BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 11 2012, pages 1463–1470
doi:10.1093/bioinformatics/bts172

Structural bioinformatics

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

Valéry Ozenne[1], Frédéric Bauer[2], Loïc Salmon[1], Jie-rong Huang[1], Malene Ringkjøbing Jensen[1], Stéphane Segard[2], Pau Bernadó[3], Céline Charavay[2] and Martin Blackledge*[,1]

[1]Protein Dynamics and Flexibility, Institut de Biologie Structurale Jean-Pierre Ebel, CEA; CNRS; UJF UMR 5075, 41 Rue Jules Horowitz, Grenoble 38027, [2]Groupe 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 [3]Centre de Biochimie Structurale, CNRS UMR 5048 - UM 1 - INSERM UMR 1054, 34090, Montpelier, France

Associate Editor: Burkhard Rost

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).

*To whom correspondence should be addressed.

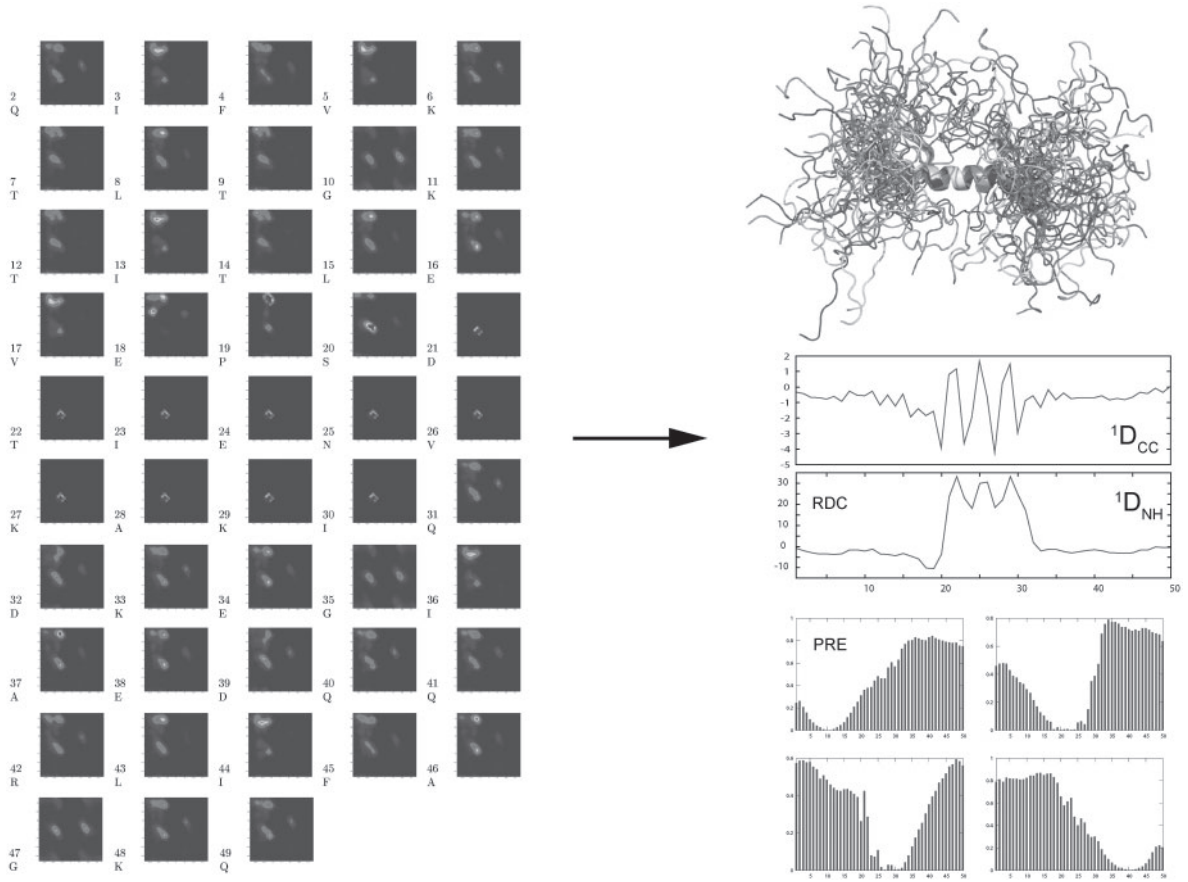

Fig. 1. Schematic representation of the *flexible-meccano* algorithm. Conformational sampling details are defined by the user (shown figuratively on the left in terms of Ramachandran sampling for each amino acid)—the default is the *flexible-meccano* statistical coil description. This can be modified 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 ϕ/ψ 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 of the unfolded state.

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 $\{\phi_{\text{target}}/\psi_{\text{target}}\}$, of width $\{\Delta\phi_{\text{target}}/\Delta\psi_{\text{target}}\}$, populated with a propensity p_{res}. 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 α-helical, β-sheet or polyproline II, with propensity p_{sec}. 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_{dist}), must contain at least one C^{α} atom from an amino acid between residues i and j, that is closer than $d_{\max}$ away from a C^{α} 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):

$$D_{IS} = -\frac{\gamma_I \gamma_S \mu_0 h}{16\pi^3 r_{IS}^3} \left\langle P_2 \cos\left(\theta_{IS}\right) \right\rangle \quad (1)$$

where γ is the gyromagnetic ratio for the two spins I and S, r_{IS} is the distance between the spins, μ_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 (2) 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 (θ, φ) with respect to the alignment tensor of each individual conformer (j) with axial and rhombic components (A_a, A_r) :

$$D_{IS}^{j} = -\frac{\gamma_I \gamma_S \hbar \mu_0}{8\pi^2 r_{IS}^3} \left[A_a (3\cos^2\theta - 1) + \frac{3}{2} A_r \sin^2\theta \cos(2\varphi) \right] \quad (2)$$

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:

$$D_{IS} = \left\langle D_{IS}^{j} \right\rangle \quad (3)$$

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 global 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 $^3J_{NHH\alpha}$ couplings Scalar couplings between amide and alpha protons ($^3J_{NHH\alpha}$) report 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:

$$^3J_{\mathrm{NHH}\alpha}(\phi) = A\cos^2(\phi - 60°) + B\cos(\phi - 60°) + C \qquad (4)$$

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 ϕ 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 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 distances. 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):

$$\Gamma_{2,H} = \frac{1}{15}\left(\frac{\mu_0}{4\pi}\right)^2 \gamma_H^2 g_e^2 \mu_B^2 s_e(s_e+1)\{4J(0) + 3J(\omega_H)\} \qquad (5)$$

where g_e is the electron g-factor, γ_H is the gyromagnetic ratio of the observed nucleus (proton), s_e is the electron spin quantum number, ω_H is the proton frequency, μ_B is the Bohr magneton and μ_0 is the permittivity of 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:

$$J(\omega) = \left\langle r_{H-e}^{-6} \right\rangle \left\{ \frac{\tau_c}{1+\omega^2\tau_c^2} \right\} \qquad (6)$$

where τ_c 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):

$$J(\omega) = \left\langle r_{H-e}^{-6} \right\rangle \left\{ \frac{S_{H-e}^2 \tau_c}{1+\omega^2\tau_c^2} + \frac{\left(1 - S_{H-e}^2\right)\tau_e}{1+\omega^2\tau_e^2} \right\} \qquad (7)$$

where the order parameter S_{H-e}^2 describes the motion of the dipolar interaction vector, $\tau_c = \tau_r\tau_s/(\tau_r + \tau_s)$ is a function of τ_s and τ_r the electron spin and rotational correlation times, respectively, and $\tau_e = 1/(\tau_i^{-1} + \tau_r^{-1} + \tau_s^{-1})$ where τ_i is the local correlation time of the spin label. r_{H-e} is the instantaneous distance between the proton and the electron spins. Order parameters can be expressed in terms of radial and angular components:

$$S_{H-e}^2 = S_{\mathrm{ang}}^2 S_{\mathrm{rad}}^2 \qquad (8)$$

where:

$$S_{\mathrm{rad}}^2 = \left\langle r_{H-e}^{-6} \right\rangle^{-1} \left\langle r_{H-e}^{-3} \right\rangle^2 \qquad (9)$$

$$S_{\mathrm{ang}}^2 = \frac{4\pi}{5} \sum_{m=-2}^{2} \left| \left\langle Y_2^m\left(\Omega^{\mathrm{mol}}\right) \right\rangle \right|^2 \qquad (10)$$

Ω^{mol} 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:

$$\Gamma_2^{\mathrm{total}} = \frac{1}{n} \sum_{i=1}^{n} \Gamma_{2,i}^{\mathrm{fm}} \qquad (11)$$

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:

$$\frac{I}{I^0} = \frac{\Gamma_2^{\mathrm{red}} \exp\left(-\Gamma_2^{\mathrm{calc}} \tau_{\mathrm{mix}}\right)}{\Gamma_2^{\mathrm{red}} + \Gamma_2^{\mathrm{calc}}} \qquad (12)$$

Here Γ_2^{red} is the intrinsic relaxation rate of the amide proton and τ_{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; Schwarzinger *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

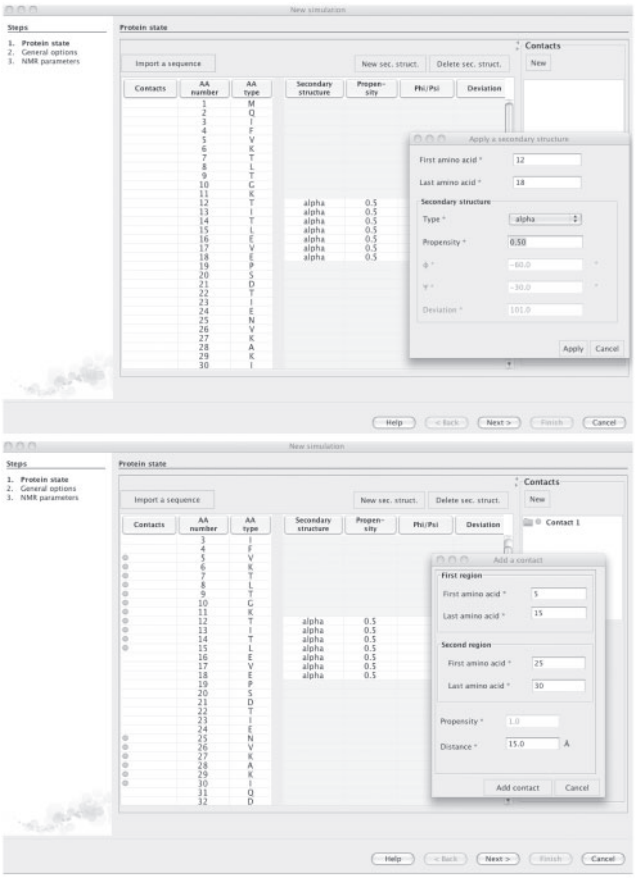

Fig. 2. Screenshot of the *flexible-meccano* program, showing the interface allowing the user to define conformational propensities [in this case an α-helical propensity between residues 12 and 18 (top), and a long-range contact between positions 5–15 and 25–30 (bottom)].

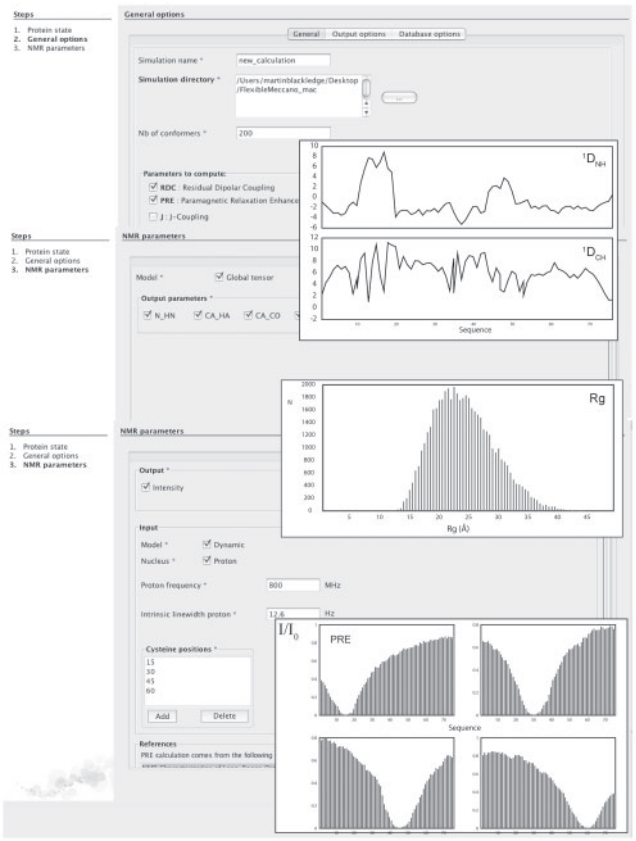

Fig. 3. Screenshot montage of the *flexible-meccano* program, showing the interface allowing the user to select which experimental and biophysical parameters to be calculated from the ensemble. As examples, $^{1}D_{NH}$ and $^{1}D_{C\alpha H\alpha}$ RDCs, a distribution of radius of gyration, and PRE 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 form residue 12–18 (giving rise to positive $^{1}D_{NH}$ RDCs in this region).

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 Hα chemical shifts depend strongly on ϕ and ψ, ^{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 ϕ/ψ 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).

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, ϕ,ψ space is divided into four quadrants as follows; α_L: $\{\phi>0°\}$, α_R : $\{\phi<0, -120°<\psi<50°\}$, β_P : $\{-100°<\phi<0°\}$, $\psi>50°$ or $\psi<-120°\}$, β_S : $\{-180°<\phi<-100°\}$, $\psi>50°$ or $\psi<-120°\}$. The population of these quadrants is indicated as $p_{\alpha L}, p_{\alpha R}, p_{\beta P}$ and $p_{\beta S}$.

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 ϕ/ψ pair. The default potentials are those provided by the *flexible-meccano* statistical coil library.

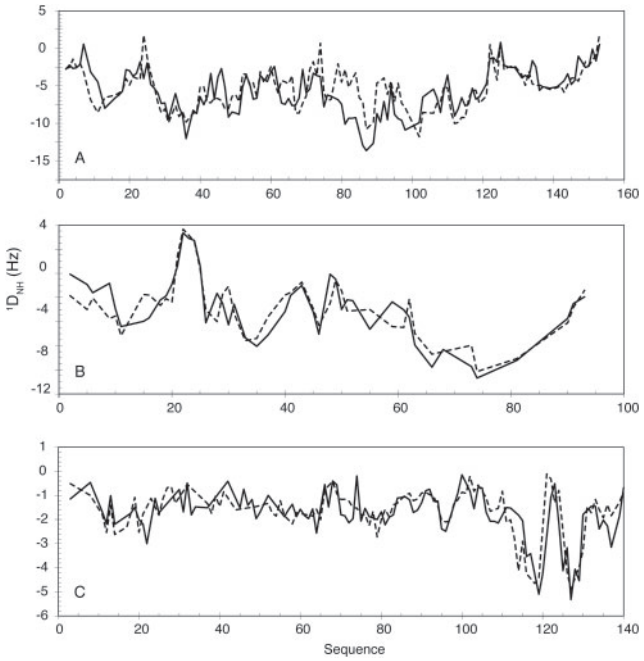

Fig. 4. RDCs calculated using *flexible-meccano* (solid lines), compared with experimental values (dashed lines) for (**A**) apomyoglobin in 8 M Urea (simulation used statistical coil for all residues; Mohana-Borges *et al*., 2004), (**B**) N-terminal disordered transactivation domain of p53 (1–93; Wells *et al*., 2008), simulation used coil sampling except for residues 22–24 which populate 30% more helix, and 58–91, which populate 20% more PPII and (**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.), ProteinDisorder—ANR JCJC 2010 (to M.R.J.), Spin-HD—ANR CHEX 2011 (to P.B.) and the GIPSE computational support group of the Commissariat à l'Energie Atomique et aux energies alternatives (CEA).

Conflict of Interest: none declared.

REFERENCES

Abragam,A. (1994) *The Principles of Nuclear Magnetism Reprint*. Clarendon Press, Oxford, UK.

Babu,M.M. (2012) Intrinsically disordered proteins. *Mol. Biosyst.*, **8**, 21.

Babu,M.M. *et al*. (2011) Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.*, **21**, 432–440.

Berjanskii,M. *et al*. (2009) GeNMR: a web server for rapid NMR-based protein structure determination. *Nucleic Acids Res.*, **37**, W670–W677.

Berlin,K. *et al*. (2009) Improvement and analysis of computational methods for prediction of residual dipolar couplings. *J. Magn. Reson.*, **201**, 25–33.

Bernadó,P. and Blackledge,M. (2010) Structural biology: proteins in dynamic equilibrium. *Nature*, **468**, 1046–1048.

Bernadó,P. and Svergun,D.I. (2012) Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol. Biosyst.*, **8**, 151–167.

Bernadó,P. *et al*. (2005a) Defining long-range order and local disorder in native alpha-synuclein using residual dipolar couplings. *J. Am. Chem. Soc.*, **127**, 17968–17969.

Bernadó,P. *et al*. (2005b) A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc. Natl Acad. Sci. USA*, **102**, 17002–17007.

Bernadó,P. *et al*. (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.*, **129**, 5656–5664.

Blackledge,M. (2005) Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. *Prog. Nucl. Magn. Reson. Spectrosc.*, **46**, 23–61.

Bouvignies,G. *et al*. (2006a) Simultaneous determination of protein backbone structure and dynamics from residual dipolar couplings. *J. Am. Chem. Soc.*, **128**, 15100–15101.

Bouvignies,G. *et al*. (2006b) Ultrahigh-resolution backbone structure of perdeuterated protein GB1 using residual dipolar couplings from two alignment media. *Angew. Chem. Int. Ed. Engl.*, **45**, 8166–8169.

Bruschweiler,R. *et al*. (1992) Influence of rapid intramolecular motion on NMR cross-relaxation rates - a molecular-dynamics study of Antamanide in solution. *J. Am. Chem. Soc.*, **114**, 2289–2302.

Cavalli,A. *et al*. (2007) Protein structure determination from NMR chemical shifts. *Proc. Natl Acad. Sci. USA*, **104**, 9615–9620.

Clore,G.M. and Iwahara,J. (2009) Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. *Chem. Rev.*, **109**, 4108–4139.

Dunker,A.K. and Uversky,V.N. (2010) Drugs for "protein clouds": targeting intrinsically disordered transcription factors. *Curr. Opin. Pharmacol.*, **10**, 782–788.

Dunker,A.K. *et al*. (2002) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.

Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**, 197–208.

Eliezer,D. *et al*. (2001) Conformational properties of alpha-synuclein in its free and lipid-associated states. *J. Mol. Biol.*, **307**, 1061–1073.

Fisher,C.K. and Stultz,C.M. (2011) Constructing ensembles for intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, **21**, 426–431.

Gillespie,J.R. and Shortle,D. (1997) Characterization of long-range structure in the denatured state of staphylococcal nuclease. II. Distance restraints from paramagnetic relaxation and calculation of an ensemble of structures. *J. Mol. Biol.*, **268**, 170–184.

Han,B. *et al*. (2011) SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR*, **50**, 43–57.

Hus,J.-C. *et al*. (2001) Determination of protein backbone structure using only residual dipolar couplings. *J. Am. Chem. Soc*, **123**, 1541–1542.

Hus,J.-C. *et al*. (2008) 16-fold degeneracy of peptide plane orientations from residual dipolar couplings: analytical treatment and implications for protein structure determination. *J. Am. Chem. Soc.*, **130**, 15927–15937.

Jensen,M.R. and Blackledge,M. *et al*. (2008) On the origin of NMR dipolar waves in transient helical elements of partially folded proteins. *J. Am. Chem. Soc.*, **130**, 11266–11267.

Jensen,M.R. *et al*. (2008) Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein. *J. Am. Chem. Soc.*, **130**, 8055–8061.

Jensen,M.R. *et al*. (2009) Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. *Structure*, **17**, 1169–1185.

Jensen,M.R. *et al*. (2010) Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *J. Am. Chem. Soc.*, **132**, 1270–1272.

Jensen,M.R. *et al*. (2011) Intrinsic disorder in measles virus nucleocapsids. *Proc. Natl Acad. Sci. USA*, **108**, 9839–9844.

Kjaergaard,M. *et al*. (2011) Random coil chemical shift for intrinsically disordered proteins: effects of temperature and pH. *J. Biomol. NMR*, **49**, 139–149.

Ludvigsen,S. *et al*. (1991) Accurate measurements of coupling constants from two-dimensional nuclear magnetic resonance spectra of proteins and determination of phi-angles. *J. Mol. Biol.*, **217**, 731–736.

Marsh,J.A. *et al*. (2006) Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. *Protein Sci.*, **15**, 2795–2804.

Marsh,J.A. *et al*. (2008) Calculation of residual dipolar couplings from disordered state ensembles using local alignment. *J. Am. Chem. Soc.*, **130**, 7804–7805.

Marsh,J.A. *et al*. (2010) Structural diversity in free and bound states of intrinsically disordered protein phosphatase 1 regulators. *Structure*, **18**, 1094–1103.

Mittag,T.*et al*. (2010) Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure*, **18**, 494–506.

Modig,K.*et al*. (2007) Detection of initiation sites in protein folding of the four helix bundle ACBP by chemical shift analysis. *FEBS Lett.*, **581**, 4965–4971.

Mohana-Borges,R. *et al*. (2004) Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. *J. Mol. Biol.*, **340**, 1131–1142.

Mukrasch,M.D. *et al*. (2007) Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. *J. Am. Chem. Soc.*, **129**, 5235–5243.

Neal,S. *et al*. (2003) Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. *J. Biomol. NMR*, **26**, 215–240.

Nodet,G. *et al*. (2009) Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. *J. Am. Chem. Soc.*, **131**, 17908–17918.

Pardi,A. *et al*. (1984) Calibration of the angular dependence of the amide proton-C alpha proton coupling constants, 3JHN alpha, in a globular protein. Use of 3JHN alpha for identification of helical secondary structure. *J. Mol. Biol.*, **180**, 741–751.

Salmon,L. *et al*. (2009) Protein conformational flexibility from structure-free analysis of NMR dipolar couplings: quantitative and absolute determination of backbone motion in ubiquitin. *Angew. Chem. Int. Ed. Engl.*, **48**, 4154–4157.

Salmon,L. *et al*. (2010) NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.*, **132**, 8407–8418.

Schneider,R. *et al*. (2012) Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. *Mol. Biosyst.*, **8**, 58–68.

Schwarzinger,S. *et al*. (2001) Sequence-dependent correction of random coil NMR chemical shifts. *J. Am. Chem. Soc.*, **123**, 2970–2978.

Shen,Y. and Bax,A. (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J. Biomol. NMR*, **38**, 289–302.

Shen,Y. *et al*. (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl Acad. Sci. USA*, **105**, 4685–4690.

Smith,L.J. *et al*. (1996) Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations. *J. Mol. Biol.*, **255**, 494–506.

Svergun,D. *et al*. (1995) CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.*, **28**, 768–773.

Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.

Tompa,P. (2011) Unstructural biology coming of age. *Curr. Opin. Struct. Biol.*, **21**, 419–425.

Uversky,V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*, **11**, 739–756.

Vendruscolo,M. and Dobson,C.M. (2007) Chemical biology: more charges against aggregation. *Nature*, **449**, 555.

Vuister,G. and Bax,A. (1993) Quantitative J correlation - a new approach for measuring homonuclear 3-bond J(H(N)H(alpha) coupling-constants in N-15-enriched proteins. *J. Am. Chem. Soc.*, **115**, 7772–7777.

Wells,M. *et al*. (2008) Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc. Natl Acad. Sci. USA*, **105**, 5762–5767.

Wishart,D.S. and Sykes,B.D. (1994) The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J. Biomol. NMR*, **4**, 171–180.

Zweckstetter,M. and Bax,A. (2000) Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. *J. Am. Chem. Soc.*, **122**, 3791–3792.