Systematic exploration of protein conformational space using a Distance Geometry approach

Thérèse E. Malliavin,[*,†] Antonio Mucherino,[‡] Carlile Lavor,[¶] and Leo Liberti[*,§]

†*Unité de Bioinformatique Structurale, Institut Pasteur, UMR 3528, CNRS, Departement de Bioinformatique, Biostatistique et Biologie Intégrative, Institut Pasteur, USR 3756, CNRS, Paris, (75015) France*

‡*IRISA, Rennes, France*

¶*Applied Math Department, IMECC - University of Campinas, Campinas, SP, Brazil*

§*LIX CNRS, Ecole Polytechnique, Institut Polytechnique de Paris, Route de Saclay, 91128 Palaiseau, France*

E-mail: therese.malliavin@pasteur.fr; leo.liberti@lix.polytechnique.fr

Short title

Distance Geometry and protein conformational space

Corresponding authors

Thérèse E Malliavin

Institut Pasteur and CNRS UMR 3528, Unité de Bioinformatique Structurale, 25 rue du Dr Roux, 75015 Paris, France E-mail address: therese.malliavin@pasteur.fr

Leo Liberti, CNRS LIX (UMR 7161), Ecole Polytechnique, 91128 Palaiseau, France. E-mail address: leo.liberti@lix.polytechnique.fr

Abstract

The optimisation approaches classically used during the determination of protein structure encounter various diffculties, specially when the size of the conformational space is large. Indeed, in such case, algorithmic convergence criteria are more difficult to set up. Moreover, the size of the search space makes it difficult to achieve a complete exploration. The interval Branch-and-Prune (iBP) approach, based on the reformulating of the Distance Geometry Problem (DGP) provides a theoretical frame for the generation of protein conformations, by systematically sampling the conformational space. When an appropriate subset of inter-atomic distances is known exactly, this worst-case exponential-time algorithm is provably complete and fixed-parameter tractable. These guarantees, however, immediately disappear as distance measurement errors are introduced. Here we propose an improvement of this approach: the threading-augmented interval Branch-and-Prune (TAiBP), where the combinatorial explosion of the original iBP approach arising from its exponential complexity is alleviated by partitioning the input instances into consecutive peptide fragments and by using Self-Organizing Maps (SOMs) to obtain clusters of similar solutions. A validation of the TAiBP approach is presented here on a set of proteins of various sizes and structures. The calculation inputs are: a uniform covalent geometry extracted from force field covalent terms, the backbone dihedral angles with error intervals, and a few long-range distances. For most of the proteins smaller than 50 residues and interval widths of 20°, the TAiBP approach yielded solutions with RMSD values smaller than 3 Å with respect to the initial protein conformation. The efficiency of TAiBP approach for proteins larger than 50 residues will require the use of non-uniform covalent geometry, and may have benefits from the recent development of residue-specific force-fields.

Introduction

Since the early days of structural biology, optimization techniques have been at the heart of biomolecular structure calculation. Indeed, most of the experimental information is only

indirectly related to protein structure. In addition, this information is noisy. Furthermore, the sparsity of data is made even bigger as most of biophysical techniques concentrates on time-average or space-average data in order to obtain large enough signal-to-noise ratio.

Several optimization schemes have been used for Nuclear Magnetic Resonance (NMR) structure determination, such as simulated annealing[1] and genetic algorithms.[2] Nowadays, several approaches exist for protein structure determinations by NMR.[3–7] A Bayesian approach,[8] using a Markov chain Monte Carlo (MCMC) scheme for the conformational space sampling,[9,10] allowed the increase of the convergence radius for problems of protein structure determination by NMR. Furthermore, the use of a log-harmonic shape for distance restraint potential,[11] along with a Bayesian approach for the restraint weighting,[12] allowed an improvement of the quality of NMR protein structures.[13–15] Log-harmonic restraints defined using Bayesian inference have been also recently[16] proposed for back-mapping from coarse-grained models to atomic structures.

Most of the optimization methods used so far provide no guarantees of optimality, although they are commonly used in the hope of obtaining the global minimum or several global minima of the optimization problem. This, however, depends on the choice of a starting point for the computation. Consequently, calculations of protein conformations under NMR restraints are repeated several times during the procedure of structure determination,[17] and the convergence of these calculations is generally required in order to accept a set of conformations as a solution. This iterative frame,[6,18] however, encounters difficulties when the problem has many local minima that are far apart. Such cases started to occur more frequently in the field of structural biology with the growing interest to disordered regions of biomolecules.[19–21] Monte Carlo approaches have been proposed for intrinsically disordered proteins[22–24] and molecular dynamics simulations[25] are also used on all kind of biomolecular polymers, but they do not provide a definitive answer to the problem of finding all minima.

Since NMR studies biomolecules in solution, and due to the large number of various parameters it can measure, it is particularly sensitive to the effect of internal mobility. NMR

measures are inter-atomic distances and angles, which are closely related parameters. The problem of protein structure determination by NMR can be thus considered as a Distance Geometry Problem (DGP).[26,27] The interval Branch-and-Prune (iBP) approach has been developed[28] for solving the DGP in the framework of the calculation of protein conformations. In this approach, an atom re-ordering[29,30] ensures that there is a restricted and manageable locus for the spatial position of every atom. This is achieved by using a "relaxed form" of trilateration with respect to the three preceding atoms in the order. More precisely, two out of three of the distances involved in trilateration must be known exactly, and one may be subject to uncertainty and represented by an interval. Any atom, together with its three reference predecessors, give rise to a 4-clique in the protein graph: the iBP approach mimics the approach of exploring protein conformation in torsion angle space.[31–34] In the clique, exact distances are provided by covalent bond lengths and bond angle values, using the cosine law. Note that applying this framework using generic information from a force field instead of measured distances makes the implicit assumption of a uniform covalent geometry within the protein structure. Analyses of high-resolution crystallographic structures,[35,36] however, have shown that this assumption is not necessarily verified. Independent parallel work has conducted to the development of residue-specific force field.[37–40]

Basing on the atom reordering, it is possible to describe a tree exploration algorithm in order to find all solutions of a DGP instance. Each tree node represents a spatial position for an atom. The level of a node in the tree is the index of the atom in the reordering. This implies that all positions on the same tree level are the possible spatial positions for the atom having as rank this level. The width of the tree increases exponentially in the worst case, but it can be bounded to more manageable levels[41] by choosing specific atomic orders. This yields a fixed-parameter tractable behavior (at least with exact distances). We note that the exploration of this tree is complete but implicit, in the sense that certain sub-trees are pruned because the atomic positions at their root nodes are not consistent with long range distances to preceding atoms. Naturally, each pruned node induces the pruning of the

sub-tree rooted at that node. It was demonstrated[27,28] that, starting from a set of exact distances measured in a given PDB structure, the search tree can be completely explored in a relatively small amount of CPU time.

Here, we employ the iBP algorithm in a setting which is considerably closer to the protocols of protein structure determination than the mathematical setting in which it was initially conceived. Instead of exact distances measured on a given PDB structure, this requires the use of a mixed set of distance intervals and of exact distances arising from a covalent geometry defined through a force field. Several attempts have been made in this direction in the recent past. A significant exploration of the conformational space of some α-helical 15 to 51-residues proteins was performed in Ref.,[42] and more recently, the iBP approach was re-implemented[43] in order to allow its application to real-life cases of protein structure determination. First, the number of tree branches was reduced[44] by taking into account the information from improper angles. Second, a parser and a grammar have been defined to convert the topology, parameter and atom type information used in molecular modeling to the distance information which is the main input of iBP. Third, a syntax has been defined to make the atom reordering information a user-defined input of the calculation. This new implementation makes it possible to perform tree branching on intervals determined on ϕ and ψ backbone angles, which may be obtained through chemical shift measurements.[45] Nevertheless, no systematic exploration of the protein conformational space has been previously attempted.[43]

In the present work, we employ the implementation of Ref.[43] to develop a new strategy which allows a systematic sampling of the conformational space of small proteins and we validate this strategy using a set of various protein structures. The expected combinatorial explosion is prevented by several ingredients: (i) the division of the protein into fragments which are sampled independently and then assembled, (ii) the extensive use of signed improper angle values to reduce the tree size of each fragment, (iii) the use of self-organizing maps to cluster conformations of intermediate fragments.

The geometrical information used for input calculation corresponds to relevant NMR measurements on proteins. Indeed, NMR chemical shifts are easily measurable parameters. The relationship between chemical shifts and atomic coordinates is not straightforward, but several methods, as the neural network TALOS-N[45] or chemical shift prediction approaches,[46–48] exist for relating chemical shifts and atomic coordinates. Among them, TALOS-N[45] predicts from chemical shift values, (ϕ, ψ) likelihood distributions. The existence of such distributions supports the use of intervals on ϕ and ψ values as inputs for the TAiBP approach. In addition to the ϕ and ψ intervals, distance restraints with interval widths of 6 and 10 Å and defining qualitatively the protein global shape were used as inputs.

The proposed approach is called threading-augmented interval Branch-and-Prune (TAiBP) approach, as it intends to generate conformations of peptide fragments using iBP, as well as to thread these fragments in 3D space in order to build protein conformations. The name was coined in analogy to the threading approach[49] used in protein 3D structure prediction. We point out that the idea to separate iBP instances in sub-instances is not completely new, but it was explored so far only in the context of parallel[50] and distributed[51] computing. Also, the idea of constructing protein conformations from fragment assembly was proposed initially[52–54] in the Rosetta approach for protein structure modeling.

The proposed methodology is innovative with respect to the state of the art because it is designed to find all possible configurations compatible with a given set of angle and distance restraints on a given protein. This is in contrast to classical methods for structure determination,[1] which might at best produce different protein conformations. The approach is different with respect to the more recently proposed methods aiming at determining the global minimum configuration of the system[4,55–59] or at determining all relative positions of monomers within a protein homo-oligomer.[60] On the contrary, the exhaustive list of conformations generated by TAiBP provides solutions for a larger range of problems.

It is important to note that our purpose is beyond finding a conformation close to the target one, since we aim instead to the much more ambitious goal of finding many (and

hopefully all) incongruent but geometrical consistent conformations. Moreover, because our algorithm approach is not iterative but based on branching, we have no need for considering "convergence to a local optimum" a requirement for accepting a conformation. The results of our computational experiments, however, have been validated by detecting whether conformations close to the target PDB structure have been sampled during the tree exploration, by Root Mean Square Distance (RMSD) of atomic coordinates to the target structure. The proposed approach allows us to explore the tree for proteins up to 50 residues. The non-uniform covalent geometry, prevents (by now) our method from being successful on proteins larger than 50 residues.

Materials and Methods

Test case database

The database of protein structures was built in the following way. The protein structures contained in `kinemage.biochem.duke.edu/databases/top100.php`[61] have been downloaded. This database was chosen as high resolution X-ray crystallographic structures on which hydrogens have been added with rotational optimization of OH, SH and NH^{3+} positions,[61] thus producing objects corresponding to those iBP is designed to calculate.

Protein structures with number of residues between 21 and 107 have been selected, containing only trans peptidic bonds and corresponding to the following list of 24 proteins: 1aacH, 1benABH, 1bkfH, 1bpiH, 1ckaH, 1cnrH, 1ctjH, 1difH, 1edmBH, 1fxdH, 1igdH, 1iroH, 1isuAH, 1mctIH, 1ptfH, 1ptxH, 1rroH, 256bAH, 2bopAH, 3b5c, 3ebxH, 451cH, bio1rpoH and bio2wrpH. In 3b5c, the N terminal residue T88 was removed because of missing backbone atoms. On each structure, the conformation of chain A was selected for preparing the iBP input, and in the case multiple conformations have been observed for a residue, the A conformation was selected.

Input values for the calculation

The parameters defining the covalent and improper geometries were taken from the geometric force field PARALLHDG (version 5.3)[62] (Table 1). One should notice that, although these parameters were proposed more than two decades ago, they still correspond to the state-of-art of molecular force fields with fixed charges, as the covalent bond lengths and bond angles of most of fixed-point force fields were determined at the same time or earlier.[63] The atom re-ordering is the same proposed than in the most recent implementation of iBP[43] (Table 2).

Two sets of values were used for the backbone angles ϕ and ψ in order to evaluate the effect of assuming a uniform covalent geometry on the TAiBP results.

(i) the ϕ_{angl}, ψ_{angl} angles of residue i measured on the X-ray crystallographic structures as the angles between planes $\mathrm{C}^{i-1}\mathrm{N}^{i}\mathrm{C}\alpha^{i}$ and $\mathrm{N}^{i}\mathrm{C}\alpha^{i}\mathrm{C}^{i}$ and between planes $\mathrm{N}^{i}\mathrm{C}\alpha^{i}\mathrm{C}^{i}$ and $\mathrm{C}\alpha^{i}\mathrm{C}^{i}\mathrm{N}^{i+1}$ using VMD.[64] The plane ABC is defined as the plane passing through the positions of the atoms A, B and C.

(ii) the ϕ_{dist}, ψ_{dist} angles calculated from the distances $\mathrm{d}(\mathrm{N}^{i},\mathrm{N}^{i+1})$ and $\mathrm{d}(\mathrm{C}^{i},\mathrm{C}^{i+1})$ between N and C atoms of successive residues, assuming the covalent geometry uniform and described in Table 1. The dihedral or pseudo-dihedral angle Ω between ordered atoms i-3,i-2,i-1 and i, is determined using the cosine law from a trihedron:[44]

$$\cos\Omega = \frac{\cos\gamma - \cos\alpha\cos\beta}{\sin\alpha\sin\beta}, \tag{1}$$

where α is the angle between atoms (i-3,i-2,i-1), β is the angle between atoms (i-1,i-2,i), and γ is the angle between atoms (i-3,i-2,i). For Ω angles being ϕ or ψ, the angles α, β and γ are calculated from the bond lengths and bond angles among heavy backbone atoms, as well as the distances $\mathrm{d}(\mathrm{N}^{i},\mathrm{N}^{i+1})$ and $\mathrm{d}(\mathrm{C}^{i},\mathrm{C}^{i+1})$ between successive residues along the protein sequence. The calculation is described in details in the Supporting Information.

The input restraints for iBP processing of peptide fragments are: (a) the restraints

corresponding to the bond lengths and bond angles of the force field PARALLHDG (version 5.3);[62] (b) the backbone angles ϕ and ψ, determined as described previously; (c) the distances between Cα atoms located at the two extremities residues of each peptide fragment. The input restraints for the fragment assembly are: (i) the long-range distances between Cα atoms of the residues located at the middle of each fragment; (ii) pruning devices avoiding that Cα atoms belonging to different fragments are closer than 1 Å. The following error bounds have been used: errors of $\pm$ 10°, 20° and 30° for the angles ϕ and ψ, an error of $\pm$ 3 Å for the Cα-Cα distance between extremities of peptide fragments, and error of $\pm$ 5 Å for the long-range Cα-Cα distance between peptides fragments. Examples of inputs in tbl format are given in the Supporting Information.

Interval Branch-and-Prune calculation of peptide fragments

As the TAiBP approach intends to explore the conformation of protein backbone, the processed protein is initially converted to a poly-Alanine chain. The protein is then divided in 15-residues peptide fragments, two successive fragments having a sequence of 5 superimposed residues. This fragment size was determined as it permits to obtain tree sizes which are manageable to explore in the reasonable amount of time, as it will be shown below in the subsection "Exploring the conformational space of fragments using iBP" of Results. The number of superimposed residues was chosen to avoid artifacts due to superimposition. The peptide fragments are assembled together to produce protein conformations, as it will be described in the next subsection.

For each fragment, the iBP tree of possible conformations is systematically explored. We employ the most recent implementation of iBP[43] (in the C programming language), which is tuned for the calculation of protein conformations based on the force field knowledge for the covalent geometry. The tree branching is performed on the ϕ and ψ backbone angles. No branching was performed on the peptidic angle ω. Indeed, analyses of variations of ω angles in the X-ray crystallographic structures[65–67] show that the angles ω mostly vary in

intervals of $\pm 8^\circ$ around -180 and 180° which are smaller than the intervals sampled in the present work for ϕ and ψ.

The position of the atoms were determined as described in Ref.[44] For any atom i in the order, we seek the atomic coordinates $\mathbf{x}_i$, given distances between the three preceding atoms $i-1$, $i-2$ and $i-3$. As described above, the distances $d_{i,i-1}$, $d_{i,i-2}$ and $d_{i,i-3}$ between atoms i,$i-1$, atoms i,$i-2$ and atoms i,$i-3$ are known, where distance $d_{i,i-3}$ between atoms i and $i-3$ is potentially an interval. The variables d_i, θ_i and τ_i, where d_i denotes $d_{i,i-1}$, permes to determine the position of atom i by the following equation:

$$\mathbf{x}_i = \mathbf{p}_1 + \tau_i\,\mathbf{p}_2 + \sigma_i\sqrt{1-\tau_i^2}\,\mathbf{p}_3$$

where $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3 \in \mathbb{R}^3$ depend only on $\mathbf{x}_{i-1}$, $\mathbf{x}_{i-2}$, $\mathbf{x}_{i-3}$, d_i and θ_i,

$$\begin{aligned}
\mathbf{p}_1 &= -\left(\frac{d_i}{\|\mathbf{r}_{12}\|}\right)\left(\left(\cos(\theta_i) - \frac{\|\mathbf{r}_{12}\|}{d_i}\right)\mathbf{x}_{i-1} - \cos(\theta_i)\,\mathbf{x}_{i-2}\right)\\
\mathbf{p}_2 &= -\left(\frac{d_i}{\|\mathbf{r}_{12}\|}\right)\left(\frac{\sin(\theta_i)}{\|\mathbf{r}_{12}\times\mathbf{r}_{23}\|}\right)\left(\|\mathbf{r}_{12}\|^2\,\mathbf{r}_{23} - (\mathbf{r}_{12}\cdot\mathbf{r}_{23})\,\mathbf{r}_{12}\right)\\
\mathbf{p}_3 &= -\left(\frac{d_i}{\|\mathbf{r}_{12}\|}\right)\left(\frac{\sin(\theta_i)}{\|\mathbf{r}_{12}\times\mathbf{r}_{23}\|}\right)\|\mathbf{r}_{12}\|\,(\mathbf{r}_{12}\times\mathbf{r}_{23})
\end{aligned}$$

and we have introduced the scalars $\mathbf{r}_{12}, \mathbf{r}_{23}$ for notational simplicity,

$$\mathbf{r}_{12} = \mathbf{x}_{i-1} - \mathbf{x}_{i-2}$$

$$\mathbf{r}_{23} = \mathbf{x}_{i-2} - \mathbf{x}_{i-3}$$

The angle θ_i is obtained from the cosine law using the relevant distances,

$$\theta_i \triangleq \cos^{-1}\left(\frac{d_i^2 + d_{i-1,i-2}^2 - d_{i,i-2}^2}{2d_i\,d_{i-1,i-2}}\right) \tag{2}$$

The pseudo-dihedral angle ω_i formed by the atoms $i-3$,$i-2$,$i-1$ and i is partially determined by its cosine value $\cos\omega_i = \tau_i$, which is calculated using the cosine law for a

trihedron:[44]

$$\tau_i = \frac{2d_{i-2,i-1}^2\left(d_{i-3,i-2}^2 + d_{i-2,i}^2 - d_{i-3,i}^2\right) - d_{i-3,i-2,i-1}d_{i-2,i-1,i}}{\sqrt{4d_{i-3,i-2}^2 d_{i-2,i-1}^2 - d_{i-3,i-2,i-1}^2}\sqrt{4d_{i-2,i-1}^2 d_{i-2,i}^2 - d_{i-2,i-1,i}^2}} \tag{3}$$

where

$$d_{i-3,i-2,i-1} \triangleq d_{i-3,i-2}^2 + d_{i-2,i-1}^2 - d_{i-3,i-1}^2$$
$$d_{i-2,i-1,i} \triangleq d_{i-2,i-1}^2 + d_{i-2,i}^2 - d_i^2$$

The determination of the pseudo-dihedral ω_i is completed by the sign $\sigma_i \in \{-1, +1\}$ of $\sin\omega_i$. When ω_i is known from either protein chemistry or measurement, we may directly compute τ_i, as well as the sign $\sigma_i \in \{-1, +1\}$. This is the case when the angle ω_i corresponds to an improper angle (Table 1) and this allows to reduce the branching to one branch.

The number of saved conformations is reduced by applying a RMSD filter of 3 Å between two successively saved conformations. In order to avoid pruning due to slight discrepancy between distance restraints, a tolerance of 0.05 Å has been added to the bounds of distance intervals. The minimum discretization factor, which is the minimum ratio between each distance interval to the number of tree branches generated within the interval, was set to 0.05 Å, in order that the branching does not over-sample small intervals. No pruning due to the van der Waals radii of the force field protein-allhdg5-4 PARALLHDG (version 5.3)[62] was applied. A maximum number of saved conformations of 10^9 was permitted for each iBP run. The solutions are stored in a multiframe dcd format.[68]

Assembling the peptide fragments and clustering

The generated conformations of neighbouring peptide fragments in the protein sequence are then assembled by superimposing the five last and initial residues of the fragments located first and second in the sequence. The conformations of fragments are assembled by root-

mean-square superimposition of backbone atoms located in the five superimposed residues. For each superimposition, the residue number for which the smallest distance was observed between corresponding atoms in the two peptides is used to decide where to stop with the first peptide and to continue with the second one. The assembled conformation is then submitted to two pruning devices: (i) a device checking whether there is no clash between the two fragments, i.e. no Cα atoms closer than 1 Å, (ii) a device checking that long-range Cα-Cα distance restraints between peptide middle residues are verified. The fragment assembly is implemented using python scripting based on the MDAnalysis[69,70] and numpy[71] python packages.

To scale down the combinatorial explosion of the calculation, a clustering approach, the **S**elf-**O**rganizing **M**aps (SOM),[72–75] which is an artificial neural network (ANN) trained using unsupervised learning, were used to reduce the number of conformations.[76] The SOM displays the advantage with respect to the k-means clustering approach that it does not require the predetermined knowledge of the number of clusters. The SOM approach was used after a iBP calculation or after an assembly step as soon as the number of saved conformations was larger than 1000. The conformations sampled by iBP were encoded from the distances d_{ij} calculated between the n C_α atoms of the fragment, by diagonalizing the covariance matrix C:

$$C_{i,j} = \frac{1}{n} \sum_{k=1}^{n} \sum_{l=1}^{n} (d_{i,k} - \bar{d}_i)(d_{l,j} - \bar{d}_j) \tag{4}$$

where $\bar{d}_i = \frac{1}{n} \sum_{j=1}^{n} d_{i,j}$. The information contained in the matrix C is equivalent to its four largest eigenvalues along with the corresponding eigenvectors. The eigenvalue and eigenvector descriptors are used to train a periodic Euclidean 2D self-organizing map (SOM), defined by a three-dimensional matrix. The first two dimensions were chosen to be 100×100 and define the map size.

The self-organizing maps were initialized with a random uniform distribution covering the range of values of the input vectors. At each step, an input vector is presented to the map, and the neuron closest to this input is updated. The maps are trained in two phases.

During the first phase, the input vectors are presented to the SOM in random order to avoid mapping bias with a learning parameter of 0.5, and a radius parameter of 36.[77] During the second phase, the learning and radius constants are decreased exponentially from starting values 0.5 and 36, respectively, during 10 cycles of presentation of all the data in random order. Once the calculation of the SOM has been realized, the conformations corresponding to local maxima of homogeneity, are detected and the total set of conformation is replaced by these representative conformations.

Results

Probing the hypothesis of uniform covalent geometry

The 24 structures extracted from the database of Word et al[61] have been processed to analyze the geometry of covalent angles (Figure 1a). The distributions of covalent angles between C-N-Cα (blue curve), N-Cα-C (magenta curve) and Cα-C-N (green curve) (Figure 1a) are centered on 121.3°, 110.6° and 116.8°, with standard deviations of 2.2°, 3.0° and 2.2°. These distributions agrees with the ones observed by Hinsen et al:[78] C-N-Cα (121.4° $\pm$ 1.6°), N-Cα-C (111.1° $\pm$ 2.9°), Cα-C-N (116.6° $\pm$ 1.3°), and in agreement to this work, the largest variability is observed for the bond angle N-Cα-C. The bond angle values were compared to the B factor values averaged on the corresponding residues (Figure 1b). The lack of correlation between the values of bond angles and B factors shows that the variations of covalent geometry cannot be assigned to differences in protein internal mobility.

The variations in covalent geometry were then plotted (Figure 1c-e) along the positions of protein residues in the Ramachandran diagram, by coloring the point describing the (ϕ, ψ) angle values of a given residue, according to the values of the residue bond angles. The Ramachandran plots are multi-colored according to the values of bond angles C-N-Cα (Figure 1c), N-Cα-C (Figure 1d) and Cα-C-N (Figure 1e). All α-helix regions, around (-60°,-45°), display a quite monochromatic pattern, with values mostly in the range 100°-105° for angle

C-N-Cα (Figure 1c), in the range 125°-130° for angle N-Cα-C (Figure 1d) and in the range 120°-125° for angle Cα-C-N (Figure 1e). On the contrary, the β-strand region and the loops region of each diagram display a larger heterogeneity in bond angle values than the α-helix region. This heterogeneity has certainly a strong influence on the overall tertiary structures. Indeed, the β strands are extended structures in which local variations can have strong influence on the orientation at long distance. Similarly, the change of direction of protein backbone can be also very sensitive to local loop structure variation. In that way, both β strand orientations and loop directions have a strong impact on the protein tertiary structure.

In order to investigate the relevance of the uniform geometry hypothesis for the iBP calculation, the ϕ_{angl} and ψ_{angl} values measured on the top100 conformations have been compared to the ϕ_{dist} and ψ_{dist} values obtained as described in the Materials and Methods subsection "Input values for the calculation". For each residue K, the cumulative sums of the differences between *angl* and *dist* backbone angles for residues i, i varying from 1 to K, were calculated:

$$\Phi_K = \sum_{i=1}^{K} (\phi^i_{\mathsf{angl}} - \phi^i_{\mathsf{dist}}) \quad (5)$$

$$\Psi_K = \sum_{i=1}^{K} (\psi_{\mathsf{angl}} - \psi^i_{\mathsf{dist}}) \quad (6)$$

In Figure 2, the variations of Φ_K (green curves) and of Ψ_K (magenta curves) have been plotted along K, for the 24 studied proteins. The most important observation from these curves is that Φ_K and Ψ_K display extraordinary large variations along protein primary sequence. These variations extend from about 100° for the proteins 1benABH, 2bopAH, 1okAaH, bio1rpoH, up to several hundreds of degrees. The drift of Φ_K and Ψ_K depends of course on the total number of residues in the protein. Another observation is that Φ_K and Ψ_K curves do not display the same features. Φ_K curves are positive and increase along K, whereas Ψ_K curves are mostly negative and decrease along K. In addition, for most of the

proteins, the comparison of the absolute values of Φ_K and Ψ_k reveals that one absolute value is larger than the other one, which induces a partial compensation between Ψ_k and Φ_K drift.

To summarize, the analysis of protein structures involved in the present validation reveals that the hypothesis of uniform covalent geometry is far from being verified even in high-resolution crystallographic structures as the ones selected from the database top100[61] with resolutions in the range 1.0-1.5 Å. Consequently, the differences between angles ϕ_{angl}, ϕ_{dist} and ψ_{angl}, ψ_{dist} display large cumulative drifts along the protein sequence.

Exploring the conformational space of fragments using iBP

iBP calculations were performed on individual peptides spanning the analyzed proteins and the obtained results are presented in Figure 3. The run durations, plotted with decimal logarithmic scale, are centered around 10^2-10^3s, for error intervals of 20° (blue curve) and 40° (magenta curve), and jump to the 10^3-10^4s range for an error of 60° (green curve). The maximum run duration thus corresponds to about one day, which is not prohibitive.

The tree sizes were reduced using the signed values of improper angles (Table 1), and are in the range 10^5-10^9. For each run, a maximal number of 10^9 conformations to generate was required as input. The number of conformations saved during each run (Figure 3c) is in the range of 1 to 10^6 which is several order of values smaller than this maximal number. All trees have thus been completely parsed during the iBP calculations.

The number of generated (Figure 3b) and saved (Figure 3c) conformations increase linearly along the run duration. The number of generated conformations is in the range 10^4-10^8, whereas the number of saved conformations is in the range 1-10^5. For the largest error intervals (40° and 60°, magenta and green dots), similar numbers of conformations are generated (Figure 3b) and these numbers depend mainly on the duration of the run. In the case of the smallest interval width (20°, blue dots in Figure 3b), few runs display much smaller numbers of generated conformations. The number of saved conformations (Figure 3c) is of course smaller by two or three orders of magnitude from the number of generated conformations,

but is also more dispersed. These numbers sample superimposed ranges for error intervals of 20° and 40° (blue and magenta dots), whereas they sample larger values for error interval of 60° (green dots). The result of tree parsing thus depends only slightly on the interval width, but the number of saved conformations vary qualitatively for error interval larger than 40° (Figure 3c, green points), and this difference is also visible in the run duration (Figure 3a, green curve).

The loss of information due to interval discretization during iBP calculations was analyzed (Figure 3d) through the discretization factor, which is the ratio between each distance interval to the number of tree branches generated within the interval. The standard deviation of this factor is plotted along its average value, both being calculated for the same individual iBP tree. Overall, one should notice that the largest discretization factors are smaller than 0.25 Å. According to a recent work,[79] the ensemble-average pairwise backbone RMSD for the microscopic ensemble underlying a typical protein X-ray structure is about 1 Å. The discretization factor of 0.25 Å is thus in the range of uncertainty of typical X-ray structures, and does not induce major loss of information in the iBP calculations.

For the various error intervals, the couples of average and standard deviation values for the discretization factor (Figure 3d) are clustered around different points: (0.11, 0.11) Å for the intervals of 20°, (0.18, 0.08) Å for the intervals of 40°, (0.23, 0.07) Å for the intervals of 60°. As expected, the average value increases with the interval width on ϕ and ψ angles. More surprisingly, the standard deviations decrease with the interval width: this is due to the discretization inputs. Indeed, the maximum number of branches is limited by 4 in all calculations, but the discretization factor should be always larger than a threshold of 0.05 Å. These two parameters induce the saturation of the number of branches for large widths as 60°. At the contrary, for smaller interval widths, the maximum number of branches is not attained for all distance intervals due to the required threshold. This induces a larger variability between the number of tree branches as well as a larger standard deviation.

During each iBP calculation, the conformations generated by branching on the ϕ and ψ

intervals are then pruned or not according to the violation or to the verification of the distance interval between the Cα atoms located at the N and C terminal residues. Percentages of pruned conformations (Figure 3e) are observed up to 100%. In the case of interval widths of 20° and 40°, three and one runs do not provide any solutions. As all these runs were performed using as input ϕ_{dist} and ψ_{dist} backbone angles, the pruning of all solutions is due to the inconsistency between the ϕ and ψ angle restraints and the extremities distance restraint. This inconsistency arises directly from the non-uniform covalent geometry described in the first section and is amplified by the use of a small error on backbone angle restraints.

Similarly to the number of runs without solutions, contrasted distributions are observed (Figure 3e) for the percentages of pruned conformations, depending on the width of intervals on backbone angles. For the smallest width (20°: blue curve), the percentage of pruned conformations displays a weak maximum at around 25%, but a non-negligible number of runs display percentages of pruned conformations in the 50-90%. Such high pruning percentages arise because in the case of narrow intervals on backbone angles, the hypothesis of uniform covalent geometry made by iBP has much more chances to induce solutions which do not verify the distance restraint between peptide extremities. For the larger interval widths on backbone angles (40°: magenta curve, 60°: green curve), the distribution is much more focused on larger percentages with respective ranges of 40-70% and 60-80%. The percentages larger than 80% are nevertheless vanishing for the largest interval widths. The global picture is that the increase of intervals on backbone restraints induces more pruning, but the percentage peaks at 50 and 60% obtained for interval widths of 40° and 60° are promising for the application of iBP to cases with error on restraints at the level of experimental cases.

After generating peptide conformations using iBP, a procedure based on the self-organizing map[75] is used to cluster the conformations and to extract representative ones. The distributions of the number of representative conformations (Figure 3f) are centered on the 0-100 and 0-50 range for the widths of 20° and 40°. Unsurprisingly, in the case of the larger width 60°, much larger numbers of representative conformations can be obtained, up to 250. The

average number of representative conformations extracted from the SOM clustering of an iBP run on peptide fragment, is of the order of 10^2, which makes the number of maximum combinations of peptides during the step of fragments assembly to be about 10^4, and permits to overcome the combinatorial explosion, as it will be shown in the following.

Efficiency of the TAiBP assembly strategy

Starting from iBP results, the individual peptide conformations were superimposed on the backbone atoms of their last and initial five residues, in order to grow the protein structure incrementally from the N terminal to the C terminal extremity. The proposed fragment assembly is then conserved or pruned according to two successively applied criteria: (i) the clashing criterion tests whether Cα atoms of each fragment are farther apart from a given threshold (1 Å), (ii) the pruning distance criterion tests whether distance between the central Cα of all inserted peptides is within 5Å of the distances observed in the initial PDB structure. Several assembly strategies have been used: (a) the fragments are added one by one from the N terminal to the C terminal extremities of the protein, (b) all possible assemblies of two fragments are formed along the sequence, and then assembled together successively from N to C terminal, (c) all possible assemblies of three fragments are formed along the sequence, and then assembled together successively from N to C terminal. Depending on the protein target, one approach can be more efficient than the others, but no general trend of efficiency for one strategy was found during the analysis, so the results of the three strategies are presented together.

Some statistics on the assembly steps of TAiBP are presented in Figure 4. The numbers, plotted in decimal logarithmic scale, of distance pruning events (#DistPruning), of clash pruning events (#ClashPruning) and of processed conformations (#Processed) are plotted along the number of peptide residues (Figures 4a-c). The numbers of distances (Figure 4a) and clash pruning (Figure 4b) events are mostly in the range 10^2-10^4 for all fragment sizes and interval of backbone restraints. In addition, these numbers increase of more than one

order of value when the fragment size changes from 25 residues (assembly of two initial iBP fragments) to larger values. Larger pruning experienced in the case of larger fragments is probably induced by the excluded volume effect arising from the construction of the protein fold.

For fragment sizes larger than 25 residues, the numbers of pruning events (#DistPruning and #ClashPruning) are mostly around 10^3-10^5 (Figures 4a,b), in a range similar to the number of processed (#Processed) conformations (Figure 4c) which proves a large efficiency of pruning events for reducing the ensemble of solutions. Interestingly, distance pruning (Figure 4a) and clash pruning (Figure 4b) events are in similar range, displaying thus similar efficiency to reduce the number of solutions.

In the case of widths of 40 and 60° (magenta and green curves), the shifts observed for the parameters #DistPruning and #ClashPruning when increasing the interval widths on backbone angles, are steeper for the distance pruning events (Figure 4a) than for the clash pruning events (Figure 4b), whereas similar shifts are obtained for the interval width of 20° (blue dots). The widening of intervals has thus a stronger effect on the distance restraints between the peptide fragments than on the clash level. Finally, for fragments larger than 50 residues, the assembled fragments vanish except for the smaller intervals of 20° (blue dots), due to a pruning of all solutions. This pruning is the consequence of the discrepancy between non-uniform covalent geometry observed in the PDB structures and of the hypothesis of uniform covalent geometry made in the frame of iBP calculations.

The two by two comparisons of the events of distance and clash pruning, and of the number of processed conformations reveal the following trends (Figure 4d-f). The numbers of pruning events by clashes or by distances do not display any correlation (Figure 4d). At the contrary, #ClashPruning displays a quite strong correlation with #Processed (Figure 4e), specially for the largest angle interval (60°: green points). A similar tendency is observed for #DistPruning with two superimposed behaviors (Figure 4f): a correlation similar to the one observed for #ClashPruning, and other points with relatively smaller numbers of distance

pruning events. This second set of points corresponds mostly to the case of the fragments of 25 residues. Indeed, as these fragments are much smaller than the full protein, they have less chance to be rejected by pruning distance information.

Each assembled fragment has been compared to the corresponding region in the top100[61] target structure. This comparison was performed using RMSD (Å) between coordinates of heavy backbone atoms (Figure 5a,b) and not the TM score.[80,81] Indeed, the statistical validation of TM score was performed on protein structures larger than 80 residues,[81] which do not correspond to the set of proteins studied here. For each fragment, maximum and minimum RMSD values are plotted in Figure 5a,b with respect to the fragment size.

For each assembly calculation, the RMSD values to the target conformation were calculated for all TAiBP conformations. The maximum and minimum values of the RMSD distribution were then analyzed. In the case of narrow intervals (20°: blue points) on backbone angle restraints, minimum RMSD values are mostly smaller than 3.0 Å for all fragments up to 65 residues (Figure 5a). The increase of width in backbone angle intervals induces a drift of RMSD toward larger values: the RMSD drift is limited to 2-4 Å up to 35 residues, but jumps up to 5-6 Å for larger fragments. The threading-augmented iBP procedure proposed here thus allows one to obtain fragment conformations close to the PDB conformations for fragment sizes smaller than 65 residues. In that case, the Φ_K and Ψ_K drifts previously described (Eqs. 5-6 and Figure 2) have thus been overcome.

The maximum RMSD values (Figure 5b) are located in the 5-20 Å range. These maximum values were put in perspective with a previous analysis[82] in which protein structures were compared to a representative set of protein-like alternative structures generated using threading. Most of the RMSD R values for an N-residue protein fall in the interval: $3.333N^{1/3} - 2.0 \leq R \leq 3.333N^{1/3} + 2.0$, producing distributions of values smaller than 20 Å. This upper limit of 20 Å is similar to the one observed in the present calculation, which means that the TAiBP approach was able to mostly span the possible range of RMSD values.

At the end of TAiBP calculation, the poly-Ala sequence was replaced by the protein spe-

cific sequence and the residue sidechains have been added using the relax tool of the Rosetta suite.[54] The relax protocol consists of five cycles with rotamer repacking and minimization with progressively higher repulsive contributions within each cycle.[83] The obtained conformations have been analyzed (Table 3) and compared (Figure 6) to the conformation of the protein present in the database top100. From the 24 top100 structures initially processed, the 29 calculations realized on the 7 proteins smaller than 50 residues (Figure 6) display conformations calculated with TAiBP close to top100 conformations. Indeed, 18 runs display RMSD to initial top100 conformation smaller than 3 Å and 23 runs display RMSD smaller than 3.7 Å. Negative Rosetta total scores calculated according to Alford et al.[84] were obtained for all calculations, except the calculations on 1mctIH with an interval width of 60°.

For a given protein, the origin of ϕ and ψ restraints: angles or distances, introduced in the subsection "Probing the hypothesis of uniform covalent geometry" display various influences (Table 3) on the coordinate RMSD between the TAiBP and top100 conformations. For 1ben1BH and 1bpiH, the RMSD is mostly smaller if the ϕ and ψ target values were extracted from the distances d(C^i,C^{i+1}) and d(N^i,N^{i+1}), assuming uniform covalent geometry. The entries 1fxdH, 1mctIH, 1cnrH and 1edmBH display an opposite trend.

The distribution of RMSD values, calculated on the whole sets of conformations obtained for a given TAiBP calculation (Figure 7) span values up to 10 Å. Due to the pruning events during the latest step of fragment assembly, this upper bound is smaller than the one observed for the maximum RMSD in Figure 5b. For the angles ϕ_{angl}, ψ_{angl} measured on the initial top100 conformation, the increase of interval width induces a drift of RMSD toward larger values (magenta, brown and orange curves). The pattern is less clear for the angles ϕ_{dist} and ψ_{dist} calculated from measured distances in top100 conformations: in that case, many RMSD curves (blue, green and cyan curves) are more or less superimposed whatever is the interval width. The different proteins display quite different RMSD distributions which some distributions quite centered to a narrow interval and other much wider. These contrasted

features arise from the various efficiencies of pruning long-range distances in the frame of different 3D protein topologies.

Influence of uniform covalent geometry

The covalent geometry of the 1benABH, 1cnrH, 1edmBH, 1fxdH, 1igdH, 1isuAH, 1mctlH, bio1rpoH and 1bpiH conformations obtained using TAiBP and then relaxed using Rosetta displays some characteristics (Figure 8) quite different with respect to the ones analyzed on the top100 conformations at the beginning of the present work (Figure 1). Similar trends in covalent geometry were observed in conformations obtained by TAiBP before adding sidechains and relaxing (data not shown). In the relaxed conformations, the distribution of covalent angles (Figure 8a) is much thinner, although some individual deviations are observed (Figure 8b). These large drifts are observed in the protein regions in which two neighboring peptide fragments were superimposed. The ω dihedral values (Figure 8c) are distributed around 180° and -180°, in a similar way than in the initial structures (data not shown). These variations of ω dihedral angles have been already observed from various analyses of X-ray crystallographic structures.[65–67]

The comparison of the Ramachandran plots between Figures 1 and 8 reveal two differences. First, the TAiBP Ramachandran plots display a narrower range of colors than the top100 ones in agreement with the thinner distribution of covalent angles. Second, the (ϕ,ψ) distributions are fuzzier in Figures 8d-f than in Figures 1c-e. The convergence toward a uniform covalent geometry is thus accompanied by a expansion of the regions sampled in the Ramachandran diagram. Such expansion of allowed regions has been also observed in a recent analysis of the Ramachandran diagram.[85] In agreement with covalent geometry close to uniformity, the plots of cumulative sum of the differences: Φ_K (Eq. 5) and Ψ_K (Eq. 6) display much smaller drifts on the TAiBP conformations with a large majority of the values ranging between -50° and 50° (data not shown).

The protein conformations generated by the TAiBP approach and relaxed with Rosetta

(Table 3 and Figure 6), have been used as target conformations for a new run of the TAiBP approach, in order to investigate whether, in the case of mostly uniform covalent geometry, different results could be obtained. In a way similar to the previous TAiBP run, the coordinate RMSD (Å) between the new target and TAiBP fragments display a drift toward larger values for increasing fragment size (data not shown). The minimal and maximal RMSD distributions calculated for the reconstructed full chains of protein targets (Figure 9) show that for all targets except 1fxdH, the distribution of minimal values are mostly in the 1-4 Å for the interval width of 20° (blue curves). For all targets except 1bpiH and on a lesser extend 1cnrH, the increase of interval width do not have strong impact on the minimal RMSD distribution (full lines). Using target conformations closer to the hypothesis of uniform covalent geometry thus reduces the impact of increased intervals for ϕ and ψ angles. The distributions of maximal RMSD values (dashed lines) displays more variability than minimum RMSD distributions. Unsurprisingly, these distributions shift toward larger values and/or become broader in the case of increased interval width for ϕ and ψ values.

Discussion-Conclusion

Two main conclusions can be derived from the present work. The most important one is the design of an approach allowing to systematically parse all conformations of a protein up to 100 residues, using low precision input restraints. The other outcome of the manuscript is that the variability of covalent local geometry is an essential parameter for building protein conformations. Until now, this aspect has been discussed only very little in the literature.[78]

A method has been described to generate the protein structure by systematically exploring all possible conformations of the protein. This method is based on a threading-augmented interval Branch-and-Prune (TAiBP) approach in which the interval Branch-and-Prune (iBP)[28,44,86] is first used to systematically explore the conformations of 15-residues peptides fragments of the protein, followed by the construction of protein structure by sys-

tematically assembling fragment's conformations, and by pruning conformations displaying atom steric clashes and violations of few long-range distance restraints.

This two steps approach along with clustering using self-organized maps[75] allows to overcome the combinatorial explosion arising from the exponential complexity of the iBP algorithm. The duration of a total calculation is of the order of tenths of hours. This could be even speed up by using compiled language in place of python scripts used for fragments assembly and clustering.

The calculations performed using the TAiBP were validated by detecting whether this approach provides at least one solution close to the target solution. This detection was performed using the coordinate RMSD between backbone atoms. For most of the proteins smaller than 50 residues and interval widths of 20° for backbone angles, solutions were obtained with RMSD values smaller than 4 Å. Larger protein sizes and/or larger interval widths induce drift in the obtained conformations, which usually conduct to a pruning of all conformations because the long-range distance restraints are no more verified.

The largest problem to which the proposed approach faces is the non-uniformity of covalent geometry among the PDB structures. This aspect of PDB structures looks very minor as it involves only few degrees variations among the covalent angles, but, as this non-uniformity induces biases in the direction of extended and loop parts of protein structures, it is obvious that it may have big consequences in fragment assembly. Also, as this non-uniformity is present since the first days of structural biology and is probably deeply related to amino-acid type and to the position in the Ramachandran diagram,[66,87–92] it is quite difficult to sort it out. Nevertheless, it should be noticed that attempts have been made[93] to explore the relationship between backbone conformations and covalent geometry.

On the other hand, the iBP approach on which we based the conformational sampling of protein fragments, was developed using the initial hypothesis of a uniform covalent geometry. This hypothesis permitted to set up an algorithm which displays good scaling properties in case of a sufficient number of exactly known inter-atomic distances.[28] Modifying the

algorithm to take into account possible variations in the covalent geometry would increase enormously its complexity. Nevertheless, the very recent development of residue-specific force fields[37–40] opens new avenues for taking into account these aspects.

The present work opens the way to the use of iBP approach for exploring systematically the conformational space of proteins, using geometric restraints analogous to those experimentally measured by NMR. The validation of this systematic exploration of protein conformational space deeply changes the perspectives on protein structure calculation.

Acknowledgments

Thérèse Malliavin thanks the Pasteur Foundation for postdoctoral support of Dr Bradley Worley and thanks Dr Bradley Worley for its implementation of iBP and Dr Guillaume Bouvier for his support in python scripting. The authors wish to thank Institut Pasteur, CNRS, Ecole Polytechnique, FAPESP, CNPq and the program Infinity of CNRS for financial support.

References

(1) Nilges, M.; Clore, G. M.; Gronenborn, A. M. Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms. Circumventing problems associated with folding. *FEBS Lett.* **1988**, *239*, 129–136.

(2) Pearlman, D. A. FINGAR: A new genetic algorithm-based method for fitting NMR data. *J. Biomol. NMR* **1996**, *8*, 49–66.

(3) Guerry, P.; Herrmann, T. Comprehensive automation for NMR structure determination of proteins. *Methods Mol. Biol.* **2012**, *831*, 429–451.

(4) Huang, Y. J.; Mao, B.; Xu, F.; Montelione, G. T. Guiding automated NMR structure determination using a global optimization metric, the NMR DP score. *J. Biomol. NMR* **2015**, *62*, 439–451.

(5) Mareuil, F.; Malliavin, T. E.; Nilges, M.; Bardiaux, B. Improved reliability, accuracy and quality in automated NMR structure calculation with ARIA. *J. Biomol. NMR* **2015**, *62*, 425–438.

(6) Wurz, J. M.; Kazemi, S.; Schmidt, E.; Bagaria, A.; Guntert, P. NMR-based automated protein structure determination. *Arch. Biochem. Biophys.* **2017**, *628*, 24–32.

(7) Schwieters, C. D.; Bermejo, G. A.; Clore, G. M. Xplor-NIH for molecular structure determination from NMR and other data sources. *Protein Sci.* **2018**, *27*, 26–40.

(8) Rieping, W.; Habeck, M.; Nilges, M. Inferential structure determination. *Science* **2005**, *309*, 303–306.

(9) Habeck, M.; Nilges, M.; Rieping, W. Bayesian inference applied to macromolecular structure determination. *Phys Rev E Stat Nonlin Soft Matter Phys* **2005**, *72*, 031912.

(10) Habeck, M.; Nilges, M.; Rieping, W. Replica-exchange Monte Carlo scheme for bayesian data analysis. *Phys. Rev. Lett.* **2005**, *94*, 018105.

(11) Rieping, W.; Habeck, M.; Nilges, M. Modeling errors in NOE data with a log-normal distribution improves the quality of NMR structures. *J. Am. Chem. Soc.* **2005**, *127*, 16026–16027.

(12) Habeck, M.; Rieping, W.; Nilges, M. Weighting of experimental evidence in macromolecular structure determination. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 1756–1761.

(13) Nilges, M.; Bernard, A.; Bardiaux, B.; Malliavin, T.; Habeck, M.; Rieping, W. Accurate NMR structures through minimization of an extended hybrid energy. *Structure* **2008**, *16*, 1305–1312.

(14) Bernard, A.; Vranken, W. F.; Bardiaux, B.; Nilges, M.; Malliavin, T. E. Bayesian estimation of NMR restraint potential and weight: a validation on a representative set of protein structures. *Proteins* **2011**, *79*, 1525–1537.

(15) Rosato, A; Aramini, J. M.; Arrowsmith, C.; Bagaria, A.; Baker, D.; Cavalli, A.; Doreleijers, J. F.; Eletsky, A.; Giachetti, A.; Guerry, P. et al. Blind testing of routine, fully automated determination of protein structures from NMR data. *Structure* **2012**, *20*, 227–236.

(16) Peng, J.; Yuan, C.; Ma, R.; Zhang, Z. Backmapping from Multiresolution Coarse-Grained Models to Atomic Structures of Large Biomolecules by Restrained Molecular Dynamics Simulations Using Bayesian Inference. *J Chem Theory Comput* **2019**, *15*, 3344–3353.

(17) Habeck, M.; Rieping, W.; Linge, J. P.; Nilges, M. NOE assignment with ARIA 2.0: the nuts and bolts. *Methods Mol. Biol.* **2004**, *278*, 379–402.

(18) Linge, J. P.; Habeck, M.; Rieping, W.; Nilges, M. ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* **2003**, *19*, 315–316.

(19) Mohan, A.; Oldfield, C. J.; Radivojac, P.; Vacic, V.; Cortese, M. S.; Dunker, A. K.; Uversky, V. N. Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* **2006**, *362*, 1043–1059.

(20) Cumberworth, A.; Lamour, G.; Babu, M. M.; Gsponer, J. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochem. J.* **2013**, *454*, 361–369.

(21) Schneider, R.; Blackledge, M.; Jensen, M. R. Elucidating binding mechanisms and dynamics of intrinsically disordered protein complexes using NMR spectroscopy. *Curr. Opin. Struct. Biol.* **2018**, *54*, 10–18.

(22) Ozenne, V.; Bauer, F.; Salmon, L.; Huang, J. R.; Jensen, M. R.; 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*, 1463–1470.

(23) Krzeminski, M.; Marsh, J. A.; Neale, C.; Choy, W. Y.; Forman-Kay, J. D. Characterization of disordered proteins with ENSEMBLE. *Bioinformatics* **2013**, *29*, 398–399.

(24) Estana, A.; Sibille, N.; Delaforge, E.; Vaisset, M.; Cortes, J.; Bernado, P. Realistic Ensemble Models of Intrinsically Disordered Proteins Using a Structure-Encoding Coil Database. *Structure* **2019**, *27*, 381–391.

(25) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129–1143.

(26) Havel, T. F. An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Prog. Biophys. Mol. Biol.* **1991**, *56*, 43–78.

(27) Liberti, L.; Lavor, C.; Maculan, N.; Mucherino, A. Euclidean Distance Geometry and Applications. *SIAM Rev* **2014**, *56*, 3–69.

(28) Lavor, C.; Liberti, L.; Mucherino, A. The interval Branch-and-Prune algorithm for the discretizable molecular distance geometry problem with inexact distances. *J Glob Optim* **2013**, *56*, 855–871.

(29) Lavor, C.; Lee, J.; Lee-St.John, A.; Liberti, L.; Mucherino, A.; Sviridenko, M. Discretization orders for distance geometry problems. *Optimization Letters* **2012**, *6*, 783–796.

(30) Lavor, C.; Liberti, L.; Donald, B.; Worley, B.; Bardiaux, B.; Nilges, M. Minimal NMR

distance information for rigidity of protein graphs. *Discrete Applied Mathematics* **2019**, *256*, 91–104.

(31) Braun, W.; Go, N. Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. *J. Mol. Biol.* **1985**, *186*, 611–626.

(32) Abagyan, R.; Totrov, M.; Kuznetsov, D. CMA new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J Comput Chem* **1994**, *15*, 488–506.

(33) Guntert, P. Automated NMR structure calculation with CYANA. *Methods Mol. Biol.* **2004**, *278*, 353–378.

(34) Bardiaux, B.; Malliavin, T. E.; Nilges, M.; Mazur, A. K. Comparison of different torsion angle approaches for NMR structure determination. *J. Biomol. NMR* **2006**, *34*, 153–166.

(35) Berkholz, D. S.; Shapovalov, M. V.; Dunbrack, R. L.; Karplus, P. A. Conformation dependence of backbone geometry in proteins. *Structure* **2009**, *17*, 1316–1325.

(36) Moriarty, N. W.; Tronrud, D. E.; Adams, P. D.; Karplus, P. A. A new default restraint library for the protein backbone in Phenix: a conformation-dependent geometry goes mainstream. *Acta Crystallogr D Struct Biol* **2016**, *72*, 176–179.

(37) Zhou, C.; Jiang, F.; Wu, Y. Residue-specific force field based on protein coil library. RSFF2: modification of AMBER ff99SB. *J Phys Chem B* **2015**, *119*, 1035–1047.

(38) Kang, W.; Jiang, F.; Wu, Y. Universal Implementation of a Residue-Specific Force Field Based on CMAP Potentials and Free Energy Decomposition. *J Chem Theory Comput* **2018**, *14*, 4474–4486.

(39) Liu, H.; Song, D.; Lu, H.; Luo, R. Intrinsically disordered protein-specific force field CHARMM36IDPSFF. *Chem Biol Drug Des* **2018**, *92*, 1722–1735.

(40) Choi, J.; Pappu, R. Improvements to the ABSINTH Force Field for Proteins Based on Experimentally Derived Amino Acid Specific Backbone Conformational Statistics. *J Chem Theory Comput* **2019**, *15*, 1367–1382.

(41) Liberti, L.; Lavor, C.; Mucherino, A. The discretizable molecular distance geometry problem seems easier on proteins. *Distance Geometry: Theory, Methods and Applications. Mucherino, Lavor, Liberti, Maculan (eds.)* **2014**, 47–60.

(42) Cassioli, A.; Bardiaux, B.; Bouvier, G.; Mucherino, A.; Alves, R.; Liberti, L.; Nilges, M.; Lavor, C.; Malliavin, T. An algorithm to enumerate all possible protein conformations verifying a set of distance constraints. *BMC Bioinformatics* **2015**, *16*, 23–37.

(43) Worley, B.; Delhommel, F.; Cordier, F.; Malliavin, T.; Bardiaux, B.; Wolff, N.; Nilges, M.; Lavor, C.; Liberti, L. Tuning interval Branch-and-Prune for protein structure determination. *Journal of Global Optimization* **2018**, *72*, 109–127.

(44) Lavor, C.; Alves, R.; Figueiredo, W.; Petraglia, A.; Maculan, N. Clifford Algebra and the Discretizable Molecular Distance Geometry Problem. *J Glob Optim* **2015**, *25*, 925–942.

(45) Shen, Y.; Bax, A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J. Biomol. NMR* **2013**, *56*, 227–241.

(46) Kohlhoff, K. J.; Robustelli, P.; Cavalli, A.; Salvatella, X.; Vendruscolo, M. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J. Am. Chem. Soc.* **2009**, *131*, 13894–13895.

(47) Shen, Y.; Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **2010**, *48*, 13–22.

(48) Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR* **2011**, *50*, 43–57.

(49) Baker, D.; Sali, A. Protein structure prediction and structural genomics. *Science* **2001**, *294*, 93–96.

(50) Mucherino, A.; Lavor, C.; Liberti, L.; Talbi, E.-G. A Parallel Version of the Branch & Prune Algorithm for the Molecular Distance Geometry Problem. *IEEE Conference Proceedings, ACS/IEEE International Conference on Computer Systems and Applications (AICCSA10)* **2010**, 1–6.

(51) Fidalgo, F.; Gonçalves, D.; Lavor, C.; Liberti, L.; Mucherino, A. A Symmetry-based Splitting Strategy for Discretizable Distance Geometry Problems. *Journal of Global Optimization* **2018**, *71*, 717–733.

(52) Das, R.; Baker, D. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **2008**, *77*, 363–382.

(53) Ovchinnikov, S.; Park, H.; Kim, D. E.; Liu, Y.; Wang, R. Y.; Baker, D. Structure prediction using sparse simulated NOE restraints with Rosetta in CASP11. *Proteins* **2016**, *84 Suppl 1*, 181–188.

(54) Ovchinnikov, S.; Park, H.; Kim, D. E.; DiMaio, F.; Baker, D. Protein structure prediction using Rosetta in CASP12. *Proteins* **2018**, *86 Suppl 1*, 113–121.

(55) Alipanahi, B.; Krislock, N.; Ghodsi, A.; Wolkowicz, H.; Donaldson, L.; Li, M. Determining protein structures from NOESY distance constraints by semidefinite programming. *J. Comput. Biol.* **2013**, *20*, 296–310.

(56) Carr, J. M.; Whittleston, C. S.; Wade, D. C.; Wales, D. J. Energy landscapes of a hairpin peptide including NMR chemical shift restraints. *Phys Chem Chem Phys* **2015**, *17*, 20250–20258.

(57) Joo, K.; Joung, I.; Lee, J.; Lee, J.; Lee, W.; Brooks, B.; Lee, S. J.; Lee, J. Protein struc-

ture determination by conformational space annealing using NMR geometric restraints. *Proteins* **2015**, *83*, 2251–2262.

(58) Khoo, Y.; Singer, A.; Cowburn, D. Integrating NOE and RDC using sum-of-squares relaxation for protein structure determination. *J. Biomol. NMR* **2017**, *68*, 163–185.

(59) Hong, S. H.; Joung, I.; Flores-Canales, J. C.; Manavalan, B.; Cheng, Q.; Heo, S.; Kim, J. Y.; Lee, S. Y.; Nam, M.; Joo, K.; Lee, I. H.; Lee, S. J.; Lee, J. Protein structure modeling and refinement by global optimization in CASP12. *Proteins* **2018**, *86 Suppl 1*, 122–135.

(60) Martin, J. W.; Zhou, P.; Donald, B. R. Systematic solution to homo-oligomeric structures determined by NMR. *Proteins* **2015**, *83*, 651–661.

(61) Word, J.; Lowell, S.; LaBean, T.; Taylor, H.; Zalis, M.; Presley, B.; Richardson, J.; Richardson, D. Visualizing and Quantifying Molecular Goodness-of-Fit: Small-probe Contact Dots with Explicit Hydrogen Atoms. *J Mol Biol* **1999**, *285*, 1713–1733.

(62) Engh, R.; Huber, R. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr A* **1991**, *47*, 392–400.

(63) Riniker, S. Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *J Chem Inf Model* **2018**, *58*, 565–578.

(64) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38,27–28.

(65) Matthews, B. W. How planar are planar peptide bonds? *Protein Sci.* **2016**, *25*, 776–777.

(66) Brereton, A. E.; Karplus, P. A. On the reliability of peptide nonplanarity seen in ultra-high resolution crystal structures. *Protein Sci.* **2016**, *25*, 926–932.

(67) Berkholz, D. S.; Driggers, C. M.; Shapovalov, M. V.; Dunbrack, R. L.; Karplus, P. A. Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 449–453.

(68) Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

(69) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **2011**, *32*, 2319–2327.

(70) Richard J. Gowers,; Max Linke,; Jonathan Barnoud,; Tyler J. E. Reddy,; Manuel N. Melo,; Sean L. Seyler,; Jan Domaski,; David L. Dotson,; Sbastien Buchoux,; Ian M. Kenney,; Oliver Beckstein, MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. Proceedings of the 15th Python in Science Conference. 2016; pp 98–105.

(71) Oliphant TE, *A guide to NumPy*; Trelgol Publishing, USA, 2006.

(72) Kohonen, T. Self-Organized formation of topologically correct feature maps. *Biol Cybern* **1982**, *43*, 59–69.

(73) Kohonen, T. Self-Organizing Maps. *Springer Series in Information Sciences, Heidelberg, Germany.* **2001**,

(74) Miri, L.; Bouvier, G.; Kettani, A.; Mikou, A.; Wakrim, L.; Nilges, M.; Malliavin, T. E. Stabilization of the integrase-DNA complex by Mg2+ ions and prediction of key residues for binding HIV-1 integrase inhibitors. *Proteins* **2014**, *82*, 466–478.

(75) Bouvier, G.; Duclert-Savatier, N.; Desdouits, N.; Meziane-Cherif, D.; Blondel, A.; Courvalin, P.; Nilges, M.; Malliavin, T. Functional motions modulating VanA ligand binding unraveled by self-organizing maps. *J Chem Inf Model* **2014**, *54*, 289–301.

(76) Bouvier, G.; Desdouits, N.; Ferber, M.; Blondel, A.; Nilges, M. An automatic tool to analyze and cluster macromolecular conformations based on Self-Organizing Maps. *Bioinformatics* **2014**, *31*, 1–3.

(77) Spill, Y.; Bouvier, G.; Nilges, M. A convective replica-exchange method for sampling new energy basins. *J Comput Chem* **2013**, *34*, 132–140.

(78) Hinsen, K.; Hu, S.; Kneller, G. R.; Niemi, A. J. A comparison of reduced coordinate sets for describing protein structure. *J Chem Phys* **2013**, *139*, 124115.

(79) Kuzmanicâ, A.; Zagrovicâiâ, B. Determination of Ensemble-Average Pairwise Root Mean-Square Deviation from Experimental B-Factors. *Biophys J* **2010**, *98*, 861–871.

(80) Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **2004**, *57*, 702–710.

(81) Xu, J.; Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **2010**, *26*, 889–895.

(82) Reva, B. A.; Finkelstein, A. V.; Skolnick, J. What is the probability of a chance prediction of a protein structure with an rmsd of 6 A? *Fold Des* **1998**, *3*, 141–147.

(83) Tyka, M. D.; Keedy, D. A.; Andre, I.; Dimaio, F.; Song, Y.; Richardson, D. C.; Richardson, J. S.; Baker, D. Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **2011**, *405*, 607–618.

(84) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* **2017**, *13*, 3031–3048.

(85) Hollingsworth, S. A.; Karplus, P. A. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol Concepts* **2010**, *1*, 271–283.

(86) Lavor, C.; Liberti, L.; Maculan, N.; Mucherino, A. The Discretizable Molecular Distance Geometry Problem. *Computational Optimization and Applications* **2012**, *52*, 115–146.

(87) Hollingsworth, S. A.; Lewis, M. C.; Karplus, P. A. Beyond basins: ϕ,ψ preferences of a residue depend heavily on the ϕ,ψ values of its neighbors. *Protein Sci.* **2016**, *25*, 1757–1762.

(88) Moriarty, N. W.; Tronrud, D. E.; Adams, P. D.; Karplus, P. A. A new default restraint library for the protein backbone in Phenix: a conformation-dependent geometry goes mainstream. *Acta Crystallogr D Struct Biol* **2016**, *72*, 176–179.

(89) Moriarty, N. W.; Tronrud, D. E.; Adams, P. D.; Karplus, P. A. Conformation-dependent backbone geometry restraints set a new standard for protein crystallographic refinement. *FEBS J.* **2014**, *281*, 4061–4071.

(90) Hollingsworth, S. A.; Lewis, M. C.; Berkholz, D. S.; Wong, W. K.; Karplus, P. A. $(\phi,\psi)_2$ motifs: a purely conformation-based fine-grained enumeration of protein parts at the two-residue level. *J. Mol. Biol.* **2012**, *416*, 78–93.

(91) Tronrud, D. E.; Karplus, P. A. A conformation-dependent stereochemical library improves crystallographic refinement even at atomic resolution. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 699–706.

(92) Berkholz, D. S.; Shapovalov, M. V.; Dunbrack, R. L.; Karplus, P. A. Conformation dependence of backbone geometry in proteins. *Structure* **2009**, *17*, 1316–1325.

(93) Berkholz, D. S.; Krenesky, P. B.; Davidson, J. R.; Karplus, P. A. Protein Geometry Database: a flexible engine to explore backbone conformations and their relationships to covalent geometry. *Nucleic Acids Res.* **2010**, *38*, D320–325.

Table 1

Table 1: Geometric parameters for covalent and improper bonds and angles taken from the force field PARALLHDG (version 5.3).[62]

Atom name	Definition of atom type	Bond atoms	Bond length (Å)
CH1E	α carbon	C-CH1E	1.525
C	carbonyl carbon	CH1E-HA	1.080
O	carbonyl oxygen	CH1E-NH1	1.458
OC	C-terminal carbonyl oxygen	CH1E-NH2	1.486
HA	Hα hydrogen	C-NH1	1.329
NH1	amide nitrogen	C-O	1.231
NH2	N-terminal amide nitrogen	C-OC	1.249
		H-NH1	0.980
H	amide hydrogen	H-NH2	0.980
Bond angle atoms	**Bond angle value (°)**	**Improper angle atoms**	**Improper angle value (°)**
C-CH1E-HA	108.9914	C-CH1E-HA-HA	-70.4072
C-CH1E-NH1	111.1396	CH1E-C-NH1-HA	66.2535
C-CH1E-NH2	106.9610	C-NH1-HA-HA	-70.8745
C-NH1-CH1E	121.6541	HA-CH1E-HA-HA	-66.5692
CH1E-C-NH1	116.1998	CH1E-C-NH1-CH1E	178.0
CH1E-C-O	120.8258	O-C-NH1-H	178.0
CH1E-NH1-H	119.2367	C-CH1E-NH1-O	0.0
C-NH1-H	119.2489	NH1-CH1E-C-HA	119.0
C-NH2-H	118.1853	NH2-CH1E-C-HA	119.0
HA-CH1E-NH1	108.0508	NH1-HA-CH1E-C	121.0
H-NH2-CH1E	109.5000	NH2-HA-CH1E-C	116.0
H-NH2-H	107.3000	C-NH1-CH1E-H	180.0
NH1-C-O	122.9907	CH1E-C-NH1-O	180.0
NH2-CH1E-HA	108.4800	NH2-H-H-CH1E	41.0
NH2-C-O	122.6277	CH1E-OC-C-OC	178.0
CH1E-C-OC	118.0611		
OC-C-OC	123.3548		

Table 2

Table 2: Atom re-ordering used during the iBP calculation step within the first, the last and the inner residues of the peptide fragment. The order is described by the list of atoms names, the signs "-" and "+" describing atoms located in the previous and the next residues in the primary sequence.

Residue position	order
first	N, H1, H2, CA, N, HA, CA, C
inner	N, -O, -CA, -C, N, CA, C, +N, -C, N, CA, H1, N, CA, C, HA, C, CA
last	N, -O, -CA, -C, N, CA, C, -C, N, CA, H1, N, CA, C, HA, C, CA, O, C, O2

Table 3

Table 3: Analysis of the TAiBP conformations further relaxed using the Rosetta suite.[54] The Rosetta energy terms are defined from the ref2015 score function.[84] Origin of ϕ and ψ angles describes whether their target values have been directly measured from the initial top100 conformation (angles) or whether they have been determined from distances measured on the top100 conformation using the hypothesis of a uniform covalent geometry (distances).

Protein	origin ϕ,ψ	interval width on ϕ,ψ	Number of residues	Minimal RMSD value (Å)	Rosetta fa_atr (ref2015)	Rosetta total (ref2015)
1benABH	angles	20°	21	1.1	-55.0	-9.49
1benABH	distances	20°	21	1.0	-62.0	-12.8
1benABH	angles	40°	21	2.2	-59.6	-13.6
1benABH	distances	40°	21	3.1	-58.5	-7.0
1benABH	angles	60°	21	3.7	-50.0	-4.8
1benABH	distances	60°	21	3.0	-52.7	-4.7
1bpiH	angles	40°	55	6.5	-142.2	-16.0
1bpiH	distances	40°	55	5.0	-151.4	-22.4
1bpiH	distances	60°	55	4.7	-168.6	-16.3
bio1rpoH	angles	20°	61	1.4	-207.7	-70.6
1fxdH	angles	20°	55	1.7	-168.2	-39.0
1fxdH	distances	40°	55	5.5	-139.1	-27.8
1fxdH	angles	40°	55	4.5	-150.8	-26.6
1mctIH	angles	20°	25	0.7	-52.7	-6.8
1mctIH	distances	20°	25	2.9	-50.8	-1.9
1mctIH	angles	40°	25	0.8	-53.4	-7.6
1mctIH	distances	40°	25	2.6	-45.9	-1.9
1mctIH	angles	60°	25	2.7	-49.5	9.1
1mctIH	distances	60°	25	2.2	-53.4	0.4
1cnrH	angles	20°	45	2.0	-133.5	-37.6
1cnrH	distances	20°	45	2.0	-123.4	-30.8
1cnrH	angles	40°	45	3.4	-121.9	-24.8
1cnrH	distances	40°	45	5.0	-123.2	-23.6
1cnrH	distances	60°	45	3.4	-115.7	-23.7
1edmBH	angles	20°	35	1.2	-93.9	-25.9
1edmBH	distances	20°	35	2.0	-80.2	-16.6
1edmBH	angles	40°	35	2.4	-78.2	-19.4
1edmBH	distances	40°	35	3.3	-82.6	-18.1
1edmBH	distances	60°	35	2.7	-93.2	-10.6

Figures

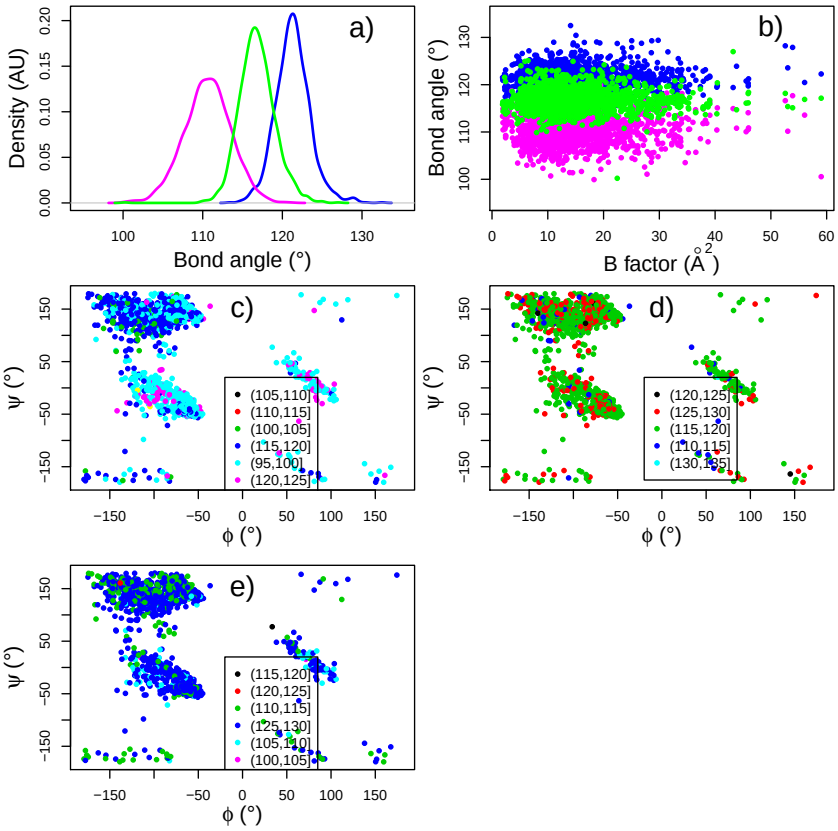

Figure 1: Variability of the covalent geometry within the 24 protein structures analyzed from the database top100.[61] (a) Distribution of bond angles (°) between atoms C-N-Cα (blue), N-Cα-C (magenta) and Cα-C-N (green). (b) Comparison between bond angles and B factors averaged on each residue. The color code is the same in (a) and (b). (c-e) Ramachandran plots of the protein structures colored according to the value of bond angles (°) between atoms C-N-Cα (c), N-Cα-C (d) and Cα-C-N (e). On each plot (c-e), the color scale is defined in the inset legend.

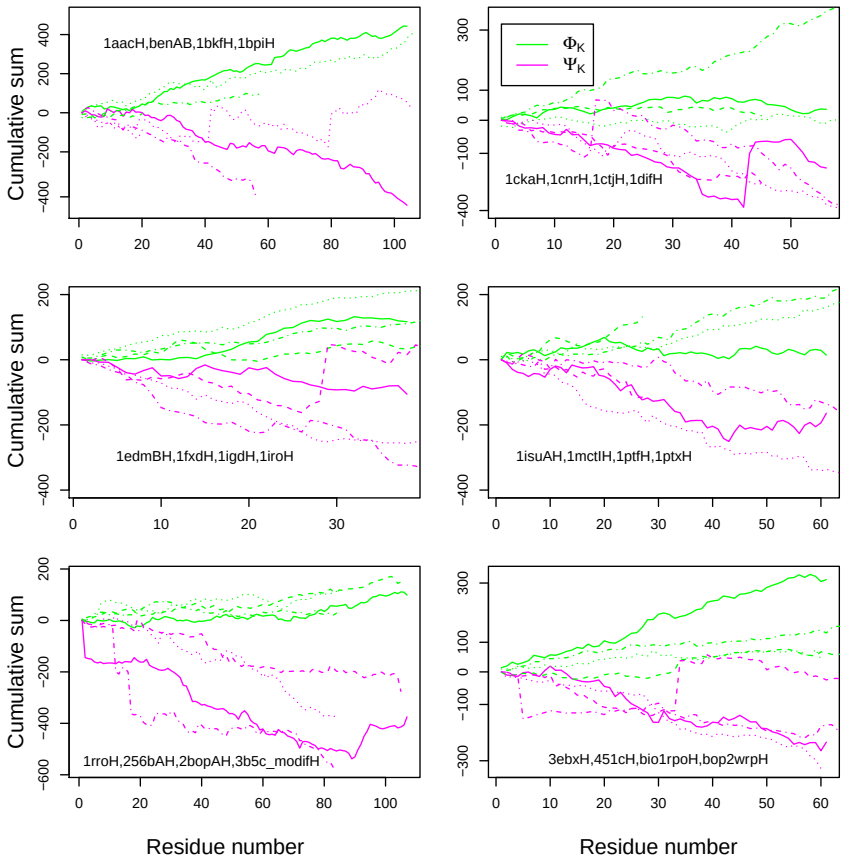

Figure 2: Variations of the Φ_K (green curves) and Ψ_K (magenta curves) (°) parameters (Eqs. 5 and 6) plotted along the primary sequences of target proteins. The entries of the analyzed proteins in the database top100[61] are given on each plot.

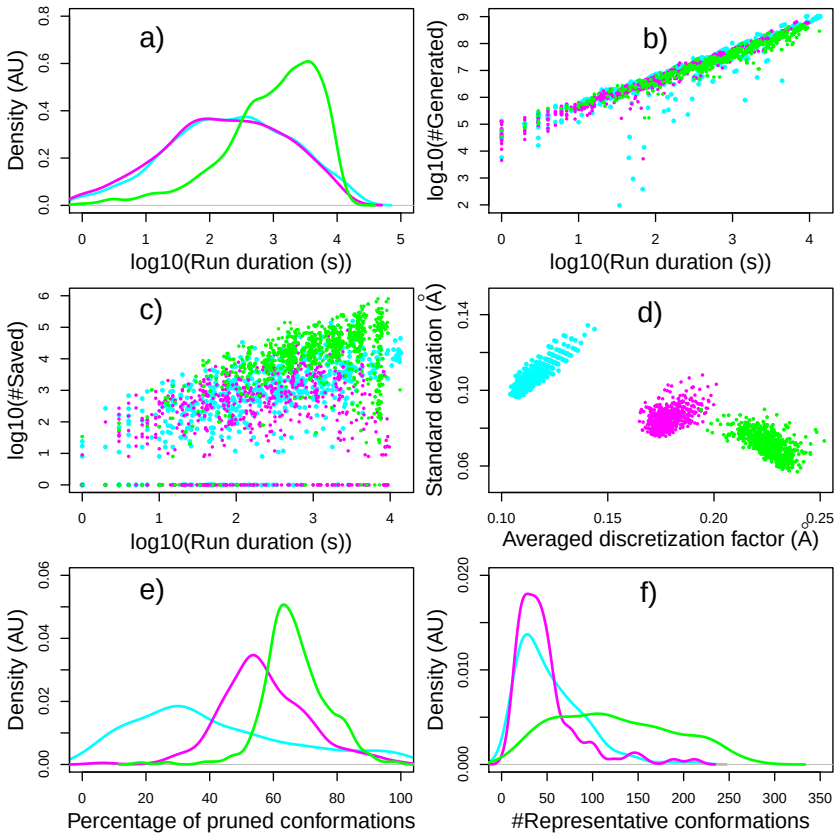

Figure 3: Analysis of the iBP runs on peptide fragments. The colors of lines/points correspond to the error interval on backbone angles: 20° (blue), 40° (magenta) and 60° (green). (a) Distribution of the run duration (s) plotted on logarithmic scale. Number of generated (b) and saved (c) conformations plotted along the run duration (s). Both axes are in decimal logarithmic scale. (d) Standard deviation (Å) of discretization factor along the average discretization factor (Å). (e) Distribution of the percentage of pruned conformations. (f) Distribution of the number of representative conformations obtained by clustering through self-organizing maps.

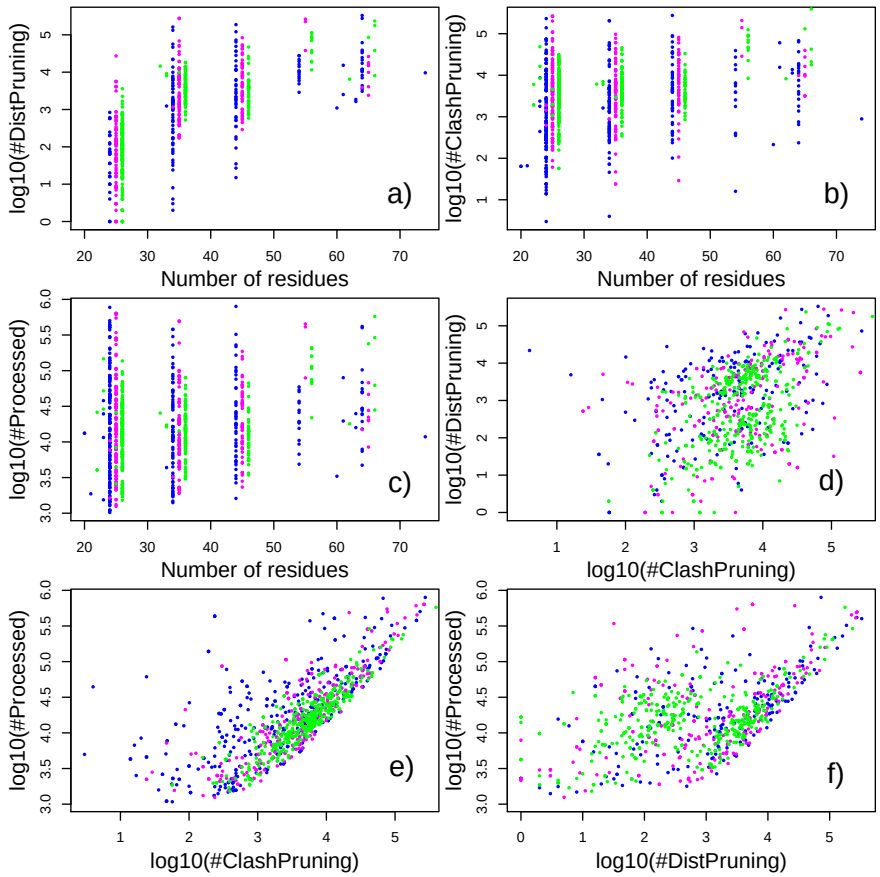

Figure 4: Analysis of the assembly of protein fragments to build the protein folds. All parameters except the number of residues in the assembled fragment are expressed in decimal logarithmic scale. The colors of points correspond to the error interval on backbone angles: 20° (blue), 40° (magenta) and 60° (green). The following plots are displayed: (a) number of distance pruning versus number of residues, (b) number of clash pruning versus number of residues, (c) number of processed conformations versus number of residues, (d) number of distance pruning versus number of clash pruning, (e) number of processed conformations versus number of clash pruning, (f) number of processed conformations versus number of distance pruning.

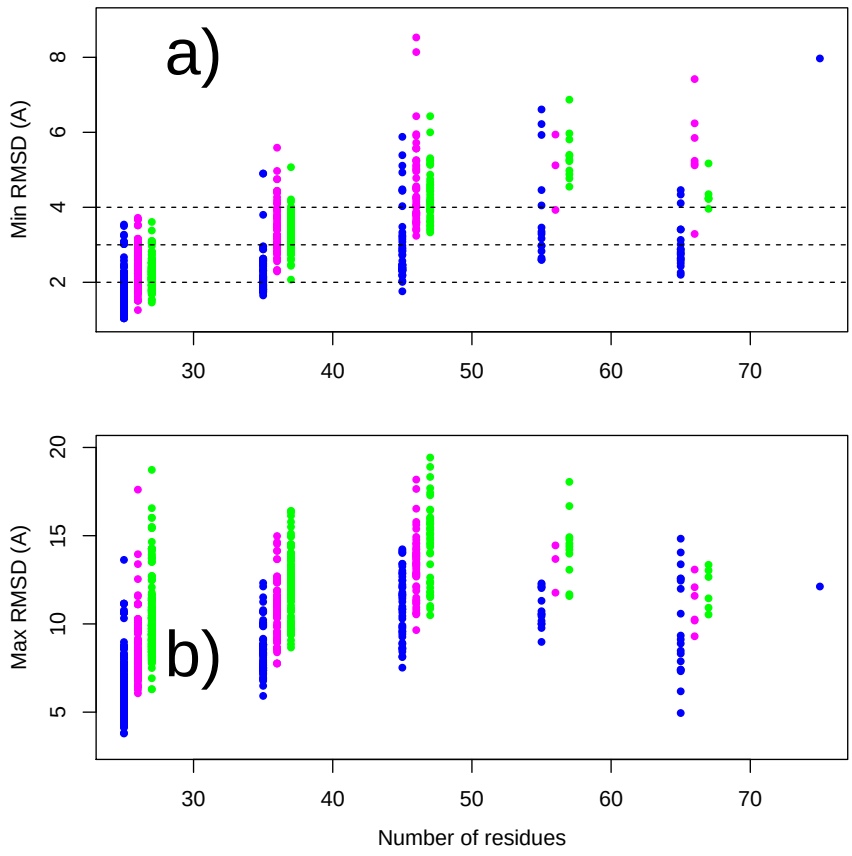

Figure 5: Comparison of the assembled fragment conformations to the corresponding regions in the top100 target structure. For each processed fragment, the minimum (a) and maximum (b) values of the coordinates RMSD (Å) between heavy backbone atoms are plotted along the number of residues. Dashed lines have been added at the levels of 2, 3 and 4 Å in minimum RMSD plot.

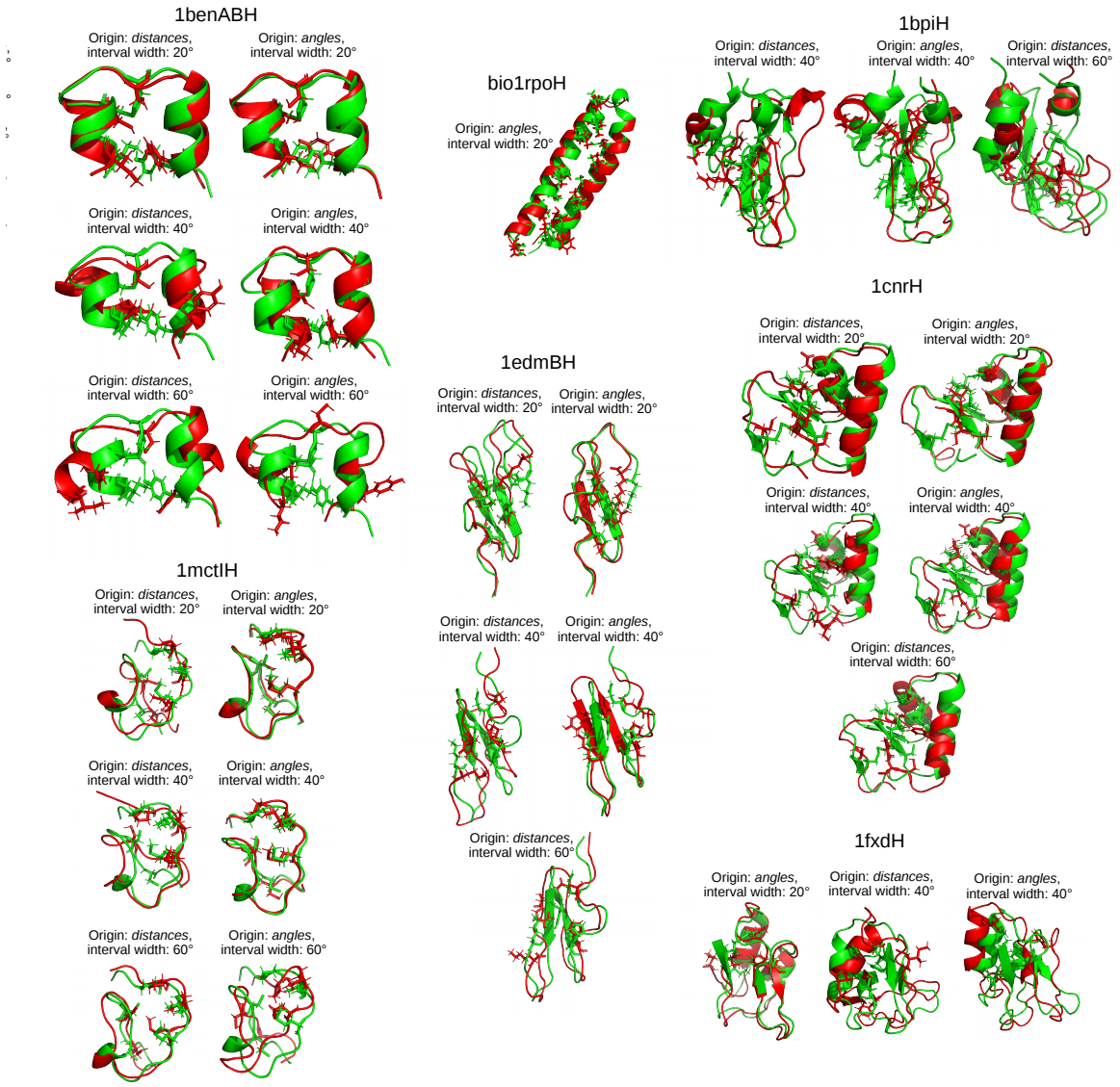

Figure 6: Comparison of the TAiBP conformations to the initial top100 conformations. The TAiBP conformations processed using the relax tool from Rosetta[54] are drawn in red cartoon and superimposed to the target top100 conformations, drawn in green cartoon. Some residues of the protein core are drawn in licorice. Close to each superimposed structures, the origin of the backbone angles restraints ϕ, ψ is given as *distances* or *angles*. If the angles ϕ, ψ are measured on the initial top100 conformation, the restraints are of *angles* origin, whereas if the angles are calculated from measured distances in top100 conformations, assuming a uniform covalent geometry, the restraints are of *distances* origin. The interval widths of ϕ, ψ restraints are also given.

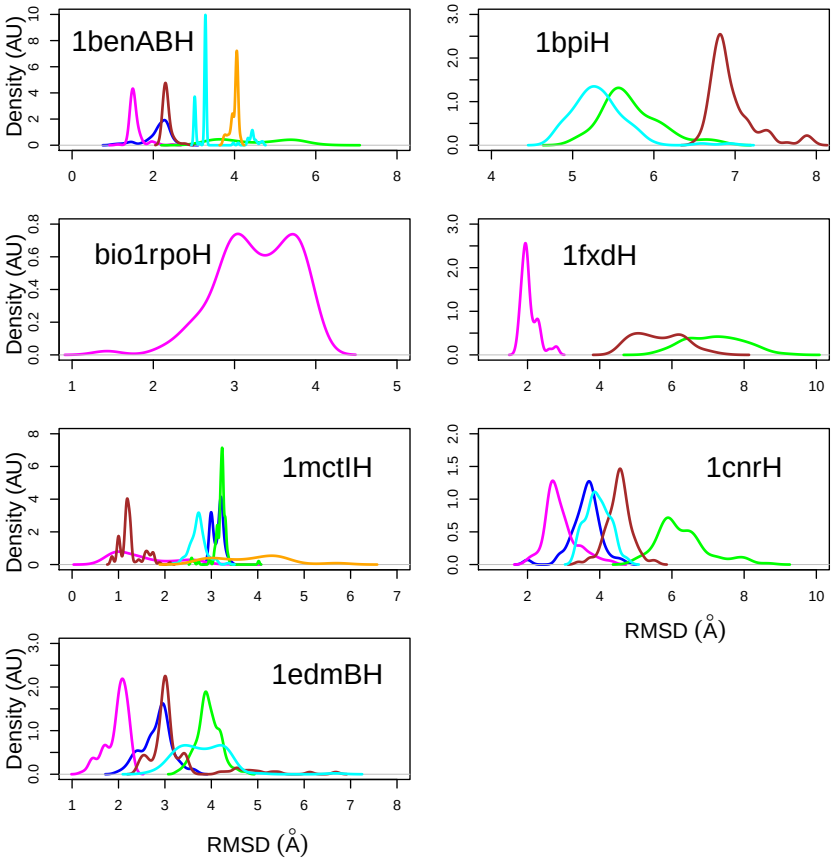

Figure 7: Distribution of RMSD values (Å) of each conformation obtained at the end of TaiBP approach with respect to the conformation of the corresponding protein in the database top100.[61] The curves are colored according to the origin of the backbone angle restraints (*distances* or *angle*) introduced in the subsection "Input values for the calculation" in Materials and Methods, and to the interval of ϕ, ψ restraints. The color used are: blue (*distances*, 20°), magenta (*angles*, 20°), green (*distances*, 30°), brown (*angles*, 30°), cyan (*distances*, 40°), orange (*angles*, 40°).

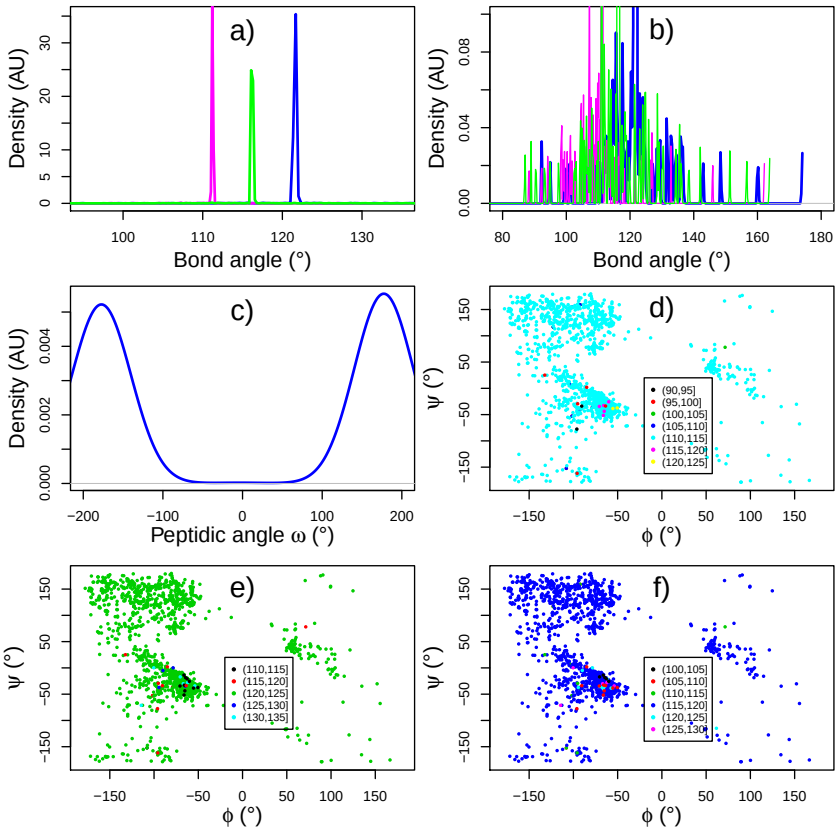

Figure 8: Variability of the covalent geometry within the set of the conformations obtained on 1benABH, 1cnrH, 1edmBH, 1fxdH, 1igdH, 1isuAH, 1mctlH, bio1rpoH, 1bpiH using TAiBP and then relaxed using Rosetta. (a) Distribution of bond angles (°) between atoms C-N-Cα (blue), N-Cα-C (magenta) and Cα-C-N (green). (b) Distribution of bond angles (°) between atoms C-N-Cα (blue), N-Cα-C (magenta) and Cα-C-N (green) with increased scale to detect the outlier peaks, (c) Distribution of the ω dihedral angle (°). (d-f) Ramachandran plots of the protein conformations colored according to the value of bond angles (°) between atoms C-N-Cα (d), N-Cα-C (e) and Cα-C-N (f). On each plot, the color scale is defined in the inset legend.

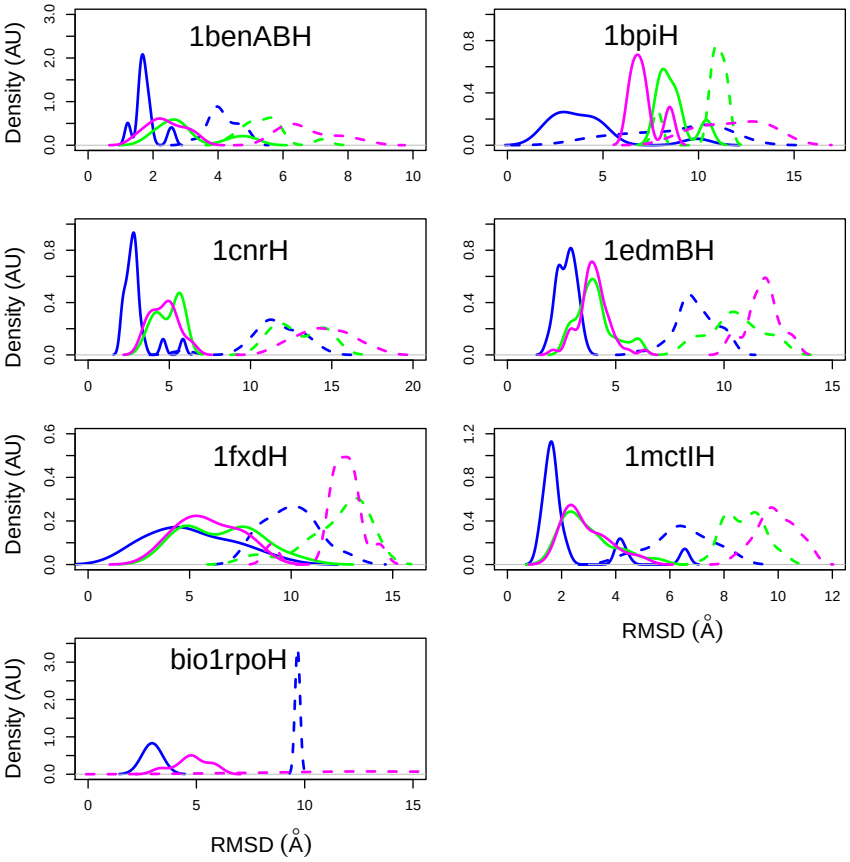

Figure 9: Distribution of minimum (full line) and maximum (dotted line) RMSD (Å) values between the target and the TAiBP protein conformations. These RMSD were obtained during a TAiBP run using as targets protein conformations generated by the TAiBP approach and relaxed with Rosetta (Table 3 and Figure 6). The calculations were realized with interval of 20° (blue), 40° (green), and 60° (magenta) widths on the ϕ and ψ values.