magnitude. Combined with appropriate ensemble-based computational methodology, it could allow for the characterization of IDP and IDR flexibility not accessible through NMR spectroscopy or X-ray crystallography alone.[11,12]

Current computational methods aim to recover a representative ensemble as a subset of conformations from a large pool of candidate structures, based on experimental SAXS data.[11–14] The initial pool of structures is generated from either knowledge- or physics-based models. A common assumption in these approaches is that the structural ensemble can be represented accurately by a weighted average of discrete conformations. Small sets of conformers are typically used as an approximation,[15] in order to avoid overfitting and to reduce the computational load. The Ensemble Optimization Method (EOM) uses a genetic algorithm with a predefined number of structures of equal weight for ensemble selection,[16] while the improved EOM 2.0 optimizes individual weights together with an ensemble size within a customizable range.[12] Minimal Ensemble Search (MES) uses a genetic algorithm on a population of ensembles of sizes between 2 and 5 structures.[17] In the Basis-Set Supported SAXS (BSS-SAXS) approach, conformations are assigned to a small number of clusters, first by RMSD and then by scattering pattern similarity, after which a Bayesian MC algorithm is used to determine the cluster weights.[18] The Ensemble Refinement of SAXS (EROS) method similarly uses RMSD clustering followed by maximum entropy[19] cluster weight optimization.[20] In the program ENSEMBLE, a predetermined number of conformations is employed, with either equal or varied weights, and the ensemble

[a] *Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, DK-2200 Copenhagen N, Denmark. E-mail: lubo.antonov@gmail.com, thamelry@binf.ku.dk*
[b] *Laboratory of Physical Chemistry, Swiss Federal Institute of Technology, ETH-Hönggerberg, Vladimir-Prelog-Weg 2, CH-8093 Zürich, Switzerland*
[c] *Institute for Research in Biomedicine, Università della Svizzera Italiana, Via Vincenzo Vela 6, CH-6500 Bellinzona, Switzerland*
[d] *Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, DK-2200 Copenhagen N, Denmark*

is optimized using axial descent or simulated annealing algorithms.[21–24] The Sparse Ensemble Selection (SES) method reformulates the ensemble selection problem as a linear least-squares problem that optimizes the weights of all structures in the initial pool, yielding a sparse ensemble of conformations.[25] Many of these approaches limit the ensemble size explicitly while others, *e.g.* BSS-SAXS and SES, use sparsity-inducing algorithms. However, in flexible systems, such as IDPs and IDRs, a small number of conformations may not adequately explain the data.[25]

In contrast, a number of methodologies that have been applied to NMR data eschew reweighing of structures in favor of probabilistic sampling according to the maximum entropy principle.[15,26–32] In this manner, an ensemble-based description is obtained that balances the experimental data with prior information, typically encoded in a force field.

Here, we approach SAXS data in a similar manner, resulting in a new method for inference of structural ensembles, called Bayesian Ensemble SAXS (BE-SAXS). BE-SAXS combines a generative, fine-grained (*i.e.* atomic-level) model of protein structure with experimental SAXS data. Through an iterative expectation maximization (EM) algorithm the method adapts a prior distribution concerning protein structure in atomic detail to match the SAXS ensemble average, within the experimental uncertainty. The resulting posterior distribution takes the ensemble nature of the data into account and correctly balances information present in both the force field and the experimental data. The number of model parameters depends only on the number of experimental observables and representative structures can be sampled *a posteriori*. Furthermore, since conformations are not restricted to a subset of an initial pool of structures, bias attributable to the initial selection process and limited sampling is avoided.

We apply the BE-SAXS method to SAXS data for the flexible antitoxin PaaA2 and show substantial agreement between the recovered distribution of conformations and the published structural ensemble of the protein. These results illustrate the utility of the method in elucidating the flexibility of partially- or fully-disordered proteins.

Theory and methods

Inferential structural ensemble determination

In probabilistic inferential structure determination (ISD) the goal is to establish a posterior distribution $p(\mathbf{x}|\mathbf{d},\boldsymbol{\sigma}^2)$ of protein conformations $\mathbf{x}$, given some experimental data $\mathbf{d}$ with experimental errors $\boldsymbol{\sigma}^2$.[33] The classic ISD approach assumes that the experimental data represent a single conformation. Consequently, application of the method to disordered systems, which are characterized by highly heterogeneous ensembles, may give misleading results.[27] Such flexible systems require an ensemble-based inference method.

SAXS experiments measure the temporal (*i.e.* over the measurement duration) and ensemble average of the X-ray scattering from all orientations and conformations of the proteins in a solution. Therefore, $\mathbf{d}$ is a noisy observation of the true ensemble average $\mathbf{e}$ of the scattering $\mathbf{f}$ for each individual conformation of a protein. $\mathbf{f}$ is a lower-dimensional projection, or coarse-grained representation, of the fine-grained variable $\mathbf{x}$, through a deterministic function, $\mathbf{f} \equiv h(\mathbf{x})$. A model for such ensemble-averaged data was previously expressed as a Bayesian network and applied in the context of NMR data.[27,28] It gives rise to the following posterior distribution over the coarse-grained variables:

$$p(\mathbf{e},\mathbf{f}|\mathbf{d},\boldsymbol{\sigma}^2) \propto p(\mathbf{d}|\mathbf{e},\boldsymbol{\sigma}^2)p(\mathbf{f}|\mathbf{e})p(\mathbf{e}). \quad (1)$$

This coarse-grained probabilistic model is then combined with the prior distribution of the fine-grained variable $\mathbf{x}$, according to an appropriate probabilistic prior model M, using the reference ratio method (RRM).[34] The RRM is based on the principles of probability kinematics, a variant of Bayesian updating that can be used to modify a given probability distribution in the light of new evidence regarding partitions of the distribution's sample space.[35] The updated posterior is:

$$p(\mathbf{e},\mathbf{f},\mathbf{x}|\mathbf{d},\boldsymbol{\sigma}^2,M) \propto p(\mathbf{d}|\mathbf{e},\boldsymbol{\sigma}^2)\frac{p(\mathbf{f}|\mathbf{e})}{p(\mathbf{f}|M)}p(\mathbf{e})p(\mathbf{x}|M). \quad (2)$$

This combined posterior is the distribution with minimum Kullback–Leibler divergence from the fine-grained prior $p(\mathbf{x}|M)$, under the requirement that the marginal distribution of the coarse-grained variables follows eqn (1).[36]

SAXS ensembles

In the case of SAXS, the experimental data $\mathbf{d}$ and the ensemble average $\mathbf{e}$ constitute vectors of scattering intensities, while the structures $\mathbf{x}$ are represented as vectors of atomic coordinates. A force field or a fragment library could be used to sample from the prior distribution $p(\mathbf{x}|M)$; here, we use the PROFASI force field.[37] A coarse-grained vector $\mathbf{f}$ is generated through a forward model by approximating the scattering function $h(\mathbf{x})$ with the Debye formula, which holds for spherical scatterers:[38]

$$\mathbf{f} \equiv g(\mathbf{x},q) = \sum_{i=1}^{K}\sum_{j=1}^{K} F_i(q)F_j(q)\frac{\sin(qr_{ij})}{qr_{ij}}, \quad (3)$$

where $q = (4\pi \sin\theta)/\lambda$ is the momentum transfer, with scattering angle 2θ and wavelength of the X-ray beam λ. $F_i(q)$ is the atomic form factor for atom i, r_{ij} is the distance between atoms i and j, and K is the number of atoms in the structure. The X-ray scattering factors are calculated using a linear combination of Gaussians fit to empirical data.[39]

Posterior distribution. We use a Gaussian distribution for the likelihood, $p(\mathbf{d}|\mathbf{e},\boldsymbol{\sigma}^2)$, to relate the data to the ensemble average $\mathbf{e}$. For the ratio of the two unknown distributions $p(\mathbf{f}|\mathbf{e})$ and $p(\mathbf{f}|M)$ in eqn (2) we use a log-linear model $\mathcal{G}(\mathbf{f}|\mathbf{e},\mathbf{B})$ with a link function $l(\mathbf{B},\mathbf{e}) = \mathbf{B}\mathbf{e}^{-1}$,[40]

$$\mathcal{G}(\mathbf{f}|\mathbf{e},\mathbf{B}) = \frac{\exp(\mathbf{f}^T\mathbf{B}\mathbf{e}^{-1})}{\mathcal{Z}}, \quad (4)$$

where $\mathbf{B}$ is a diagonal matrix and $\mathcal{Z}$ is a normalization constant. The matrix $\mathbf{B}$ serves to match the first moment, $\langle\mathbf{f}\rangle$, of the

coarse-grained prior represented by the PROFASI force field to the ensemble average **e**. This model is scale-invariant when **f** and **e** are scaled together, *i.e.* $\mathcal{G}(\mathbf{f}|\mathbf{e},\mathbf{B}) = \mathcal{G}(c\mathbf{f}|c\mathbf{e},\mathbf{B})$ for any constant c. This is required due to the arbitrary scale of SAXS data.

Assuming a uniform prior for **e**, the joint posterior distribution from eqn (2) for SAXS ensembles becomes:

$$p(\mathbf{e},\mathbf{f},\mathbf{x}|\mathbf{d},\boldsymbol{\sigma}^2,\mathbf{B}) \propto \mathcal{N}(\mathbf{d}|\mathbf{e},\boldsymbol{\sigma}^2)\mathcal{G}(\mathbf{f}|\mathbf{e},\mathbf{B})\exp(-\beta E_{\text{prof}}(\mathbf{x})). \quad (5)$$

In the last term, E_{prof} is the energy of the PROFASI force field and $\beta \equiv 1/kT$, where T is the temperature and k is the Boltzmann constant.

Determining B. We modify the EM algorithm described by Olsson *et al.*,[28] to estimate the matrix **B** (Fig. 1). This corresponds to adopting an empirical Bayes strategy for the prior distribution of the ensemble posterior.

In the E-stage of iteration k of the algorithm, a Markov chain Monte Carlo (MCMC) simulation, as implemented in the PHAISTOS framework,[41] produces N samples $\mathcal{S}_{(k)} = \{\mathbf{f}_{1..N}, \mathbf{e}_{1..N}, \mathbf{x}_{1..N}\}$ from the posterior $p(\mathbf{e},\mathbf{f},\mathbf{x}|\mathbf{d},\boldsymbol{\sigma}^2,\mathbf{B}_{(k)})$. The result is a conformational ensemble of structures together with their forward-computed SAXS profiles, whose average optimally matches the experimental data. The iterative algorithm is initialized with the zero matrix, $\mathbf{B}_{(0)}= 0$, resulting in an unrestrained simulation with the structural prior, $\exp(-\beta E_{\text{prof}}(\mathbf{x}))$.

A new scaling matrix $\mathbf{B}_{(k+1)}$ is estimated in the M-stage, by minimizing a χ^2_{EM} objective function:

$$\mathbf{B}_{(k+1)} = \arg\min_{\mathbf{B}^*_{(k+1)}} \chi^2_{\text{EM}}, \quad (6)$$

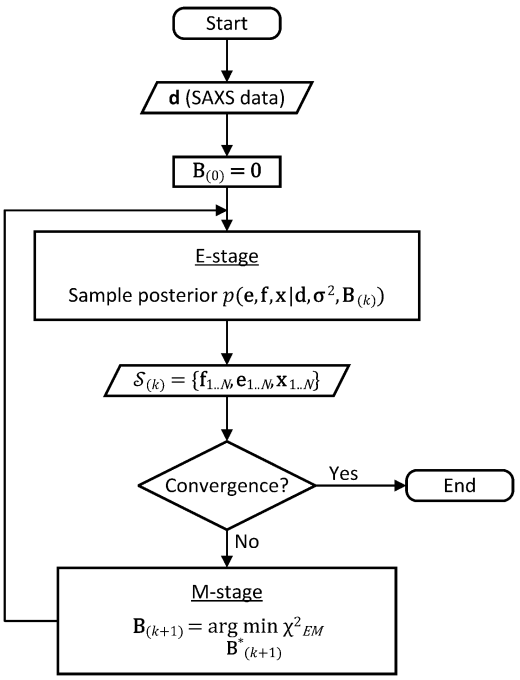

Fig. 1 Flow chart of the BE-SAXS algorithm. The method ensures that the ensemble average of the posterior distribution matches the experimental SAXS data, through an empirical Bayes procedure, formulated as an iterative EM algorithm.

with:

$$\chi^2_{\text{EM}} \equiv \left\| \frac{\left\langle \mathbf{e}_{\mathbf{B}^*_{(k+1)}} \right\rangle - \left\langle \mathbf{f}_{\mathbf{B}^*_{(k+1)}} \right\rangle}{\boldsymbol{\sigma}} \right\|^2 + \left\| \mathbf{D}_{(k+1)} \right\|^2, \quad (7)$$

where $\mathbf{D}_{(k+1)} \equiv \mathbf{B}^*_{(k+1)} - \mathbf{B}_k$.

Conceptually, the M-stage aims to ensure that a given ensemble average **e** and the matching coarse-grained average of the sampled structures $\langle \mathbf{f} \rangle$ coincide. It is necessary to normalize by the experimental errors in eqn (7), since SAXS data ranges over several orders of magnitude across the scattering profile. The role of the second term is to use Tikhonov regularization to avoid overfitting.[42] Here, it is utilized specifically to avoid excessive changes to the matrix **B** due to finite sampling issues, allowing for monotonous convergence of the parameters.

The expectation of the coarse-grained variable, $\left\langle \mathbf{f}_{\mathbf{B}^*_{(k+1)}} \right\rangle$, is estimated from the N samples using importance sampling:[43]

$$\left\langle \mathbf{f}_{\mathbf{B}^*_{(k+1)}} \right\rangle \approx \sum_{i=1}^{N} \mathbf{f}_i \frac{\exp\left(\mathbf{f}_i^T \mathbf{D}_{(k+1)} \mathbf{e}_i^{-1}\right)}{\sum_{j=1}^{N} \exp\left(\mathbf{f}_j^T \mathbf{D}_{(k+1)} \mathbf{e}_j^{-1}\right)}. \quad (8)$$

It is notable that the importance weights in eqn (8) do not change when **f** and **e** are scaled together. In practice, both the coarse-grained vector **f** and the ensemble average **e** are brought to scale with the experimental data **d** – the former through a scaling coefficient determined at initialization, and the latter through the Gaussian ensemble likelihood. Therefore, the matrix $\mathbf{B}_{(k+1)}$ and the associated structural ensemble produced by the algorithm remain invariant, regardless of the absolute magnitude of **d**.

The expectation of the ensemble average is approximated by the sample average:

$$\left\langle \mathbf{e}_{\mathbf{B}^*_{(k+1)}} \right\rangle \approx \frac{1}{N} \sum_{i=1}^{N} \mathbf{e}_i. \quad (9)$$

For further details see the work of Olsson *et al.*[28]

We use the basin hopping stochastic global optimization algorithm[44] for the minimization of the objective function in eqn (6); however, other optimization techniques such as genetic algorithms or parallel tempering may be utilized. In principle, because the function is convex, gradient descent algorithms are also applicable but we found that they can be unstable due to finite statistical sampling. Convergence can be considered achieved once the objective function falls below 0.5, indicating incremental improvements within the experimental uncertainty of the data.

Simulations

Experimental data. We utilized the published conformational ensemble of the disordered protein PaaA2 in order to test the BE-SAXS ensemble method.[45] PaaA2 is an antitoxin that is encoded by a toxin–antitoxin module in *Escherichia coli* O157.[46] In the absence of its binding partner, the toxin ParE2, PaaA2 behaves like an IDP. However, it contains two stable α-helical regions that are flanked by highly disordered stretches of amino acids.[45]

The published structural ensemble of PaaA2 consists of 50 conformations and is available from the PDB database under the code 3ZBE. The structures were selected by the application of a jackknife procedure to EOM-derived SAXS ensembles from a pool of NMR-restrained conformers.[45] Following the Reference Ensemble Method,[47] in order to validate the BE-SAXS algorithm we used a SAXS forward model to create a synthetic data set from the reference ensemble of 50 conformations. This allows controlling for all sources of uncertainty in the evaluation. We constructed the SAXS ensemble average data $\mathbf{d}$ for the protein by generating SAXS profiles $\mathbf{d}_i$ for each conformation, using the FoXS program,[48] and averaging the individual profiles:

$$\mathbf{d} = \frac{1}{50}\sum_{i=1}^{50}\mathbf{d}_i. \tag{10}$$

Experimental errors $\boldsymbol{\sigma}^2$ were assigned as the population variance of the data.

Computation. The EM algorithm ran for a total of 21 iterations. In each E-stage, the PHAISTOS framework was used to run 64 independent MCMC chains for 10^6 steps.[41] Samples $\mathcal{S}_{(k)}$ were saved every 10^3 steps to be used in the M-stage, after a 40% burn-in. The global optimization algorithm of the M-stage was run for up to 20 independent iterations, or until a stable solution was found. The algorithm reached convergence at iteration 10, as judged from the change in fit between EM steps, χ^2_{EM}, from the ensemble SAXS profile fit, χ^2_{SAXS}, and from the magnitude of the changes in the scaling matrix $\mathbf{B}$. The measure of fit to the experimental data was defined as:

$$\chi^2_{\mathrm{SAXS}} \equiv \frac{1}{N}\left\|\frac{\mathbf{d}-\langle\mathbf{f}\rangle}{\boldsymbol{\sigma}}\right\|^2, \tag{11}$$

where $\langle\mathbf{f}\rangle$ is the ensemble average:

$$\langle\mathbf{f}\rangle \equiv \frac{1}{N}\sum_{i=1}^{N}\mathbf{f}_i \tag{12}$$

The generative probabilistic models TorusDBN and BASILISK were used as proposal distributions during the MCMC simulation for main chain and side chain moves, respectively.[49,50] The introduced bias was subsequently removed. The PROFASI force field at T = 300 K was used as the prior distribution of the structures $\mathbf{x}$.[37]

GPU calculations. The forward calculation of the SAXS profile is the most compute-intensive part of the BE-SAXS ensemble method. We used our GPU Parallel Page-Tile SAXS algorithm with atomic form factors to accelerate the computation of eqn (3).[51,52] We utilized a 16-core Intel Xeon E5-2660 server with 2 NVIDIA GeForce GTX 690 GPU cards (4x1536 GPU cores), which allowed us to run the 64 MCMC chains in parallel.

To accelerate the M-stage, we implemented an OpenCL kernel that calculates eqn (8) on the GPU.[53] The efficiency of this approach depends on the number of samples used; for this simulation, the GPU acceleration reduced the stage time by a factor of 3.

Ensembles. The structural ensembles for each EM iteration (EM_i, for i = 0,...,20) were generated by uniformly sampling conformations from the 64 independent MCMC chains at 10^4 MC-step intervals, after a 40% burn-in. This resulted in 3904 structures per iteration. 128 structures were sampled uniformly from EM_0 and EM_9 in order to visualize the ensembles.

Results and discussion

Algorithm convergence for PaaA2

In the E-stage of the first iteration of the BE-SAXS algorithm, the conformational ensemble EM_0 of the protein PaaA2 was effectively sampled from an unrestrained PROFASI force field. The resulting ensemble average does not fit the SAXS scattering profile well, as evidenced by the high value of the χ^2_{SAXS} measure (Fig. 2). This suggests that PROFASI alone, as a minimalistic force field, does not accurately capture the details of the flexibility of PaaA2 represented in the calculated ensemble-averaged SAXS data. In subsequent iterations, however, the fit improves rapidly and reaches a stable region. The objective function, χ^2_{EM}, also reaches a low value quickly and falls below 0.5 in iteration 9 (Fig. 2). At this level, by the nature of χ^2_{EM}, modifications to the matrix $\mathbf{B}$ produce changes in the importance sampling approximating distribution that are within the experimental uncertainty of the data. The individual coefficients of $\mathbf{B}$ also stabilize at iteration 9, further indicating convergence. The equilibrium reached thereby is dynamic, due to the stochastic nature of the basin hopping global optimization algorithm used in the M-stage, combined with the underdetermined optimization problem in eqn (6).

Convergence in the BE-SAXS algorithm has to be evaluated comprehensively, by examination of both χ^2_{EM} and χ^2_{SAXS}, since a low χ^2_{EM} does not guarantee that the conformational ensemble provides a good fit to the data. If there is an insufficient number of steps in the E-stage to allow for the MCMC to reach equilibrium, then the Boltzmann distribution will not be

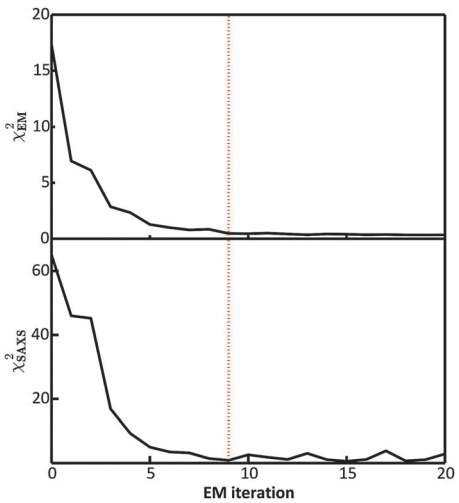

Fig. 2 Convergence of the BE-SAXS algorithm for the protein PaaA2. (top) χ^2_{EM} is a measure of the change in fit between the approximating and target distributions of the ensemble average at each iteration. (bottom) χ^2_{SAXS} measures the fit between the data and the posterior ensemble average $\langle\mathbf{f}\rangle$ at each iteration. The dotted red line indicates the point of convergence of the algorithm at iteration 9, where χ^2_{EM} is below 0.5 and χ^2_{SAXS} is close to unity.

sampled successfully. Thus, a low χ^2_{EM} could be achieved at a specific iteration and still result in a **B** matrix that does not produce an ensemble average matching the experimental data. Furthermore, it is necessary to examine the behavior of the χ^2 statistics and the **B** coefficients over a range of EM iterations, to determine if an equilibrium has in fact been reached. Because the optimization problem in eqn (6) is underdetermined, fluctuations in both the matrix **B** and χ^2_{SAXS} are expected. However, in order to assume convergence, these fluctuations should be confined to a stable and relatively narrow region.

BE-SAXS restrains the PaaA2 ensemble

We examined and compared the EM_0 and EM_9 structural ensembles of the protein PaaA2, in order to evaluate the performance of the BE-SAXS method. The scattering average for the initial, unrestrained ensemble EM_0 exhibits a poor fit to the SAXS profile, **d**, (χ^2_{SAXS} = 65.0) while the average for the restrained ensemble EM_9 shows good agreement with the data (χ^2_{SAXS} = 0.9), within the margins of error (Fig. 3). The high q range of the SAXS profile contains atomic-level data and the larger deviation observed there could be due to the stronger influence of the PROFASI force field on the local structure of the simulated IDP protein than on the overall shape. While the deviation is within the error bounds, it may be desirable to further penalize discrepancies within this range during the M-stage optimization. Alternatively, better sampling of the local structure could be achieved by a longer simulation that emphasizes local and side chain moves. This may allow for a more accurate assessment of the agreement between the ensemble averages of the target and approximating distributions in the M-stage.

To further characterize the EM_0 and EM_9 ensembles, we compared their radius of gyration (R_g) distributions to the R_g distribution of the published PaaA2 reference ensemble (Fig. 4). The 50-structure 3ZBE ensemble is relatively compact, while the unrestrained PROFASI-driven EM_0 exhibits a wider variation of R_g with two prominent modes. On the other hand, the SAXS-restrained EM_9 closely matches the original ensemble in both its mean and sample error, suggesting that BE-SAXS is able to extract ensemble-level R_g information from the SAXS profile.

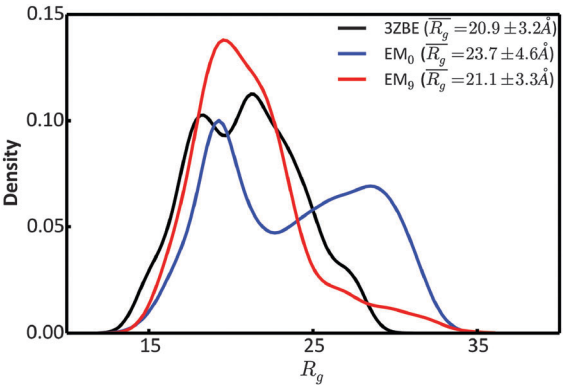

Fig. 4 Comparison of the distributions of the radius of gyration, R_g, for the 3ZBE ensemble reported by Sterckx *et al.*[45] (black) and the ensembles at EM iterations 0 (blue) and 9 (red). The distribution for 3ZBE was derived through kernel density estimation, due to the limited number of conformations.

Due to the low information content of SAXS data, it is not possible to summarize the ensemble using only a few representative conformations, despite the presence of a force field. However, the scattering profile can inform about the general shape of the protein. Taking advantage of the stable α-helical regions in PaaA2, we defined a shape descriptor, K_{sh}, as a proxy to the 3-dimentional shape. The K_{sh} measure is calculated as the ratio of the distances between the distal and proximal ends of the two helices (the Cα atoms of residue pairs (16, 57) and (28, 42), respectively); thus K_{sh} is an indicator of the "openness" of the overall structure. We compared the distributions of the descriptor for the EM_0, EM_9, and reference ensembles (Fig. 5). The unrestrained EM_0 gives rise to a bimodal distribution for K_{sh} and favors open structures. The shape descriptor distributions for the reference ensemble and the SAXS-restrained EM_9 show substantial similarity to each other, and share a propensity for more compact structures.

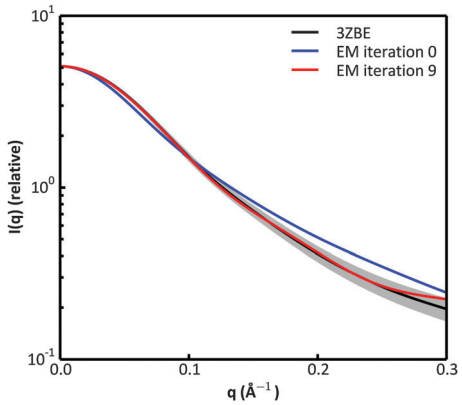

Fig. 3 Scattering curves for the protein PaaA2. The original data calculated from the published structural ensemble are shown in black, with error margins in grey. The fit of the unrestrained ensemble at iteration 0 of the EM algorithm is shown in blue. The fit of the optimized ensemble at iteration 9 of the EM algorithm is shown in red.

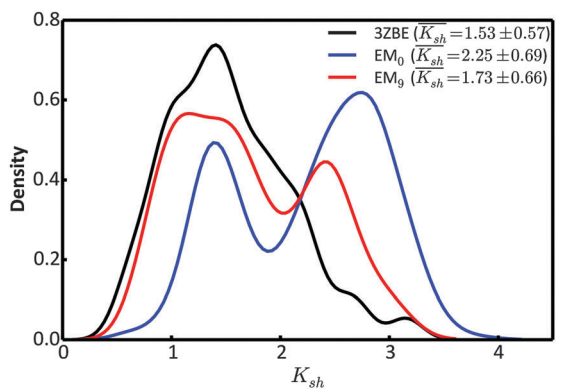

Fig. 5 Comparison of the distributions of the shape descriptor, K_{sh}, for the 3ZBE ensemble reported by Sterckx *et al.*[45] (black) and the ensembles at EM iterations 0 (blue) and 9 (red). The distribution for 3ZBE was derived through kernel density estimation, due to the limited number of conformations.

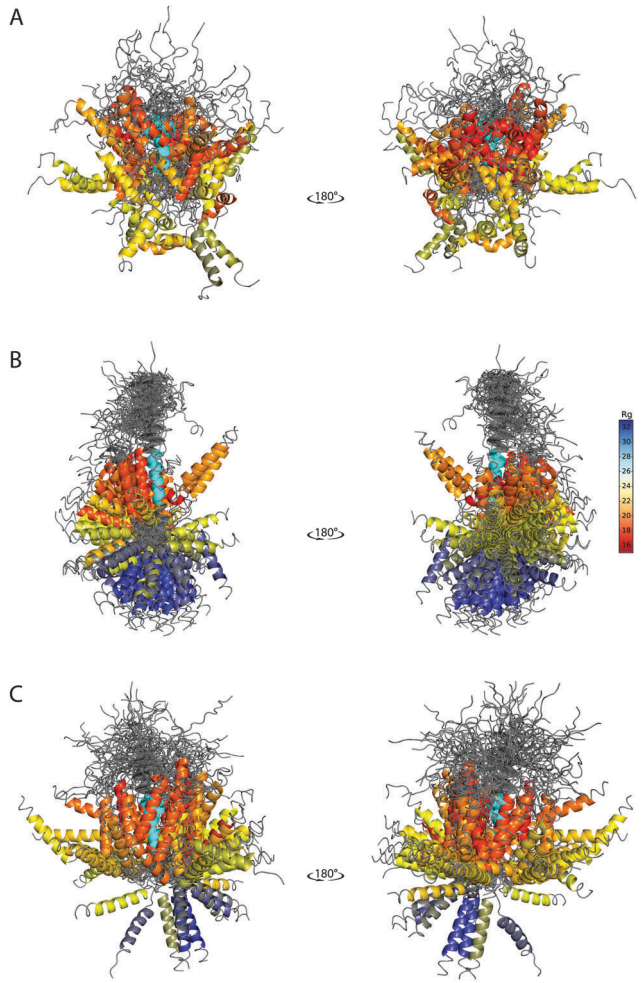

Fig. 6 SAXS-derived conformational ensembles of PaaA2. (A) The published 50-member ensemble of PaaA2 (PDB 3ZBE), derived from NMR and SAXS data. (B) Subsample of 128 conformations from EM_0, the unrestrained ensemble at iteration 0 of BE-SAXS. (C) Subsample of 128 conformations from EM_9, the SAXS-restrained ensemble at iteration 9 of BE-SAXS. All structures are aligned on the first helix (colored in cyan). The color of the second helix corresponds to the R_g of the structure in Å (indicated in the color bar).

The ability of the BE-SAXS method to restrict the solution space to areas consistent with the experimental data is further evident in the visualized ensembles (Fig. 6). EM_9 exhibits characteristics similar to the reference ensemble – it favors conformations in which the two α-helices are packed closely together, while maintaining significant overall flexibility. At the same time, the unrestrained EM_0 comprises structures that are consistent with uniform rotation around the disordered linker. The linker flexibility is greater in EM_9 than in EM_0, with more diversity in the relative orientations of the two helices, as in the original ensemble.

The peripheral disordered regions in both EM_0 and EM_9 exhibit much more helical structure than the 3ZBE ensemble. This is likely the effect of the PROFASI force field on local structure and it helps explain the larger deviation of the scattering profile at high q values. The main advantage of PROFASI is efficiency, but a more sophisticated force field would presumably produce a better fit with the data.

Conclusions

A novel method for inference of protein ensembles from SAXS data, which we call Bayesian Ensemble SAXS, was described and demonstrated here as a proof of principle. BE-SAXS proceeds through successive expectation maximization steps and uses a Bayesian probabilistic model for ensemble-averaged SAXS data to modify a probabilistic model of protein structure, in agreement with an experimental scattering profile. This results in a generative model that can be used directly to characterize a protein's conformational ensemble, or that can be further restrained with other types of experimental data, such as NMR. The generative approach offers a particular advantage for flexible systems, such as intrinsically disordered proteins and proteins with long disordered regions, since it does not impose restrictions on the ensemble size and allows sampling of the full conformational space allowed by the data. The number of parameters of the generative probabilistic model only depends on the number of experimental observables, and not on the size of the ensemble. This stands in contrast to many existing SAXS ensemble methods that fit a set of structures to the data and where each replica results in a linear increase in the number of parameters.

To illustrate the BE-SAXS method, we applied it to the ensemble-averaged SAXS data for the published conformational ensemble of the highly flexible antitoxin PaaA2. We showed that our approach restrains the conformational space accessible to the protein simulation and yields ensembles with characteristics consistent with the original set of structures. The ability of the method to model protein flexibility suggests its utility in characterizing other IDPs and multi-domain proteins. The Bayesian probabilistic formulation used here can be complemented by other probabilistic models based on experimental observables. In particular, NMR residual dipolar couplings (RDCs) and chemical shifts are commonly utilized in the context of disordered proteins.[29,54] We expect that employing BE-SAXS in concert with methods that make use of other experimental data, can greatly help elucidate the native state ensembles of flexible macromolecular systems.

Acknowledgements

S. O. is funded by an Independent Postdoc grant from The Danish Council for Independent Research for Natural Sciences (ID: DFF-4002-00151). T.H. acknowledges support from the University of Copenhagen 2016 Excellence Programme for Interdisciplinary Research (UCPH2016-DSIN). W. B. is supported by the Villum Foundation.

Notes and references

1 P. E. Wright and H. J. Dyson, *J. Mol. Biol.*, 1999, **293**, 321–331.
2 P. Tompa, *Curr. Opin. Struct. Biol.*, 2011, **21**, 419–425.
3 P. Tompa, *Nat. Chem. Biol.*, 2012, **8**, 597–600.
4 A. Mittal, N. Lyle, T. S. Harmon and R. V Pappu, *J. Chem. Theory Comput.*, 2014, **10**, 3550–3562.
5 V. N. Uversky, *Front. Aging Neurosci.*, 2015, **7**, 18.

6 V. N. Uversky and A. K. Dunker, *Biochim. Biophys. Acta*, 2010, **1804**, 1231–1264.
7 L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradović and A. K. Dunker, *J. Mol. Biol.*, 2002, **323**, 573–584.
8 P. Tompa, P. Buzder-Lantos, A. Tantos, A. Farkas, A. Szilágyi, Z. Bánóczi, F. Hudecz and P. Friedrich, *J. Biol. Chem.*, 2004, **279**, 20775–20785.
9 K. Uéda, H. Fukushima, E. Masliah, Y. Xia, A. Iwai, M. Yoshimoto, D. A. Otero, J. Kondo, Y. Ihara and T. Saitoh, *Proc. Natl. Acad. Sci. U. S. A.*, 1993, **90**, 11282–11286.
10 K. K. Dev, K. Hofele, S. Barbieri, V. L. Buchman and H. Van Der Putten, *Neuropharmacology*, 2003, **45**, 14–44.
11 D. Schneidman-Duhovny, S. J. Kim and A. Sali, *BMC Struct. Biol.*, 2012, **12**, 17.
12 G. Tria, H. D. T. Mertens, M. Kachala and D. I. Svergun, *IUCrJ*, 2015, **2**, 207–217.
13 M. Hammel, *Eur. Biophys. J.*, 2012, **41**, 789–799.
14 S. Yang, *Adv. Mater.*, 2014, **26**, 7902–7910.
15 A. Cavalli, C. Camilloni and M. Vendruscolo, *J. Chem. Phys.*, 2013, **138**, 094112.
16 P. Bernadó and D. I. Svergun, *Mol. BioSyst.*, 2012, **8**, 151–167.
17 M. Pelikan, G. L. Hura and M. Hammel, *Gen. Physiol. Biophys.*, 2009, **28**, 174–189.
18 S. Yang, L. Blachowicz, L. Makowski and B. Roux, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 15757–15762.
19 E. T. Jaynes, *Phys. Rev.*, 1957, **106**, 620–630.
20 B. Różycki, Y. C. Kim and G. Hummer, *Structure*, 2011, **19**, 109–116.
21 W. Y. Choy and J. D. Forman-Kay, *J. Mol. Biol.*, 2001, **308**, 1011–1032.
22 J. A. Marsh, C. Neale, F. E. Jack, W.-Y. Choy, A. Y. Lee, K. A. Crowhurst and J. D. Forman-Kay, *J. Mol. Biol.*, 2007, **367**, 1494–1510.
23 J. A. Marsh and J. D. Forman-Kay, *Proteins*, 2012, **80**, 556–572.
24 M. Krzeminski, J. A. Marsh, C. Neale, W.-Y. Choy and J. D. Forman-Kay, *Bioinformatics*, 2013, **29**, 398–399.
25 K. Berlin, C. A. Castañeda, D. Schneidman-Duhovny, A. Sali, A. Nava-Tudela and D. Fushman, *J. Am. Chem. Soc.*, 2013, **135**, 16595–16609.
26 W. Boomsma, J. Ferkinghoff-Borg and K. Lindorff-Larsen, *PLoS Comput. Biol.*, 2014, **10**, e1003406.
27 S. Olsson, J. Frellsen, W. Boomsma, K. V. Mardia and T. Hamelryck, *PLoS One*, 2013, **8**, e79439.
28 S. Olsson, B. R. Vögeli, A. Cavalli, W. Boomsma, J. Ferkinghoff-Borg, K. Lindorff-Larsen and T. Hamelryck, *J. Chem. Theory Comput.*, 2014, **10**, 3484–3491.
29 S. Olsson, D. Ekonomiuk, J. Sgrignani and A. Cavalli, *J. Am. Chem. Soc.*, 2015, **137**, 6270–6278.
30 J. W. Pitera and J. D. Chodera, *J. Chem. Theory Comput.*, 2012, **8**, 3445–3451.
31 B. Roux and J. Weare, *J. Chem. Phys.*, 2013, **138**, 084107.
32 S. Olsson and A. Cavalli, *J. Chem. Theory Comput.*, 2015, **11**, 3973–3977.
33 W. Rieping, M. Habeck and M. Nilges, *Science*, 2005, **309**, 303–306.
34 T. Hamelryck, M. Borg, M. Paluszewski, J. Paulsen, J. Frellsen, C. Andreetta, W. Boomsma, S. Bottaro and J. Ferkinghoff-Borg, *PLoS One*, 2010, **5**, e13714.
35 P. Diaconis and S. L. Zabell, *J. Am. Stat. Assoc.*, 1982, **77**, 822–830.
36 *Bayesian Methods in Structural Bioinformatics*, ed. T. Hamelryck, K. Mardia and J. Ferkinghoff-Borg, Springer, 2012.
37 A. Irbäck, S. Mitternacht and S. Mohanty, *PMC Biophys.*, 2009, **2**, 2.
38 P. Debye, *Ann. Phys.*, 1915, **351**, 809–823.
39 D. Waasmaier and A. Kirfel, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 1995, **51**, 416–431.
40 P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd edn, Chapman & Hall, 1989.
41 W. Boomsma, J. Frellsen, T. Harder, S. Bottaro, K. E. Johansson, P. Tian, K. Stovgaard, C. Andreetta, S. Olsson, J. B. Valentin, L. D. Antonov, A. S. Christensen, M. Borg, J. H. Jensen, K. Lindorff-Larsen, J. Ferkinghoff-Borg and T. Hamelryck, *J. Comput. Chem.*, 2013, **34**, 1697–1705.
42 A. N. Tikhonov, *Dokl. Akad. Nauk SSSR*, 1943, **39**, 195–198.
43 C. M. Bishop, *Pattern Recognition and Machine Learning*, 2006.
44 D. J. Wales and J. P. K. Doye, *J. Phys. Chem. A*, 1997, **101**, 5111–5116.
45 Y. G. J. Sterckx, A. N. Volkov, W. F. Vranken, J. Kragelj, M. R. Jensen, L. Buts, A. Garcia-Pino, T. Jové, L. Van Melderen, M. Blackledge, N. A. J. van Nuland and R. Loris, *Structure*, 2014, **22**, 854–865.
46 Y. G. J. Sterckx, A. Garcia-Pino, S. Haesaerts, T. Jové, L. Geerts, V. Sakellaris, L. Van Melderen and R. Loris, *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.*, 2012, **68**, 724–729.
47 C. K. Fisher and C. M. Stultz, *Curr. Opin. Struct. Biol.*, 2011, **21**, 426–431.
48 D. Schneidman-Duhovny, M. Hammel and A. Sali, *Nucleic Acids Res.*, 2010, **38**, W540–W544.
49 W. Boomsma, K. V Mardia, C. C. Taylor, J. Ferkinghoff-Borg, A. Krogh and T. Hamelryck, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 8932–8937.
50 T. Harder, W. Boomsma, M. Paluszewski, J. Frellsen, K. E. Johansson and T. Hamelryck, *BMC Bioinf.*, 2010, **11**, 306.
51 L. Antonov, C. Andreetta and T. Hamelryck, *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS 2012)*, 2012, pp. 102–108.
52 L. D. Antonov, C. Andreetta and T. Hamelryck, in *Biomedical Engineering Systems and Technologies SE – 15*, ed. J. Gabriel, J. Schier, S. Huffel, E. Conchon, C. Correia, A. Fred and H. Gamboa, Springer, Berlin, Heidelberg, 2013, vol. 357, pp. 222–235.
53 J. E. Stone, D. Gohara and G. Shi, *Comput. Sci. Eng.*, 2010, **12**, 66–72.
54 J. M. Krieger, G. Fusco, M. Lewitzky, P. C. Simister, J. Marchant, C. Camilloni, S. M. Feller and A. De Simone, *Biophys. J.*, 2014, **106**, 1771–1779.