Ensemble models of proteins and protein domains based on distance distribution restraints

Journal Article

Author(s):
Jeschke, Gunnar

Publication date:
2016-04

Permanent link:
https://doi.org/10.3929/ethz-a-010745315

Originally published in:
Proteins: Structure, Function and Bioinformatics 84(4), https://doi.org/10.1002/prot.25000

Funding acknowledgement:
157034 - Nanometer-range distance constraints for biomolecular complexes - Measurement, uncertainty estimates, and use in modeling (SNF)

Ensemble models of proteins and protein domains based on distance distribution restraints

Short title: Ensembles by distance distribution restraints

Keywords: intrinsically disordered proteins, intrinsically disordered domains, EPR spectroscopy, NMR spectroscopy, membrane proteins, mixed resolution, statistical coil

Gunnar Jeschke

ETH Zürich, Laboratory of Physical Chemistry, Vladimir-Prelog-Weg 2, 8093 Zürich, Switzerland

Institution: ETH Zürich, Laboratory of Physical Chemistry

Corresponding author:

Gunnar Jeschke

ETH Zürich, Laboratory of Physical Chemistry, Vladimir-Prelog-Weg 2, 8093 Zürich, Switzerland

e-mail: gjeschke@ethz.ch,

Tel.: +41 44 632 5702

Fax.: +41 44 633 1448

Abstract

Conformational ensembles of intrinsically disordered peptide chains are not fully determined by experimental observations. Uncertainty due to lack of experimental restraints and due to intrinsic disorder can be distinguished if distance distributions restraints are available. Such restraints can be obtained from pulsed dipolar electron paramagnetic resonance (EPR) spectroscopy applied to pairs of spin labels. Here, we introduce a Monte Carlo approach for generating conformational ensembles that are consistent with a set of distance distribution restraints, backbone dihedral angle statistics in known protein structures, and, optionally, secondary structure propensities or membrane immersion depths. The approach is tested with simulated restraints for a terminal and an internal loop and for a protein with 69 residues by using sets of sparse restraints for underlying well-defined conformations.

Introduction

Intrinsic disorder is one of the fundamental design principles in proteins. It conveys functional advantages in the context of posttranslational modifications and binding to different partners[1-3] and is probably a key feature of cell cycle regulation.[4] Much progress has been made in structural and dynamical characterization of intrinsically disordered proteins (IDPs) and domains (IDDs),[5-8] yet the problem of deriving statistically reliable conformational ensembles from experimental data is not solved. This problem was summarized in a study on the unfolded state of proteins.[9] Statistical coil models derived from different libraries for the preferences of backbone dihedrals were found to be in similar good agreement with ^{15}N-^{1}H^{N} residual dipolar couplings from NMR experiments on apo-myoglobin in 8 M urea in a compressed 10% polyacrylamide gel,[10] although the content of α, β, and polyproline II secondary structure varied significantly between them.

Essentially, the problem is the one of statistical mechanics. Because of an unavoidable lack of experimental information, the observable macrostate is consistent with a large number of microstates. In statistical mechanics observables are computed from the most probable macrostate, which is representative for a sufficiently large ensemble. For IDPs and IDDs it is usually impossible to find this macrostate by generating and analysing conformational ensembles of sufficient size. The problem can be avoided by generating the smallest possible ensemble that is consistent with experimental evidence. However, such a minimal ensemble might broaden when further experimental data is added, so that uncertainty of observables would paradoxically increase. Here we follow the alternative approach of specifying a maximal ensemble that includes any tested conformation which is consistent with all experimental data. In practice, ensemble size is usually limited to a few ten to a few hundred conformations.[11] However, by varying size or by repeating Monte Carlo simulation it can be assessed whether expectation values of observables have converged. If this is the case and restraints are correct, the distribution width of an expectation value in the simulated ensemble is an upper bound for the width in the true distribution of conformers.

The maximal ensemble approach encounters the problem that most experimental data are ensemble averages. During generation of the ensemble these averages are not yet available. It is thus necessary to first generate a large ensemble and later reduce it to an ensemble that is consistent with the restraints.[12] Since conformational space is vast, even the largest ensemble that can be generated and stored may not sufficiently sample the populated region of conformational space. Accessible space can be sampled more densely if inaccessible subspace can be recognized early during ensemble generation. For this, experimental restraints must be available as distributions rather than only as mean values.

Pulsed dipolar EPR spectroscopy techniques, such as double electron electron resonance (DEER, also called PELDOR)[13,14] or double-quantum coherence (DQC) build-up,[15] can provide distance distribution information[16-18] on length scales between about 10 Å by DQC[19] and 140 Å with deuterated proteins.[20] These length scales match the dimensions of IDPs and IDDs. In most cases the experiments require site-directed spin labelling[21] and thus provide label-to-label distance distribution restraints (DDRs). Rotamer library[22,23] and accessible space[24] approaches can model the distribution of spin label coordinates for a given structural context. It has been demonstrated that a small number of spin label DDRs is sufficient to localize cofactors[25,26] in ordered domains or residues in disordered domains.[27] Spin labels, which are also used for deriving paramagnetic relaxation enhancement (PRE) restraints by NMR,[28] can potentially bias the actual conformational distribution. However, tests on well-structured proteins in solution and in membranes as well as crystal structures of spin-labelled proteins[29] indicate that such bias is usually weak and accounted for by the uncertainty of predicting spin label side chain conformations that we assume in this work.

Here we present an approach for generating conformational ensembles that are consistent with preferences for backbone dihedral angles derived from the PDB, with spin label DDRs, with optional secondary structure propensities, as they can be obtained by NMR experiments,[30,31] and with optional membrane immersion depth restraints for membrane proteins. Preliminary versions of the algorithm were used with experimental DDRs to model the linker between FnIII-3,4 domains of integrin α6β4,[32] a moderately disordered region of the pro-apoptotic protein Bax in its activated form,[27] and a disordered section of the N-terminal domain of major plant light harvesting complex LHCII.[33] Here, the approach is tested with simulated DDRs for a terminal domain, an internal domain, and a peptide with DDRs that were derived *in silico* from well-defined template structures.

Materials and Methods

General

The program was implemented in Matlab® (The MathWorks Inc., Natick, MA) and is freely available as source code at www.epr.ethz.ch/software as part of the MMM[22] package. Template structures were downloaded from the Protein Data Base (PDB) or the Protein Ensemble Database.[11] Visualization was performed in MMM.

Simplified flow charts of the ensemble generator and of the backbone generator for a single conformer are given in Fig. 1A and B, respectively. The processes and tests in these flow charts are explained in more detail in the following Sections. For internal IDDs, the last residue of each simulated conformer must be joined to the C-terminal anchor residue in the well-structured part of the protein. This requires a modification of the flow chart shown in Fig. 1B, which is also described below. The basic idea of the approach is maximization of sampling of conformational space at given computational effort. Hence, the computationally most efficient steps and consistency tests are performed first and more elaborate steps are performed later only for the few conformers that have passed the simpler tests.

Generation of the unrestrained ensemble

The unrestrained conformational ensemble should correspond to a statistical coil that reproduces the dihedral angle distribution in unfolded proteins. It has been shown that distributions from different libraries of peptide segments extracted from the PDB lead to similar results, as long as α-helical segments are excluded.[9] We rely on residue-specific distributions extracted by Hovmoller et al.[34] by excluding α-helix and β-strand sections. We further assume an incidence of *cis*-peptide bonds between residues *i* and *i*+1 of 0.03% if residue *i*+1 is not Pro and of 6.5% if it is.[35] Dihedral angles conforming to these statistics are randomly selected and the backbone geometry is generated by the Sugeta-Miyazawa algorithm,[36] for IDDs using expressions by Shimanouchi and Mizushima[37] for bootstrapping from an anchor residue. The anchor residue immediately precedes C-terminal and internal domains or immediately follows N-terminal domains.

After the whole backbone has been generated it is tested for self-clashes, which are defined as an approach of two atoms of non-consecutive residues within less than 2 Å. Conformations with self-clashes are discarded. For internal IDDs, the clash test is performed twice, once after half of the backbone has been generated and once at the end. Backbone conformations of IDDs are subjected to the same clash test with respect to non-hydrogen atoms of the structurally resolved core of the protein.

Conformations that have passed the clash tests are decorated with side chains using SCWRL4.[38] Self-avoidance and protein clash tests are then repeated with an adjustable clash threshold. We found that it is computationally more efficient in terms of sampling of conformational space to add side chains only after the entire backbone is generated and all

DDRs have been tested, even though a significant fraction of backbone models is rejected at this stage. Since SCWRL4 works with a modified repulsion potential, this clash threshold must be lower than the sum of van-der-Waals radii. Our own experience with spin label side chain modelling[22] and preliminary tests indicated that a threshold of 1.2 Å is appropriate.

Closure of internal loops

An internal IDD has a C-terminal anchor residue to which it must be attached. A random walk in conformational space has too low probability to end up sufficiently close to this anchor. Therefore, after generating half of the backbone, the loop is gently steered towards the anchor. To that end we define a mean distance reduction $\Delta R = R/n$ required per residue, where R is the distance between the Cα atom of the current residue and the Cα atom of the anchor and n is the number of residues yet to be modelled. A proposed Monte Carlo step by one further residue is associated with a distance reduction Δr. The step is accepted for an approach ratio $a = |\Delta r - \Delta R|/\Delta m < 0.5$, where $\Delta m = 2.65$ Å is a mean backbone progression between Cα atoms. Mean backbone progression per residue in a given direction cannot exceed the Cα-Cα distance of 3.8 Å and is rarely smaller than the progression of 1.5 Å along the helix axis in α-helices. Our choice of Δm is the mean of these two limiting values. Note however, that only the product $a\ \Delta m$ is relevant and that this product is an empirical parameter of the algorithm that balances success rate *versus* potential conformational bias.

If 100 attempts in a row fail to turn up $a < 0.5$, the conformation is discarded. This threshold was empirically optimized to guarantee a sufficient success rate for reaching a convergence radius of 3 Å around the Cα coordinate of the anchor. The remaining coordinate difference vector is distributed evenly over all backbone atoms, which leads to relative shifts of neighbouring backbone atoms of less than 0.05 Å and thus to changes in bond lengths, bond angles, and dihedral angles that are smaller than the variation of these parameters in well-resolved protein structures. Conformations that do not end up within the convergence radius are discarded. In such cases backbone generation is restarted at the loop midpoint. If 100 attempts for generating the second half loop fail, the first half loop is discarded.

Steering of the second half of the loop could potentially introduce conformational bias that would result in an asymmetry of the spatial distribution width of backbone atoms between the first and second half loop. We have tested for such asymmetry by extracting Cα coordinate RMSD from ensembles of 100 unrestrained conformations each for two internal loops. No consistent and significant asymmetry was found.

Dihedral angles of the terminal residue (index *k*) result from attachment to the C-terminal anchor and may fall outside allowed Ramachandran regions. This problem can often be fixed by rotating the peptide plane around the Cα_k-Cα_{k-1} vector until an allowed pair of angles ϕ_k, ψ_k is encountered. Such rotation alters dihedral angles (ϕ_{k-1}, ψ_{k-1}) of the preceding residue. If they have left the Ramachandran-allowed region, the procedure is repeated with the Cα_{k-1}-Cα_{k-2} vector. Rotation angles required for consecutive corrections decay fast and it is rarely necessary to proceed beyond the Cα_{k-2}-Cα_{k-3} vector. If the procedure fails, it is repeated with opposite sense of rotation. If it fails again, the conformation is discarded. This happens occasionally, in particular, when the last residue is Pro. Since the correction changes side chain orientations of the affected residues, the test of the conformation against restraints is repeated.

Secondary structure propensities

Propensities between 0 and 1 can be specified for α-helical, β-strand, and polyproline II helix secondary structure. The set of propensities is processed into a secondary structure vector before backbone generation. Repeated calls of the backbone generator result in different secondary structure vectors, with the statistics of all these vectors conforming to the propensities. The vector is further processed to assign weak α-helical restraints to the two residues at each terminus of an α-helix and strong restraints to residues within the helix.[34] For weak restraints, any value pair with ϕ between -89 and -39°, ψ between -66 and -16°, and $\phi+\psi$ between -115 and -95° is accepted. For strong restraints, uniform distributions are assumed within ±2° around a residue-specific ideal α-helix pair (ϕ_α,ψ_α). This pair is centred at (-61,-36.5)° for Pro, at (-59.1,-42.4)° for Gly, and (-63.8,-41.1)° for any other residue.

For β-strand residues, values of ϕ between -130 and -105° and ψ between 128 and 145° are accepted. If a Pro residue is specified as a β-strand residue, the upper bound for ϕ is raised to -80°. For polyproline II helix residues, Gaussian distributions with standard deviations of 5° are assumed for both ϕ and ψ with mean values ϕ_{PPII} = 75° and ψ_{PPII} = 160°.

Beacon restraints

A site within an IDD can be restrained by distances to several sites in the protein core, which we call beacons. At least four beacons are required for unambiguous site localization. Additional beacons reduce uncertainty. The unknown site can be localized by distance matrix geometry[25] or multilateration.[26,27] Distance matrix geometry allows for relaxing the beacon coordinates in cases where these are uncertain. For site localization in an IDD with respect to the core multilateration is better suited, since coordinate uncertainty is usually much larger for the unknown site than for the beacons.

Multilateration has received much attention since it underlies the global positioning system, where the beacons are satellites. Here we are interested in choosing *n* out of *N* available beacons so that localization error is minimal.[39,40] The *N* available beacons are potential labelling sites. In order to reduce the effort in mutation, protein production, and labelling, DDR collection is restricted to *n* sites. This problem can be posed in terms of the ‘visibility matrix’

$$\boldsymbol{H} = \begin{pmatrix} a_{x1} & a_{y1} & a_{z1} & 1 \\ a_{x2} & a_{y2} & a_{z2} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ a_{xn} & a_{yn} & a_{zn} & 1 \end{pmatrix},$$

which is set up for any set of *n* beacons with the elements

$$a_{xi} = \frac{x_i - x_u}{r_i}$$

and analogous definitions for the a_{yi} and a_{zi}. Where x_i is the *x* coordinate of beacon *i*, x_u an estimate for the *x* coordinate of the unknown site, and r_i a distance estimate between beacon *i* and the unknown site. Expected uncertainties of distance measurements can be considered by a diagonal matrix $\boldsymbol{R}$ with elements $R_{ii} = \sigma_i^2$, where σ_i^2 is the variance expected for the distance between the unknown site and beacon *i*. The best set of beacons minimizes position dilution of precision (PDOP), which is given by

$$P = \sqrt{Trace\{(\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H})^{-1}\}},$$

where $\boldsymbol{H}^\mathrm{T}$ is the transpose of $\boldsymbol{H}$. In our case, variances σ_i^2 are dominated by the uncertainty of predicting the spin label position relative to the backbone (*vide infra*) and are all the same. Thus, we can redefine *P* as

$$P = \sqrt{Trace\{(\boldsymbol{H}^T\boldsymbol{H})^{-1}\}}.$$

Beacon selection is based on a spin labelling site scan[41] and pre-selection of sites that are not too tight and have estimated distances r_i in a favourable range for pulsed dipolar EPR measurements (*vide infra*). Among all sets of *n* out of the *N* available sites the set with minimum *P* is chosen.

As an estimate for the coordinates of the unknown site we use the coordinates of a spin label attached to the anchor residue. For internal IDDs we combine the PDOP values of both anchor residues, $P = \sqrt{P_1^2 + P_2^2}$.

Internal restraints

DDRs between two sites within an IDD or IDP are internal restraints. They are less valuable than beacon restraints, since both sites have uncertain backbone coordinates and the uncertainty in predicting of label position relative to the backbone is larger than for beacons. For a beacon label, the distribution of spin label coordinates can be computed by a rotamer library approach[22,23] or by accessible space approaches[24] with prediction uncertainty for mean distances between two sites varying between 2.5 and 3.5 Å.[29]

If internal DDRs are tested during backbone generation, structural context is unknown, i.e., coordinates of neighbouring side chains and of the downstream backbone are not available. The best estimate for label position is then the mean position of an unrestrained label, i.e., the population-weighted mean coordinate for the whole rotamer library. We have tested this approach against the same data set used for the other approaches[29] and found mean distance uncertainties of 3.7 Å for the 32 solvent-exposed sites in T4 lysozyme and of 4.4 Å for the whole data set. Larger uncertainty for the whole set is probably related to inferior quality of some of the experimental data and to deviations between solution and crystal structure for proteins other than T4 lysozyme. Taking this into account, we estimate the uncertainty of relating internal restraints to backbone coordinates as 3.7 Å whereas it is 3.1 Å for beacon restraints.

Oligomer restraints

Some membrane proteins form homooligomers with a symmetry axis that is perpendicular to the membrane. When a single label is attached to the protein, a polygon of o spin labels is formed for an oligomer with o protomers. In the limit of coinciding conformations of the IDDs, the polygon is regular with side length s_i. This side length corresponds to the only peak in the distance distribution for $o = 2$ and $o = 3$ and to the modal distance for $o > 3$. With the coordinates of the core expressed in a frame whose z axis is the symmetry axis, the distance R of a label from the symmetry axis is $R_i = \sqrt{x_i^2 + y_i^2}$, where x_i and y_i are label coordinates. The predicted side length results as $s_i = 2R_i \sin(\pi/o)$. The same approach can be used for symmetric dimers of soluble proteins.[42]

Collecting oligomer restraints requires lower effort than collecting beacon or internal restraints, as only single rather than double mutants are required. However, oligomer restraints do not provide information on the z_i coordinate. Hence, if they are the only available DDRs, additional membrane insertion depth restraints are required.

Membrane insertion depth restraints

Assume that the coordinates of the core of a membrane protein are given in a frame where the membrane normal is the z axis. This frame is readily found for homooligomeric proteins (*vide supra*). For other cases, a bilayer modelling approach exists[43] that also provides the bilayer centre ($z = 0$) and thickness. A similar, somewhat less sophisticated approach is implemented in MMM. Membrane insertion depth of a residue corresponds to a restraint on the z_i coordinate in such a frame. Such insertion depths can be obtained, for instance, by continuous-wave EPR power saturation measurements,[44] electron spin echo envelope modulation (ESEEM) measurements of water accessibility,[33,45] or solid-state NMR spin diffusion measurements.[46] They can be specified as restraints on either the Cα atom coordinate or the spin label coordinate.

Use of DDRs for efficient sampling

Conformational space grows exponentially with the number of residues. Monte Carlo sampling can compete with such growth only if the search is restricted to a decreasing fraction of conformational space. To achieve this, DDRs are tested as soon as all information for their evaluation is available, thus excluding part of conformational space with each additional DDR. It is hard to assess whether such subspace reduction is weaker or stronger than growth of conformational space. In practice, the generated ensemble is used for visualization or for computation of expectation values of observables. In both cases one can test for convergence by repeating the computation a few times with different random number seeds. By default, we derive the random number seed from the starting time of the computation, thus ensuring that consecutive runs produce different ensembles. The user can select a fixed seed in the graphical user interface or specify a positive integer seed value in the restraint file to allow for reproducible runs.

Beacon restraints, homooligomer restraints, and membrane insertion depth restraints are evaluated as soon as the backbone coordinates of the corresponding residue are available. Internal restraints are evaluated as soon as the coordinates of the second residue are available. Efficient sampling of conformational space requires that the DDRs are distributed evenly over the peptide chain.

Occasionally, restraints must be defined by lower and upper bounds. For such restraints a conformation is discarded if the distance falls outside the bounds. Whenever possible, DDRs should be defined in a probabilistic manner. We assume Gaussian distributions with mean $\langle r_i \rangle$ and standard deviation σ_i, but any other distribution could be implemented. For Gaussian DDRs, the probability that the current conformation is consistent with the restraint is $p_i = \exp[-(R_i - \langle r_i \rangle)^2/\sigma_i^2]$, where is R_i is the predicted distance.

A conformation is discarded if its total probability $P = \prod_i p_i$ falls below a threshold P_{thr}. Since probabilities associated with independent random numbers multiply, it is plausible to choose $P_{\mathrm{thr}} = p_{\mathrm{thr}}^{m}$, where m is the total number of probabilistic restraints and p_{thr} is a threshold for a single restraint. The threshold

$$p_{thr} = \exp[-erf^{-1}(f)] \ ,$$

where erf^{-1} is the inverse of the error function, roughly corresponds to an ensemble that resides in fraction f of the total subspace that is consistent with the DDR. As a default value we take $f = 0.5$.

Once P falls below P_{thr} a conformation is discarded, but often part of the computation can be reused. To see this, consider the case where until residue number b DDRs are better fulfilled than expected, i.e., $P(b) = \prod_{i=1}^{m_b} p_i > p_{thr}^{m_b}$, where m_b is the number of DDRs evaluated until residue b. Assume further that residue b is the last residue where this relation holds and that at residue w, P falls below P_{thr}. The conformation between residues b and w is inconsistent with the DDRs, but the conformation until residue b may be consistent with them for other choices of dihedrals between b and w. Hence, on average, it is advantageous to restart from residue b instead of residue 1. Such a restart may turn up a valid conformation or generate a longer section that fulfils DDRs better than expected, i.e., lead to an increase of b.

How often should restarts be repeated from a given b? By computations on a test case we find a broad range for the number of restarts from the same b that optimizes efficiency (Fig. 2). The final implementation uses 100 restarts, corresponding to the lower end of this range. Compared to a larger number of restarts, this improves statistics for secondary structure vectors and for diagnostics information (*vide infra*).

Diagnostics

Experimental DDRs may be inconsistent if spin labelling has caused structural changes or if peaks in a distance distribution were wrongly assigned. Such inconsistencies may lead to a situation where no three-dimensional structure can fulfil all DDRs. Some of these cases can be recognized by violation of triangle inequalities. To this end we perform triangular bound smoothing[47] with lower and upper bounds for DDRs defined as $\langle r_i \rangle \pm 2\sigma_{r,i}$. Computation is cancelled if the triangular bound smoothing algorithm detects an inconsistency. Triangular bound smoothing may provide auxiliary lower/upper bound restraints for some site pairs where no experimental data exist. We did not find any efficiency improvement by including such auxiliary restraints.

DDRs can be inconsistent without violating triangle inequalities if they encode a polyhedron of sites that cannot be connected by a peptide chain. A certain DDR may then lead to rejection of all conformations. To help in diagnosing such cases, some information on the running computation is displayed after every 500 trials. For each residue *i* the probability p_r is plotted that a backbone conformation, which is acceptable until residue *i*-1, is rejected at residue *i* (Fig. 2). Since in our examples, labelling sites are periodically distributed along the chain (vide infra), a periodicity of rejection probability is observed in Fig. 2. With beacon restraints, rejection probabilities should have a nearly uniform distribution along the peptide chain [Fig. 2(A)]. Variations arise mainly from different standard deviations $\sigma_{r,k}$ of the DDRs. If only internal restraints are present, p_r usually increases with increasing *i* [Fig. 2(B)]. Rejection probabilities p_r close to unity for residues with small *i* indicate inconsistency of a DDR at this residue. For internal loops, larger rejection probability is observed at the central residue, where the algorithm checks for clashes and at the terminal residue, where loop closure may fail.

The diagnostics panel also reports percentages of trials that failed because of restraint violations, because of internal clashes of the backbone, and because of clashes of the backbone with the core of the protein. For valid conformations, it reports percentages of internal side chain clashes and side chain clashes with the core. For internal loops, it reports percentages of trials failed at loop closure or during correction of Ramachandran-forbidden dihedral angles at the linkage to the C-terminal anchor.

Conformations that passed all tests are displayed in the MMM main window and enter the ensemble. Distance distributions corresponding to all probabilistic restraints are computed for the whole ensemble using a rotamer library approach[22] after backbone decoration by side chains. For each DDR, the difference in mean distance between the restraint and the distance distribution computed for the whole ensemble is reported in the graphical user interface and in the log file of the computation. Values up to 4.5 Å are no cause of concern if they occur only in rare cases. A stricter quality measure is the overlap of the two normalized probability distribution vectors $P_{\mathrm{DDR}}(r)$ and $P_{\mathrm{ensemble}}(r)$, which is sensitive also to the distribution width. Overlap is defined as 1 - $\|\, P_{\mathrm{DDR}}(r) - P_{\mathrm{ensemble}}(r)\|$, where $\|\ldots\|$ denotes the vector norm.

Representation of ensembles

Ensembles for which a template structure exists assume a compact form if each conformation is transformed to the frame where coordinate root mean square deviation (RMSD) from the template structure is minimal. A standard frame can be defined by computing the inertia tensor of the template structure and assigning the x, y, and z axes of the frame as the principal axes of this tensor corresponding to the smallest, intermediate, and largest eigenvalue, respectively. In this work the inertia tensor is computed from atom coordinates excluding hydrogen atoms and assuming that all ‘heavy’ atoms have the same mass. The command `inertiatensor` in MMM transforms a structure into this frame.

If no template structure exists, it can be replaced by the most representative conformer in the ensemble. This conformer is the one that has minimal mean square coordinate deviation from all other conformers, in other words, the central conformer of the ensemble. Transformation of all other conformers to their respective frames where they have minimal RMSD from the most representative conformer provides the most compact ensemble representation. The command `compact` in MMM computes the most representative conformer.

Results

C-terminal domain of LHCII

Preliminary versions of this approach have provided plausible and consistent ensemble descriptions from experimental data.[27,32,33] However, stringent tests require that the correct result is known beforehand. Our maximal ensemble approach is expected to perform best for broad ensembles, where a small number of DDRs is expected to provide sufficient information. Furthermore, the broader the ensemble, the easier it is to find representative conformations in conformational space. In contrast, our approach is expected to run into problems for the other limiting case of a single, well-defined conformation. Therefore, we decided to first test the approach for this limiting case for a terminal domain, an internal domain, and a moderately sized protein using DDRs simulated from known structures.

As an example for a terminal domain we use residues 202-232 of major plant light harvesting complex LHCII. Chain A was extracted from the PDB file of the crystal structure with PDB ID code 2BHW[48] and residues 202-232 were removed. Optimal beacon residues for quadrilateration (n = 4) and pentalateration (n = 5) were selected by minimizing PDOP for the N-terminal anchor residue 201 under the constraints that all mean distances were within the optimum DEER measurement range between 25 and 40 Å and that none of the labelling sites was tight. This distance range for the anchor ensures that all distances to residues in the domain fall within the safely measurable range from 15 to 60 Å. The minimum PDOP was $P = 2.844$ for beacon residues 28, 73, 89, and 105 ($n = 4$), and $P = 2.489$ for residues 28, 77, 89, 107, and 132 ($n = 5$). DDRs were simulated by rotamer computations for methanethiosulfonato spin label (MTSL) based on the crystal structure PDB file. Furthermore, we specified α-helical propensities of 1 for the two short helices 206-214 and 221-224 seen in the crystal structure. Ensembles of 20 conformations were used for visualization.

In the unrestrained ensemble [Fig. 4(A,B)] most conformations point away from the lipid bilayer, because the protein and the cofactors block space in the membrane region. Apart from this, the loop samples space uniformly. This computation took less than a minute on a single processor core of an Intel® Xeon® CPU E3-2141 v3 @ 3.50 GHz, with most of the time spent by SCWRL4 for side chain attachment. The backbone RMSD of the ensemble with respect to its mean coordinates (ensemble RMSD), RMSD of the ensemble with respect to the template structure, and computation time are listed in Table 1.

Consideration of secondary structure propensities does not significantly increase simulation time. Helix 206-214 clearly makes the ensemble more compact close to the N-terminal anchor, but most conformations still point away from the bilayer and they are still distributed over the whole range of angles within the bilayer plane [Fig. 4(C,D)].

In order to confine the ensemble, we used 44 Gaussian DDRs from beacon residues 28, 73, 89, and 105 to every third residue in the C-terminal domain, starting at residue 202, together with the helix propensities. Computation time increased drastically since only a fraction of $5 \cdot 10^{-5}$ of all trials was successful. Ensemble RMSD with 44 DDRs is 4.4 Å, whereas backbone RMSD with respect to the template structure (RMSD to template) is 7.3 Å.

The restrained ensemble contains conformations very similar to the template conformation and reasonably reproduces the two nearly straight sections of the domain and the angle between them [Fig. 4(E,F)]. Yet, it is biased towards conformations that extend further away from the rigid core of LHCII than the template. Such conformations are less likely to be rejected by clashes with the protein.

In order to obtain a narrower ensemble, we generated 55 pentalateration DDRs between beacon residues 28, 77, 89, 107, and 132 and the same 11 residues in the C-terminal domain. With these DDRs all backbone conformations were rejected already at the first residue 202. The problem could not be solved by removing individual DDRs. Strong localization by pentalateration either leads to inconsistencies with context-free prediction of label coordinates or to such a strong narrowing of the ensemble that the allowed conformations can no longer be found in the vast conformational space. In contrast, a narrower ensemble can be obtained with quadrilateration DDRs to every second residue (64 DDRs) [Fig. 4(G,H)], albeit at the cost of a success rate of only $5\cdot10^{-6}$ for backbone generation.

Internal loop 71-87 in CaiT

As an example of an internal loop we selected residues 71-87 in the secondary carnitine transporter CaiT. Chain A was extracted from the CaiT crystal structure with PDB ID 2WSX[49] and residues 71-87 were deleted. Conformational space of such internal loops is strongly constrained by the closure condition. Therefore, residues in the loop cannot have much longer distances from beacon residues than the two anchor residues 70 and 88. We thus included all core residues with distances between 25 and 45 Å to both anchor residues in the search for an optimum set of four beacons. This search provided a minimum PDOP of 2.635 for beacon residues 107, 257, 280, and 501. Sets of DDRs between the beacons and five labelling sites 73, 76, 79, 82, and 85 (20 DDRs) or eight labelling sites 71, 73, 75, 77, 80, 82, 84, and 86 in the loop (32 DDRs) were generated.

Here, generation of the unrestrained ensemble required significant computation time, since conformational space had to be searched for backbones that fulfil the closure constraint (see Table 2). Among all trials 33.5% failed to produce a closed loop despite restarts and about 0.5% failed because Ramachandran-disallowed dihedrals at the last residue could not be corrected. Most valid conformations [crimson coil models in Fig. 5(A,B)] occupy a spatial region far from the one of the template loop (green coil model), but two of them come reasonably close (blue coil models).

Non-clashing backbone models that fulfilled 20 DDRs resemble the template conformation quite well, but again some bias away from the protein is observed [Fig. 5(C,D)]. The situation does not change significantly with 32 DDRs [Fig. 5(E,F)]. At this level, uncertainty of the conformation appears to be dominated by uncertainty in the prediction of label coordinates.

p27Kip1 bound to Cdk2 complex

As an example for an IDP we selected the kinase inhibitory domain of p27 (residues 25-93), which becomes structured when it binds to the phosphorylated cyclin A-cyclin-dependent kinase 2 (Cdk2) complex.[50] We first consider the case of well-structured p27 and then the case of unbound intrinsically disordered p27 in solution. In the former case, DDRs were derived from the crystal structure of this complex with PDB ID 1JSU. For IDPs it is more difficult than for IDDs to select site pairs that are likely to provide mean distances in the range accessible by DEER measurements. In order to solve this problem, we performed a statistical analysis of mean label-to-label distances for all site pairs within the peptide for an unrestrained ensemble of 200 models. To that end, we labelled all sites *in silico* by the rotamer library approach, computed the mean distances between the rotamer distributions for each pair (*i*,*k*) of sites, and binned them into a histogram as a function of the difference of their residue numbers |*k*-*i*| (segment length). The histogram was then analysed for the fraction of residues at given segment length with a mean distance in the well accessible DEER range between 20 and 60 Å (Fig. 6A). At a success rate of slightly less than 60%, segment lengths from 9 to 57 are acceptable (dashed red lines). Mean distances outside the assumed range may still be measurable in some cases or the corresponding site pairs may at least provide lower/upper bound restraints by line shape analysis of continuous-wave EPR spectra (shorter distances than 20 Å) or by approximate analysis of DEER data (longer distances than 60 Å).

In order to study RMSD to template as a function of the number of DDRs, we compiled several DDR sets. In each case we generated forward DDRs starting at residue 25 and backward DDRs starting at residue 93. For a minimum segment length $s_{\min}$ = 14 the labelled residues in the forward set are 25, 39, 53, 67, and 81 and the ones in the backward set are 93, 79, 65, 51, and 37. After discarding pairs 25-81 and 93-37 we obtained 18 DDRs from 10 labelled residues. Further sets were created analogously with $s_{\min}$ = 11 (36 DDRs from 14 labelled residues), $s_{\min}$ = 10 (40 DDRs from 14 labelled residues), and $s_{\min}$ = 9 (54 DDRs from 16 labelled residues). Another set with $s_{\min}$ = 9 and 16 labelled residues included the two longest segments 25-88 (52.4 Å) and 93-30 (59.5 Å) to give 56 DDRs. A final set with $s_{\min}$ = 9 included a further interleaved set of residues 27, 36, … 90, providing 84 DDRs from 24 labelled residues. Secondary structure was restrained for segments 38-49, 54-60, and 86-89, where short α-helices are found in the template. Ensembles were computed without any restraints, with only secondary structure restraints, and with all DDR sets in addition to secondary structure restraints.

For the unrestrained case [Fig. 7(A)], ensemble RMSD is 12.1 Å, which reduces to 10.7 Å with secondary structure restraints [Fig. 7B]. RMSD to template is 14.3 Å for the unrestrained ensemble and 13.6 Å with secondary structure restraints. RMSD100[51] to template is 10.5 Å for the unrestrained ensemble and reduces to 10.0 Å with secondary structure restraints. Further reduction of the RMSD measures is observed when including DDRs [Fig. 8(A)]. The curve for RMSD100 (red asterisks and dotted line) closely resembles the one for ensemble RMSD (blue circles and dashed line). With 84 DDRs, where a fraction of $3.5 \cdot 10^{-6}$ of all trials is successful, both the ensemble RMSD and the RMSD100 to template are 5.8 Å. Fig. 7(C-E) shows how the ensemble progressively narrows and emulates the template structure better and better with increasing number of DDRs. Computations with up to 56 DDRs finished within a few hours or overnight on a single processor core [Fig. 8B]. The computation with 84 DDRs was running for a week in parallel on four processor cores.

Ensemble of unbound p27KID

An ensemble model for the unbound p27KID domain in solution was generated by MD simulations starting from the bound structure.[52] Secondary structure propensities extracted from the MD trajectories were found to be in qualitative agreement with ^{1}HN-^{1}HN NOE correlations from NMR measurements and some domains were recognized as intrinsically folded structural units that resembled structural features seen in the bound form. This ensemble is available from the Protein Ensemble Database[11] as entry PED2AAA. It is visualized in its compact form (semi-transparent crimson coil models) and with its most representative conformer (green coil model, see Materials & Methods) in Fig. 9A,B.

Since DDRs derived from this ensemble have larger standard deviations $\sigma_{\underline{r},i}$ than those derived from the well-defined structure of p27Kip1 bound to Cdk2 complex, a larger fraction of backbone conformations fits all DDRs (Table 2). Accordingly, computation times for the same number of DDRs are much shorter (Fig. 8B). Ensembles derived without any restraints (Fig. 9C,D) with 56 DDRs (Fig. 9E,F), and with 56 DDRs and α-helical secondary structure restraints for residues 38-58 (Fig. 9G,H) have been visualized by optimal superposition onto the most representative conformer of the template ensemble (green coil model). The secondary structure restraints are based on the observation that the corresponding residues have stable α-helical structure in all conformers of the template ensemble. Even with the secondary structure restraints, the ensemble derived by our maximal ensemble approach (Fig. 9G,H) is still somewhat broader than the template ensemble (Fig. 9A,B). Nevertheless, it is obvious that the DDRs and secondary structure restraints are successful in narrowing down the ensemble to the region in space where the template ensemble conformers are distributed.

Fig. 6B shows the mean label-to-label distance in a statistical coil ensemble as a function of segment length (black solid line) and the and the range of mean distances of DDRs for the unbound p27KID domain (crimson range qualifiers). Some of the DDRs deviate strongly from the mean distance in a statistical coil, they vary strongly at given segment length, and, on average, the ensemble is more compact than a statistical coil. All these features are mainly due to the existence of a stable α-helix between residues 38-58 in the template ensemble.

Fig. 10 shows the typical agreement between specified DDRs and the label-to-label distance distributions computed for the ensemble. For more than 50% of all DDRs the distributions have an overlap better than 0.97, corresponding to the fit quality visualized in Fig. 10A. In 90% of all cases, the overlap is better than 0.94, corresponding to the fit quality visualized in Fig. 10B. The mean distance shifts and overlaps are also reported in the log file of the computation. In no case we observed an overlap of less than 0.90. Note that in the computations in this work, the main error source are uncertainties in prediction of the conformational distribution of the label side chain. In an experimental context, some DDRs may be erroneous which could be recognized by particularly poor overlap values.

Ensemble of α-synuclein

An ensemble for the intrinsically disordered solution state of α-synuclein has been derived on the basis of paramagnetic relaxation enhancement NMR data and CHARMM MD simulations[53] and can be accessed from the Protein Ensemble Database as entry PED9AAC. From this ensemble we have simulated 110 internal DDRs at segment lengths of 10, 20, 30, 40, and 50 residues. The distribution of the mean distances $\langle r_i \rangle$ for these DDRs for the selected segment lengths is displayed in Fig. 6B (blue range qualifiers). Comparison with mean segment distances for a random coil (black solid line) and distributions for unbound p27KID (crimson range qualifiers) reveals that the α-synuclein ensemble deviates much less from a statistical coil than the p27KID ensemble. The same conclusion can be drawn from the widths of the DDRs, which is in the range of 17-19 Å for 50-residue segments in α-synuclein and in the range of 6-11 Å for 54-residue segments in p27KID. The published α-synuclein ensemble is somewhat more compact (backbone radius of gyration R_g = 32.4 Å) than an unrestrained ensemble that we compute with our approach (R_g = 36.7 Å).

The larger width of the DDRs for α-synuclein drastically reduces computational effort for our maximal ensemble approach. Although the chain is longer (149 residues) than for p27KID (69 residues) and the number of restraints is larger, 100 conformers that fulfil all 110 DDRs can be computed on a single processor in 1 h 54 min. This is because ~5.5% of all generated backbone conformations fulfil all restraints. The rejected conformers are mainly more extended conformers, so that the restrained ensemble with R_g = 33.7 Å matches the backbone radius of gyration of the template better than the unrestrained ensemble.

Discussion

If DDRs are available, representative conformational ensembles can be computed with reasonable effort by sampling complete conformational space. Computation times required for the test cases with well-defined templates can be considered as upper bounds, since the

DDRs were derived *in silico* from a single conformation. Structures of IDDs and IDPs are distributed, which leads to broader distance distributions and to a larger probability of finding conformations that fit all DDRs as was confirmed with template ensembles for unbound p27KID and α-synuclein.

In principle, existing software such as XPLOR-NIH[54] or CNS[55] could have been used to implement a maximal ensemble approach using DDRs. Although these programs can generate ensembles for distributed restraints, they were designed with the concept of a single relatively narrow energy potential minimum, and thus a single well-defined structure, in mind. The main difference between their approach and our approach is our strict priority on maximizing sampling density of conformation space in a given computation time. This is achieved by testing DDRs as early as possible during conformer generation, thus allowing to recognize inaccessible subspaces of conformation space with a minimum of computational effort. Furthermore, information on accessible subspaces is stored in restart points. Since only consistent conformers need to be stored, vast numbers of conformers can be tested against DDRs with small memory consumption. Therefore, we believe that ensemble selection by DDRs is more efficient with our approach, whereas further optimization of the ensembles by additional restraints that can be evaluated only for the whole ensemble is probably best done with established software.

For a test case of a peptide with well-defined structure (p27Kip1 bound to Cdk2 complex) we find that a sufficiently restrained ensemble [Fig. 7(H)] is not or only weakly biased towards conformations that are less compact than the template. Ensemble RMSD appears to be a good estimate for RMSD100 to template. However, more test computations for different sequence lengths are required to decide whether this is coincidence or a stable property of such ensembles. With DDRs derived from the premolten globule-like ensemble of p27KID we again find that the ensemble restrained by DDRs is less compact than the template ensemble (Fig. 9).

RMSD to template for the 69-residue protein p27Kip1 still decreases when increasing the number of DDRs from 54 or 56 to 84. In fact, the data shown in Fig. 8(A) suggests that further increase may still narrow the ensemble and reduce RMSD to template. However, such improvements would require much larger computational and experimental effort for only a moderate gain in accuracy and precision. The data indicates that, even with a much larger number of constraints, RMSD100 to template would not fall below 5 Å. This finding strongly suggests that atomistically resolved structures cannot be obtained from only spin label DDRs and secondary structure restraints. The apparent lower bounds for ensemble RMSD and RMSD to template are most likely caused by flexibility of the spin label. Hence, less flexible labels than MTSL could improve the situation, but they would also increase the risk of biasing the structure with respect to wild type. For work on IDPs, the ensemble widths achieved with 0.8-1.2 DDRs per residue [Fig. 7(F-H), Fig. 9(G,H)] should be sufficient, in particular when keeping in mind that such ensembles could be further narrowed down by taking into account additional restraints from NMR or SAXS that can be evaluated only from the entire ensemble.

Conformational bias due to spin labelling is expected if some of the generated, otherwise consistent backbone conformations would cause clashes. Such cases would be recognized in our approach during computation of the distance distributions since *in silico* spin labelling would fail. In all our computations we did not encounter a single instance of such a tight labelling site. Note however that a subtler bias may result from differences in the Ramachandran plot or in side-chain specific interactions between the wildtype residue and spin labelled cysteine. To avoid such bias as far as possible, amino acids with strongly deviating Ramachandran plots (Gly, Pro), charged amino acids (Glu, Asp, Lys) or those that can form π-cation complexes (Trp, Phe, Tyr) should be avoided as labelling sites if possible. Best matches in polarity are Ile, Leu and Val.

For both IDD test cases we find a bias of the ensemble towards conformations that interact less with the protein core than the templates. This is probably due to the fact that packing against the core is highly optimized for well-structured loops, so that slight differences in backbone conformation already cause clashes. The problem is expected to be less severe for real IDDs where packing against the core is not that well optimized, yet the potential bias should be considered in interpretation of conformational ensembles obtained by our approach.

The IDD test cases also reveal limitations on decreasing width of the conformational ensemble. For the C-terminal domain in LHCII with a length of 31 residues pentalateration of the labelled sites failed. Probably localization precision with five beacons is better than the accuracy of predicting mean spin label positions, which leads to inconsistencies. This problem may not be relevant for real IDDs, since the ensemble RMSD of 3.8 Å that we achieved by quadrilateration with 2.1 DDRs per residue corresponds to higher order than expected in IDDs.

For the internal loop in CaiT with a length of 17 residues an ensemble RMSD of 3.3 Å could be achieved with 20 quadrilateration DDRs, i.e., with less DDRs per residue than for LHCII. This can be traced back to the strong constraint imposed by anchoring both ends of the loop in the core. An increase of the number of DDRs to 32 hardly changed ensemble RMSD and did not reduce bias against conformations that interacted with the protein. Hence, for internal loops, about one DDR per residue appears to be sufficient for attaining the precision and accuracy limit of our approach.

Computation time for generating an IDP ensemble of given size appears to scale roughly exponentially with the number of internal DDRs [Fig. 8(B)]. For IDDs the situation is less clear cut. Computation time is not a monotonous function of the number of DDRs or even of the width of the ensemble (data not shown).

Even with subspace restriction and iterative prolongation of favourable segment conformations we find a success rate of less than 10^{-5} for backbone generation trials at ensemble RMSDs of 3.8 Å (C-terminal domain of LHCII) and 5.8 Å (p27Kip1). Thus, it may not be feasible to select such narrow ensembles from a pre-computed unrestrained ensemble of a size that can be computed and stored at reasonable effort. On the other hand, much larger success rates were found for the premolten globule-like template ensemble of unbound p27KID and the coil-like template ensemble of α-synuclein. For coil-like IDPs, where not much more long-range information than just the radius of gyration can be extracted our approach may entail an unnecessary effort, whereas for premolten globule-like IDPs, such as unbound p27KID, our approach can reveal long range correlations and provide a maximal ensemble model. In its compact representation, such a model should be interpretable in terms of the mechanism of interaction with binding partners.

A preliminary version of the approach has been used for modelling residues 3-13 of the disordered section N-terminal domain of major plant light harvesting complex LHCII from 6 beacon restraints obtained with heterogeneously labelled trimers (only one protomer in the trimer is doubly spin labelled), 7 trimer restraints from singly spin-labelled protomers, and 7 membrane insertion depth restraints.[33] This section was found to cover a restricted area above the superhelix of LHCII that is much smaller than the area covered by an unrestrained statistical coil model. Furthermore, a preliminary version of the approach was used for modelling the 5/6 hairpin and helix 6 in the active, membrane bound form of the proapoptotic protein Bax based on 6 beacons restraints to residue 149 near the end of helix 6, using one ambiguous dimer restraint for a consistency check. Together with 13 further DDRs with very broad distance distributions to residues 169, 186 and 193 that were used for localization of these residues by multilateration and 3 dimer restraints for these residues, the model suggested that a piercing domain downstream from residue 126 clamps Bax to the membrane while the dimerization domain up to residue 126 lines the water-accessible pore that oligomerized Bax forms in the mitochondrial membrane.[27] These examples demonstrate that a small number of DDRs may be sufficient for obtaining new structural information if auxiliary information, in these cases x-ray crystal structures and an NMR structure of the inactive form of Bax, are available.

The approach may also be useful for modelling multi-domain proteins with long flexible linkers, as we have demonstrated for the FnIII-3,4 domains of integrin α6β4.[32] In this case 13 DDRs were used to establish the relative orientation and translation of the two rigid domains and only 2 beacon restraints to the central residue of the 21-residue long flexible linker were used to restrain conformation of the internal linker loop. The ensemble of structural models was further refined using SAXS data. A generally applicable MMM module for such modelling of proteins in terms of rigid domains joined by flexible linkers is currently under development.

Conclusion

Conformational ensembles of protein domains and small proteins can be generated from DDRs between spin labels in the 10 to 100 Å range by an approach that uses the DDRs to reduce the part of conformational space that needs to be sampled. If the underlying structure is a single conformation, ensemble RMSD converges to 3.5 – 6 Å at about 1-2 DDRs per residue, with the lowest RMSDs for internal domains, intermediate values for terminal domains, and the highest RMSDs for peptide chains that are not attached to a structured protein core. RMSD to template is somewhat larger, indicating some bias to less compact structures, in particular for domains attached to a structured protein core. For IDPs, where distance distributions are broader, a larger fraction of conformation space is accessible and computational effort decreases. Premolten globule-like IDPs are a class of proteins where long-range label-to-label DDRs can potentially provide experimental information on accessible conformational subspace that cannot easily be obtained by other techniques. This expectation needs to be tested by further computational and experimental work.

Acknowledgments

I thank Ad Bax, Enrica Bordignon, Yevhen Polyhach, Inés García Rubio, and Maxim Yulikov for helpful advice. This research was supported by SNF grant 200020_157034.

References

1. Berlow RB, Dyson HJ, Wright PE. Functional advantages of dynamic protein disorder. FEBS Lett 2015; 589:2433-40.

2. Uversky, V.N. (2015). Functional roles of transiently and intrinsically disordered regions within proteins. FEBS J. *282*, 1182-9.

3. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol 2015; 16:18-29.

4. Tantos A, Kalmar L, Tompa P. The role of structural disorder in cell cycle regulation, related clinical proteomics, disease development and drug targeting. Expert Rev Proteomics 2015; 12:221-33.

5. Uversky VN, Longhi S. Instrumental Analysis of Intrinsically Disordered Proteins. Hoboken: Wiley; 2010.

6. Jensen MR, Zweckstetter M, Huang JR, Blackledge M. Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. Chem Rev 2014; 114:6632-60.

7. Kikhney AG, Svergun DI. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. FEBS Lett 2015; 589:2570-7.

8. Le Breton N, Martinho M, Mileo E, Etienne E, Gerbaud G, Guigliarelli B, Belle V. Exploring intrinsically disordered proteins using site-directed spin labeling electron paramagnetic resonance spectroscopy. Front Mol Biosci 2015; 2:21.

9. Jha AK, Colubri A, Freed KF, Sosnick TR. Statistical coil model of the unfolded state: resolving the reconciliation problem. Proc Natl Acad Sci U S A 2005; 102:13099-104.

10. Mohana-Borges R, Goto NK, Kroon GJ, Dyson HJ, Wright PE. Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. J Mol Biol 2004; 340:1131-42.

11. Varadi M, Kosol S, Lebrun P, Valentini E, Blackledge M, Dunker AK, Felli IC, Forman-Kay JD, Kriwacki RW, Pierattelli R, Sussman J, Svergun DI, Uversky VN, Vendruscolo M, Wishart D, Wright PE, Tompa P. pE-DB: a database of structural ensembles of intrinsically disordered; and of unfolded proteins. Nucleic Acids Res 2014; 42:D326-D335.

12. Ozenne V, Bauer F, Salmon L, Huang JR, Jensen MR, Segard S, Bernado P, Charavay C, Blackledge M. Flexible-meccano: a tool for the generation of explicit ensemble

descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics 2012; 28:1463-70.

13. Milov AD, Ponomarev AB, Tsvetkov YD. Electron electron double-resonance in electron-spin echo - model biradical systems and the sensiitized photolysis of decalin. Chem Phys Lett 1984; 110:67-72.

14. Pannier M, Veit S, Godt A, Jeschke G, Spiess HW. Dead-time free measurement of dipole-dipole interactions between electron spins. J Magn Reson 2000; 142:331-40.

15. Borbat PP, Mchaourab HS, Freed JH. Protein structure determination using long-distance constraints from double-quantum coherence ESR: study of T4 lysozyme. J Am Chem Soc 2002; 124:5304-14.

16. Jeschke G, Koch A, Jonas U, Godt A. Direct conversion of EPR dipolar time evolution data to distance distributions. J Magn Reson 2002; 155:72-82.

17. Jeschke G, Panek G, Godt A, Bender A, Paulsen H. Data analysis procedures for pulse ELDOR measurements of broad distance; distributions. Appl Magn Reson 2004; 26:223-244.

18. Chiang YW, Borbat PP, Freed JH. The determination of pair distance distributions by pulsed ESR using Tikhonov regularization. J Magn Reson 2005; 172:279-95.

19. Borbat PP, Freed JH. Pulse Dipolar Electron Spin Resonance: Distance Measurements. Struct Bond 2014; 152:1-82.

20. El Mkami H, Norman DG. EPR Distance Measurements in Deuterated Proteins. Methods Enzymol 2015; 564:125-52.

21. Hubbell WL, Cafiso DS, Altenbach C. Identifying conformational changes with site-directed spin labeling. Nat Struct Biol 2000; 7:735-9.

22. Polyhach Y, Bordignon E, Jeschke G. Rotamer libraries of spin labelled cysteines for protein studies. Phys Chem Chem Phys 2010; 13:2356-66.

23. Hatmal MM, Li Y, Hegde BG, Hegde PB, Jao CC, Langen R, Haworth IS. Computer modeling of nitroxide spin labels on proteins. Biopolymers 2011; 97:35-44.

24. Hagelueken G, Ward R, Naismith JH, Schiemann O. MtsslWizard: In Silico Spin-Labeling and Generation of Distance Distributions in PyMOL. Appl Magn Reson 2012; 42:377-391.

25. Gaffney BJ, Bradshaw MD, Frausto SD, Wu F, Freed JH, Borbat P. Locating a lipid at the portal to the lipoxygenase active site. Biophys J 2012; 103:2134-44.

26. Abdullin D, Florin N, Hagelueken G, Schiemann O. EPR-based approach for the localization of paramagnetic metal ions in biomolecules. Angew Chem Int Ed Engl 2015; 54:1827-31.

27. Bleicken S, Jeschke G, Stegmueller C, Salvador-Gallego R, Garcia-Saez AJ, Bordignon E. Structural model of active Bax at the membrane. Mol Cell 2014; 56:496-505.

28. Salmon L, Nodet G, Ozenne V, Yin G, Jensen MR, Zweckstetter M, Blackledge M. NMR characterization of long-range order in intrinsically disordered proteins. J Am Chem Soc 2010; 132:8407-18.

29. Jeschke G. Conformational dynamics and distribution of nitroxide spin labels. Prog Nucl Magn Reson Spectrosc 2013; 72:42-60.

30. Nodet G, Salmon L, Ozenne V, Meier S, Jensen MR, Blackledge M. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. J Am Chem Soc 2009; 131:17908-18.

31. Kragelj J, Ozenne V, Blackledge M, Jensen MR. Conformational propensities of intrinsically disordered proteins from NMR chemical shifts. ChemPhysChem 2013; 14:3034-45.

32. Alonso-Garcia N, Garcia-Rubio I, Manso JA, Buey RM, Urien H, Sonnenberg A, Jeschke G, de Pereda JM. Combination of X-ray crystallography, SAXS and DEER to obtain the structure of the FnIII-3,4 domains of integrin alpha6beta4. Acta Crystallogr D Biol Crystallogr 2015; 71:969-85.

33. Fehr N, Dietz C, Polyhach Y, von Hagens T, Jeschke G, Paulsen H. Modeling of the N-terminal Section and the Lumenal Loop of Trimeric Light Harvesting Complex II (LHCII) by Using EPR. J Biol Chem 2015; 290:26007-20.

34. Hovmoller S, Zhou T, Ohlson T. Conformations of amino acids in proteins. Acta Crystallogr D Biol Crystallogr 2002; 58:768-76.

35. Stewart DE, Sarkar A, Wampler JE. Occurrence and role of cis peptide bonds in protein structures. J Mol Biol 1990; 214:253-60.

36. Sugeta H, Miyazawa T. General method for calculating helical parameters of polymer chains from bond lengths, bond angles, and internal-rotation angles. Biopolymers 1967; 5:673-679.

37. Shimanouchi T, Mizushima S. On the Helical Configuration of a Polymer Chain. J Chem Phys 1955; 23:707-711.

38. Krivov GG, Shapovalov MV, Dunbrack RL Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 2009; 77:778-95.

39. Yarlagadda R, Ali I, Al-Dhahir N, Hershey J. GPS GDOP metric. IEE Proc-Radar Sonar Navig 2000; 147:259-264.

40. Sairo H, Akopian D, Takala J. Weighted dilution of precision as quality measure in satellite positioning. IEE Proc-Radar Sonar Navig 2003; 150:430-436.

41. Polyhach Y, Jeschke G. Prediction of favourable sites for spin labelling of proteins. Spectr-Int J 2010; 24:651-659.

42. Sen KI, Logan TM, Fajer PG. Protein dynamics and monomer-monomer interactions in AntR activation by electron paramagnetic resonance and double electron-electron resonance. Biochemistry 2007; 46:11639-49.

43. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res 2011; 40:D370-6.

44. Rufener E, Frazier AA, Wieser CM, Hinderliter A, Cafiso DS. Membrane-bound orientation and position of the synaptotagmin C2B domain; determined by site-directed spin labeling. Biochemistry 2005; 44:18-28.

45. Dzuba SA, Raap J. Spin-echo electron paramagnetic resonance (EPR) spectroscopy of a pore-forming (Lipo)peptaibol in model and bacterial membranes. Chem Biodivers 2013; 10:864-75.

46. Wang T, Widanapathirana L, Zhao Y, Hong M. Aggregation and Dynamics of Oligocholate Transporters in Phospholipid Bilayers Revealed by Solid-State NMR Spectroscopy. Langmuir 2012; 28:17071-17078.

47. Crippen GM, Havel, TF. Distance Geometry and Molecular Conformation. Taunton: Research Studies Press; 1988.

48. Standfuss J, Terwisscha van Scheltinga AC, Lamborghini M, Kuhlbrandt W. Mechanisms of photoprotection and nonphotochemical quenching in pea light-harvesting complex at 2.5 A resolution. EMBO J 2005; 24:919-28.

49. Schulze S, Koster S, Geldmacher U, Terwisscha van Scheltinga AC, Kuhlbrandt W. Structural basis of Na(+)-independent and cooperative substrate/product antiport in CaiT. Nature 2010; 467:233-6.

50. Russo AA, Jeffrey PD, Patten AK, Massague J, Pavletich NP. Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. Nature 1996; 382:325-31.

51. Carugo O, Pongor S. A normalized root-mean-square distance for comparing protein three-dimensional structures. Protein Sci 2001; 10:1470-3.

52. Sivakolundu SG, Bashford D, Kriwacki RW. Disordered p27(Kip1) exhibits intrinsic structure resembling the Cdk2/cyclin A-bound conformation. J. Mol. Biol. 2005; 353:1118-1128.

53. Allison JR, Varnai P, Dobson CM, Vendruscolo M. Determination of the free energy landscape of alpha-synuclein using spin label nuclear magnetic resonance measurements. J Am Chem Soc 2009; 131:18314-26.

54. Schwieters CD, Kuszewski JJ, Clore GM. Using Xplor-NIH for NMR molecular structure determination. Prog Nucl Magn Reson Spectrosc 2006; 48:47-62.

55. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr Sect D-Biol Crystallogr 1998; 54:905-921.

Figure legends

Figure 1

Flow charts for the ensemble generator (A) and backbone generator (B). For simplicity, loop closure for internal IDDs is not contained in the backbone generator flow chart.

Figure 2

Mean time required for generating one backbone model fulfilling all restraints as a function of the number of restarts from the end of the last favourable backbone section. Computations were performed for residues 25-93 of p27Kip1 using 54 DDRs derived from the ordered structure of p27Kip1 bound to CdK2 complex with PDB ID 1JSU.

Figure 3

Rejection probability per residue *i* expressed as the fraction of models that were accepted up to residue *i*-1 and rejected at residue *i*. (A) Ensemble for the C-terminal domain of LHCII (residues 202-232) obtained with 44 DDRs between 11 residues in the domain and 4 beacon residues in the core of the protein. DDRs were derived from the structure with PDB ID

2BHW. (B) Conformational ensemble for residues 25-93 of p27Kip1 using 56 DDRs derived from the structure with PDB ID 1JSU.

Figure 4

Conformational ensembles with 20 models each computed for the C-terminal domain (residues 202-232) of LHCII. Views along the membrane normal are shown in panels A, C, E, and G and views parallel to the membrane plane in panels B, D, F, and H. The core (residues 14-201) is shown as a grey ribbon model and the ensemble by semi-transparent crimson coil models. The template structure is shown as a green coil model. (A,B) Unrestrained ensemble. (C,D) Ensemble computed with only secondary structure constraints. (E,F) Ensemble computed with secondary structure restraints and 44 DDRs between 11 residues in the domain and 4 beacon residues in the core. (G,H) Ensemble computed with secondary structure restraints and 64 DDRs between 16 residues in the domain and 4 beacon residues. DDRs were derived from the structure with PDB ID 2BHW.

Figure 5

Conformational ensembles with 20 models each computed for the internal domain (residues 71-87) of CaiT. Views along the membrane normal are shown in panels A, C, and E and views parallel to the membrane plane in panels B, D, and F. The core (residues 8-70 and 88-503) is shown as a grey ribbon model and the ensemble by semi-transparent coil models. The template is shown as a green coil model. DDRs were derived from the structure with PDB ID 2WSX. (A,B) Unrestrained ensemble. Conformations close to the experimental one are shown in blue and other conformations in crimson colour. (C,D) Ensemble computed with 20 DDRs from loop residues 73, 76, 79, 82, and 85 to beacon residues 107, 257, 280, and 501. (E,F) Ensemble computed with 32 DDRs from loop residues 71, 73, 75, 77, 80, 82, 84, and 86 to the same beacon residues.

Figure 6

Label-to-label distances in statistical coils and IDPs. (A) Fraction of conformations with mean distances between spin labels in the favourable range for DEER measurements between 20 and 60 Å as a function of length of the peptide segment flanked by the labelled residues. Data was obtained by analysing an unrestrained ensemble (statistical coil) of 200 conformations with an amino acid sequence corresponding to residues 25-93 of the kinase inhibitory domain of p27. (B) Mean label-to-label distance in a statistical coil ensemble of p27 (solid black line) and variation of DDRs for chain segments of given length for ensembles of unbound p27KID (crimson, entry PDE2AAA of the Protein Ensemble Database) and α-synuclein (blue, entry PDE9AAC).

Figure 7

Conformational ensembles with 20 models each (semi-transparent crimson coil models) computed for the template p27Kip1 (residues 25-93) bound to Cdk2 complex (green coil model). The viewing direction is the principal axes of the inertia tensor of the template corresponding to the largest eigenvalue. (A) Unrestrained ensemble. (B) Ensemble considering only secondary structure restraints. (C) Secondary structure restraints (SSRs) and 18 DDRs from 10 labelled residues (LRs). (D) SSRs and 36 DDRs from 14 LRs. (E) SSRs and 40 DDRs from 14 LRs. (F) SSRs and 54 DDRs from 16 LRs. (G) SSRs and 56 DDRs from 16 LRs. (H) SSRs and 84 DDRs from 24 LRs.

Figure 8

Quality and efficiency measures for ensembles with 20 models each for the template p27Kip1 (residues 25-93) bound to Cdk2 complex and for unbound p27KID. Computations for p27Kip1/CDk2 include secondary structure restraints, those for p27KID do not. (A)

Ensemble RMSD (blue circles and dashed line), RMSD to template (black squares and solid line), and RMSD100 to template (red asterisks and dotted line) for p27Kip1/CDk2. (B) Computation times for ensemble generation with a single processor core of an Intel® Xeon® CPU E3-2141 v3 @ 3.50 GHz for p27Kip1/CDk2 (blue circles) and unbound p27KID (crimson asterisks).

Figure 9

Conformational ensembles with 100 models each (semi-transparent crimson coil models) computed with DDRs generated for unbound p27KID from the ensemble in entry PDE2AAA of the Protein Ensemble Database. The viewing direction for panels A, C, E, and G is the principal axes of the inertia tensor corresponding to the largest eigenvalue for the most representative conformation in the ensemble (green coil model). For panels B, D, F, and H the viewing direction corresponds to the smallest eigenvalue. (A,B) Template ensemble PDE2AAA. (C,D) Unrestrained ensemble. (E,F) Ensemble restrained by 56 DDRs from 16 LRs. (G,H) Ensemble restrained by 56 DDRs from 16 LRs and α-helical secondary structure for residues 38-58.

Figure 10

Typical plots for the agreement between Gaussian DDRs (dashed red lines) and distance distributions computed for the ensemble (solid blue lines) visualized in Fig. 9(G,H). (A) Agreement for a mean distance shift of 1.38 Å and an overlap of 0.972. Out of all DDRs, 50% have a larger overlap than that. (B) Agreement for a mean distance shift of 3.73 Å and an overlap of 0.940. Out of all DDRs, 90% have a larger overlap than that.

Table 1. Ensemble computations for the C-terminal loop 202-232 of LHCII

DDRs	Helix restraints	Ensemble RMSD [Å]	RMSD to template [Å]	Computation time /model [min]
0	no	21.8	30.4	0.02
0	yes	19.1	26.6	0.03
44	yes	4.4	7.3	50
64	yes	3.8	5.5	4000

Table 2. Ensemble computations for the internal loop 71-87 of CaiT

DDRs	Ensemble RMSD [Å]	RMSD to template [Å]	Computation time /model [min]
0	7.9	16.7	0.33
20	3.3	5.1	23.5
32	3.1	5.7	39

Figure 1

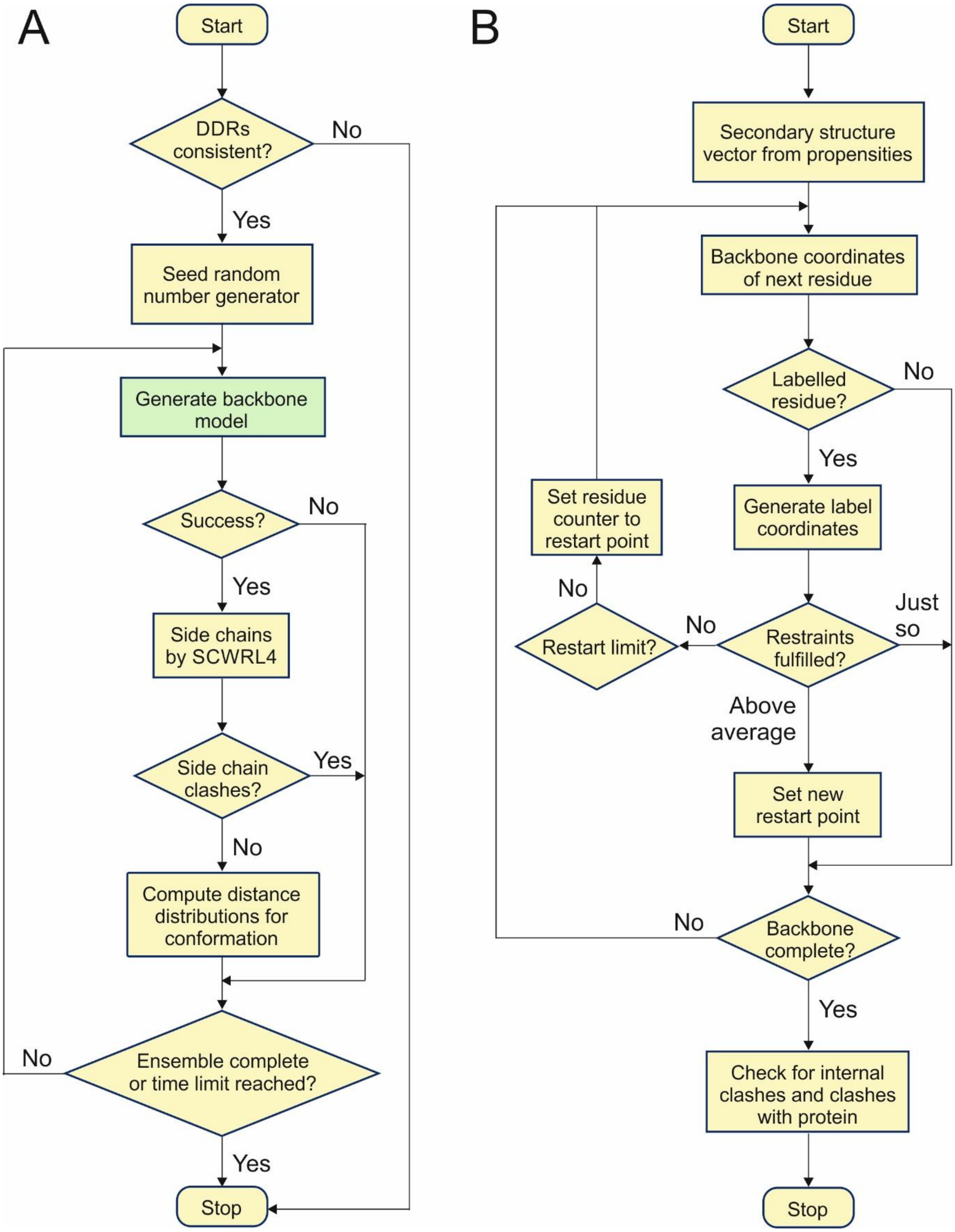

A
Start
DDRs consistent?
No
Yes
Seed random number generator
Generate backbone model
Success?
No
Yes
Side chains by SCWRL4
Side chain clashes?
Yes
No
Compute distance distributions for conformation
No
Ensemble complete or time limit reached?
Yes
Stop
B
Start
Secondary structure vector from propensities
Backbone coordinates of next residue
Labelled residue?
No
Yes
Set residue counter to restart point
Generate label coordinates
No
Restart limit?
No
Restraints fulfilled?
Just so
Above average
Set new restart point
No
Backbone complete?
Yes
Check for internal clashes and clashes with protein
Stop

Figure 2

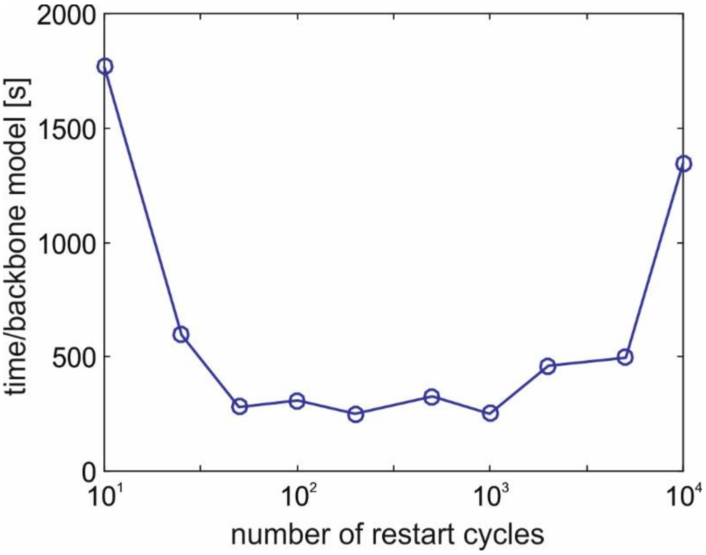

2000
1500
1000
500
0
time/backbone model [s]
10^1
10^2
10^3
10^4
number of restart cycles

Figure 3

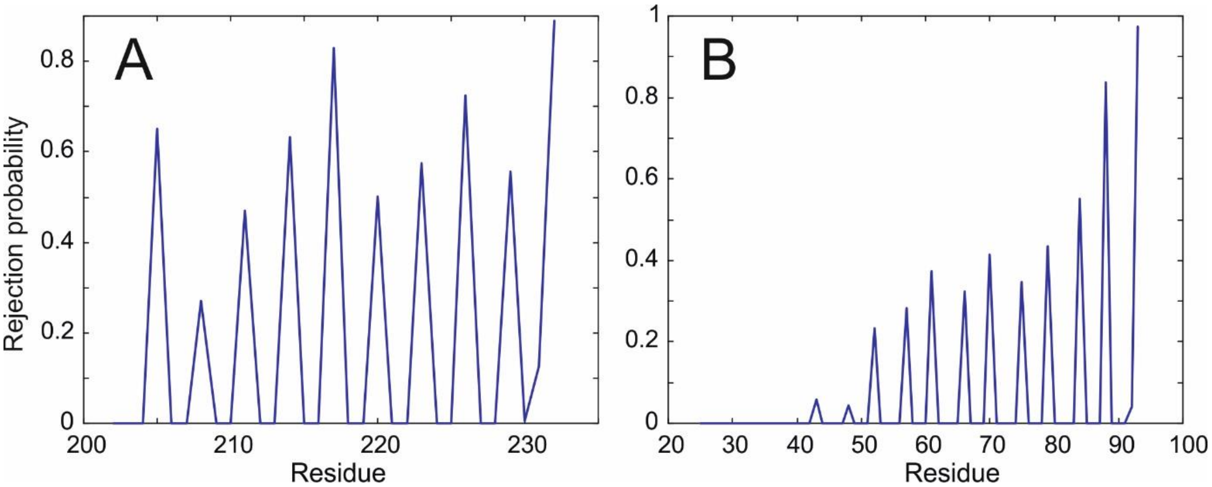

A
Rejection probability
0
0.2
0.4
0.6
0.8
200
210
220
230
Residue
B
0
0.2
0.4
0.6
0.8
1
20
30
40
50
60
70
80
90
100
Residue

Figure 4

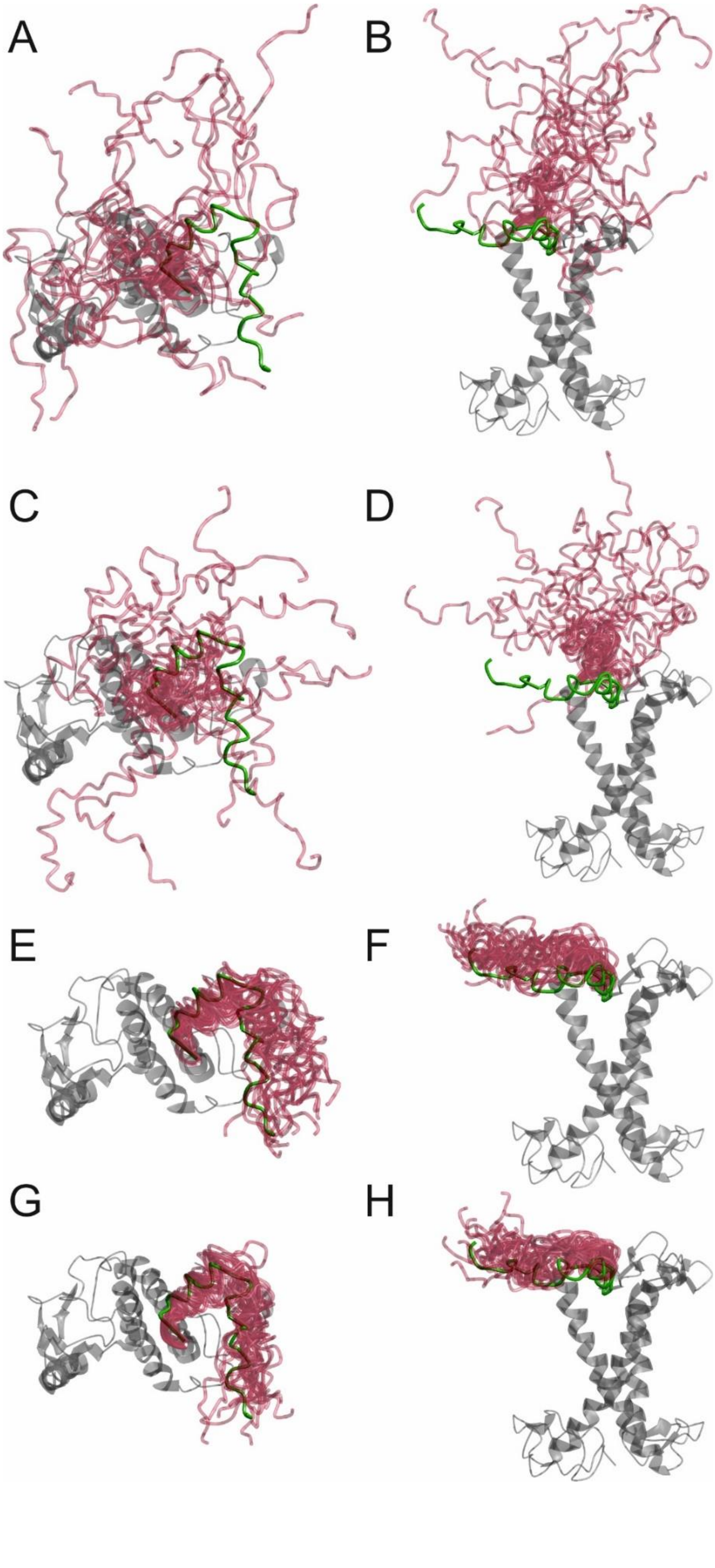


Figure 5

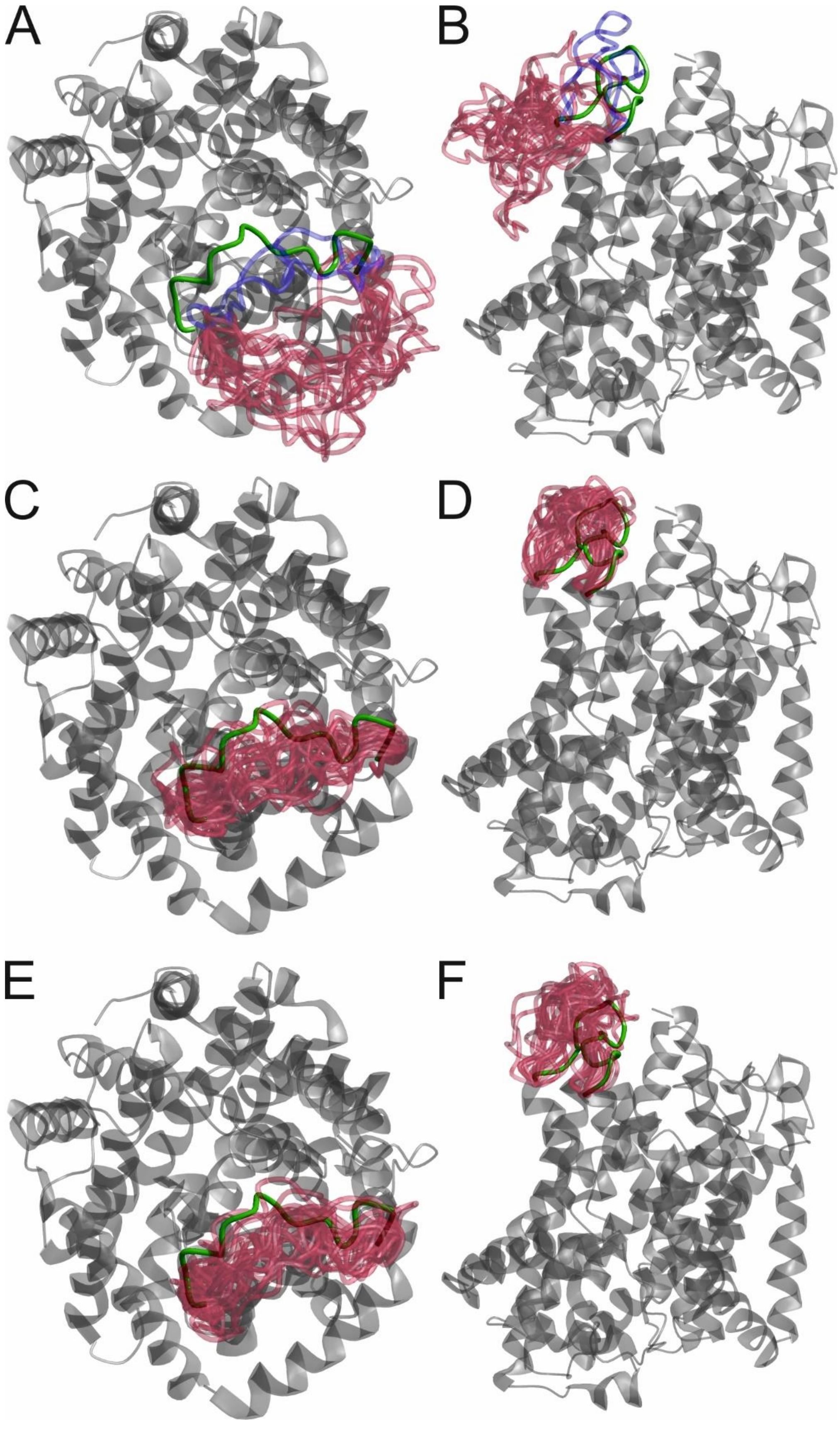


Figure 6

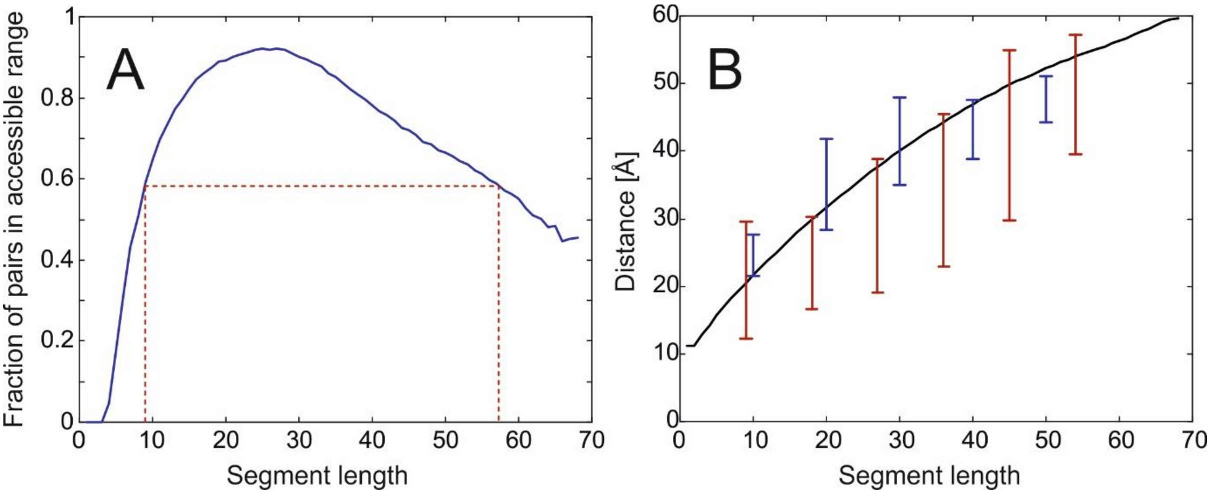

A
Fraction of pairs in accessible range
0
0.2
0.4
0.6
0.8
1
0
10
20
30
40
50
60
70
Segment length
B
Distance [Å]
0
10
20
30
40
50
60
0
10
20
30
40
50
60
70
Segment length

Figure 7

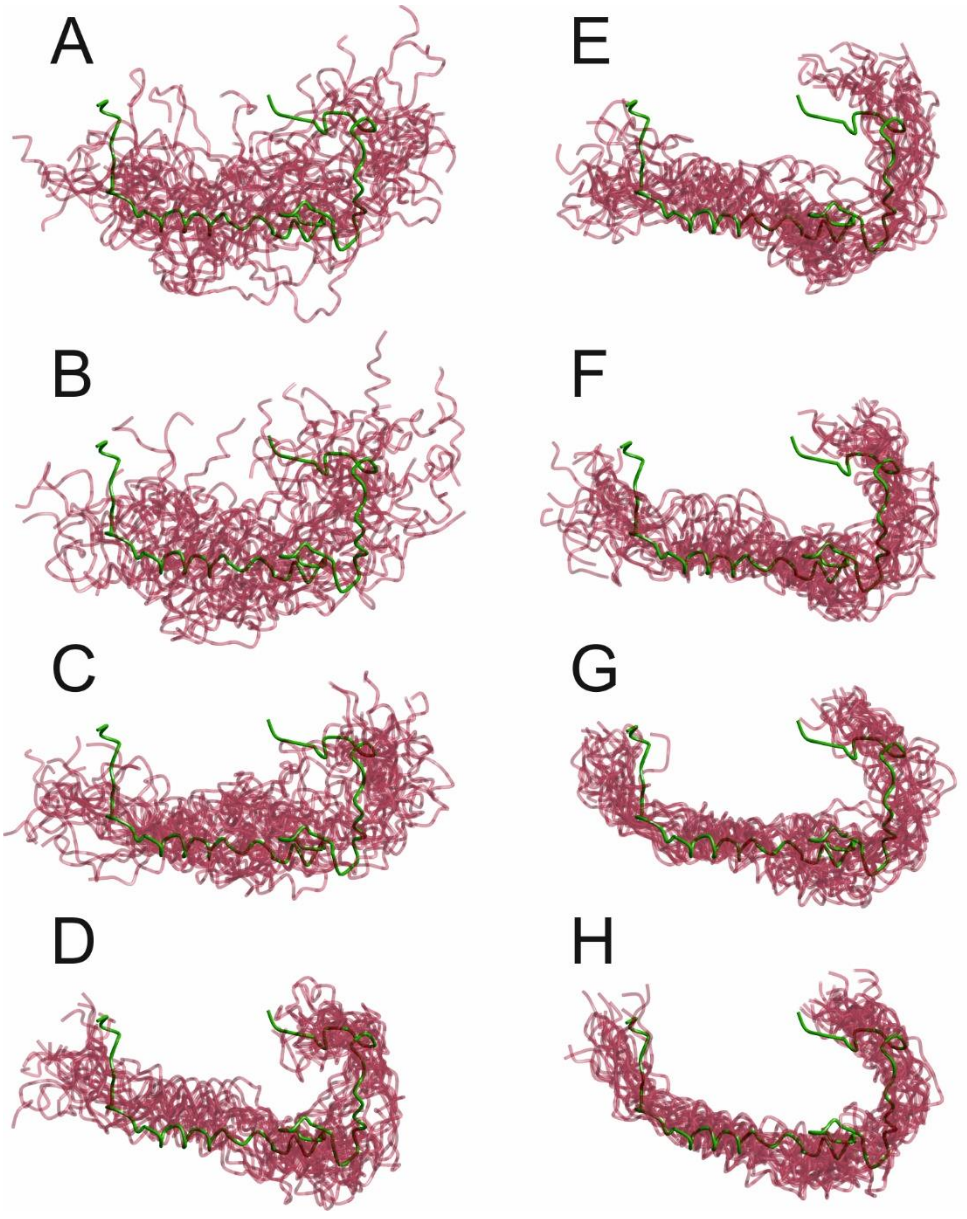


Figure 8

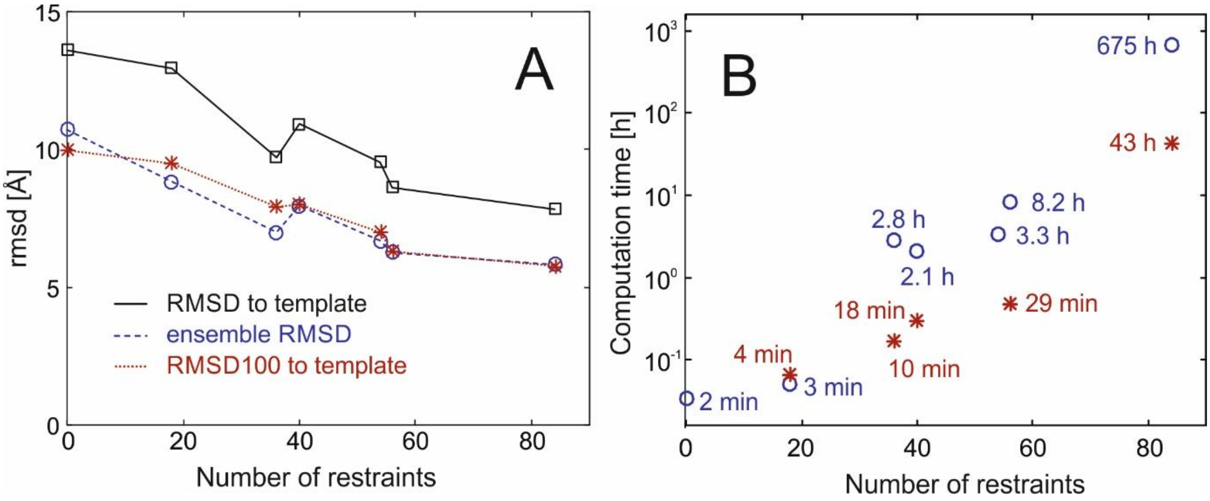

A
rmsd [Å]
15
10
5
0
0
20
40
60
80
Number of restraints
RMSD to template
ensemble RMSD
RMSD100 to template
B
Computation time [h]
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
675 h
43 h
2.8 h
8.2 h
3.3 h
2.1 h
18 min
29 min
4 min
10 min
2 min
3 min
0
20
40
60
80
Number of restraints

Figure 9

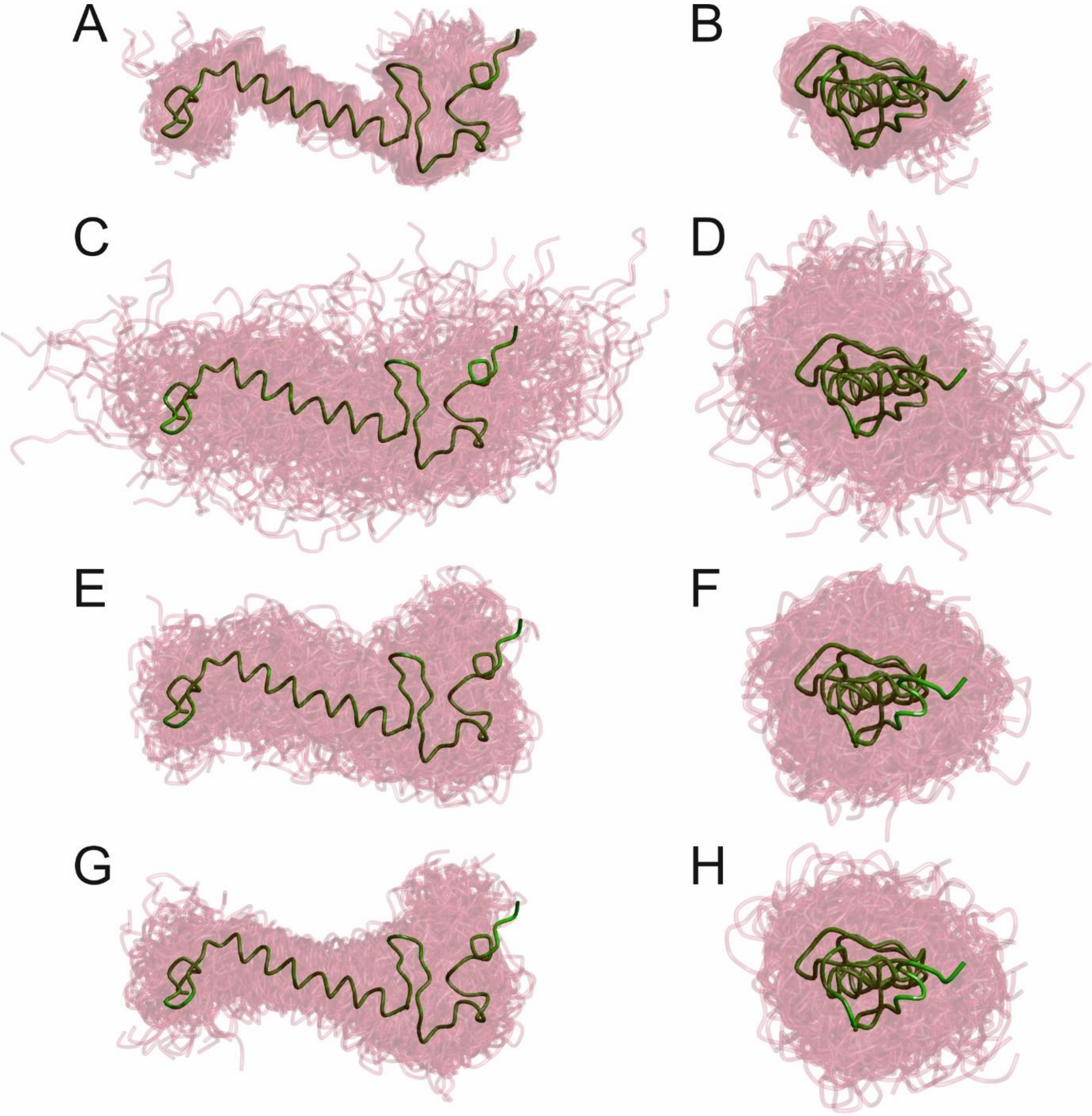
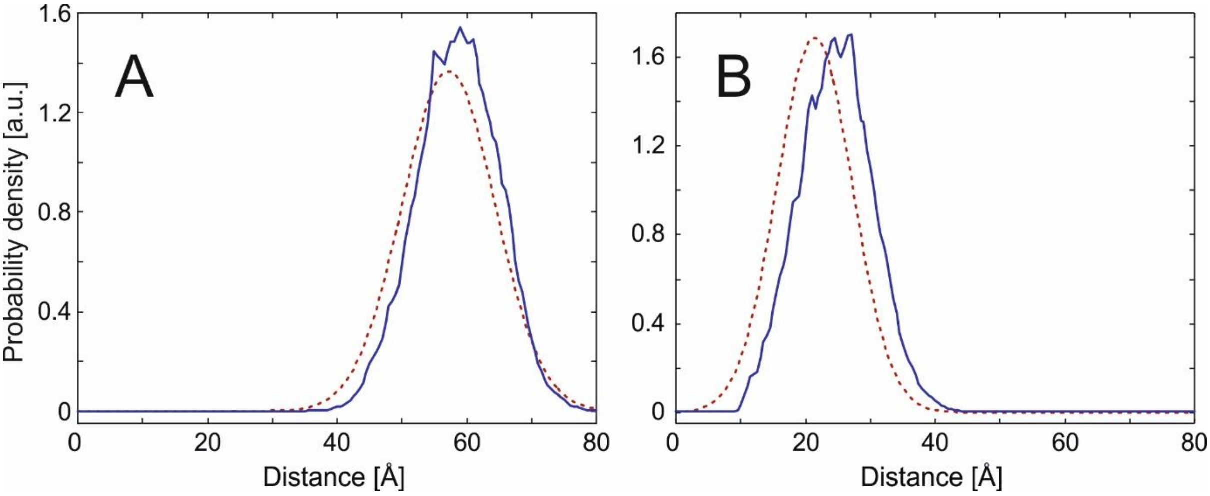


Figure 10


A
B
Probability density [a.u.]
0
0.4
0.8
1.2
1.6
0
20
40
60
80
Distance [Å]
Distance [Å]