Journal of Molecular Graphics and Modelling 55 (2015) 72–84

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI

Yuan Zhang[a,b], Celeste Sagui[a,b,*]

[a] Department of Physics, North Carolina State University, Raleigh, NC 27695, United States
[b] Center for High Performance Simulations (CHiPS), North Carolina State University, Raleigh, NC 27695, United States

ARTICLE INFO

Article history:
Accepted 8 October 2014
Available online 6 November 2014

Keywords:
Residual transient secondary structure
Intrinsically disordered proteins
gp41
Asparagine

ABSTRACT

Secondary structure assignment codes were built to explore the regularities associated with the periodic motifs of proteins, such as those in backbone dihedral angles or in hydrogen bonds between backbone atoms. Precise structure assignment is challenging because real-life secondary structures are susceptible to bending, twist, fraying and other deformations that can distance them from their geometrical prototypes. Although results from codes such as DSSP and STRIDE converge in well-ordered structures, the agreement between the secondary structure assignments is known to deteriorate as the conformations become more distorted. Conformationally irregular peptides therefore offer a great opportunity to explore the differences between these codes. This is especially important for unfolded proteins and intrinsically disordered proteins, which are known to exhibit residual and/or transient secondary structure whose characterization is challenging. In this work, we have carried out Molecular Dynamics simulations of (relatively) disordered peptides, specifically $gp41_{659-671}$ (ELLELDKWASLWN), the homopeptide polyasparagine (N_{18}), and polyasparagine dimers. We have analyzed the resulting conformations with DSSP and STRIDE, based on hydrogen-bond patterns (and dihedral angles for STRIDE), and KAKSI, based on α-Carbon distances; and carefully characterized the differences in structural assignments. The full-sequence Segment Overlap (SOV) scores, that quantify the agreement between two secondary structure assignments, vary from 70% for $gp41_{659-671}$ (STRIDE as reference) to 49% for N_{18} (DSSP as reference). Major differences are observed in turns, in the distinction between α and 3_{10} helices, and in short parallel-sheet segments.

1. Introduction

Secondary structure allows for a simple description of the complex three-dimensional structure of proteins. It maps the atomic coordinates for every atom into intuitive visual diagrams [1], and facilitates protein structural comparison and analysis [2–5]. The first and most important structural motifs, the α helix [6] and β sheet [7] were predicted by Pauling and Corey in 1951. Since then, other periodic motifs have been identified, including 3_{10} helix, π helix, β turn, γ turn, etc. Secondary structures are extremely important for methods aiming at predicting protein three-dimensional structures [8–12]. However, a precise and evidence-based secondary structure assignment is quite challenging due to the fact that real-life secondary structures are susceptible to bending, twist, fraying and other deformations that can distance them from their ideal geometrical prototypes [13–16,11].

Automatic secondary structure assignment programs were introduced starting more than three decades ago. These programs try to assign secondary structure in a consistent manner, and they are built to explore the regularities associated with periodic structures, such as the dihedral backbone angles (ϕ, ψ) used for Ramachandran plots, the α-Carbon distances, the pattern of hydrogen bonds between backbone atoms, the three-dimensional geometry of local fragments, the backbone curvature, etc. The first implementation of these methods was introduced by Levitt and Greer in 1977 [17] and was based on distances and dihedral angle profiles of α-Carbon atoms over a four-residue sliding window. Other methods include DSSP [18], DEFINE [19], P-CURVE [20], STRIDE [21], P-SEA [22], XTLSSTR [23], SECSTR [24], VoTAP [25], KAKSI [16], PROSIGN [26] and, more recently, a Bayesian method of minimum message length inference, where the protein is described as a collection of segments, with each segment belonging to one of eight potential models (the secondary structure motifs) [27]. Of these, DSSP and STRIDE use quite close criteria for the assignment

* Corresponding author at: Department of Physics, North Carolina State University, Raleigh, NC 27695, United States. Tel.: +1 919 515 3111.
E-mail address: sagui@ncsu.edu (C. Sagui).

http://dx.doi.org/10.1016/j.jmgm.2014.10.005
1093-3263/

of secondary structure and are the most widely used methods. The DSSP (Dictionary of Secondary Structure of Proteins) method, developed by Kabsch and Sander, uses hydrogen bond patterns based on an electrostatic model to assign the secondary structures. DSSP is the method of choice to annotate depositions to PDB [28]. The visualization software RASMOL [29] uses a fast DSSP-like algorithm, and both AMBER [30] and GROMACS [31] analysis tools use the DSSP software. The STRIDE (STRuctural IDEntification) method, developed by Frishman and Argos, makes use of both hydrogen-bond patterns and backbone dihedral angles to assign the secondary structures. STRIDE is used by the popular visualization tool VMD [32]. Both DSSP and STRIDE have been used in structure prediction methods, for instance in Hidden Markov Models where the secondary structure assignments are used as a basis for generalized "alphabets" of backbone geometry [15]. The more recent KAKSI method [16] uses backbone dihedral angles and α-Carbon distances. It was introduced to improve the treatment of irregularities, such as kinks in α helices, in which case STRIDE tends to assign a longer, kinked helix, while KAKSI assigns several short helices.

Recent comparisons between DSSP and STRIDE show, in general, good agreement for well-defined secondary structures [22,33,25,34,16]. Thus, for instance, the results of both assignments were compared for 29 reference crystal structures (both in their reference state and in slightly distorted states) in order to relate them with the results of Circular Dichroism (CD) spectroscopy [33]. The results showed good agreement for α helices and β strands, while displaying more discrepancies in the "turns". The overall performance of DSSP assignments was considered slightly better in the context of this CD spectra analysis. A more recent study [16] focused on how several of the assignment methods handled irregularities of the structures, such as the edges of the motifs, and various distortions [16]. The study was based on 3 X-ray datasets with, respectively, high, medium and low resolution and an NMR dataset. The authors found that the results from DSSP and STRIDE were very close for *all* the four datasets: a direct comparison between DSSP and STRIDE gave between 93.4% and 95.4% agreement using the C_3 scores; while a comparison of each these methods with respect to KAKSI based on SOV [35,36] scores yielded 89% and 92% agreement. The KAKSI method seemed to perform better than STRIDE when there were kinks in the motifs (DSSP was not used in this comparison). There seemed to be less agreement with other methods.

Clearly, there is not an exact and unique way to assign secondary structural motifs, since this depends on the definition for the secondary structure underlying the method [37,11]. The hope is that these automatic computational methods converge (e.g., to a secondary structure assignment provided by an experienced crystallographer). Most disagreements between the various methods occur in the terminal regions of the assigned structural motifs [14,38,37,34,16,39]. Conflicts also arise in determining whether the deviation from a given motif is simply a distortion of the motif or a break in the structure [14,16]. The disagreements between the methods is therefore expected to increase when the peptides exhibit irregular conformations.

In this work, we compare the predictions of these methods when applied to conformationally irregular peptides. In particular, we choose codes that give the closest results in comparisons of ordered structures: DSSP and STRIDE, based on hydrogen bonds; and KAKSI, based on α-Carbon distances. Clearly, these codes were developed to be applied mainly to periodic structures. However, the performance of these codes on more "disordered" structures is of interest for the following reasons. First, as mentioned above, these assignment programs tend to disagree in the terminal regions at the beginning and end of repetitive structures [14,38,37,34,16,39], resulting on different lengths for these structures, and further difficulties for the analysis of connecting loops [34]. Second, it is now understood that disordered proteins and peptides do not quite behave as random coils: both unfolded protein states and Intrinsically Disordered Proteins [40] (IDPs) show considerable residual secondary structure [41,42]. Indeed, the use in NMR of chemical shifts, scalar couplings and residual dipolar couplings have found residual and/or transient secondary structure in both unfolded proteins and IDPs [43–59,41,60–64]. These are particularly present when IDPs become structured when binding a partner [65], or when the IDPs are prone to aggregation [58,66,67]. Both population analysis and free energy landscapes [68–71] reveal that the conformational preferences are much less entropic than those of a random coil. Third, in clear need of a tool to analyze residual secondary structures, researchers fall back into the secondary structure assignment codes, which unfortunately in these irregular or disordered structures do not necessarily achieve consensus.

We recently came across one of such cases [72] when we studied the structural preferences of the tridecapeptide corresponding to residues 659–671 of the envelope glycoprotein gp41 of HIV-1, which spans the 2F5 monoclonal antibody epitope ELDKWA. There have been many studies to determine the structure of $gp41_{659-671}$ (ELLELDKWASLWN) in aqueous solution, giving (at least initially) conflicting results. These include NMR and CD studies [73–75], a UV Resonance Raman Spectroscopy investigation [76], and crystal studies [77,78]. Molecular Dynamics (MD) simulations also showed varying results [79,75,80,72], that can be attributed to the use of different force fields and sometimes inadequate sampling [72]. In addition, the use of different methods to characterize the secondary structure can further complicate comparisons and interpretation of the data. In this work, we compare the secondary structure assignments by DSSP, STRIDE and KAKSI when applied to $gp41_{659-671}$. In addition, we also compare these methods as applied to the disordered homopeptide polyasparagine. Asparagine-rich peptides are disordered at the monopeptide level, but are also found aggregated in highly ordered amyloids [81–83]. Since the β content of both peptides is low, further comparisons are carried on polyasparagine dimers, expressly built as β-sheet dimers.

2. Methods

In this section we briefly review the basics of the DSSP, STRIDE and KAKSI secondary structure assignment methods, as well as the Segment Overlap method (SOV) for the evaluation of secondary structure prediction methods.

2.1. DSSP

The DSSP method [18] by Kabsch and Sander carries out the helical and sheet assignments entirely based on the backbone hydrogen bonds, as defined by a electrostatic model [84]. The electrostatic interaction energy is given by:

$$E = f\, q_1 q_2 \left(\frac{1}{r(ON)} + \frac{1}{r(CH)} - \frac{1}{r(OH)} - \frac{1}{r(CN)} \right), \qquad (1)$$

where $q_1 = 0.42e$, $q_2 = 0.20e$, $f = 332$ Å kcal/(e^2 mol), and $r(AB)$ is the distance between atoms A and B. A hydrogen bond exists between the C═O group of residue i and the N─H group of residue j if the electrostatic interaction energy satisfies $E < -0.5$ kcal/mol.

The basic turn pattern is defined if a hydrogen bond exists between residue i and residue $i+n$ ($i \rightarrow i+n$). If the residues in a basic turn do not belong to a helix, the DSSP output labels them as Turn ("T") (the smallest helix has to have two consecutive hydrogen bonds). Continuous "n helices" are defined when two adjacent hydrogen bonds are at $i-1 \rightarrow i+n-1$ and at $i \rightarrow i+n$, and the residues $(i, i+1, \ldots, i+n-1)$ are labeled as "n helix". The notations for the "n helix" are "G" ($n=3$) for 3_{10} helix, "H" ($n=4$) for

α helix, and "I" ($n=5$) for π helix. This helix definition does not provide assignments to the edge residues with the initial and final hydrogen bonds in the helix.

Another basic pattern, bridge, has been defined as follows: two non-overlapping stretches of three residues each, $i-1$, i, $i+1$ and $j-1$, j, $j+1$ form a parallel bridge when there are two hydrogen bonds $i-1\rightarrow j$ and $j\rightarrow i+1$, or two hydrogen bonds $j-1\rightarrow i$ and $i\rightarrow j+1$; they form an anti-parallel bridge when there are two hydrogen bonds $i\rightarrow j$ and $j\rightarrow i$, or two hydrogen bonds $i-1\rightarrow j+1$ and $j-1\rightarrow i+1$. The residues i and j are labeled as bridge by using lower case for parallel bridge and upper case for anti-parallel bridge. The parallel and anti-parallel information are included in the sheet "structure" section, but not in the "summary" section. One or more consecutive bridges of identical type form a ladder; and one or more connected ladders form a sheet. In the "summary" section, the single bridge (ladder of length 1) is denoted as "B", and the longer ladder is denoted as "E", without distinguishing parallel and antiparallel information. A β bulge connects two perfect ladders or bridges of the same type, and the connected ladders are treated as a single ladder, with all the residues, including the bulges, being labeled as "E". There are at most one bulge residue on one strand and 4 bulge residues on the other strand.

DSSP also defines Bend, Chirality, SS bonds and Chain Breaks. Bend is defined when the angle $C^{\alpha}_{i-2}-C^{\alpha}_{i}-C^{\alpha}_{i+2}$ is smaller than 110°, and is labeled as "S". The dihedral angle $\alpha(i)=(C^{\alpha}_{i-1}, C^{\alpha}_{i}, C^{\alpha}_{i+1}, C^{\alpha}_{i+2})$ is used to define the Chirality, and only the sign of $\alpha(i)$ is reported as "+", if $0^{\circ}<\alpha<180^{\circ}$; or as "−", if $-180^{\circ}<\alpha<0^{\circ}$. SS bonds represent the disulfide bonds joining Cys residues. The chain breaks indicate covalent breaks in the peptide sequence, and are labeled as "!". The information about chirality, SS bonds and chain breaks is included in the DSSP output files but not in the summary section which only includes "H", "G", "I", "B", "E" (parallel and anti-parallel are labeled together here), "T", "S" (any bend that is not assigned to any of the six previous categories) and "C"; the last one used to indicate coil (as the default for a conformation that does not belong to any of the other categories).

2.2. STRIDE

The STRIDE method [21] by Frishman and Argos makes use of hydrogen bonds in a similar manner as DSSP, except that the hydrogen bonds are defined differently. STRIDE uses an empirical function to assign the hydrogen bond energy E_{hb}. In addition, the backbone torsion angles evaluated according to the regions of the Ramachandran plot in which the residues fall, are also involved into the definition of α helix and β strands. The hydrogen bond energy E_{hb} is ignored if the backbone torsions are unfavorable.

For the α helix assignment, STRIDE uses a similar definition as DSSP, so that when two adjacent hydrogen bonds are located at $i-1\rightarrow i+3$, then the residues $i, i+1, \ldots, i+3$ are assigned as α helix ("H"). STRIDE extends the criterion so that if the probabilities of the torsional angles of edge residues $i-1$ and $i+4$ satisfy some additional constraints, the edge residues are also included in the helix. The 3_{10} helix and π helix definitions and notation agree with those of DSSP (except for the definition of the hydrogen bond itself).

STRIDE has several definitions of bridge, which include three bridges with a similar definition to that in DSSP, and two definitions not found in DSSP. Consecutive bridges form a β sheet, with no more than one bulge between bridges and no more than four bulges on the other strand, and the participating residues are denoted as "E", except the bridges flanking the given β sheet where only internal residues are labeled as "E". An isolated bridge is labeled as "B", except for the type III bridges [21] that on one side neither of the residues in the bridge is internal, which are assigned as "b". This "b" is not to be confused with the one in the sheet structure section in DSSP (which is used to label a *parallel* bridge).

The turn assigned by STRIDE is based on Richardson's definition [1], which classifies six distinct turn types (I, I', II, II', VIa and VIb) based on the dihedral angles (ϕ, ψ); as well as the extended definition (type VIII) proposed by Wilmot and Thornton [85].

2.3. KAKSI

The KAKSI method [16] was introduced to optimize the fit to the secondary structure assignments obtained from the PDB files. The assignment is based on the C^{α} distances and (ϕ/ψ) dihedral angles, and it is done by sliding one window along the sequence for α helix, and two windows for β sheet, to ensure that the assigned β strand is involved in a β sheet. As the windows slide along the sequence, the distance criterion (C_1) for helix is checked first, this is followed by the angle criterion (C_2) for helix; all the eligible residues are assigned as helix temporarily. The kink criterions (K_1) and (K_2) are applied to detect kinks along these helix residues, and all the kinks are assigned as coil. Helices with length shorter than 5 residues are discarded. The β sheet distance (C_3) and angle (C_4) criterions are then used to detect β strands after helices have been assigned. Finally, the contiguous criterions (C_5) is used to introduce a coil between adjacent helices and β strands, a process that shortens the intervening helix by one residue. The output includes only three kinds of secondary structure: helix, β strand and coil, which are labeled as "H", "b" and "C".

2.4. SOV

The Segment Overlap measure [35,36] (SOV) by Rost et al. is usually employed to evaluate the protein secondary structure prediction assessment. SOV was developed to take into account that the correct characterization of the type and location of the secondary structure elements is frequently far more important than the assignment at the residue level. By contrast, the C_3 scores define the overall fraction of residues predicted in a given pattern, which sometimes can be quite misleading as to the accuracy of the prediction (as the SOV authors noted, assigning the entire myoglobin chain as a single helix gives a C_3 score of approximately 80% [36]). In addition, SOV is constructed so as to minimize the ambiguity in the definition of the segment ends due to the differences of criteria of the different secondary structural assignment methods.

The SOV measure is defined as follows. Let (s_1, s_2) denote a pair of overlapping segments, $S(i)$ denote the set of all the overlapping pairs of segments (s_1, s_2) in state i, and $S'(i)$ be the set of all the segments s_1 and s_2 in state i, with no overlap between s_1 and s_2. The segment overlap measure for state i is defined as [36]:

$$\mathrm{SOV}(i)=100\times\frac{1}{N(i)}\sum_{S(i)}\left[\frac{\mathrm{minov}(s_1,s_2)+\delta(s_1,s_2)}{\mathrm{maxov}(s_1,s_2)}\times\mathrm{len}(s_1)\right] \quad (2)$$

$$N(i)=\sum_{S(i)}\mathrm{len}(s_1)+\sum_{S'(i)}\mathrm{len}(s_1) \quad (3)$$

$$\delta(s_1,s_2)=\min\left\{(\mathrm{maxov}(s_1,s_2)-\mathrm{minov}(s_1,s_2));\ \mathrm{minov}(s_1,s_2);\ \mathrm{int}\left(\frac{\mathrm{len}(s_1)}{2}\right);\ \mathrm{int}\left(\frac{\mathrm{len}(s_2)}{2}\right)\right\} \quad (4)$$

In this expression, $\mathrm{len}(s_1)$ is the number of residues in segment s_1; $\mathrm{minov}(s_1,s_2)$ is the length of the actual overlap between s_1 and s_2 when they are in state i; $\mathrm{maxov}(s_1,s_2)$ is the total extent for which either of the segments s_1 or s_2 has a residue in state i; $\sum_{S(i)}$ is taken over all the segment pairs in state i which overlap by at least one residue; and $\sum_{S'(i)}$ is taken over the remaining segments in state i (that do not overlap). The SOV measurement for multi-state

secondary structure assignment is defined as

$$\mathrm{SOV} = 100 \times \frac{1}{N} \sum_{i\in[H,G,\ldots,C]} \sum_{S(i)} \left[\frac{\mathrm{minov}(s_1,s_2) + \delta(s_1,s_2)}{\mathrm{maxov}(s_1,s_2)} \times \mathrm{len}(s_1) \right], \tag{5}$$

where N is the sum of $N(i)$ over all secondary structure assignments:

$$N(i) = \sum_{i\in[H,G,\ldots,C]} N(i) \tag{6}$$

Note that since one of the two assignments is chosen as observed assignment (reference), the roles of s_1 and s_2 are not symmetric.

3. Simulation details

As example of conformationally irregular peptides we chose the $gp41_{659-671}$ peptide, polyasparagine with 18 residues, here denoted as N_{18}. In addition, we also considered polyasparagine dimers (N_8–N_8).

3.1. $gp41_{659-671}$ peptide

This simulation [72] includes one $gp41_{659--671}$ molecule with sequence Ace-ELLELDKWASLWN-NH_2. The capping groups are assumed to make the peptide more similar to the corresponding segment in the native protein. The simulation was carried out with the AMBER ff12SB version of the Cornell et al. force field [86] under explicit solvent.

Topology and parameter files along with the coordinates corresponding to the unfolded peptide were generated via the LEAP program of the AMBER v.12 [30] simulation packages. We started the simulations with an implicit solvent. The implicit water model is based on the Generalized Born approximation (GB) [87] including the surface area contributions using the LCPO model [88] (GB/SA) with the surface tension set to 0.005 kcal/mol/Å^2. For the GB model, we used the GB_{OBC} II model [89], with a cutoff of 25 Å. After a short minimization, an initial 10 ns MD simulation was used to generate equilibrium conformations of the peptide. Five different conformations from this run were chosen as initial conformations for the explicit simulations.

The TIP3P water model [90] was used for the explicit solvent simulations under periodic boundary conditions. The five initial conformations from the implicit solvent simulations were solvated in a rectangular box, with an average number of waters of approximately 5670. The peptide has two net negative charges that were neutralized by the addition of two Na^+ ions. Electrostatics were handled by the PME method [91,92], with a direct space cutoff of 9 Å and an average mesh size of approximately 1 Å for the lattice calculations. The equilibration process took place in four steps. First, we applied steepest descent followed by conjugate gradient minimization keeping the peptide atoms fixed at their initial positions. Then we carried out unrestrained steepest descent followed by conjugate gradient minimizations. This was followed by short MD runs under constant volume while the system was heated from 0 K to 300 K with weak restraints on the peptide atoms. Finally, the system was kept at 300 K via Langevin dynamics with a collision frequency $\gamma = 1.0\,\mathrm{ps}^{-1}$, and at constant pressure (1 atm) via the Berendsen barostat [93] with the isothermal compressibility $\beta = 44.6 \times 10^{-6}\,\mathrm{bar}^{-1}$ and the pressure relaxation time $\tau_p = 1.0$ ps. These NPT Langevin dynamics simulations were carried out for 2 ns, and the density of the system was found to be stable around 1.0 g/cm^3. The five equilibrated conformations provided the initial conformations for NPT simulations with Langevin dynamics, each simulation 40 ns long. One configuration was selected for a further 100 ns run. The second half of each of the 40 ns runs and the 100 ns runs were used for data analysis (for a total of 300 ns). Coordinates were sampled every 2 ps.

3.2. Polyasparagine

Polyasparagine of sequence Ace-Asn_{18}-NH_2 (N_{18}) was simulated in an implicit water model. In order to enhance the conformational sampling of this peptide we used temperature replica exchange molecular dynamics (T-REMD) [94]. The unfolded peptide was generated via the LEAP program of AMBER v.12, and the implicit water model was the same one as described for $gp41_{659-671}$. Eighteen replicas were used for the T-REMD scheme at temperatures 300, 322, 346, 371, 398, 427, 459, 493, 529, 567, 609, 654, 702, 753, 809, 868, 932, 1000 K. After a short minimization, an initial 500 ps MD simulation for each replica was used to heat the system to the respective temperature via Langevin dynamics with a collision frequency $\gamma = 1.0\,\mathrm{ps}^{-1}$, and at constant pressure (1 atm) via the Berendsen barostat [93]. Then another 500 ps MD simulation for each replica was run to generate equilibrium conformations of the peptide. The T-REMD scheme was used on the equilibrium conformations from this run. We ran 200 ns T-REMD simulations and the coordinates of the $T = 300$ K replica were sampled every picosecond. Only the second half of the sampled conformations was used for secondary structure analysis.

3.3. Polyasparagine dimers

In order to enhance the conformational sampling of β sheet, polyasparagine dimers (N_8–N_8), consisting of two Ace-Asn_8-NMe peptides, were simulated in an explicit water model with parallel and anti-parallel orientations.

The unfolded peptide was generated via the LEAP program of AMBER v.12. The dimers were assembled in parallel and antiparallel conformations with the proper symmetry and then solvated with TIP3P waters in a truncated octahedron box with 2772 waters and periodic boundary conditions. Electrostatics were handled by the PME method. The dimers tend to quickly unfold under regular MD. Since our aim is to compare how the codes assign structure to β sheet conformations that are not ideal, this system provides a good benchmark as long as the β sheet is somehow enforced. Thus, we carry out two sets of simulations: in one set, we kept the dimers restrained (allowing, of course, some range of motion); in the other set, we took out the restraints and allowed the dimers to evolve, not at room temperature but at 200 K. Initial equilibration started by minimizing the energy of the system through steepest descent followed by conjugate gradient minimization, keeping the backbone atoms fixed at their initial positions with restraints of 100.0 kcal/mol/Å^2 on the backbone atoms. This was followed by heating the systems via Langevin dynamics with a collision frequency $\gamma = 1.0\,\mathrm{ps}^{-1}$ under constant volume for 100 ps. The restrained systems were heated to 310 K and the soon-to-be unrestrained systems to 200 K. This was then followed by constant pressure (1 atm) NPT simulations with a collision frequency $\gamma = 2.0\,\mathrm{ps}^{-1}$, and the Berendsen barostat for 1 ns, with backbone atoms still restrained. The restraints on the backbone were dropped for the unrestrained systems, that were further equilibrated for 10 ns. Finally, the production runs involve 100 ns NPT simulations with no restraint for the two systems at 200 K, and with a restraint of 5.0 kcal/mol/Å^2 on the backbone atoms for the systems at 310 K. Data was stored every 10 ps, with 10,000 conformations for each case.

4. Results

In this section we present the results of the three structure prediction programs DSSP, STRIDE and KAKSI, after applied to 100,000

conformations for each gp41$_{659–671}$ and N_{18} system, as well as to 10,000 conformations of each of the four N_8–N_8 dimer systems. We discuss each result and explain the differences in terms of the criteria that these methods have used to assign the secondary structure.

4.1. Secondary structure assignment and SOV scores

Table 1 shows the average secondary structure content (over all the residues), including α helix (H), 3_{10} helix (G), π helix (I), turn (T), isolated β bridge (B), extended strand (E), and sheet (B+E) as assigned by DSSP, STRIDE and KAKSI (when it provides output). First we consider the results corresponding to the mainly disordered peptides gp41$_{659–671}$ and N_{18}. Notice that even though the strand and sheet contents represent small percentages, the absolute numbers of residues in these conformations are not negligible, as they are calculated over 100,000 conformations for each peptide. The results from DSSP and STRIDE are comparable (DSSP gives higher helical content and lower strand content) except for turn, where DSSP gives considerably lower values. KAKSI gives similar helical content, lower sheet content, and no output for turns. Fig. 1a and b show sample conformations of gp41$_{659–671}$ and N_{18} displaying secondary structure assignments by the three different methods.

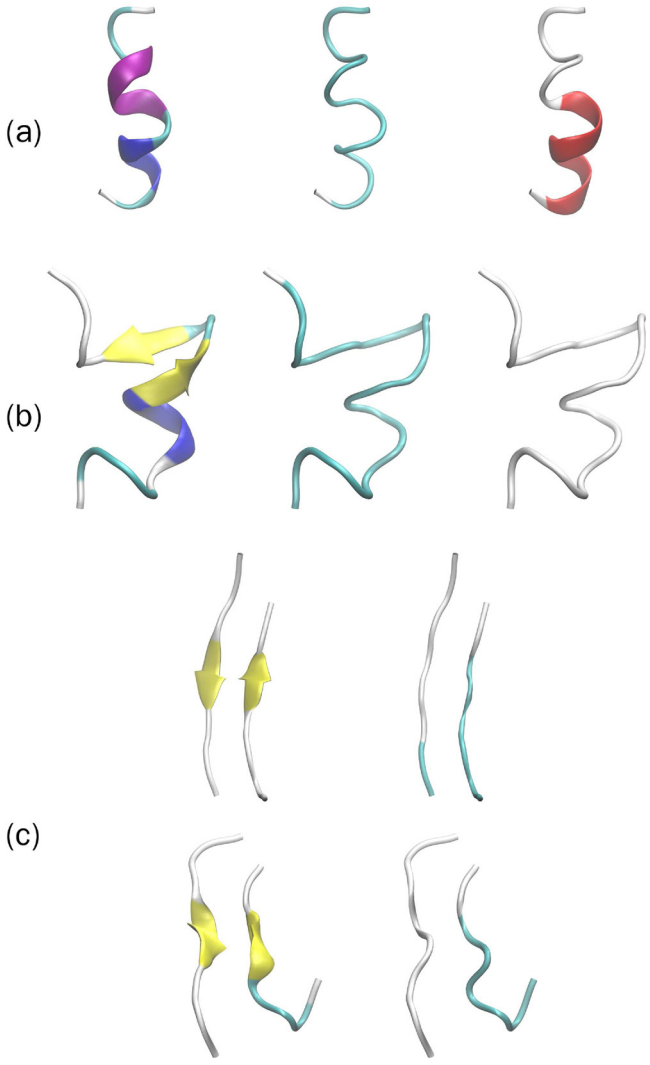

Fig. 1. Snapshots of sample conformations. (a) gp41$_{659–671}$ and (b) N_{18}, with assignments by DSSP (left), STRIDE (middle), and KAKSI (right). (c) Dimer N_8–N_8 analyzed with DSSP (left) and STRIDE (right). Colors purple, blue, green, yellow and white represent α helix (H), 3_{10} helix (G), turn (T), β sheet(E) and coil (C) for DSSP and STRIDE assignments; for KAKSI, red represents helix and white represents coil.

The β sheet dimers show good agreement for the parallel conformations (the restrained one in better agreement than the unrestrained) and less agreement on the anti-parallel conformations. In particular, the unrestrained, antiparallel dimers exhibit the smallest β sheet population, and are in rather poor agreement. Sample conformations of this situation are given in Fig. 1c.

Next, we compare SOV and C3 scores for the different assignments. The simplest residue by residue comparison would involve counting the number of residues that agree for a given category versus the total number of residues (for a full sequence comparison). There are several possibilities for evaluating this. The most commonly used is the overall C3 score, which is defined as:

$$\mathrm{C3} = \frac{N_\mathrm{H} + N_\mathrm{b} + N_\mathrm{C}}{N} \tag{7}$$

where N_H, N_b and N_C represent the number of identical residues as assigned by the secondary structure algorithms in the categories helix ("H"), β sheet ("b") and coil ("C"). Here "H" includes "H", "G", and "I" from DSSP and STRIDE; "b" includes "B", "E", and their lower cases; and "C" includes "C" and "T". In principle, one could define a "C7" score with the same definition as the C3 score, except that it includes 7 categories ("H", "G", "I", "B", "E", "T", "C"). Naturally, C7 scores would generally be lower than C3 scores. As an extreme example, given two sequences SA = GGGGHHHH and SB = HHHHIIII, the C3 score is 100% while the C7 score is 0%. One could also define a C3 score for a particular category, by counting the number of residues that coincide versus the total number of residues in that category. This would generally not be symmetric and would depend on which category is taken as reference. For instance, given the sequences SA = CHHHHHHHHHHC and SB = CCCHHHHHCCCC, then $C3_H$ = 50% when SA is taken as reference, and $C3_\mathrm{H}$ = 100% when SB is taken as reference. In general, this type of comparison is not employed and only the symmetric overall C3 score is used (C3 = 58% in this case).

The SOV measure was introduced to take into account both the type and the location of the secondary structure elements, and in general (but not always) gives more realistic scores. Its results depend on which structure is taken as reference. For instance, consider the two twelve-residue segments SA and SB with sequences SA = CHHHHHHHHHHC and SB = CCCHHHHHCCCC, for which the overall C3 score is C3 = 58%. If SA is taken as reference, the helical overlap SOV_H is 70%, and the overall overlap SOV_{all} is 63%. If SB is taken as reference, then $\mathrm{SOV_H}$ = 70% while $\mathrm{SOV_{all}}$ = 46%. As a more extreme example, consider the 10-residue sequences SA = CECECECECE and SB = EEEEEEEEEE. Taking either sequence as reference gives $\mathrm{SOV_E}$ = 10%, with an overall value $\mathrm{SOV_{all}}$ = 5% for SA as reference and $\mathrm{SOV_{all}}$ = 10% for SB as reference. The overall C3 scores instead are 50%.

Table 2 gives the average value of the SOV measures over all the conformations between DSSP and STRIDE, the latter being chosen as reference. First we look at the disordered peptides gp41$_{659–671}$ and N_{18}. The $\mathrm{SOV_{all}}$ value is larger for gp41$_{659–671}$ than N_{18}, which reflects the fact that gp41$_{659–671}$ has more helical population than N_{18}. The SOV measures between DSSP and STRIDE on α helix are in good agreement, with values greater than 95%. But the SOV measures on other secondary structure assignments are considerably less good, especially for the turn assignment, whose SOV measures are less than 45%. The overall SOV measures for the entire peptides are also lower (as low as 54% for N_{18}), indicating a relative low overall overlap between the assignments of secondary structure by DSSP and STRIDE. Since the SOV measure is not symmetric, we also show the values obtained when DSSP is taken as reference in Table 3. In this case, there is a further lowering of the overall overlap. Table 4 gives the average value of the SOV

Table 1
Secondary structure content (%) of $gp41_{659-671}$, N_{18} and dimers N_8–N_8 in parallel (P) and anti-parallel (AP) orientation, with a restrained (R) or unrestrained (U) backbone.

Peptide	Codes	α	3_{10}	π	Helix	β bridge	β strand	Sheet	Turn
$gp41_{659-671}$	DSSP	33.36	16.56	1.12	51.04	0	0	0	19.54
	STRIDE	32.89	13.19	0.87	46.95	0	0	0	31.09
	KAKSI	–	–	–	44.56	–	–	0	–
N_{18}	DSSP	14.02	6.85	0.21	21.08	1.05	0.13	1.18	27.72
	STRIDE	11.91	4.87	0.12	16.90	1.06	0.25	1.31	59.00
	KAKSI	–	–	–	20.13	–	–	0.04	–
N_8–N_8, R–P	DSSP	0.00	0.00	0.00	0.00	0.00	62.50	62.50	0.00
	STRIDE	0.00	0.00	0.00	0.00	0.44	61.11	61.55	0.00
	KAKSI	–	–	–	0.00	–	–	45.69	–
N_8–N_8, R–AP	DSSP	0.00	0.00	0.00	0.00	5.90	25.22	31.12	0.00
	STRIDE	0.00	0.00	0.00	0.00	9.40	10.71	20.11	28.77
	KAKSI	–	–	–	0.00	–	–	32.62	–
N_8–N_8, U–P	DSSP	0.00	0.00	0.00	0.00	13.53	27.99	41.52	0.00
	STRIDE	0.00	0.00	0.00	0.00	12.96	20.65	33.61	0.14
	KAKSI	–	–	–	0.00	–	–	37.71	–
N_8–N_8, U–AP	DSSP	0.00	0.00	0.00	0.00	0.10	9.26	9.36	7.45
	STRIDE	0.00	0.00	0.00	0.00	0.14	2.61	2.75	46.35
	KAKSI	–	–	–	0.00	–	–	0.00	–

Table 2
Average SOV measures (%) between DSSP and STRIDE, with STRIDE taken as reference. Values are provided for helix (H), 3_{10} helix (G), π helix (I), turn (T), isolated bridge (B), β strand (E), coil (C), as well as the overlap for the complete peptide, SOV_{all}. Data is shown for $gp41_{659-671}$, N_{18}, and dimers N_8–N_8 in parallel (P) and anti-parallel (AP) orientation, with a restrained (R) or unrestrained (U) backbone.

Peptide	SOV_H	SOV_G	SOV_I	SOV_T	SOV_B	SOV_E	SOV_C	SOV_{all}
$gp41_{659-671}$	98.06	78.11	79.31	43.53	–	–	70.90	70.35
N_{18}	95.18	61.82	52.74	41.01	86.64	45.25	72.68	53.78
N_8–N_8, R–P	–	–	–	–	0.00	98.54	98.54	98.28
N_8–N_8, R–AP	–	–	–	0.00	56.36	80.52	59.69	43.97
N_8–N_8, U–P	–	–	–	0.00	96.14	90.85	89.96	90.45
N_8–N_8, U–AP	–	–	–	22.49	6.94	95.79	67.67	48.85

Table 3
Same as Table 2, except that DSSP is taken as reference.

Peptide	SOV_H	SOV_G	SOV_I	SOV_T	SOV_B	SOV_E	SOV_C	SOV_{all}
$gp41_{659-671}$	83.98	62.72	61.69	32.77	N/A	N/A	46.31	62.20
N_{18}	65.13	46.01	30.86	44.51	87.11	89.11	44.92	49.27
N_8–N_8, R–P	–	–	–	–	–	98.54	99.95	99.07
N_8–N_8, R–AP	–	–	–	–	89.73	39.91	44.45	45.62
N_8–N_8, U–P	–	–	–	–	94.03	96.28	92.53	93.28
N_8–N_8, U–AP	–	–	–	49.43	15.79	27.02	55.34	51.89

measures over all the conformations as obtained when DSSP or STRIDE are measured against KAKSI, which is taken as reference. The apparent improvement of overall SOV measures for N_{18} is due to the fact that KAKSI combines several categories of the other two codes into one.

Now we consider the dimers. Both Tables 2 and 3 show the same trend as observed in Table 1: the parallel dimers are in good agreement, whether the backbone is restrained or not, but the predictions for the antiparallel dimers are in poor agreement, with STRIDE predicting considerably more turn content and less β strand

Table 4
Average SOV measures (%) between KAKSI and DSSP/STRIDE based on helix (including "H", "G", "I" from DSSP and STRIDE), β sheet (including "B", "E" and "b" from DSSP and STRIDE) and coil (including "T" and "C" from DSSP and STRIDE). KAKSI is chosen as reference.

Peptide	Codes	SOV_{Helix}	SOV_{Sheet}	SOV_{Coil}	SOV_{all}
$gp41_{659-671}$	DSSP	88.97	N/A	53.45	65.48
	STRIDE	85.74	N/A	60.26	69.80
N_{18}	DSSP	73.12	33.79	71.46	70.61
	STRIDE	62.56	37.58	76.95	74.37
N_8–N_8, R–P	DSSP	–	66.67	53.46	58.45
	STRIDE	–	65.79	53.34	57.83
N_8–N_8, R–AP	DSSP	–	73.33	72.63	66.31
	STRIDE	–	47.09	72.63	59.27
N_8–N_8, U–P	DSSP	–	76.11	49.13	53.86
	STRIDE	–	69.72	46.27	50.23
N_8–N_8, U–AP	DSSP	–	–	75.63	75.63
	STRIDE	–	–	92.94	92.94

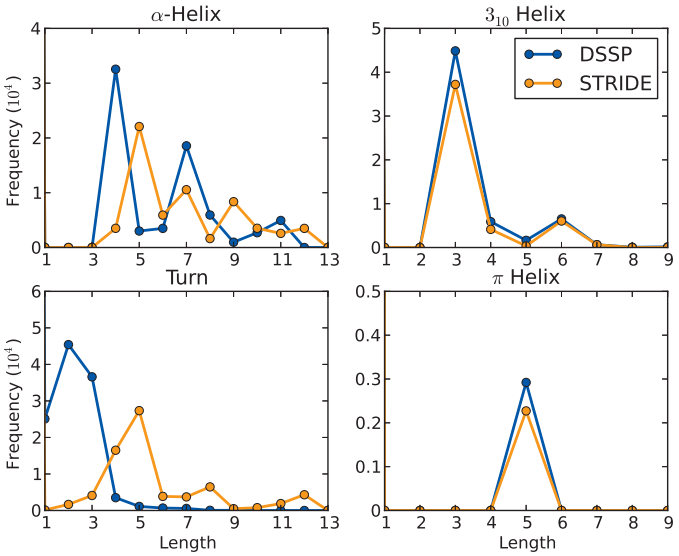

Fig. 2. Length distribution of α helix, 3_{10} helix, and π helix of gp41 as assigned by DSSP (blue) and STRIDE (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

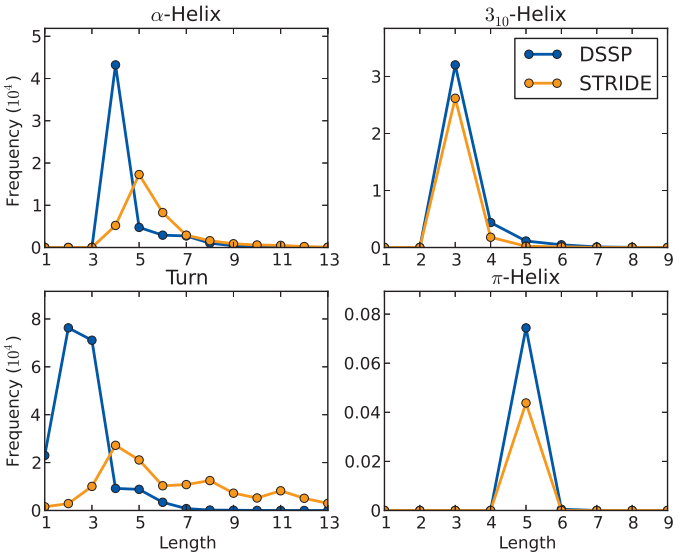

Fig. 3. Length distribution of α helix, 3_{10} helix, and π helix of N_{18} as assigned by DSSP (blue) and STRIDE (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

content. Evaluating the SOV scores of DSSP and STRIDE against KAKSI notably decreases the overall agreement for the parallel dimers, and increases the agreement for the antiparallel dimers. This is particularly the case for the results of the unrestrained parallel dimer, where both STRIDE and KAKSI predict a large amount of turns and coils.

Table 5 compares the average overall C3 scores over all the conformations. Comparisons with KAKSI give lower values, reflecting the different approach of KAKSI. C7 scores between DSSP and STRIDE are 66.82% for $gp41_{659–671}$ and 58.35% for N_{18}.

4.2. Segment length distribution

Figs. 2 and 3 show the population distribution of conformations versus the segment length, e.g., the number of adjacent residues with the same assignment. The 3_{10} helix and π helix for DSSP and STRIDE are in good agreement. The preferred lengths for the 3_{10} helix are 3 and 6, which means that the secondary structure consists of a single 3_{10} helix loop or two 3_{10} helix contiguous loops. For the π helix, the favorite length is 5, i.e., the length of a single π helix loop. STRIDE tends to give longer α helices and turns. The favorite length of the α helix from DSSP is 4, which is the length of a single α helix loop with 2 hydrogen bonds between residues $(i-1, i+3)$ as well as $(i, i+4)$; while the favorite length from STRIDE is 5, which is one typical α helix loop with one of the end residues. This is obviously due to the different criteria used by these two algorithms. In the case of turn, DSSP gives a short turn with a length around 2 residues, but STRIDE favors a longer turn with a length around 4 residues, and even longer. Fig. 4 shows how the population distribution changes when the helical secondary structures G, H and I are grouped together into the single "Helix" category. By design, the minimum helical segment length for KAKSI is 5.

Table 5
C3 scores (%) between KAKSI, DSSP and STRIDE on $gp41_{659–671}$ and N_{18}.

		STRIDE	KAKSI
$gp41_{659–671}$	DSSP	85.12	70.35
	STRIDE		75.66
N_{18}	DSSP	89.56	80.71
	STRIDE		83.72

4.3. Hydrogen bonds

Both DSSP and STRIDE base their secondary structure algorithms on the definition of hydrogen bonds, and many of the differences between their results are due to the different criteria employed for the definition of these bonds. Fig. 5 shows the population distribution on the plane determined by the alignment angle and the hydrogen-bond length. The alignment angle θ is defined as the angle N–(H)–O, and the hydrogen bond length as the distance between atoms O and N in each-hydrogen bond pair. We observe good agreement between both predictions except that there is a 3.5 Å cutoff on the hydrogen-bond length in the STRIDE prediction. However, as we show in the next figures, the hydrogen bonds predicted by DSSP are more consistent with the helical definitions than those by STRIDE.

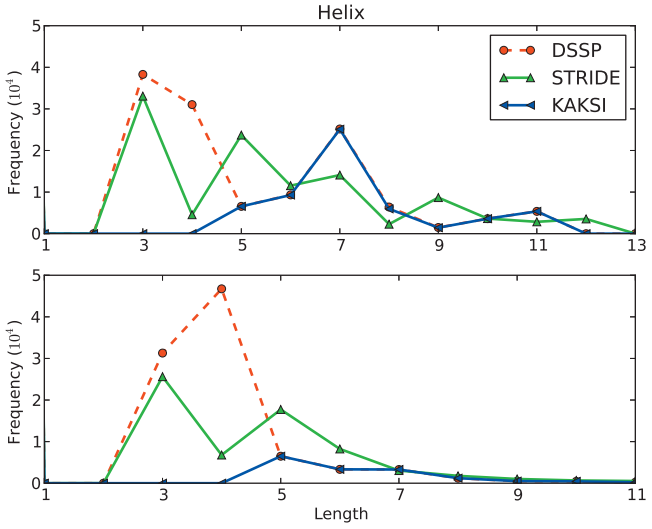

Fig. 4. Length distribution of "Helix" for $gp41_{659–671}$ (top) and N_{18} (bottom).

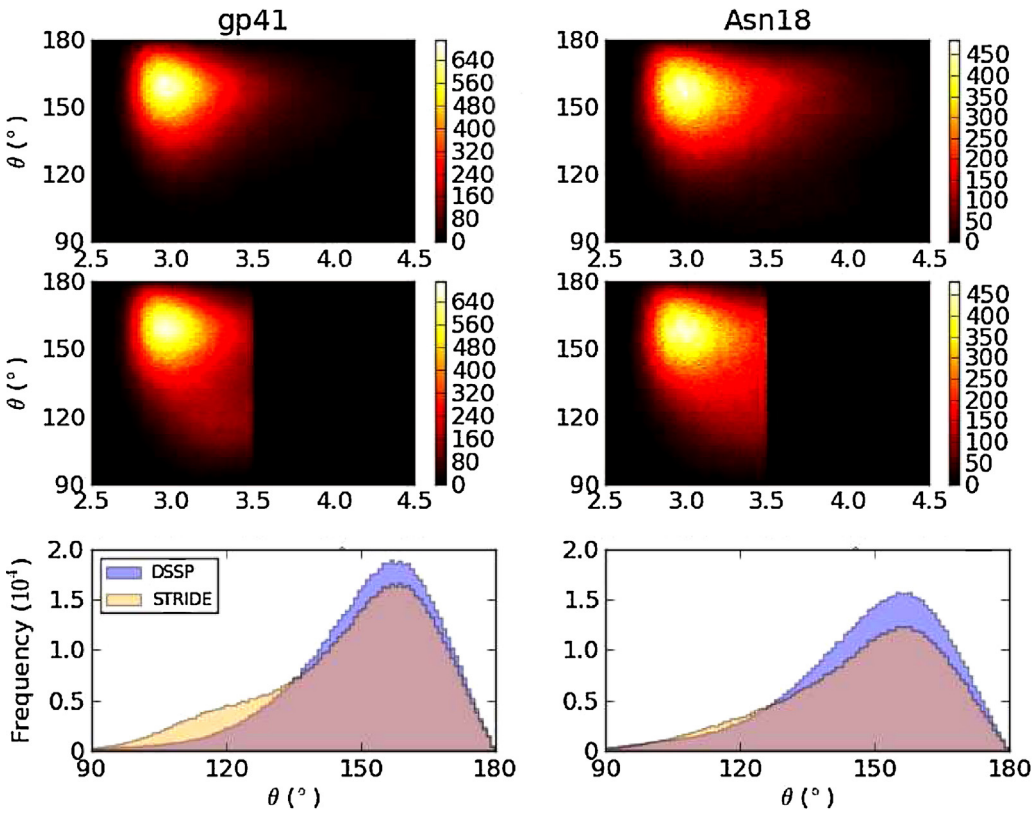

Fig. 5. The population distribution of $gp41_{659–671}$ (left) and N_{18} (right) on the plane determined by the alignment angle θ and the hydrogen bond length. The alignment angle θ is defined as the N—(H)—O angle, and the hydrogen bond length is defined as the distance between atoms N and O in each hydrogen-bond pair. The hydrogen bonds are assigned by DSSP (top) and STRIDE (middle). The two bottom pictures show the corresponding population distributions versus the alignment angle, with hydrogen bonds assigned by DSSP in blue and by STRIDE in green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.4. Helical properties

We find that there are some differences between the helical predictions by DSSP and STRIDE, especially for the 3_{10} helix. Here we show results for the dihedral angles (ϕ, ψ) and the translations of the helices as assigned by these two algorithms.

Two important characteristics of a helix are the hydrogen bond geometry and the rotation angle Ω per residue. The 3_{10} helix is defined by the repeated hydrogen bond ($i+3 \rightarrow i$) between the N—H group of an amino acid and the C═O group three residues earlier. The rotation angle of a 3_{10} helix is about 120°, resulting in 3 residues per turn. The α helix is defined by the ($i+4 \rightarrow i$) hydrogen bond, with a rotation angle of about 100°, which includes 3.6 residues per turn. A general formula for the rotation angle Ω per residue of the helix structure is given by:

$$3\cos\Omega = 1 - 4\cos^2\left(\frac{\psi_i + \phi_{i+1}}{2}\right) \tag{8}$$

where ψ_i is the backbone dihedral angle ψ in residue i, and ϕ_{i+1} is the backbone dihedral angle ϕ in the next residue. Since the 3_{10} helix and α helix adopt rotation angles at approximately 120° and 100°, we can identify the helix by checking the sum of the backbone dihedral angle $\psi_i + \phi_{i+1}$. For the 3_{10} helix, it is roughly −75° and for the α helix it is approximately −104°. The α helix has 1.5 Å translation length along the axis and its backbone dihedral angles are distributed around (−60°, −45°). The 3_{10} helix translational length is about 2.0 Å. The generally accepted values of the dihedral angles are (−49°, −26°) [95,96]. The ideal 3_{10} helix has angles (−74°, −4°) [97,98]. Sometimes values of (−49°, −18°) (with rise per residue of 1.7 Å) and (−71°, −18°) (with 3.2 residues per turn) are also cited [99,1].

Figs. 6 and 7 show the Ramachandran plot and the population distribution as function of translation length for the 3_{10} and α helices. The residues are assigned by DSSP and STRIDE as "p" or "n", where p (positive) indicates that the residue is assigned as 3_{10} or α helix by the respective algorithm (based on the H-bond pattern), and n (negative) indicates that the residue is not assigned to that state by the algorithm. Thus, for instance, "DpSn" denotes the set of residues found in a given state by DSSP (DSSP positive) but found not belonging to that state by STRIDE (STRIDE negative). The three possible combinations, DpSp, DpSn and DnSp are shown for the 3_{10} helix in Fig. 6 and for the α helix in Fig. 7. For $gp41_{659–671}$, the conformation ensemble DpSp assigned as 3_{10} helix both by DSSP and STRIDE has a translation length distribution located around 1.8 Å and the Ramachandran plot shows the backbone dihedral distributed around (−64°, −23°); both metrics are similar to the α helix properties. The DpSn ensemble assigned as 3_{10} helix by DSSP but not STRIDE has a better translation located exactly at 2 Å but the DnSp ensemble has a 1.5 Å translational length with dihedral angles located at (−64°, −29°), which are closer to the α helix values. On the other hand, the conformation ensembles DpSp and DpSn corresponding to the α helix have the expected α helix rise of 1.5 Å per residue and dihedral angles located at (−65°, −34°). But the DnSp ensemble has a 1.8 Å helical rise, and dihedral angles (−65°, −18°), which is very close to the DnSp ensemble in the 3_{10} helix case. These results show that STRIDE tends to give tighter 3_{10} helices and looser α helices, which suggests that STRIDE cannot clearly distinguish between 3_{10} helices and α helices. Similar results are observed for N_{18}.

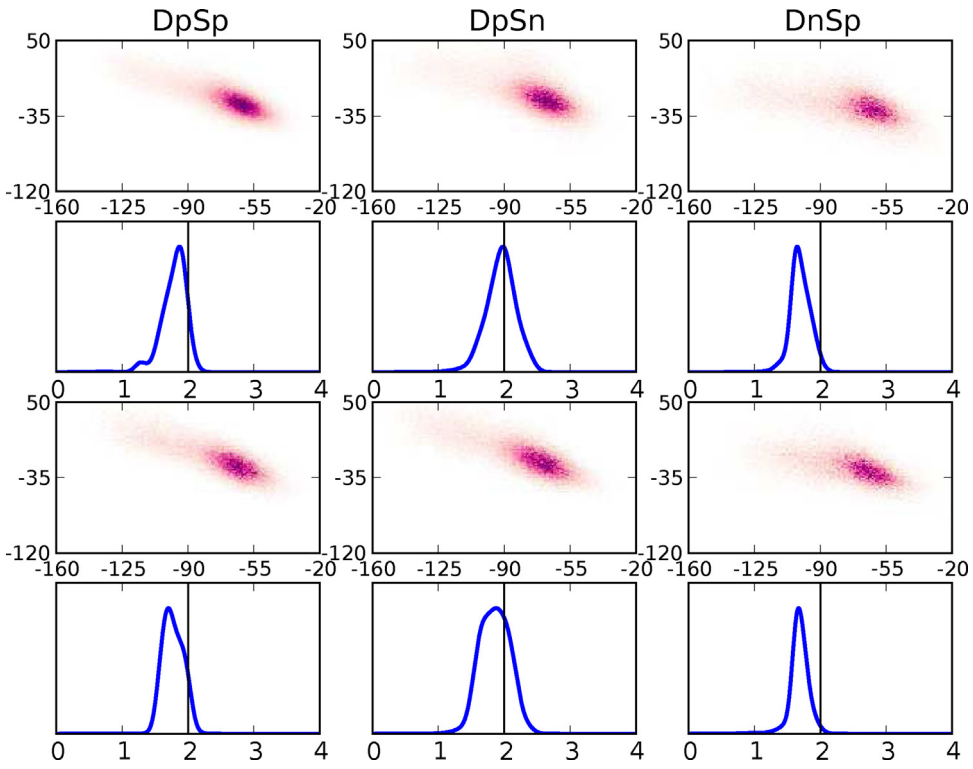

Fig. 6. Odd rows show Ramachandran plots (y-axis is ψ angle in degrees, x-axis is ϕ angle in degrees) and even rows show translation length distribution (y-axis) as a function of helical rise (x-axis, in Å) for 3_{10} helix residues in $gp41_{659–671}$ (top) and N_{18} (bottom). The DpSp (left column), DpSn (middle column) and DnSp (right column) are residues assigned by DSSP and STRIDE as pp, pn and np, where p indicates that the residue is assigned as 3_{10} helix by the respective algorithm, and n indicates that the residue is not assigned to that state by the algorithm.

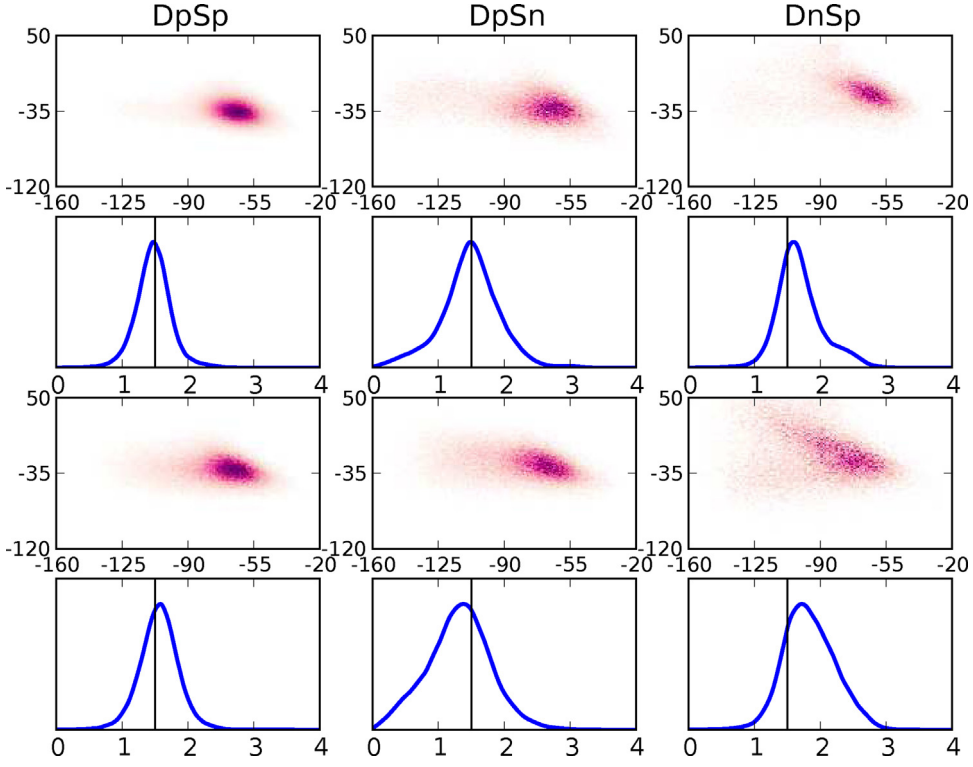

Fig. 7. Odd rows show Ramachandran plots (y-axis is ψ angle in degrees, x-axis is ϕ angle in degrees) and even rows show translation length distribution (y-axis) as a function of helical rise (x-axis, in Å) for α helix residues in $gp41_{659–671}$ (top) and N_{18} (bottom). The DpSp (left column), DpSn (middle column) and DnSp (right column) are residues assigned by DSSP and STRIDE as pp, pn and np, where p indicates that the residue is assigned as α helix by the respective algorithm, and n indicates that the residue is not assigned to that state by the algorithm.

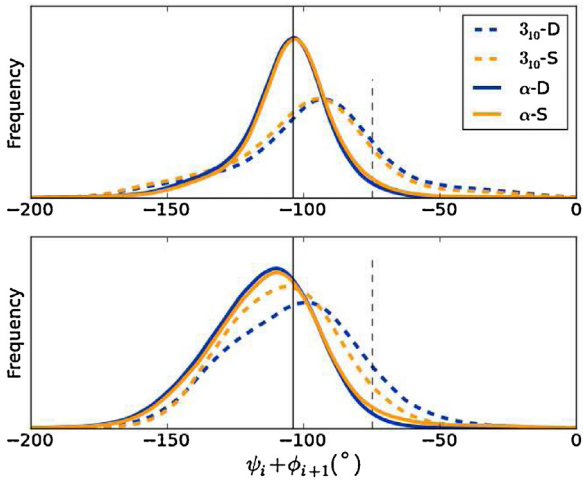

Fig. 8. The dihedral angle sum $\psi_i+\phi_{i+1}$ distribution of 3_{10} helix residues (dashed line) and α helix residues (solid line) as assigned by DSSP (blue) and STRIDE (red). The standard dihedral angles sum values are marked by the black lines, located at -75° (dashed line) for the 3_{10} helix and at -104° (solid line) for the α helix. Top: $gp41_{659-671}$; bottom: N_{18}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

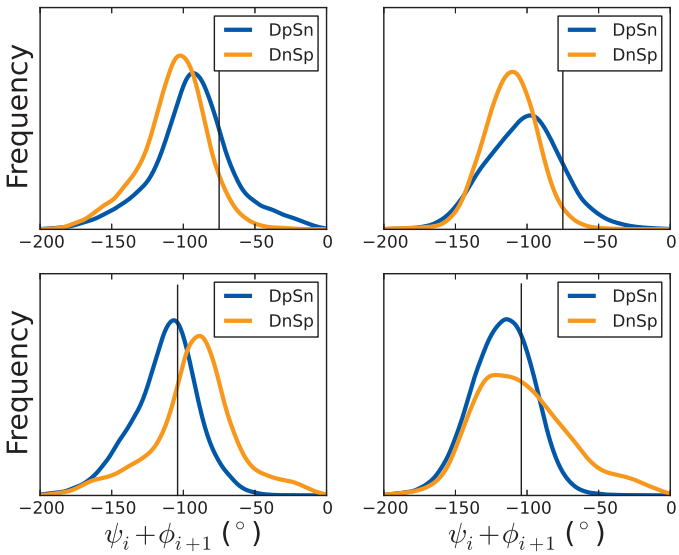

Fig. 9. The dihedral angle sum $\psi_i+\phi_{i+1}$ distribution of 3_{10} helix residues (top) and α helix residues (bottom) as assigned by DSSP and STRIDE: DpSn (DSSP positive, STRIDE negative) is in blue, DnSp (DSSP negative, STRIDE positive) is in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 8 shows the dihedral angle sum $\psi_i+\phi_{i+1}$ distribution of 3_{10} helix and α helix as assigned by DSSP and STRIDE. The standard 3_{10} helix is expected to show values of the sum at around -75° (marked by the vertical dashed line), however both DSSP and STRIDE assignments are shifted 20° towards -95°, which is closer to the α helix. The α helix assignments by these two algorithms have a better agreement with the theoretical dihedral angle sum, -104°. This suggests that the 3_{10} and α helices identified by DSSP or STRIDE for these peptides tend to be quite close to each other in their helical properties. Fig. 9 shows the $\psi_i+\phi_{i+1}$ distribution of 3_{10} helix and α helix residues when the assignments by DSSP and STRIDE do not agree (i.e., the DpSn and DnSp sets). The positive DSSP assignments (blue lines) have better agreement with the standard value of the dihedral angle sum for both the α helix (-104°) and the 3_{10} helix (-75°). This, together with the better helical rise prediction shown in Figs. 6 and 7, points out to a better helix recognition by DSSP.

4.5. Turn and bend

DSSP and STRIDE use different definitions for turn. The n-turn pattern is defined as a single H bond from CO(i) to NH($i+n$) by DSSP. If there is no helix structure detected at the n-turn, the DSSP output identifies these residues as turn, i.e.,"T". STRIDE predicts turn patterns by using Richardson's definition [1], plus the extended turn structure definition by Wilmot and Thornton [85]. The β turn "T" from STRIDE is given by the following criterion: the atoms $C_i^\alpha - C_{i+3}^\alpha$ distance is <7 Å and the central residues are not helix. The extended turn type such as βI, βII, etc. are given by using the backbone dihedral informations. Even though these two definitions are equivalent in many cases, there may still be considerable disagreement between the predictions.

Fig. 10 shows the secondary structure distribution predicted by DSSP on all the residues which are predicted as turn type by STRIDE. Only 36.7% and 40.5% of turn type residues from STRIDE are predicted as "T" by DSSP; 38.7% and 48.1% residues are predicted as either bend ("S") or coil ("C") by DSSP (the summary file reports the sum of bend and coil as coil). Finally, we dump the remaining 24.5% and 11.3% turn-type residues identified by STRIDE into an "other" category, mainly composed of helical residues. Checking the H-bonds detected by DSSP on these helices reveals that 46% ($gp41_{659-671}$) and 49% (N_{18}) of these H-bonds are longer than 3.5 Å, which is forbidden by STRIDE. Conversely, for the residues identified as turn by DSSP, STRIDE gives 58.4% ($gp41_{659-671}$) and 86.2% (N_{18}) turn; 22.0% ($gp41_{659-671}$) and 5.5% (N_{18}) coil; and 19.6% ($gp41_{659-671}$) and 8.3% (N_{18}) "other" (which is mainly helix).

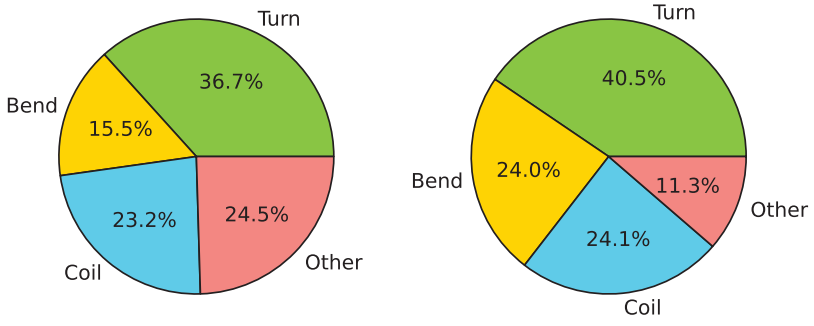

Fig. 10. Pie chart of secondary structure distribution assigned by DSSP for all the residues identified as "turn" by STRIDE (i.e., the STRIDE graphs in either case are a full green "turn" circle). Turn: green; bend: yellow; coil: blue; other: red. Left: $gp41_{659-671}$; right: N_{18}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Discussion and conclusions

Secondary structure analysis codes were originally developed to be applied mainly to periodic structures. Even though there is not an exact and unique way to assign secondary structural motifs [37,11], these automatic computational methods generally converge when applied to regular structures. However, as several comparison studies in the past have shown, not all codes converge equally. DSSP and STRIDE, based on hydrogen bonds and also on dihedral angles for STRIDE, are the most popular codes in current use, and they give the closest results in comparisons of ordered structures [22,25,34,33,16]. KAKSI, using a different criterion based on α-Carbon distances and backbone dihedral angles, also gives relatively close results [16].

As the protein conformations become more distorted, the agreement between the secondary structure assignment codes is expected to deteriorate. Indeed, it has been observed that most disagreements between the various methods occur in the terminal regions of the assigned structural motifs [14,38,37,34,16,39], which results on different lengths for these structures, and further difficulties for the analysis of connecting loops [34]. Conflicts also arise in determining whether the deviation from a given motif is simply a distortion of the motif or a break in the structure [14,16]. Conformationally irregular peptides offer therefore a great test bench to explore the limits and differences between these codes. Molecular dynamics simulations are optimal to generate large number of conformations (there are 100,000 conformations for each peptide presented in this work), especially when one considers the dearth of experimental structural data for disordered peptides.

Secondary structure analysis codes, or perhaps suitable extensions of them, are also needed to help characterize residual or transient secondary structure in unfolded proteins and IDPs [43–59,41,60–64]. Indeed, it is recognized that disordered proteins and peptides do not quite behave as random coils at the residue level [100,41,42]: yes, many of these chains (especially if they are long) satisfy global properties and scaling laws typical of random coils [101,102], such as radius of gyration $R_G \propto N^{0.6}$, where N is the number of residues, and a Gaussian distribution for end-to-end distances. However, they show conformational preferences at the residue level, and thus do not satisfy Flory's model for a polymer chain, where each monomer is randomly oriented with respect to its neighboring monomers. This of course is by no means contradictory. For instance, Fitzkee and Rose [103] modified proteins of known structures by varying backbone torsion angles for ~8% of the residues, while the remaining 92% stayed fixed in their native conformations. By generating ensembles of these "disordered" structures, they were able to recover the typical random-coil statistics. This is simply because the residual structure can be grouped into Kuhn segments, and a chain of Kuhn segments can be treated as a random coil.

In experiments, the use in NMR of chemical shifts, scalar couplings and residual dipolar couplings has been particularly useful to find residual and/or transient secondary structure in both unfolded proteins and IDPs. Residual structures in the unfolded states introduce conformational biases that greatly enhance the folding of the protein [104,105]. In addition, many unfolded states seem to prefer PPII conformations [68] (PPII is not a state recognized by these codes, where PPII conformations are simply characterized as "coils"). Transient and in some cases even residual secondary structures are particularly present when IDPs become structured upon binding a partner [65], or when the IDPs are prone to aggregation [58,66,67]. Both population analysis and free energy landscapes [68–71] reveal that the conformational preferences are much less entropic than those of a random coil. For instance, a polyglutamine peptide in aqueous solution is intrinsically disordered, but it still exhibits relatively large percentages of secondary structure motifs [71]. Moreover, Q_{40}, a polyglutamine of 40 residues just above the threshold of ≃36 for Huntington's disease, has been found to exhibit long-range (over 20 residues) structural correlations [71], that are not present in shorter, non-pathological polymer lengths. These correlations may underlie the aggregation phenomena related to polyglutamine diseases, and are destroyed when a C-terminal hexaproline is added to Q_{40}.

Given the difficulties to characterize the structure of IDPs experimentally, MD has become a valuable tool for this purpose. For instance, our recent work [72] on the $gp41_{659-671}$ peptide was spurred by its role as HIV antibody. However the original experimental data of the monomeric peptide in aqueous solution was highly contradictory. An early study based on NMR indicated that the conformation of the monomeric peptide in water was an amphiphilic 3_{10} helix with minor random coil representation [73]. A second NMR study found no major population of 3_{10} helical conformers, but a mixture of various conformers [74]. Later, a UV Resonance Raman Spectroscopy (UVRRS) investigation [76] concluded that there is a rough energy landscape with a wider variety of conformations than found in the NMR studies and previous Circular Dichroism (CD) studies [73]. The authors suggested that $gp41_{659-671}$ exhibits a broad distribution of conformations that includes significant population of β turns, as well as 3_{10} helix and π helix motifs but little α helix. Recently, a far UV CD spectroscopy study [75] also revealed the conformational plasticity of $gp41_{659-671}$, with no stable α helical, 3_{10} helical, or turn motifs. On the other hand, the crystal structure of the peptide bound to the 2F5 antibody shows an extended conformation [77,78]. MD simulations also showed varying results [79,75,80,72], due to different force fields and sometimes inadequate sampling [72]. In addition, the use of different methods to characterize the secondary structure motifs can further complicate comparisons and interpretation of the data.

In this work we have compared DSSP, STRIDE and KAKSI as applied to mainly disordered $gp41_{659-671}$, N_{18} and N_8–N_8 peptides. As expected, the agreement between the codes is considerably poorer than that for regular structures. The SOV scores between STRIDE and DSSP for the full peptide comparison, SOV_{all} are 70.4% ($gp41_{659-671}$) and 53.8% (N_{18}) when STRIDE is taken as reference; and 62.2% ($gp41_{659-671}$) and 49.3% (N_{18}) when DSSP is taken as reference. For the dimers that show some measure of (non-ideal) β sheet, SOV_{all} scores are good for parallel dimers, in the range 90–98% when STRIDE is taken as reference (93–99% with DSSP as reference). However, for antiparallel dimers, for which STRIDE assigns considerably more turns and fewer strands, the SOV_{all} scores are in the range 44–49% when STRIDE is taken as reference (46–52% with DSSP as reference). Segment length distributions are the same for 3_{10} and π helices, but are different for α helices and turns, due to the difference in definitions as discussed in the Results section. STRIDE and DSSP define more secondary structure elements than KAKSI, and thus can provide more detail when this information is needed. When it comes to distinguishing between 3_{10} and α helices, DSSP performs better than STRIDE, as shown in Figs. 6–9. In particular, the DpSn ensembles (the subset of residues found in a given state by DSSP but not by STRIDE) correctly produce a helical rise distribution centered at 2 Å for 3_{10} helices and at 1.5 Å for α helices; while the DnSp ensembles exhibit distributions centered at shorter helical rises for 3_{10} helices and longer helical rises for α helices. Distributions for the dihedral angle sum $\psi_i + \phi_{i+1}$ also confirm a better assignment of 3_{10} and α helices by DSSP. The largest source of differences is due to the definition of turns: in DSSP their definition is based on the hydrogen bonds, and include α turns, β turns, γ turns; in STRIDE only β turns and γ turns are recognized based on the distance $C_i^\alpha \rightarrow C_{i+3}^\alpha$ and the backbone dihedral angles (ϕ, ψ). This results in a great disparity of these conformations: of all the turns defined by STRIDE, only approximately 40%

are recognized as "turns" by DSSP, while the remaining residues are classified as bends, coils and other structures.

Acknowledgments

We gratefully acknowledge help from Dr. Viet Hoang Man, especially for the preparation of the N_8–N_8 dimers. We thank NSF (Grants MCB-1021883 and SI2/SEE-1148144) for funding. We also thank the NC State HPC Center for extensive computational support.

References

[1] J.S. Richardson, The anatomy and taxonomy of protein structure, Adv. Protein Chem. 34 (1981) 167–339.
[2] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, Scop: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (4) (1995) 536–540.
[3] J.F. Gibrat, T. Madej, S.H. Bryant, Surprising similarities in structure comparison, Curr. Opin. Struct. Biol. 6 (3) (1996) 377–385.
[4] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton, Cath - a hierarchic classification of protein domain structures, Structure 5 (8) (1997) 1093–1109.
[5] A.P. Kamat, A.M. Lesk, Contact patterns between helices and strands of sheet define protein folding patterns, Proteins: Struct., Funct., Bioinform. 66 (4) (2007) 869–876.
[6] L. Pauling, R.B. Corey, H.R. Branson, The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain, Proc. Natl. Acad. Sci. U. S. A. 37 (1951) 205–211.
[7] L. Pauling, R.B. Corey, The pleated sheet, a new layer configuration of polypeptide chains, Proc. Natl. Acad. Sci. U. S. A. 37 (5) (1951) 251–256.
[8] A. Šali, T.L. Blundell, Comparative protein modelling by satisfaction of spatial restraints, J. Mol. Biol. 234 (3) (1993) 779–815.
[9] G. Pollastri, D. Przybylski, B. Rost, P. Baldi, Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles, Proteins: Struct., Funct., Bioinform. 47 (2) (2002) 228–235.
[10] P. Bradley, D. Chivian, J. Meiler, K.M. Misura, C.A. Rohl, W.R. Schief, W.J. Wedemeyer, O.S. Furman, P. Murphy, J. Schonbrun, C.E. Strauss, D. Baker, Rosetta predictions CASP5: successes, failures, and prospects for complete automation, Proteins: Struct., Funct., Bioinform. 53 (6) (2003) 457–468.
[11] C.A. Andersen, B. Rost, Structural Bioinformatics, 2nd ed., Wiley-Blackwell, 2009, pp. 459–484, Ch. 19. Secondary Structure Assignment.
[12] H. Zhang, T. Zhang, K. Chen, K.D. Kedarisetti, M.J. Mizianty, Q. Bao, W. Stach, L. Kurgan, Critical assessment of high-throughput standalone methods for secondary structure prediction, Briefings Bioinform. 12 (6) (2011) 672–688.
[13] G.E. Schulz, C.D. Barry, J. Friedman, P.Y. Chou, G.D. Fasman, A.V. Finkelstein, V.I. Lim, O.B. Ptitsyn, E.A. Kabat, T.T. Wu, M. Levitt, B. Robson, K. Nagano, Comparison of predicted and experimentally determined secondary structure of adenyl kinase, Nature 250 (1974) 140–142.
[14] B. Robson, J. Garner, Introduction to Proteins and Protein Engineering, Elsevier Science, Amsterdam, 1986.
[15] R. Karchin, M. Cline, Y. Mandel-Gutfreund, K. Karplus, Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry, Proteins: Struct., Funct., Genet. 51 (4) (2003) 504–514.
[16] J. Martin, G. Letellier, A. Marin, J.-F. Taly, A.G. de Brevern, J.-F. Gibrat, Protein secondary structure assignment revisited: a detailed analysis of different assignment methods, BMC Struct. Biol. 5 (2005).
[17] M. Levitt, J. Greer, Automatic identification of secondary structure in globular proteins, J. Mol. Biol. 114 (2) (1977) 181–293.
[18] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (12) (1983) 2577–2637.
[19] F.M. Richards, C.E. Kundrot, Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure, Proteins: Struct., Funct., Bioinform. 3 (2) (1988) 71–84.
[20] H. Sklenar, C. Etchebest, R. Lavery, Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis, Proteins: Struct., Funct., Bioinform. 6 (1) (1989) 46–60.
[21] D. Frishman, P. Argos, Knowledge-based secondary structure assignment, Proteins: Struct., Funct., Genet. 23 (1995) 566–579.
[22] G. Labesse, N. Colloc'h, J. Pothier, J.P. Mornon, P-sea: a new efficient assignment of secondary structure from c alpha trace of proteins, Comput. Appl. Biosci. 13 (3) (1997) 291–295.
[23] S.M. King, W.C. Johnson, Assigning secondary structure from protein coordinate data, Proteins: Struct., Funct., Bioinform. 35 (3) (1999) 313–320.
[24] M.N. Fodje, S. Al-Karadaghi, Occurrence, conformational features and amino acid propensities for the π-helix, Protein Eng. 15 (5) (2002) 352–358.
[25] F. Dupuis, J.-F. Sadoc, J.-P. Mornon, Protein secondary structure assignment through voronoï tessellation, Proteins: Struct., Funct., Bioinform. 55 (3) (2004) 519–528.
[26] S.-R. Hosseini, M. Sadeghi, H. Pezeshk, C. Eslahchi, M. Habibi, Prosign: a method for protein secondary structure assignment based on three-dimensional coordinates of consecutive c_α atoms, Comput. Biol. Chem. 32 (6) (2008) 406–411.
[27] A.S. Konagurthu, A.M. Lesk, L. Allison, Minimum message length inference of secondary structure from protein coordinate data, Bioinformatics 28 (12) (2012) 97–105.
[28] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (1) (2000) 235–242.
[29] R.A. Sayle, E.J. Milner-White, RASMOL: biomolecular graphics for all, Trends Biochem. Sci. 20 (9) (1995) 374–376.
[30] D.A. Case, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, et al., AMBER 12, University of California, San Francisco, 2012.
[31] D. van der Spoel, E. Lindahl, B. Hess, The GROMACS Development Team, GROMACS User Manual, 4th ed., 2013.
[32] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, J. Mol. Graph. 14 (1996) 33–38.
[33] N. Sreerama, S.Y. Venyaminov, R.W. Woody, Estimation of the number of α-helical and β-strand segments in proteins using circular dichroism spectroscopy, Protein Sci. 8 (2) (1999) 370–380.
[34] L. Fourrier, C. Benros, A.G. de Brevern, Use of a structural alphabet for analysis of short loops connecting repetitive structures, BMC Bioinform. 12 (5) (2004) 58.
[35] B. Rost, C. Sander, R. Schneider, Redefining the goals of protein secondary structure prediction, J. Mol. Biol. 235 (1) (1994) 13–26.
[36] A. Zemla, Česlovas Venclovas, K. Fidelis, B. Rost, A modified definition of sov, a segment-based measure for protein secondary structure prediction assessment, Proteins: Struct., Funct., Genet. 34 (2) (1999) 220–223.
[37] J.A. Cuff, G.J. Barton, Evaluation and improvement of multiple sequence methods for protein secondary structure prediction, Proteins: Struct., Funct., Bioinform. 34 (4) (1999) 508–519.
[38] N. Colloc'h, C. Etchebest, E. Thoreau, B. Henrissat, J.-P. Mornon, Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment, Protein Eng. 6 (4) (1993) 377–382.
[39] W. Zhang, A.K. Dunker, Y. Zhou, Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks, Proteins: Struct., Funct., Bioinform. 71 (2008) 61–67.
[40] V.N. Uversky, A.K. Dunker, Protein and Peptide Folding, Misfolding, and Non-Folding, 1st ed., John Wiley & Sons, Inc., Hoboken, NJ, USA, 2012, Ch. 1. Why are we Interested in the Unfolded Peptides and Proteins?
[41] D. Eliezer, Biophysical characterization of intrinsically disordered proteins, Curr. Opin. Struct. Biol. 19 (1) (2009) 23–30.
[42] N. Rezaei-Ghaleh, M. Blackledge, M. Zweckstetter, Intrinsically disordered proteins: from sequence and conformational properties toward drug discovery., ChemBioChem 12 (7) (2012) 930–950.
[43] K. Chandrasekhar, A.T. Profy, H.J. Dyson, Solution conformational preferences of immunogenic peptides derived from the principal neutralizing determinant of the hiv-1 envelope glycoprotein gp120, Biochemistry 30 (38) (1991) 9187–9194.
[44] K.A. Dill, D. Shortle, Denatured state of proteins, Annu. Rev. Biochem. 60 (1991) 795–825.
[45] R. Brüschweiler, M. Blackledge, R.R. Ernst, Multi-conformational peptide dynamics derived from NMR data: a new search algorithm and its application to antamanide, J. Biomol. NMR 1 (1) (1991) 3–11.
[46] H.J. Dyson, P.E. Wright, Peptide conformation and protein folding, Curr. Opin. Struct. Biol. 3 (1) (1993) 60–65.
[47] K.M. Fiebig, H. Schwalbe, M. Buck, L.J. Smith, C.M. Dobson, Toward a description of the conformations of denatured states of proteins. Comparison of a random coil model with NMR measurements, J. Phys. Chem. 100 (7) (1996) 2661–2666.
[48] L.J. Smith, K.A. Bolin, H. Schwalbe, M.W. MacArthur, J.M. Thornton, C.M. Dobson, Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations, J. Mol. Biol. 255 (3) (1996) 494–506.
[49] H. Schwalbe, K.M. Fiebig, M. Buck, J.A. Jones, S.B. Grimshaw, A. Spencer, S.J. Glaser, L.J. Smith, C.M. Dobson, Structural and dynamical properties of a denatured protein. Heteronuclear 3D NMR experiments and theoretical simulations of lysozyme in 8 M urea, Biochemistry 36 (29) (1997) 8977–8991.
[50] O. Zhang, J.D. Forman-Kay, NMR studies of unfolded states of an SH3 domain in aqueous solution and denaturing conditions, Biochemistry 36 (13) (1997) 3959–3970.
[51] Y. Bai, J. Chung, H.J. Dyson, P.E. Wright, Structural and dynamic characterization of an unfolded state of poplar apo-plastocyanin formed under nondenaturing conditions., Protein Sci. 10 (5) (2001) 1056–1066.
[52] J. Yao, J. Chung, D. Eliezer, P.E. Wright, H.J. Dyson, NMR structural and dynamic characterization of the acid-unfolded state of apomyoglobin provides insights into the early events in protein folding, Biochemistry 40 (12) (2001) 3561–3571.
[53] D. Shortle, The expanded denaturated state: an ensemble of conformations trapped in a locally encoded topological space, Adv. Protein Chem. 62 (2002) 1–23.
[54] S. Ohnishi, A.L. Lee, M.H. Edgell, D. Shortle, Direct demonstration of structural similarity between native and denatured eglin C, Biochemistry 43 (14) (2004) 4064–4070.
[55] H.J. Dyson, P.E. Wright, Unfolded proteins and protein folding studied by NMR, Chem. Rev. 104 (8) (2004) 3607–3622.

[56] P. Bernadó, L. Blanchard, P. Timmins, D. Marion, R.W.H. Ruigrok, M. Blackledge, A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 17002–17007.

[57] P. Bernadó, C.W. Bertoncini, C. Griesinger, M. Zweckstetter, M. Blackledge, Defining long-range order and local disorder in native α-synuclein using residual dipolar couplings, J. Am. Chem. Soc. 127 (2005) 17968–17969.

[58] M.D. Mukrasch, P. Markwick, J. Biernat, M. von Bergen, P. Bernadó, C. Griesinger, E. Mandelkow, M. Zweckstetter, M. Blackledge, Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation, J. Am. Chem. Soc. 129 (16) (2007) 5235–5243.

[59] P. Bernadó, E. Mylonas, M.V. Petoukhov, M. Blackledge, D.I. Svergun, Structural characterization of flexible proteins using small-angle X-ray scattering, J. Am. Chem. Soc. 129 (17) (2007) 5656–5664.

[60] C. Gerum, R. Silvers, J. Wirmer-Bartoschek, H. Schwalbe, Unfolded-state structure and dynamics influence the fibril formation of human prion protein, Angew. Chem. Int. Ed. Engl. 48 (50) (2009) 9452–9456.

[61] M.K. Cho, G. Nodet, H.Y. Kim, M.R. Jensen, P. Bernadó, C.O. Fernandez, S. Becker, M. Blackledge, M. Zweckstetter, Structural characterization of alpha-synuclein in an aggregation prone state, Protein Sci. 18 (9) (2009) 1840–1846.

[62] M.R. Jensen, L. Salmon, G. Nodet, M. Blackledge, Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts, J. Am. Chem. Soc. 132 (4) (2010) 1270–1272.

[63] R. Silvers, F. Sziegat, H. Tachibana, S. Segawa, S. Whittaker, U.L. Günther, F. Gabel, J.R. Huang, M. Blackledge, J. Wirmer-Bartoschek, H. Schwalbe, Modulation of structure and dynamics by disulfide bond formation in unfolded states, J. Am. Chem. Soc. 134 (15) (2012) 6846–6854.

[64] F. Sziegat, R. Silvers, M. Hähnke, M.R. Jensen, M. Blackledge, J. Wirmer-Bartoschek, H. Schwalbe, Disentangling the coil: modulation of conformational and dynamic properties by site-directed mutation in the non-native state of hen egg white lysozyme, Biochemistry 51 (16) (2012) 3361–3372.

[65] M. Fuxreiter, I. Simon, P. Friedrich, P. Tompa, Preformed structural elements feature in partner recognition by intrinsically unstructured proteins, J. Mol. Biol. 338 (5) (2004) 1015–1026.

[66] H.-Y. Kim, H. Heise, C.O. Fernandez, M. Baldus, M. Zweckstetter, Correlation of amyloid fibril β-structure with the unfolded state of α-synuclein, ChemBioChem 8 (14) (2007) 1671–1674.

[67] N. Rezaei-Ghaleh, E. Andreetto, L.-M. Yan, A. Kapurniotu, M. Zweckstetter, Interaction between amyloid beta peptide and an aggregation blocker peptide mimicking islet amyloid polypeptide, PLOS ONE 6 (2011) e20289.

[68] R. Schweitzer-Stenner, Conformational propensities and residual structures in unfolded peptides and proteins, Mol. BioSyst. 8 (1) (2012) 122–133.

[69] M. Moradi, V. Babin, C. Sagui, C. Roland, A statistical analysis of the PPII propensity of amino acid guests in proline-rich peptides, Biophys. J. 100 (2011) 1083–1093.

[70] M. Moradi, V. Babin, C. Sagui, C. Roland, PPII propensity of multiple-guest amino acids in a proline-rich environment, J. Phys. Chem. B 115 (2011) 8645–8656.

[71] M. Moradi, V. Babin, C. Roland, C. Sagui, Are long-range structural correlations behind the aggregation phenomena of polyglutamine diseases? PLoS Comput. Biol. 8 (2012) e1002501.

[72] Y. Zhang, C. Sagui, The $gp41_{659-671}$ hiv-1 antibody epitope: a structurally challenging small peptide, J. Phys. Chem. B 118 (1) (2014) 69–80.

[73] Z. Biron, S. Khare, A.O. Samson, Y. Hayek, F. Naider, J. Anglister, A monomeric 3_{10}-helix is formed in water by a 13-residue peptide representing the neutralizing determinant of hiv-1 on gp41, Biochemistry 41 (42) (2002) 12687–12696.

[74] G. Barbato, E. Bianchi, P. Ingallinella, W.H. Hurni, M.D. Miller, G. Cilibertol, R. Cortese, R. Bazzo, J.W. Shiver, A. Pessi, Structural analysis of the epitope of the anti-HIV antibody 2F5 sheds light into its mechanism of neutralization and HIV fusion, J. Mol. Biol. 330 (2003) 1101–1115.

[75] P.R. Tulip, C.R. Gregor, R.Z. Troitzsch, G.J. Martyna, E. Cerasoli, G. Tranter, J. Crain, Conformational plasticity in an HIV-1 antibody epitope, J. Phys. Chem. B 114 (23) (2010) 7942–7950.

[76] Z. Ahmed, S.A. Asher, UV resonance Raman investigation of a 3_{10}-helical peptide reveals a rough energy landscape, Biochemistry 45 (30) (2006) 9068–9073.

[77] G. Ofek, M. Tang, A. Sambor, H. Katinger, J.R. Mascola, R. Wyatt, P.D. Kwong, Structure and mechanistic analysis of the anti-human immunodeficiency virus type 1 antibody 2F5 in complex with its gp41 epitope, J. Virol. 78 (2004) 10724–10737.

[78] G. Ofek, F.J. Guenaga, W.R. Schief, J. Skinner, D. Baker, R. Wyatt, P.D. Kwong, Elicitation of structure-specific antibodies by epitope scaffolds, Proc. Natl. Acad. Sci. U. S. A. 107 (42) (2010) 17880–17887.

[79] A.M.T. Martins Do Canto, A.J.P. Carvalho, J.P.P. Ramalho, L.M.S. Loura, T-20 and T-1249 HIV fusion inhibitors' structure and conformation in solution: a molecular dynamics study, J. Pept. Sci. 14 (4) (2008) 442–447.

[80] C.R. Gregor, E. Cerasoli, P.R. Tulip, M.G. Ryadnov, G.J. Martyna, J. Crain, Autonomous folding in the membrane proximal HIV peptide $gp41_{659-671}$: pH tuneability at micelle interfaces, Phys. Chem. Chem. Phys. 13 (2011) 127–135.

[81] M.F. Perutz, B.J. Pope, D. Owen, E.E. Wanker, E. Scherzinger, Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of sup35 and of the amyloid β-peptide of amyloid plaques, Proc. Natl. Acad. Sci. U. S. A. 99 (8) (2002) 5596–5600.

[82] V.N. Uversky, Amyloidogenesis of natively unfolded proteins, Curr. Alzheimer Res. 5 (3) (2008) 260–287.

[83] S. Alberti, R. Halfmann, O. King, A. Kapila, S. Lindquist, A systematic survey identifies prions and illuminates sequence features of prionogenic proteins, Cell 137 (1) (2009) 146–158.

[84] S. Lifson, A.T. Hagler, P. Dauber, Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 1. Carboxylic acids, amides, and the c═o···h– hydrogen bonds, J. Am. Chem. Soc. 101 (18) (1979) 5111–5121.

[85] C.M. Wilmot, J.M. Thornton, β-Turns and their distortions: a proposed new nomenclature, Protein Eng. 3 (6) (1990) 479–493.

[86] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, Comparison of multiple amber force fields and development of improved protein backbone parameters, Proteins: Struct., Funct., Bioinform. 65 (3) (2006) 712–725.

[87] A. Onufriev, D. Bashford, D.A. Case, Modification of the generalized born model suitable for macromolecules, J. Phys. Chem. B 104 (15) (2000) 3712–3720.

[88] J. Weiser, P.S. Shenkin, W.C. Still, Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO), J. Comput. Chem. 20 (2) (1999) 217–230.

[89] A. Onufriev, D. Bashford, D.A. Case, Exploring protein native states and large-scale conformational changes with a modified generalized born model, Proteins: Struct., Funct., Bioinform. 55 (2) (2004) 383–394.

[90] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, Comparison of simple potential functions for simulating liquid water, J. Chem. Phys. 79 (1983) 926–935.

[91] T. Darden, D. York, L. Pedersen, Particle mesh Ewald: an N log (N) method for Ewald sums in large systems, J. Chem. Phys. 98 (1993) 10089–10092.

[92] U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee, L.G. Pedersen, A smooth particle mesh Ewald method, J. Chem. Phys. 103 (1995) 8577–8593.

[93] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. Di Nola, J.R. Haak, Molecular dynamics with coupling to an external bath, J. Chem. Phys. 81 (1984) 3684–3690.

[94] Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding, Chem. Phys. Lett. 314 (1999) 141.

[95] G.N. Ramachandran, V. Sasisekharan, Conformation of polypeptides and proteins, Adv. Protein Chem. 23 (1968) 283–438.

[96] T.E. Creighton, Proteins: Structures and Molecular Properties, 2nd ed., W. H. Freeman, New York, 1992.

[97] Jena Library of Biological Macromolecules, Structural Elements of Proteins, 2014 http://jenalib.flileibniz.de/ImgLibDoc/prot_struc/IMAGE_PROT_ELEMENTS.html (accessed 14.05.14).

[98] J. Pitts, C. Sansom, The Principles of Protein Structure, 2014 http://www.cryst.bbk.ac.uk/PPS2/course/section8/ss-960531_7.html#HEADING6 (accessed 14.05.14).

[99] L. Banci, F. Cantini, M. Cevec, H.R.A. Jonker, S. Nozinovic, C. Richter, H. Schwalbe, NMR of Biomolecules: Towards Mechanistic Systems Biology, Wiley-Blackwell, 2012, pp. 7–32, Ch. 2. Structure of Biomolecules: Fundamentals.

[100] R.V. Pappu, R. Srinivasan, G.D. Rose, The flory isolated-pair hypothesis is not valid for polypeptide chains: Implications for protein folding, Proc. Natl. Acad. Sci. U. S. A. 97 (23) (2000) 12565–12570.

[101] P.J. Flory, Statistical Mechanics of Chain Molecules, 1st ed., Interscience, New York, 1969.

[102] C. Tanford, Protein denaturation, Adv. Protein Chem. 23 (1968) 121–283.

[103] N.C. Fitzkee, G.D. Rose, Reassessing random-coil statistics in unfolded proteins, Proc. Natl. Acad. Sci. U. S. A. 101 (34) (2004) 12497–12502.

[104] R. Zwanzig, A. Szabo, B. Bagchi, Levinthal's paradox, Proc. Natl. Acad. Sci. U. S. A. 89 (1) (1992) 20–22.

[105] R. Srinivasan, G.D. Rose, Methinks it is like a folding curve, Biophys. Chem. 101-102 (2002) 167–171.