Structural bioinformatics

Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables

Valéry Ozenne1, Frédéric Bauer2, Loïc Salmon1, Jie-rong Huang1, Malene Ringkjøbing Jensen1, Stéphane Segard2, Pau Bernadó3, Céline Charavay2 and Martin Blackledge4,1

1Protein Dynamics and Flexibility, Institut de Biologie Structurale Jean-Pierre Ebel, CEA; CNRS; UJF UMR 5075, 41 Rue Jules Horowitz, Grenoble 38027, 2Groupe Informatique pour les Scientifiques du Sud Est (GIPSE), IRTSV / Laboratoire Biologie à Grande Echelle, CEA - INSERM U1038 - UJF, 17 avenue des Martyrs, 38054 Grenoble Cedex 9 and 3Centre de Biochimie Structurale, CNRS UMR 5048 - UM 1 - INSERM UMR 1054, 34090, Montpellier, France

ABSTRACT

Motivation: Intrinsically disordered proteins (IDPs) represent a significant fraction of the human proteome. The classical structure function paradigm that has successfully underpinned our understanding of molecular biology breaks down when considering proteins that have no stable tertiary structure in their functional form. One convenient approach is to describe the protein in terms of an equilibrium of rapidly inter-converting conformers. Currently, tools to generate such ensemble descriptions are extremely rare, and poorly adapted to the prediction of experimental data.

Results: We present flexible-meccano—a highly efficient algorithm that generates ensembles of molecules, on the basis of amino acid-specific conformational potentials and volume exclusion. Conformational sampling depends uniquely on the primary sequence, with the possibility of introducing additional local or long-range conformational propensities at an amino acid-specific resolution. The algorithm can also be used to calculate expected values of experimental parameters measured at atomic or molecular resolution, such as nuclear magnetic resonance (NMR) and small angle scattering, respectively. We envisage that flexible-meccano will be useful for researchers who wish to compare experimental data with those expected from a fully disordered protein, researchers who see experimental evidence of deviation from ‘random coil’ behaviour in their protein, or researchers who are interested in working with a broad ensemble of conformers representing the flexibility of the IDP of interest.

Availability: A fully documented multi-platform executable is provided, with examples, at http://www.ibs.fr/science-213/scientific-output/software/flexible-meccano/

Contact: martin.blackledge@ibs.fr

Received on February 29, 2012; revised on March 23, 2012; accepted on April 2, 2012

1 INTRODUCTION

The realization that a significant percentage of the functional proteins encoded in eukaryotic genomes are fully or partially disordered in their functional state has revolutionized our understanding of structural and molecular biology (Babu, 2012; Dunker et al., 2002; Dyson and Wright, 2005; Tompa, 2002; Uversky, 2002). Intrinsically disordered proteins (IDPs), or proteins containing long intrinsically disordered regions, do not adopt a stable 3D fold, and therefore fall beyond the scope of classical structural biology. IDPs are biologically functional in the disordered state imposing a very different perspective on the relationship between primary protein sequence and function compared with the standard structure/function relationship that underpins our understanding of molecular biology. IDPs are implicated in a large number of human pathologies, and the development of pharmacological solutions to these problems awaits a molecular description of the role of flexibility in a number of diseases (Babu et al., 2011; Dunker and Uversky, 2010; Vendruscolo and Dobson, 2007).

In order to understand the conformational behaviour of IDPs it is essential to develop a molecular description of the disordered state. The structural biology paradigm shifts when we consider disordered proteins, so that the determination of a single structure has no real physical relevance, or at best can only describe isolated sub-states on a vast potential energy landscape. A structural description of IDPs rather aims to determine rules that define the behaviour of the flexible protein in terms of probabilities of populating different regions of conformational space, and to correlate these probabilities with the function of the protein. The description of conformational propensities can be conveniently achieved by evoking an explicit ensemble description of inter-converting structures in equilibrium. Due to the very large number of degrees of freedom available to such a disordered system, the problem of defining conformational space is highly underdetermined, requiring extensive experimental data to delimit the structural propensities of a given protein. Novel analytical tools are required to exploit the specific conformational sensitivity of different experimental parameters. Each experimental NMR parameter for example, is sensitive to different aspects of the structural and dynamic behaviour of the disordered state and requires specific consideration of the relevant averaging properties of the physical interaction (Schneider et al., 2012).
Fig. 1. Schematic representation of the flexible-meccano algorithm. Conformational sampling details are defined by the user (shown figuratively on the left in terms of Ramachandran sampling for each amino acid)—the default is the flexible-meccano statistical coil description. This can be modified in terms of additional propensities to form secondary structure, by modifying the amino acid-specific potentials or by introducing long-range contacts. Biophysical parameters (radius of gyration and \(\phi/\psi\) sampling), and expected experimental parameters (shown figuratively on the right, for example NMR RDCs, PREs etc.) are calculated for this conformational regime, as well as explicit conformational ensemble of backbone conformers.

2 SYSTEM AND METHODS

The protocol that we present here describes the generation of explicit ensemble descriptions of proteins, using the program flexible-meccano that has been explicitly developed to describe the behaviour of IDPs and denatured proteins. Expected experimental parameters, such as NMR and small angle scattering data, are calculated from the conformational ensembles (Bernadó et al., 2005a, b; Jensen et al., 2008; Jensen et al., 2011; Mukrasch et al., 2007; Nodet et al., 2009; Salmon et al., 2010; Wells et al., 2008). These parameters can then be compared with experimental measurements. Amino acid-specific statistical coil sampling is used to describe the unfolded state on the basis of the primary sequence with the possibility of introducing additional local conformational propensities, or long-range tertiary contacts. The flexible-meccano approach was exploited for the first time to study the C-terminal domain of Sendai virus phosphoprotein, and has been gradually refined, extended and tested on a number of different experimental systems, including the disordered proteins α-synuclein, Tau and p53, involved in Parkinson’s and Alzheimer’s diseases and human cancer, respectively. Currently, no alternative tools exist that can provide ensembles and ensemble-averaged parameters for user-defined conformational sampling regimes. The program is interfaced to a fully interactive and robust graphical interface as described below.

3 ALGORITHM

3.1 Flexible-meccano ensemble generation

Flexible-meccano uses a highly efficient minimization algorithm to build multiple, different copies of the same polypeptide chain by randomly sampling amino acid-specific backbone dihedral angle \(\phi/\psi\) potential wells. The population-weighted amino acid-specific potentials are derived from a compilation of non-secondary structural elements of high-resolution X-ray crystallographic protein structures. The peptide chain is constructed by using the selected \(\phi/\psi\) pairs to sequentially connect peptide planes.

The algorithm is based on tools developed for Meccano and Dynamic Meccano approaches (Bouvignies et al., 2006a, b; Hus et al., 2001; Hus et al., 2008; Salmon et al., 2009). These algorithms were used to determine the average orientation of peptide planes, and their associated dynamics, in folded proteins on the basis of optimization of parameters defining the orientation of each place on
the basis of NMR residual dipolar couplings (RDCs). The difference in this case is that, rather than determining the orientation of the peptide plane on the basis of experimental data, the unique constraint used to orient each peptide unit is the backbone dihedral angle pair, that is randomly selected from the database potential for each amino acid. Each tetrahedral junction is constructed with optimal geometry.

The coordinates of the generic peptide plane were derived from high-resolution X-ray crystallographic structures (Salmon et al., 2009). Amino acid-specific hard-spheres are used to avoid steric clashes, to provide an efficient, but physically reasonable model of repulsive interatomic forces. No attractive forces are explicitly used. A total of 23 potential energy wells are sampled: one for each of the 20 different amino acid types, and specific potentials accounting for the particular backbone conformational propensities of residues that precede prolines, prolines that precede prolines, and glycines that precede prolines. The simplicity of the model makes the structure ensemble generation highly efficient (100 000 structures of a 100-amino acid protein can be created in 30 min on a single processor, although the time increases with the number of experimental parameters that are simultaneously predicted). The complete absence of experimental constraints in this sampling phase avoids distortions due to additional potential energy terms such as those used in restrained MD calculations. Although this statistical coil model of the unfolded state has been tested with respect to its predictive power of diverse experimental parameters (for example RDCs and chemical shifts, and small angle scattering curves), it is simple for the user to replace the statistical coil potentials by an alternative description by specifying that a certain percentage of residues that precede prolines, prolines that precede prolines, and glycines that precede prolines.

Additional conformational propensities can be added to influence the sampling of the protein in the following ways (Figure 1):

1. Local conformational propensities can be modified on an amino acid-specific basis to include an additional potential, centred on a specific Gaussian shaped region of backbone dihedral angle space \( \Delta \phi_{\text{target}}/\Delta \psi_{\text{target}} \), of width \( \Delta \phi_{\text{target}}/\Delta \psi_{\text{target}} \), populated with a propensity \( p_{\text{oes}} \). The widths of the additional Gaussian shaped potentials and their propensities can be set by the user, and are simply added to, or replace, the existing potential.

2. Conformational propensities of regions of the primary sequence can be modified to include the presence of continuous secondary structure—either \( \alpha \)-helical, \( \beta \)-sheet or polyproline II, with propensity \( p_{\text{oes}} \). Propensities can be introduced in a cooperative, or independent manner, for example a complete helix can be constructed for 20% of conformers, (cooperative) or 20% additional helical sampling can be introduced randomly in the same sequence (non-cooperative).

3. Long-range contacts can be included in the conformational description by specifying that a certain percentage of structures \( p_{\text{galakh}} \), must contain at least one C\(^\alpha\) atom from an amino acid between residues \( i \) and \( j \), that is closer than \( \phi_{\text{max}} \) away from a C\(^\alpha\) atom in any amino acid between residues \( k \) and \( l \). The calculation can be used to generate explicit coordinates of each conformer in the ensemble, or simply to calculate appropriately averaged observables that would be expected in the presence of such a conformational regime (see below). At present the introduction of folded domains into the ensemble is restricted to the presence of individual secondary structural elements, or regions that can be uniquely encoded by their \( \phi/\psi \) values.

### 3.2 Application to the prediction of experimental parameters

After construction of the conformational ensemble, expected experimental values can be calculated from this ensemble.

#### 3.2.1 Residual dipolar couplings

RDCs report on local conformational sampling of each amino acid in the sequence, and are exquisitely sensitive to the presence of even weakly populated secondary structural elements (Jensen and Blackledge, 2008; Jensen et al., 2009). The dipolar coupling of a given magnetic moment with any other magnetic moment in its surroundings is given by (Blackledge, 2005):

\[
D_{ij} = -\frac{1}{\hbar} \frac{1}{2\pi^2 r_{ij}^3} \left[ A_\text{IS}(3\cos^2\theta - 1) + \frac{3}{2} A_\text{IS} \sin^2\theta \cos(2\phi) \right]
\]

(1)

where \( \gamma \) is the gyromagnetic ratio for the two spins \( I \) and \( S \), \( r_{ij} \) is the distance between the spins, \( \mu_0 \) is the permeability of free space, and \( h \) is Planck’s constant. In the fast exchange regime, the measured RDC reports on the arithmetic average over all conformations sampled up to the millisecond timescale. It has been shown that averaging of expected RDCs from each conformer in the ensemble using the expression in Equation (1) gives reasonable reproduction of the distribution of experimental RDCs measured in IDPs and denatured proteins. The RDCs from each internuclear vector \( IS \) are calculated from the orientation \( (\theta, \phi) \) with respect to the alignment tensor of each individual conformer \( j \) with axial and rhombic components \( (\lambda, \alpha) \) :

\[
D_{ij} = -\frac{1}{\hbar} \frac{1}{2\pi^2 r_{ij}^3} \left[ A_\text{IS}(3\cos^2\theta - 1) + \frac{3}{2} A_\text{IS} \sin^2\theta \cos(2\phi) \right]
\]

(2)

The alignment tensors are calculated using an in-house routine, based on previous published algorithms (Berlin et al., 2009, Zweckstetter and Bax, 2000). The average shown in Equation (1) is then approximated by the calculation of the mean of the RDCs over all structures in the ensemble:

\[
D_{IS} = \langle D_{ij} \rangle
\]

(3)

As discussed in previous publications, this average has highly unfavourable convergence characteristics (Nodet et al., 2009). As a rule of thumb, \( n \times 1000 \) structures are required in the ensemble, where \( n \) is the number of amino acids in the sequence, before convergence of the RDC of a specific amino acid has been achieved. It is possible to alleviate the convergence problem by calculating the alignment characteristics of uncorrelated 'local alignment windows’ (LAWs) (Marsh et al., 2008), but this approach requires knowledge of the explicit modulation of the underlying baseline of the RDC profile. The baseline profile normally exhibits a bell-shaped distribution of RDCs, however, this profile may be significantly modulated in the presence of persistent long-range contacts between regions of the chain that are distant in the primary sequence (Nodet et al., 2009; Salmon et al., 2010). When selecting ensembles of structures in agreement with experimental data, it is useful to average over a smaller number of structures, and in this case the combination
of LAWs and explicit modulation of the underlying baseline has been shown to provide a means of combining RDCs and paramagnetic relaxation enhancements (PREs) for ensemble selection (Salmon et al., 2010). Here, we concentrate on the prediction of values of experimental parameters that would be expected under given conformational sampling regimes, so we have chosen to retain the explicit molecular description for all parameters. The algorithm allows for the definition of long-range contacts between different regions of the primary sequence at a given propensity (vide supra), alone, or in combination with given populations of secondary structural motifs. This allows the prediction of expected RDC profiles even in the presence of complex levels of local and long-range structure.

3.2.2 Scalar \(^3\)J_{HNα} couplings Scalar couplings between amide and alpha protons (\(^3\)J_{HNα}) reflect on the conformationally averaged \( φ \) dihedral angle (Pardi et al., 1984; Ludvigsen et al., 1991; Vuister and Bax, 1993). The following Karplus relationship is used to calculate the values for each conformer:

\[ \text{J}_{\text{HNα}}(\phi) = A \cos^2(\phi - 60^\circ) + B \cos(\phi - 60^\circ) + C \]

(4)

which are then averaged over the ensemble. \( A, B \) and \( C \) have been optimized using coupling constants measured in several proteins of known structure and therefore provide a constraint on the distribution of \( \phi \) angles in conformational ensembles of IDPs (Mukrasch et al., 2007; Smith et al., 1996).

3.2.3 Paramagnetic relaxation enhancements A coherent picture of the conformational behaviour of IDPs and partially folded proteins requires not only a mapping of local structure but also long-range order. Long-range interactions in IDPs are often transient in nature and their detection, therefore, requires a strong probe that is active over a few nanometers such as that provided by an unpaired electron. One of the most efficient ways of introducing an unpaired active over a few nanometers such as that provided by an unpaired electron is by attaching a thiol reactive methanethiosulfonate (MTSL) spin label to the protein through a cysteine residue. The dipolar interaction between the unpaired electron and the protein nuclei induces PREs that strongly depend on the electron-nucleus distance. By introducing spin labels at several different positions in the protein, a mapping of long-range interactions in the disordered state becomes possible.

The transverse relaxation rate due to the presence of the unpaired electron can be expressed as follows (Abragam, 1994; Gillespie and Shortle, 1997):

\[ \Gamma_{2,H} = \frac{1}{I^2} \left[ \frac{1}{\gamma_H} \left( \gamma_B \mu_B e B \right)^2 \right] \left( \gamma_B \mu_B e B \right) \left( I(0) + 3I(\omega_0) \right) \]

(5)

where \( \gamma_H \) is the gyromagnetic ratio of the observed nucleus (proton), \( \gamma_B \) is the Bohr magneton, and \( I(0) \) is the intrinsic relaxation rate of the amide proton. \( \omega_0 \) is the sufficient frequency in free space. PREs can be calculated over structural ensembles by considering a fixed position of the MTSL side-chain (for example on the Cβ atom of the cysteine) and by invoking the spectral density function:

\[ J(\omega) = \left[ \frac{\omega_0}{I(\omega)} \right] \left[ \frac{\tau_e}{1 + \omega_0^2 \tau_e^2} \right] \]

(6)

where \( \omega_0 \) is the correlation time of the relaxation active interaction. This simple description does not account for the potentially high level of flexibility of the spin label itself (the electron spin label is attached to the molecule via MTSL attached to a cysteine side-chain). In order to address this, in the flexible-meccano approach, MTSL conformations are built explicitly for each backbone conformer by randomly sampling available rotamers and retaining only conformations that are sterically allowed. \( J(\omega) \) can be described using a model-free expression of the order parameter, comprising the orientational and distance-dependent components of the internal motion that both strongly depend on the motion of the spin label with respect to the observed nuclear spin (Bruschweiler et al., 1992; Clore and Iwahara, 2009).

\[ J(\omega) = \left( \frac{\omega_0}{I(\omega)} \right) \left[ \frac{\tau_e}{1 + \omega_0^2 \tau_e^2} \right] \]

(7)

where the order parameter \( S_{2,\omega} \) describes the motion of the dipolar interaction vector, \( \tau_e = \tau_e/\omega_0 \), \( \tau_e \) is a function of \( \tau_e \) and \( \tau_e \) the electron spin and rotational correlation times, respectively, and \( \tau_e = 1/(\tau_e^3 + \tau_e^3 + \tau_e^3) \) where \( \tau_e \) is the local correlation time of the spin label. \( \tau_e \) is the instantaneous distance between the proton and the electron spins. Order parameters can be expressed in terms of radial and angular components:

\[ S_{2,\omega} = S_{2,\omega}^2 S_{2,\omega}^2 \]

(8)

where:

\[ S_{2,\omega} = \left( \frac{-\omega_0}{I(\omega)} \right) \left[ \frac{\tau_e}{1 + \omega_0^2 \tau_e^2} \right] \]

(9)

\[ \Omega_{\omega} = \left( \frac{-\omega_0}{I(\omega)} \right) \left[ \frac{\tau_e}{1 + \omega_0^2 \tau_e^2} \right] \]

(10)

\( \Omega_{\omega} \) describes the orientation of the interaction vector in the frame of each conformer. The above expressions are used to calculate the transverse relaxation rate for each backbone conformation produced with the flexible-meccano algorithm, and the effective relaxation rate for each amide proton is then averaged over all retained conformers:

\[ \Gamma_{\text{total}} = \frac{1}{I^2} \sum_{i=1}^{I_{\text{dim}}} \Gamma_{\text{mix}} \]

(11)

PREs are often described in terms of the ratio between the intensity measured in a standard HSQC experiment in the presence of the reduced and oxidized forms of the MTSL label:

\[ \frac{I_{\text{red}}}{I_{\text{ox}}} = \exp \left( -\frac{1}{2} \frac{1}{\tau_{\text{mix}}} \right) \]

(12)

Here \( \frac{I_{\text{red}}}{I_{\text{ox}}} \) is the intrinsic relaxation rate of the amide proton and \( \tau_{\text{mix}} \) is the mixing time during which relaxation occurs in the HSQC pulse sequence (typically 10 ms).

3.2.4 Chemical shift prediction Chemical shifts are the most accessible NMR parameter, and provide powerful probes of conformation, in particular of secondary structural propensity (Eliezer et al., 2001; Kjaergaard et al., 2011; Marsh et al., 2006; Modig et al., 2007; Schwarczinger et al., 2001). Remarkable progress has been made in recent years in the prediction of chemical shifts from protein conformation (Wishart and Sykes, 1994), and
Flexible-meccano

Fig. 2. Screenshot of the flexible-meccano program, showing the interface allowing the user to define conformational propensities [in this case an α-helical propensity between residues 12 and 18 (top), and a long-range contact between positions 5–15 and 25–30 (bottom)].

vice-versa (Berjanskii et al., 2009; Cavalli et al., 2007; Shen et al., 2008). We have previously investigated the possibility of combining flexible-meccano with chemical shift prediction to analyze experimental chemical shifts in IDPs (Jensen et al., 2010).

While $^{13}$C$_\alpha$, $^{13}$C$_\beta$, $^{13}$C' and H$_\alpha$ chemical shifts depend strongly on $\phi$ and $\psi$, $^{15}$N and $^1$H$_N$ chemical shifts show a more or less uniform dependence on the two dihedral angles. In addition, the $^{13}$C$_\alpha$ and $^{13}$C$_\beta$ chemical shifts display an almost inverse dependence on the $\phi/\psi$ distribution and therefore report on the populations of α-helix and β-sheet in the disordered state. Scripts are provided with the flexible-meccano program that can be directly interfaced to well-known chemical shift prediction algorithms such as Sparta (Shen and Bax, 2007) and ShiftX (Han et al., 2011; Neal et al., 2003).

3.2.5 SAXS prediction Small angle X-ray scattering provides complementary information about the extent of conformational sampling of the unfolded protein. Scripts are again provided that allow the user to interface the flexible-meccano program with the SAXS program Crysol (Svergun et al., 1995) and calculate expected SAXS curves that would be associated with the given ensemble (Bernadó and Blackledge, 2010; Bernadó and Svergun, 2012; Bernadó et al., 2007).

4 IMPLEMENTATION

Examples of the implementation of flexible-meccano are shown in Figures 2 and 3. The following steps are shown:

(1) Reading of the primary sequence of the protein. This provides a scrollable table describing the conformational potentials to be used in the statistical sampling for each $\phi/\psi$ pair. The default potentials are those provided by the flexible-meccano statistical coil library.

Fig. 3. Screenshot montage of the flexible-meccano program, showing the interface allowing the user to select which experimental and biophysical parameters to be calculated from the ensemble. As examples, $^1$D$_{NH}$ and $^1$D$_{RDC}$ RDCs, a distribution of radius of gyration, and PRE profiles, are shown in black, light grey and dark grey, respectively. This calculation corresponds to the top screen in Figure 2, with 50% of conformers containing a helical element from residue 12–18 (giving rise to positive $^1$D$_{NH}$ RDCs in this region).

3.2.6 Ramachandran segment division In order to describe the sampling of conformational space in the different ensembles, the flexible-meccano program provides a statistical analysis of the Ramachandran space sampled by the ensemble. In order to do this, $\phi, \psi$ space is divided into four quadrants as follows; $\alpha_L$: \{ $\phi > 0^\circ$ \}, $\alpha_R$: \{ $\phi < 0^\circ$, $-120^\circ < \psi < 50^\circ$ \}, $\beta_P$: \{ $-100^\circ < \phi < 0^\circ$, $\psi > 50^\circ$ or $\psi < -120^\circ$ \}, $\beta_S$: \{ $-180^\circ < \phi < -100^\circ$, $\psi > 50^\circ$ or $\psi < -120^\circ$ \}. The population of these quadrants is indicated as $p_{\alpha_L}$, $p_{\alpha_R}$, $p_{\beta_P}$ and $p_{\beta_S}$.
of a long range contact between residues 1–10 and 130–140. (populate 30% more helix, and 58–91, which populate 20% more PPII and et al. 2004), (alpha-synuclein. Simulation used coil sampling except for the presence (of a long range contact between residues 1–10 and 130–140). (C) alpha-synuclein. Simulation used coil sampling except for the presence of a long range contact between residues 1–10 and 130–140.

(2) It is possible to modify the potentials by adding additional propensities. In the example shown in Figure 2 the strand from residue 12–18 populates an α-helix for 50% of the conformers. The remaining conformers follow the amino acid-specific potentials. It is also possible to introduce specific sampling for each amino acid, focusing on specific regions of Ramachandran space. All modifications can be introduced ‘by-hand’ in the input file containing the sequence information, or via the graphical interface.

(3) Long-range contacts between different regions of the primary sequence are also specified using the graphical interface, where the range of amino acids involved in the contact, and the maximum distance between any residues in the two ranges, are introduced.

(4) The user can choose between the following output options: the number of structures in the ensemble to be calculated, whether or not explicit structural coordinates should be written for each conformer in the ensemble, and which data types (RDCs, PREs, J-couplings etc.) should be predicted from the ensemble. In the case of RDC prediction, the user can select the RDCs to be calculated, whereas for PRE prediction the position of the cysteine mutants must be specified, the magnetic field strength and the intrinsic (diamagnetic) linewidth in the proton dimension.

(5) The ensemble calculation and data prediction algorithm runs entirely as a background calculation, so that the user can either run numerous calculations from the same interface, or analyze the results of one calculation while running another. All results are robustly classified in terms of date and time of initialization of each calculation.

(6) Data output is provided in text format, and in graphical form (postscript). In addition to the specified data types, the program also provides a distribution of the radius of gyration over the ensemble.

(7) Robust scripts are provided that interface the resulting coordinate files with protocols for calculating ensemble averaged chemical shifts and small angle scattering curves.

Examples of RDCs calculated using flexible-meccano in comparison to experimental values from apomyoglobin (Mohana-Borges et al., 2004), alpha-synuclein (Bernadó et al., 2005a) and p53 (Wells et al., 2008) are shown in Figure 4 along with the conformational sampling regime used to reproduce the data.

5 DISCUSSION

The classical structure function paradigm that has successfully underpinned our understanding of molecular biology breaks down when considering proteins that have no stable tertiary structure in their functional form. The determination of a 3D structure can provide only a single snapshot of such a highly flexible system, and alternative methods are essential to study the behaviour of these disordered proteins (Fisher and Stultz, 2011; Marsh et al., 2010; Mittag et al., 2010; Tompa, 2011). One promising approach is to describe the protein in terms of an equilibrium of rapidly inter-converting conformers. Currently, tools to generate such ensemble descriptions are rare, and poorly adapted to the prediction of experimental data. In this article, we present an algorithm that generates ensembles of molecules, on the basis of amino acid-specific conformational potentials and volume exclusion. Conformational sampling depends uniquely on the primary sequence, with the possibility of introducing additional local or long-range conformational propensities. We show how the algorithm can be used to calculate expected values of experimental NMR parameters measured at atomic or molecular resolution, for a broad range of user-defined conformational sampling regimes.

We envisage three generic levels of interest: (i) Researchers who have measured experimental data from a particular IDP of interest and are motivated to compare their data with those expected from a fully disordered protein with this specific primary sequence. No such tool exists at this time. (ii) Researchers who see evidence of deviation from ‘random coil’ behaviour in their protein, and who would like to determine whether these data are in agreement with a particular conformational sampling regime (e.g. a particular propensity of helical sampling in a given region of the chain, or the presence of weak long-range contacts between parts of the chain that are distant in primary sequence; Nodet et al., 2009; Salmon et al., 2010; Schneider et al., 2012). (iii) Researchers who are interested in working with a broad ensemble of conformers representing the flexibility of the IDP of interest, either to use ‘sample and select’ approaches to develop a sub-ensemble in agreement with experimental data, or to seed molecular dynamics or restrained
ensemble molecular dynamics simulations of their protein. We add a final note of caution: any analysis that involves detailed inspection of the structure of specific conformers must of course be performed in the knowledge that no ensemble of highly disordered proteins can ever be considered to be unique. Agreement with experiment only confirms that a given ensemble does not violate a specific data type, which itself is only sensitive to particular aspects of the conformational sampling.

Funding: Agence National de Recherche for financial support from TAUSTRUCT—ANR MALZ 2010 (to M.B.), ProteinDisease—ANR JCJC 2010 (to M.R.J.), Spin-HD—ANR CHEX 2011 (to P.B.) and the GIP computational support group of the Commissariat à l’Energie Atomique et aux energies alternatives (CEA).

Conflict of Interest: none declared.

REFERENCES


