Journal of Magnetic Resonance 202 (2010) 223–233

Contents lists available at ScienceDirect

Journal of Magnetic Resonance

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

DANGLE: A Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure

Ming-Sin Cheung, Mahon L. Maguire, Tim J. Stevens, R. William Broadhurst *

Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK

ARTICLE INFO

Article history:
Received 12 August 2009
Revised 4 November 2009
Available online 16 December 2009

Keywords:
Bioinformatics
Chemical shifts
Dihedral angle prediction
NMR spectroscopy
Protein structure
Secondary structure prediction

ABSTRACT

This paper introduces DANGLE, a new algorithm that employs Bayesian inference to estimate the likelihood of all possible values of the backbone dihedral angles ϕ and ψ for each residue in a query protein, based on observed chemical shifts and the conformational preferences of each amino acid type. The method provides robust estimates of ϕ and ψ within realistic boundary ranges, an indication of the degeneracy in the relationship between shift measurements and conformation at each site, and faithful secondary structure state assignments. When a simple degeneracy-based filtering procedure is applied, DANGLE offers an ideal compromise between accuracy and coverage when compared with other shift-based dihedral angle prediction methods. In addition, per residue analysis of shift/structure degeneracy has potential to be a useful new approach for studying the properties of unfolded proteins, with sufficient sensitivity to identify regions of residual structure in the acid denatured state of apomyoglobin.

1. Introduction

Early solution state NMR studies demonstrated that resonance frequencies are profoundly influenced by the local environments created by protein secondary, tertiary and quaternary structure [1]. Although isotropic chemical shift measurements promise to reveal much, their dependence on structure is not straightforward, complicated by the influence of many atoms in the protein system on the electronic environment around each nucleus. Recent advances have begun to elucidate the complex relationship between shift and conformation [2], from low resolution attempts to classify elements of secondary structure [3–10], through prediction of the backbone dihedral angles ϕ and ψ [11–14], to the generation of high resolution protein structures solely from chemical shift and primary sequence information [15–20]. Blind determination of full three dimensional structures is currently restricted to smaller proteins (<150 amino acids), is less accurate than traditional NOE-based NMR methods [15–17] and can fail in regions where backbone chemical shift measurements are not available, for example due to the effects of conformational exchange broadening in exposed loops. Such gaps can be addressed by additional modelling procedures [20], but the resolution of *de novo* methods is also limited by the accuracy of algorithms for predicting backbone dihedral angles from chemical shift measurements [15–17] and for back-calculating shifts from atomic coordinates [21,22].

Conventional structure calculation protocols typically combine inter-atomic distance measurements with secondary structure and dihedral angle restraints, aiming to improve the convergence and resolution of the final ensemble. If sufficient long-range distance constraints have been collected, the additional information contributed by dihedral angle restraints may be small [23]. However, when experimental data is sparse or distance restraints are highly ambiguous (e.g. for solid state NMR studies), shift-based dihedral angle restraints have been shown to increase the precision and accuracy of the resulting structures [24]. Constraints that can define local backbone conformations therefore remain valuable in many day-to-day applications. In addition, the main shift-based structure determination methods, CHESHIRE [15], CS-ROSETTA [16] and CS23D [17], all construct tertiary folds using fragments that are initially identified using dihedral angle prediction algorithms.

The development of methods for using chemical shift data to predict protein conformation has been facilitated by access to extensive databases of experimental shift measurements [25,26] and of protein structures [27,28]. TALOS, the most popular technique for estimating dihedral angles, searches for tripeptide fragments with amino acid sequence and secondary shift patterns that are similar to the query protein, assuming that close matches from a database will possess related backbone conformations [11]. The PREDITOR approach supplements a fragment-matching algorithm with information derived from homologous protein structures [12,13]. Both methods use Ramachandran plots [29] to analyse the backbone conformations of the 10 closest matching fragments, deriving shift-based predictions of ϕ and ψ from the

* Corresponding author. Fax: +44 1223 776002.
E-mail addresses: r.w.broadhurst@bioc.cam.ac.uk (R.W. Broadhurst).

1090-7807/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2009.11.008

mean values of hits within the major cluster, while ignoring contributions from outliers. These procedures bias the final predictions towards regions of Ramachandran space that are highly populated. In our hands both TALOS and PREDITOR are regularly inaccurate for residues in non-canonical structures, such as 3_{10} helices and conformations with positive values of ϕ, in part because neither approach can reliably handle glycine or residues that precede prolines. The recently updated TALOS+ package can make acceptable estimates for a greater proportion of residues [14], but all three methods return boundary ranges for ϕ and ψ that typically fail to reflect the accuracy of the prediction.

Steric hindrance and electrostatic interactions restrict the sampling of conformational space so that the backbone dihedral angles measured in protein structures cluster together in distinct regions of the Ramachandran plot [30]. Densely populated regions correspond to the low energy conformations found in common elements of secondary structure, most significantly right-handed α-helices, left-handed α_L-turns and extended β-strands. TALOS, TALOS+ and PREDITOR make scant use of this predefined stereochemistry, which also dictates that different amino acid types possess distinct population distributions in Ramachandran space. Analysis of high resolution X-ray structures has shown that it is convenient to consider these distributions in four classes [31,32]: glycines, prolines, residues that precede proline, and the most frequently encountered “generic” class (Fig. 1A–D).

To address the problems encountered with popular dihedral angle prediction methods, here we introduce the DANGLE (Dihedral ANgles from Global Likelihood Estimates) algorithm, which uses Bayesian inference to estimate the likelihood of conformations throughout Ramachandran space, paying explicit attention to the population distributions expected for different residue types. We describe a straightforward method for identifying residues with chemical shifts that are consistent with multiple conformations. Filtering out the estimates from such sites yields predictions of ϕ and ψ that are more realistic than those made by TALOS, TALOS+ or PREDITOR, with significant improvements for glycine and pre-proline residues. We also introduce and assess a new tool for assigning secondary structure states to protein residues by analysis of conformations in similar database fragments.

2. Experimental

2.1. The fragment and prior data sets

A database of protein structure fragments was compiled using chemical shift and PDB files for 186 proteins taken from the TALOS database (version 2007.068.09.07) [11]. The secondary chemical shifts of ^{1}H$^\alpha$, ^{15}N, ^{13}C′, ^{13}C$^\alpha$ and ^{13}C$^\beta$ nuclei were obtained by subtracting sequence-corrected random coil shifts from deposited values [33,34]. 27,014 five-residue fragments were generated, of which 772 were discarded due to sequence discrepancies between the original PDB and chemical shift files. The fragment database also stores the primary sequence of each pentapeptide, the experimental ϕ and ψ angles and a three-state classification of the secondary structure of the central residue. Each residue was assigned using output from the DSSP program [35]: α-helix, 3_{10} helix and π-helix states were grouped into the “H” class; β-bridge and extended strand states into the “E” class; and all remaining states into the “C” class.

Prior information about experimental backbone conformations was collated from a data set of 500 high resolution X-ray structures assembled by Lovell and colleagues [31]. Residues were grouped into four classes (generic, glycine, proline and pre-proline) and atoms with B-factors >30 were excluded. Normalised dihedral angle distributions were stored in the form of population frequencies within 36×36 10° square bins spanning all of Ramachandran space; to ensure that uncommon (ϕ, ψ) combinations could be sampled, a small pseudo-count was assigned to each bin before the experimental populations were introduced.

The fragment database was shown to comprise conformations that are representative of high-quality crystal structures, as judged by comparison with the prior data set using pairwise symmetrized Jensen-Shannon divergence scores [36,37] (see Supplementary Information for more details).

2.2. The query scatter pattern

After reading in the chemical shift data and amino acid sequence of the query protein, DANGLE computes sequence-corrected secondary shifts for all measured ^{1}H$^\alpha$, ^{15}N, ^{13}C′, ^{13}C$^\alpha$ and ^{13}C$^\beta$ nuclei [33,34] and then makes dihedral angle predictions for each residue. First, the fragment database is searched to find close matches to the shifts and sequence of each five-residue window along the query polypeptide chain, using a scoring function adapted from TALOS [11] (see Supplementary Information for more details). If no chemical shift information is available within the query window, similar fragments are identified by sequence alone. For each query window, the 10 lowest scoring matches from the fragment database define a scatter pattern of (ϕ, ψ) coordinates, which is transformed into a frequency matrix of 36×36 10° bins and then smoothed (see Supplementary Information for more details). This contrasts with the approach taken by TALOS and TALOS+, which use shorter tripeptide segments for fragment matching and different cluster analysis protocols [11,14].

2.3. The posterior probability scoring function

According to Baye’s theorem [38], the posterior probability that a particular backbone conformation (ϕ, ψ) can be deduced from a given query scatter pattern (QSP) is $P(\phi, \psi|QSP)$, where:

$$P(\phi, \psi|QSP) \propto P(QSP|\phi, \psi) \times P(\phi, \psi).$$

The prior probability $P(\phi, \psi)$, defining the chance that the central residue in the query window possesses ϕ and ψ angles within a particular 10° square bin in Ramachandran space, is estimated using the four dihedral angle prior distributions described above.

$P(QSP|\phi, \psi)$ represents the conditional probability that a conformation within a particular 10° square bin could produce the observed query scatter pattern. To estimate $P(QSP|\phi, \psi)$, a set of “scattergrams” was assembled from shift and sequence information in the fragment database. For each occurrence of experimentally determined ϕ and ψ angles within a particular bin, a similarity search identified the 10 lowest scoring pentapeptide matches to construct a scatter pattern distribution for the central residue; the distributions for each occurrence were then summed together and the final distribution was normalised. The resulting scattergram defines the shape of the scatter patterns that can be produced by conformations from within the bin. Scattergrams were generated for each of 36×36 bins in Ramachandran space.

A query scatter pattern will be representative of an angle located within a given 10° square bin if it resembles the scattergram associated with that bin. DANGLE quantifies the difference between the query frequency matrix and scattergram distributions for bin (ϕ, ψ) using a likelihood ratio known as the Kullback–Leibler divergence [39], $D_{\mathrm{KL}}(\phi, \psi)$:

$$D_{\mathrm{KL}}(\phi, \psi) = \sum_{\phi'} \sum_{\psi'} Q(\phi', \psi') \times \ln\left(\frac{Q(\phi', \psi')}{S(\phi', \psi'|\phi, \psi)}\right),$$

where $Q(\phi', \psi')$ is the value of cell (ϕ', ψ') in the query frequency matrix and $S(\phi', \psi'|\phi, \psi)$ is the value of cell (ϕ', ψ') in the scatter-

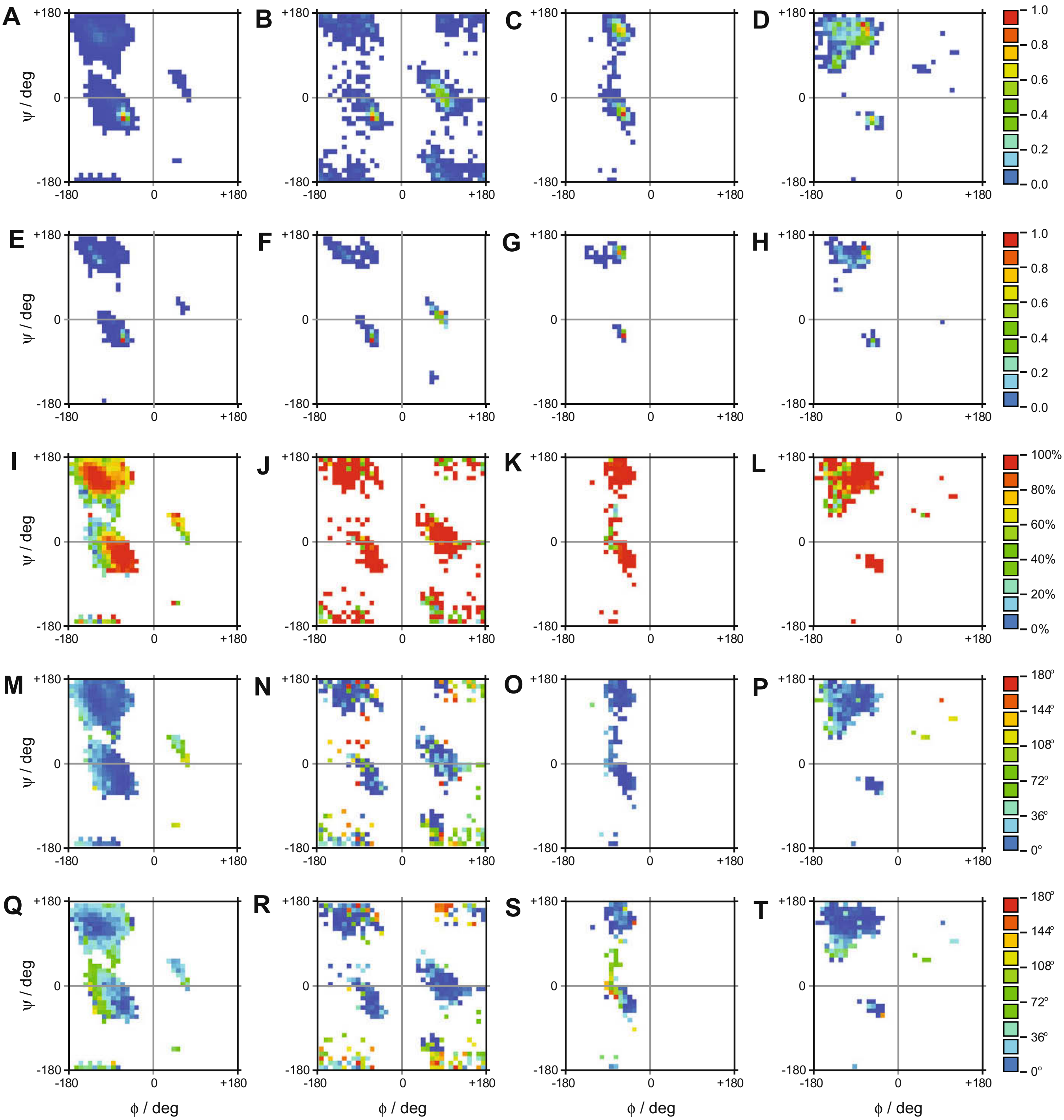

Fig. 1. Ramachandran space distribution plots are shown for generic (A, E, I, M and Q), glycine (B, F, J, N and R), proline (C, G, K, O and S) and pre-proline (D, H, L, P and T) sites in the 186 proteins that constitute the fragment database. Panels (A–D) are population frequency distributions for reference structure backbone angles, colour coded from blue (least populous) to red (most populous). Panels (E–H) display population frequency distributions for the angles predicted by DANGLE after filtering out ambiguous sites. Panels (I–L) display the percentage of predictions of ϕ and ψ that are both within 30° of their reference angles ($A_{30}(\phi,\psi)$). Panels (M–P) display the mean RMS errors for predictions of ϕ. Panels (Q–T) display the mean RMS errors for predictions of ψ. Bins with population $\leqslant 0.2\%$ of the most populous within the plot are shown as white background. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

gram matrix associated with bin (ϕ,ψ). $D_{\mathrm{KL}}(\phi,\psi)$ is zero if the query frequency matrix and the current scattergram are a perfect match; large positive values indicate that the two distributions are highly dissimilar.

Because $D_{\mathrm{KL}}(\phi,\psi)$ behaves as an inverse function of similarity, the conditional probability $P(QSP|\phi,\psi)$ is derived using a normalised similarity score:

$$P(QSP|\phi,\psi) = \frac{\exp(-D_{\mathrm{KL}}(\phi,\psi))}{\sum_{\phi'}\sum_{\psi'}\exp(-D_{\mathrm{KL}}(\phi',\psi'))},$$

such that

$$\sum_{\phi}\sum_{\psi}P(QSP|\phi,\psi) = 1.$$

Finally, $B(\phi, \psi)$, the posterior probability score for bin (ϕ, ψ) is determined from the equation:

$$B(\phi,\psi) = \frac{\exp(-D_{\mathrm{KL}}(\phi,\psi))}{\sum_{\phi'}\sum_{\psi'}\exp(-D_{\mathrm{KL}}(\phi',\psi'))} \times P(\phi,\psi).$$

2.4. Global likelihood estimate diagrams

A global likelihood estimate (GLE) diagram is assembled from the $B(\phi, \psi)$ values of each bin in Ramachandran space, normalised with respect to the largest score in the diagram, $B_{\max}$. A cluster of adjacent bins with $B(\phi, \psi)/B_{\max}$ greater than an empirically optimised threshold of 1.5×10^{-5} is designated an "island", signifying a range of similar possible conformations. The island that contains the $B_{\max}$ value is termed the principal island. DANGLE determines the B-weighted angular means of ϕ and ψ within the principal island and reports these as the prediction of the backbone conformation (see Supplementary Information for more details). The locus of 10° square bins adjacent to the boundary of the principal island defines upper and lower limits for the predicted values of ϕ and ψ. DANGLE can be configured to reject predictions from GLE diagrams that contain more than a user-defined number of islands; the default approach is to use estimates from single-island sites only.

2.5. Jury-based predictions of secondary structure class

DANGLE also predicts conformational information at lower resolution by attributing a secondary structure class to the query residue. The DSSP-derived secondary structure classifications [35] of the 10 lowest scoring pentapeptides from a similarity search are analysed using a simple jury system: if $\geqslant 6$ of the fragments are in the H state, the query residue is assigned to the H (helix) class; if $\geqslant 6$ are in the E state, the E (strand) class is predicted; otherwise, the C (coil) class is returned.

2.6. Implementation and availability

The DANGLE prediction program and its graphical user interface (GUI) are coded in Python and tested with Python versions 2.4–2.6. Both have been publicly released in two forms: as a stand-alone package (http://dangle.sourceforge.net) and as an integrated tool within version 2.0 (and above) of the CcpNmr Analysis spectrum visualisation and assignment program (http://www.ccpn.ac.uk/) [40]. For details of input and output file formats, see Supplementary Information. All development procedures were performed with a machine running SuSe Linux version 10.0 with a 1.6 GHz AMD processor and 512 MB of memory, on which DANGLE took 2 s per residue to make predictions for a query protein.

2.7. Self-assessment tests and comparison with other algorithms

Self-assessment tests were performed by analysing the 186 protein entries in the fragment database, omitting the query protein from the database in each case. $A_{30}(\phi)$, $A_{30}(\psi)$ and $A_{30}(\phi, \psi)$ denote the percentage of predictions of ϕ, ψ and (ϕ, ψ) pairs, respectively, that deviate from the structure-derived reference by less than 30°. Similarly, $A_{\mathrm{range}}(\phi)$, $A_{\mathrm{range}}(\psi)$ and $A_{\mathrm{range}}(\phi, \psi)$ represent the percentage of estimates for which the reference angle lies within the predicted boundary range. The percentage of predictions of three-state secondary structure class (H, E or C) that are identical to the class in the reference structure is represented by A_3.

To compare the performance of DANGLE with other algorithms, we assembled an independent test set of geometric and shift data for 2870 residues from 29 non-redundant proteins (see Supplementary Information for more details). Two sets of DANGLE output were investigated: DANG-A, containing predictions for every site, and DANG-B, retaining predictions from sites that return single-island GLE plots only. Version 2007.068.09.07 of TALOS [11] was assessed in two modes: TALOS-A, representing all predictions made by the algorithm; and TALOS-B, including only estimates that are classed as "good". Version 1.2009.0618.13 of TALOS+ [14] was evaluated using two similarly constructed data classes, TPLUS-A (all) and TPLUS-B ("good" predictions only). The performance of the PREDITOR server (database version DB1.0, accessed in September 2007) was also judged in two modes: predictions based on shift and sequence data only (PRED-A); or with additional dihedral angle information from homologous structures (PRED-B) [13]. Values of ϕ and ψ were also estimated on the basis of sequence alone using the Real-SPINE 3.0 server [41].

To study the effects of incorrect referencing, for each of the 29 proteins in the test set the ^{1}H$^{\alpha}$, ^{15}N and the entire set of ^{13}C chemical shift values were separately increased or reduced by intervals of 0.05, 0.5 and 0.2 ppm, respectively; a wide range of offsets was investigated, so we opted to compare the results using this smaller test group rather than the full set of 186 proteins.

Five shift-based secondary structure prediction routines were evaluated, also using the 29 protein test set: the consensus CSI approach (conCSI) [4]; PSSI (version 2) [5]; psiCSI [6]; PECAN (version 0.1 beta) [7] and TALOS+ [14]. Three-state prediction methods based on secondary shift differences between ^{13}C$^{\alpha}$ and ^{13}C$^{\beta}$ (ΔCAB) [8] and between ^{13}C' and ^{13}C$^{\beta}$ (ΔCOB) [9] nuclei were also considered, along with sequence-only prediction results from the Jpred3 web server [42]. For further details about the implementation of these methods, see Supplementary Information.

The output of DANGLE is illustrated using data for the origin binding domain of the SV40 T-antigen, using an X-ray structure (PDB entry 2FUF) [43], a re-calculated solution structure ensemble from the RECOORD database (1TBD) [28,44], and shifts from the RefDB database (Accession Number 4127). Chemical shift data for the acid denatured state of apomyoglobin was taken from BMRB entry 4676 [45]; the results from DANGLE were compared with random coil index (RCI) values returned by the PREDITOR server [17,48].

3. Results and discussion

3.1. Interpretation of GLE diagrams

The DANGLE algorithm operates by searching a database to find short peptide fragments with local amino acid sequence and chemical shifts that are similar to those of each residue in the query protein. The ϕ and ψ backbone dihedral angles of the 10 best matches for the residue are considered together, defining a "query scatter pattern" for comparison with a library of "scattergram" distributions, each of which constrains the shape of query scatter patterns that can be produced by a particular protein backbone conformation. A systematic comparison across all values of ϕ and ψ throughout Ramachandran space is then reported in the form of a grid of posterior probability scores (B-scores), termed a "global likelihood estimate" (GLE) diagram. If the distributions of the query scatter pattern and the relevant scattergram resemble each other closely, the B-score will be large; otherwise, its value will be close to zero. In the GLE diagram, an "island" describes a cluster of adjacent ϕ and ψ values with high B-scores, representing a range of backbone conformations that the residue is likely to adopt.

Fig. 2 displays examples of typical GLE diagrams obtained for the origin binding domain (OBD) of the simian virus 40 (SV40) large T-antigen [43,44]. Residues in elements of regular secondary structure usually possess a single island, as shown for Tyr-162 (located in a β-strand, Fig. 2A) and Ala-169 (in an α-helix, Fig. 2B). The

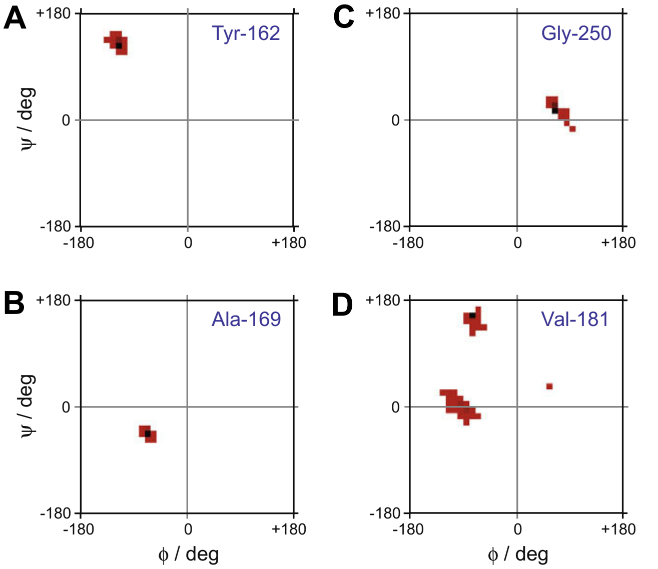

Fig. 2. Representative GLE diagrams for residues from the origin binding domain of the simian virus 40 large T-antigen: (A) Tyr-162, a β-strand conformation; (B) Ala-169, an α-helical conformation; (C) Gly-250, a left-handed turn conformation; and (D) Val-181, an ambiguous site showing three possible conformations. Bins with *B*-scores below the threshold value are in white; other bins are shaded from red (threshold value) to black (B_{max}). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

islands in both panels are compact, meaning that the experimental chemical shifts are consistent with a tightly defined set of conformations, leading to dihedral angle estimates that are both accurate and precise. For example, Tyr-162 has (ϕ, ψ) values of (−108°, +122°) in the reference crystal structure [43] and (−93 ± 5°, +112 ± 2°) in the ensemble of 25 NMR solution structures [44]. In the corresponding GLE diagram (Fig. 2A), the bin that contains the largest *B*-score is centred at (−115°, +125°). DANGLE estimates backbone dihedral angles by determining *B*-weighted mean values for ϕ and ψ across all bins within the island, which in this case yields −115° and +128°, respectively, indicating a conformation within the bin that contains the highest score. Conservative upper and lower limits for these predictions are taken from the locus of 10° square bins adjacent to the island boundary, here dictating ranges of −150° to −90° for ϕ and +100° to +160° for ψ. The experimental angles usually fall well within these boundaries.

The GLE diagram for Gly-250 also contains a single island, but in the positive ϕ region of Ramachandran space, consistent with an α_L-turn conformation (Fig. 2C). According to the NMR ensemble, the dihedral angles of Gly-250 are close to (+74°, −5°), with standard deviations of (±78°, ±37°). This is consistent with the elongated shape of the island, which results in a prediction of (+66°, +18°) within broader boundary ranges of +40° to +110° in ϕ and −30° to + 50° in ψ.

3.2. Number of islands and conformational flexibility

The three islands in the GLE diagram for Val-181 (Fig. 2D) highlight distinct regions of the (ϕ, ψ) plot that create secondary shift patterns similar to those found in the query fragment. The number of islands reflects the degeneracy of mapping between the query scatter pattern and scattergram distributions from different locations in Ramachandran space. Degeneracy of this sort may be accidental, implying that several conformations are capable of inducing similar electronic environments at multiple sites within the fragment. Alternatively, degeneracy could be related to an averaging of chemical shifts due to conformational flexibility. For Val-181, the former explanation is most likely: in the solution structure ensemble the standard deviations of ϕ and ψ are small, (−65 ± 9°, +158 ± 9°), indicating that the backbone geometry is relatively well defined, so that the residue samples a single conformation. As is often the case for multi-island plots, the experimental reference angles lie within the cluster that contains the highest *B*-score; the *B*-weighted mean values of ϕ and ψ within this principal island suggest a conformation of (−74°, +154°), similar to that found in the NMR ensemble [44].

The SV40 OBD protein is relatively well structured throughout, and consequently only 22 % of its backbone sites possess multi-island GLE diagrams. To illustrate the effects of more extensive conformational flexibility, we also analysed the chemical shifts of apomyoglobin in a predominantly unfolded denatured state, under low-salt conditions at pH 2.3 [45]. In this case, 60% of residues produce GLE plots that contain two or more islands (Fig. 3A), presumably because more nuclei sense time-averaged electronic environments due to frequent rearrangements of the polypeptide chain. Five runs of consecutive single-island GLE diagrams were detected (labelled P–T in Fig. 3), each containing (ϕ, ψ) values consistent with helix conformations. All five sections detected by DANGLE correspond to regions previously implicated in residual helical structure by a wide range of biophysical probes [45].

An alternative chemical shift-based probe of protein flexibility, the Random Coil Index (RCI) [46–49], appears to be less sensitive to the presence of these residual structures (Fig. 3B). Four of the five single-island regions identified by DANGLE correspond to residues with small RCI values, indicating greater rigidity, but the lowest RCI scores appear elsewhere in the polypeptide chain. As a result, when applied to unfolded proteins the RCI approach is likely to produce both false positives (low scoring regions that correspond to portions of the polypeptide that are flexible, such as residues 33 to 37) and false negatives (high scores in regions of residual structure, e.g. 124–126). These discrepancies probably occur because the RCI method was optimised for folded proteins using molecular dynamics simulations; it is therefore best suited for quantifying low amplitude sub-nanosecond motions, corresponding to RCI scores <0.05 [48,49]. The acid denatured state of apomyoglobin is considerably more flexible throughout, yielding no RCI scores <0.10, and likely experiences extensive high amplitude motions on the millisecond timescale [45]. This data set clearly

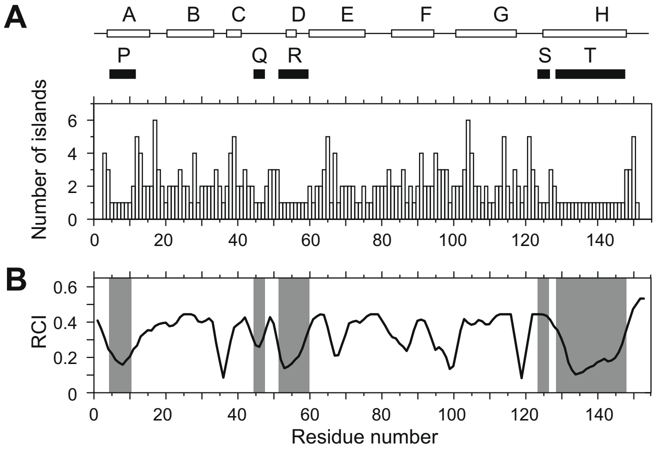

Fig. 3. Analysis of chemical shifts measured in the acid denatured state of apomyoglobin. (A) Histogram showing the number of islands in the GLE diagram of each residue. Above, helical regions of the native protein are indicated by white boxes, labelled A–H; stretches that contain $\geqslant 3$ consecutive residues with single-island GLE diagrams are indicated by black boxes, labelled P–T (B) Random coil index (RCI) values for each residue. Single island stretches identified by DANGLE are indicated by grey shading.

pushes the RCI approach beyond its limits of application but, by contrast, DANGLE remains able to provide useful insights. It should be noted that relaxation experiments provide very useful complementary data for studies of flexible protein states.

3.3. Accuracy and the rejection of ambiguous predictions

The performance of DANGLE was evaluated by analysing chemical shift data for the 186 proteins in the fragment database, using reference angles from crystal structures. The approach is broadly successful, generating $A_{30}(\phi)$ and $A_{30}(\psi)$ values (percentage of predictions within 30° of the reference) of 88% and 85%, respectively, over all cases (see Supplementary Table S5). The two dihedral angles have different chemical shift dependencies, so the $A_{30}(\phi,\psi)$ joint accuracy metric provides a more stringent test, in this case yielding a value of 80%.

To improve the accuracy of the method, it was important to identify which estimates are likely to be unreliable. Considering the four amino acid classes separately reveals that predictions are significantly more accurate for generic and proline sites, both producing $A_{30}(\phi,\psi)$ values of 82%, than for pre-proline (71%) or glycine residues (59%). These type-specific effects are correlated with the proportion of residues that returned GLE plots containing multiple islands: 16% for generic sites, and 22% for prolines, 24% for pre-proline residues and 41% for glycines.

Potentially incorrect predictions can therefore be filtered by systematic exclusion of sites with GLE diagrams that contain more than a specified number of islands (Supplementary Table S5). When all estimates derived from multi-island GLE diagrams are discarded, the proportion of remaining predictions accurate to within 30° of the reference increases to 88% (Fig. 4A). Across all residue types, application of this filter causes the root mean square (RMS) error to reduce from 30.7° to 22.1° for ϕ and from 45.8° to 31.4° for ψ. The improvement for glycine residues is particularly notable, with the $A_{30}(\phi,\psi)$ value increasing from 59% to 76% (Fig. 4A).

The cost of filtering out multi-island predictions is that 18% of all estimates are rejected. Glycine and proline residues are frequently found in flexible regions of the protein backbone [50,51], so it is not surprising that a relatively large portion of these sites should produce ambiguous predictions, most likely due to the effects of conformational averaging on chemical shifts. Some well-structured sites also produce multi-island GLE diagrams due to accidental degeneracies, as found for Val-181 in the SV40 OBD (Fig. 2D), so the rejection process is bound to throw away some valuable facts about conformation. However, for the purpose of protein structure determination, the most conservative approach is to avoid introducing false information, making it appropriate to cull all restraints that are likely to be ambiguous.

Another option would be to recognise that a multi-island GLE diagram presents information about possible alternative conformations, making it reasonable to consider the use of ambiguous dihedral angle restraints. Predictions could then be defined as being accurate if both reference angles appeared within the boundary ranges of any of the islands in the relevant GLE diagram, not just the principal cluster. In this case, more than 90% of all estimates would be classified as being accurate for both ϕ and ψ, irrespective of the level of filtering (see Supplementary Table S6). The two dimensional probability distribution provided by a multi-island GLE diagram should prove to be useful input data for inferential structure determination (ISD) procedures, which are designed to make best use of uncertain or incomplete information [52].

The Ramachandran space population distributions for post-filter predictions of ϕ and ψ (Fig. 1E–H) are somewhat contracted towards the "most favourable" conformations defined by PROCHECK [53]. From the viewpoint of the reference structure, however, the backbone angles of generic residues (Fig. 1I) are seen to be estimated with high confidence in the α-helix region near (−65°, −45°), the 3_{10}-helix region near (−80°, −30°) and the β-sheet region near (−135°, +135°). Predictions have medium reliability when the reference angles are in "additionally allowed" portions of the Ramachandran plot, such as the α_L-turn region near (+55°, +45°), but are inaccurate in the "generously allowed" and "disallowed" sections [53]. If only single-island estimates are considered, accurate predictions are generally obtained for all accessible conformations at the remaining glycine, proline and pre-proline sites (Fig. 1J–L). The mean RMS error distributions in Fig. 1M–P demonstrate that all negative values of ϕ are estimated with high accuracy and precision. In comparison, predictions of ψ have low RMS error values only in the most favoured regions and show medium reliability elsewhere (Fig. 1Q–T).

For DANGLE, the reliability of dihedral angle estimation is not dictated predominantly by the population of conformations in the fragment database. For example, the bin centred at (−65°, −45°) in the α-helix region of Fig. 1I contains information from 2318 query fragments, while the (−135°, +135°) bin in the β-sheet region summarises results from only 176 predictions. Both sets of predictions are highly accurate, giving $A_{30}(\phi,\psi)$ values of 99.4% and 97.2%, respectively. This behaviour contrasts with the secondary structure dependence of the PREDITOR approach, which reportedly performs better for residues in α- rather than β-structure [13].

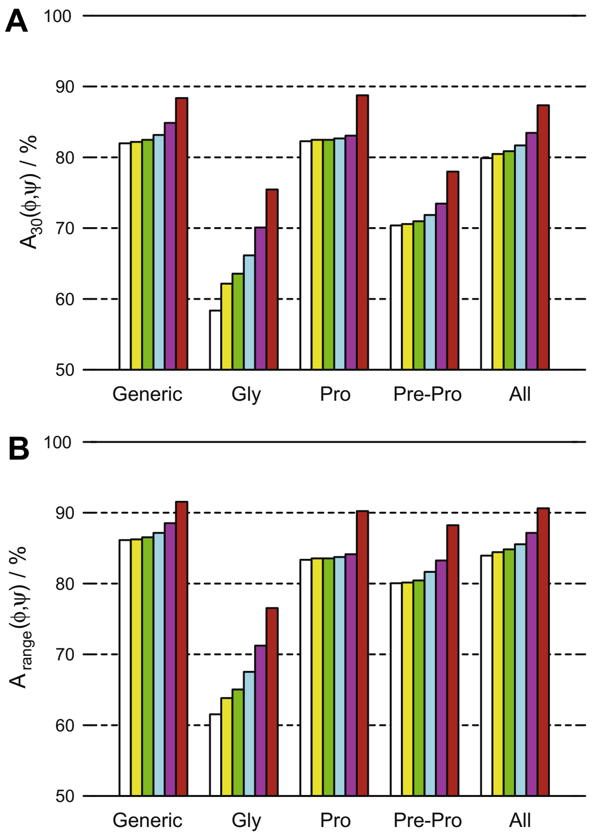

Fig. 4. Histograms of accuracy indices, showing generic, glycine, proline, pre-proline and all residue types together. (A) $A_{30}(\phi,\psi)$, the percentage of predictions of ϕ and ψ both within 30° of their reference angles; and (B) $A_{\rm range}(\phi,\psi)$, the percentage reference angles that lie within the predicted range for both ϕ and ψ. Bars shaded in white: all predictions are considered; in yellow: predictions from GLE diagrams with >5 islands are rejected; in green: rejecting >4 islands; in cyan: rejecting >3 islands; in magenta: rejecting >2 islands; in red: rejecting all multi-island predictions. Results are from a self-assessment test on 186 proteins from the fragment database. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

3.4. Boundaries for dihedral angle predictions

All conformations within an island have significant probability of producing a scatter pattern that matches the query frequency matrix. The width of the principal island in the ϕ (or ψ) dimension should therefore be related to the uncertainty of the dihedral angle prediction. An empirical search led us to set upper and lower boundaries for each dihedral angle based on the maximum width of the principal island extended by 10° at either side, limiting the minimum boundary range to 30°. Across all residue types and when no predictions were discarded, this approach yields values of 92%, 88% and 84% for $A_{\mathrm{range}}(\phi)$, $A_{\mathrm{range}}(\psi)$ and $A_{\mathrm{range}}(\phi, \psi)$, respectively (see Supplementary Table S7). When multi-island estimates are rejected, these accuracy indices increase to 95%, 93% and 91%, respectively (Fig. 4B).

A major objective for the development of DANGLE was to provide each prediction with a measure of uncertainty that bears a realistic relationship to the error in the estimate. After filtering the list of predictions, the ranges returned for ϕ and ψ are typically similar (with mean values of $62 \pm 20°$ and $61 \pm 13°$, respectively) and are only weakly dependent on conformation. On average, boundary ranges are narrower for prolines (51° for ϕ and 56° for ψ), but broader for glycines (61° and 64°). Fig. 5 shows clear correlations between the mean deviation of predicted angles from the reference structure and the boundary ranges selected for ϕ and ψ, confirming that a broad range indicates a realistic degree of uncertainty in the estimate. Outliers are found at the narrowest range (30°) for both dihedral angles (Fig. 5), but these are probably sampling errors due to infrequent occurrence.

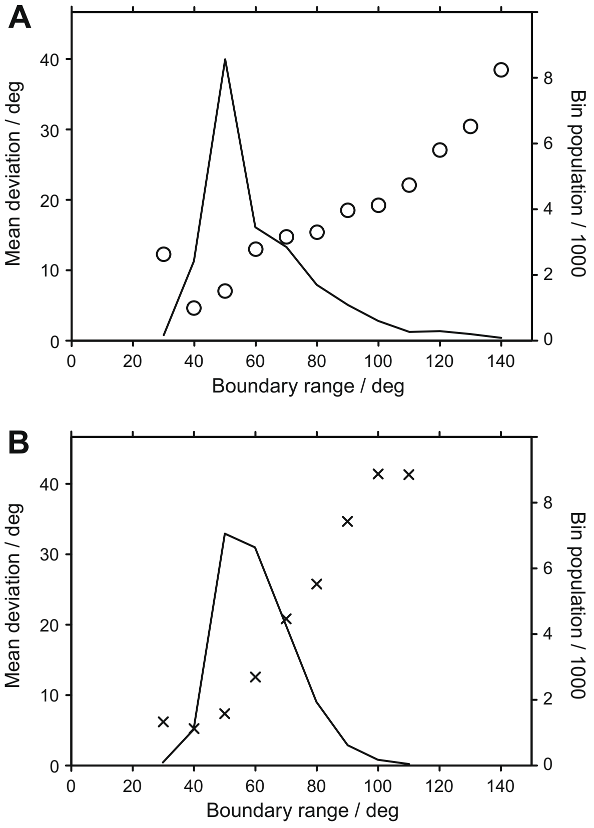

Fig. 5. Scatter plots illustrating the relationship between the boundary range predicted by DANGLE and the mean deviation of estimates from experimental reference angles for (A) ϕ (circles) and (B) ψ (crosses), after filtering out ambiguous sites. Data points were grouped into 10° bins according to the boundary range and results for groups that contained <100 predictions are not shown. The solid line indicates the population frequency of each bin. Results are from a self-assessment test on 186 proteins from the fragment database.

Until recently, ensembles of NMR protein structures had a reputation for being more precise than they were accurate. This problem has largely been resolved, mainly due to improved structure calculation procedures, such as a final refinement step with a realistic potential in the presence of explicit solvent molecules [28]. To sustain this improvement, it is important that experimental data should be applied cautiously, within boundary ranges that are wide enough to avoid distorting favourable geometries. Our results show that DANGLE estimates the majority of dihedral angles with acceptable precision, particularly those in the most favoured and additionally allowed regions of the Ramachandran plot, and that these restraints are defined within realistic boundary ranges.

3.5. Referencing errors and systematic absences

Chemical shift measurements may be incorrectly referenced for varied experimental reasons, such as omitting a suitable internal reference compound, interactions between reference compounds and sample constituents, uncompensated Bloch-Siegert shifts, or secondary isotope shifts in deuterated proteins [26,54–56]. We assessed the impact of systematic referencing errors on the accuracy of predictions by separately adjusting the shifts of ^{1}H, ^{15}N or all ^{13}C nuclei for our 29 protein test set. Dihedral angle predictions were relatively insensitive to ^{1}H referencing errors of up to 0.5 ppm or ^{15}N shift errors $\leqslant$5.0 ppm; by contrast, systematic errors in ^{13}C shift measurements $\geqslant$0.5 ppm degraded the performance of DANGLE significantly (see Supplementary Figure S5). In part, this sensitivity may be because systematic errors in ^{13}C shifts were introduced for all C$^{\alpha}$, C$^{\beta}$ and C$'$ sites simultaneously, producing a larger effect than the incorrect referencing of just a single nucleus type. Nevertheless, our results illustrate the importance of correct ^{13}C referencing for shift-based structure prediction methods.

The consequences of excluding all shift measurements for selected types of nucleus were also investigated, using the 186 protein entries in the fragment database (see Supplementary Information for more details). When all predictions are considered, systematic omission of data for any single type lowers the $A_{30}(\phi, \psi)$ index slightly: by only 0.2% if ^{1}H$^{\alpha}$ nuclei are ignored; to 3.3% in the absence of ^{13}C$^{\alpha}$ shifts. DANGLE is therefore well suited for studies of proteins that lack ^{1}H shift measurements, for example due to uniform deuteration at methylene and methine sites, or when using ^{13}C-based detection for solid state experiments. When additional information is ignored, the accuracy continues to degrade: using only ^{13}C$^{\alpha}$ shifts gives an $A_{30}(\phi, \psi)$ of 73%, while other single nucleus types produce values between 65% and 67%. These results suggest that ^{13}C$^{\alpha}$ shifts encode the most information about backbone structure and are key to making predictions that are as accurate as possible.

Once again, the reliability of DANGLE improves considerably when the output is filtered, even when few nucleus types were taken into account. For example, the use of only ^{1}H$^{\alpha}$ data gives an $A_{30}(\phi, \psi)$ accuracy index of 65%. When the fraction of these that possess GLE diagrams with multiple islands (38%) are excluded, 75% of the remaining estimates are within 30° of the reference for both ϕ and ψ. If ^{15}N assignments are also available, the $A_{30}(\phi, \psi)$ index increases to 71% when all predictions are retained, rising to 77% after multi-island filtering. Some screening protocols utilise unlabelled or ^{15}N-labelled protein samples; if sequence-specific assignments are also at hand, DANGLE still has the potential to be a useful tool.

The algorithm is remarkably successful even when all chemical shift information is discarded, yielding an $A_{30}(\phi, \psi)$ accuracy index of 54.7%. This level of performance is equivalent to that of the Real-SPINE 3.0 server (54.6%), which uses a sophisticated two-layer neural network to predict dihedral angles from the primary sequence of a query protein [41]. However, under these conditions, GLE diagram analysis indicates that 47% of the predictions made by DANGLE are ambiguous; when these were culled, the $A_{30}(\phi, \psi)$ value for the remaining estimates is boosted to 67.5%. These observations illustrate the power of the five-residue fragment matching approach when it is coupled to a filtering procedure.

3.6. Comparison with other angle prediction algorithms

Using the test set of 29 proteins, we compared the performance of DANGLE with shift-based predictions by TALOS [11], TALOS+ [14] and PREDITOR [13] and with sequence-only estimates from the Real-SPINE 3.0 server [41] (Table 1). To ensure a realistic field-test, all sites in query proteins were assessed, including residues in flexible regions or with incomplete chemical shift sets.

All of the NMR-based methods perform better than the Real-SPINE 3.0 sequence-only approach (with an $A_{30}(\phi, \psi)$ accuracy index of 55%), confirming that chemical shift measurements do indeed convey useful additional information (Fig. 6A). If both "good" and "ambiguous" sites are considered together, the $A_{30}(\phi, \psi)$ rating increases in the order PRED-A (71%) < PRED-B (74%) < DANG-A (75%) < TALOS-A (76%) < TPLUS-A (77%). Methods that are able to discard ambiguous predictions are significantly more accurate, leading to the surprising result that the original implementation of TALOS performs best: DANG-B (82%) < TPLUS-B (83%) < TALOS-B (86%). The explanation for this becomes apparent on viewing Fig. 7, which plots the RMS errors for estimates of ϕ and ψ against the degree of completeness, i.e. the percentage of sites at which predictions are accepted. TALOS-B yields dihedral angle values with very low errors, but at the expense of making predictions for only 72% of residues in the query protein. By contrast, TPLUS-B returns predictions for a larger proportion of sites (85%), but at lower accuracy. From this perspective, DANG-B produces a good compromise, offering an RMS error rate close to that of TALOS-B, but achieving this while approaching the degree of coverage observed for TPLUS-B.

The tiny errors returned by TALOS and TALOS+ reflect details of the cluster analysis procedure, rather than true confidence intervals, and so are rarely used as realistic upper and lower boundaries in structure calculations. We therefore also assessed the percentage of predictions for which the reference values of both ϕ and ψ fall within the stated boundary range (i.e. $A_{\mathrm{range}}(\phi, \psi)$) for the competing methods (Fig. 6B). In this test, both of the DANGLE modes clearly lead the field. Taken together, these results demonstrate that the self-guided filtering process implemented in DANGLE yields dihedral angle predictions that are more realistic than other available methods, achieving an optimal balance between accuracy and completeness within pragmatic, well defined limits.

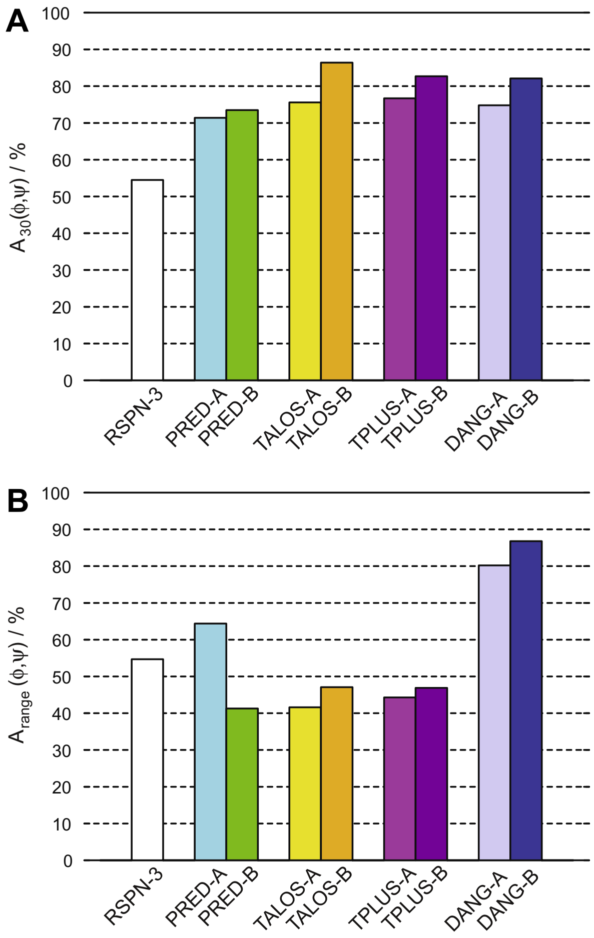

Fig. 6. Histograms comparing accuracy indices for various dihedral angle prediction methods for all residue types. (A) $A_{30}(\phi, \psi)$, the percentage of predictions of ϕ and ψ both within 30° of their reference angles. (B) $A_{\mathrm{range}}(\phi, \psi)$, the percentage of reference angles that fall within the predicted boundary range for both ϕ and ψ. RSPN-3 indicates results from Real-SPINE 3.0; PRED-A: all results from PREDITOR using shift information only; PRED-B: all results from PREDITOR using both shift information and homologous structures; TALOS-A: all results from TALOS; TALOS-B: results classed as "good" by TALOS; TPLUS-A: all results from TALOS+; TPLUS-B: results classed as "good" by TALOS+; DANG-A: all results from DANGLE; DANG-B: all results from DANGLE after rejecting all multi-island predictions. Predictions were made for the 29 protein test set.

Table 1
Accuracy comparison for dihedral angle prediction methods.[a]

Method	$A_{30}(\phi, \psi)$/%[b]					Rejection level/%
	Generic	Glycine	Proline	Pre-proline	All	
Real-SPINE 3.0	56.4	32.0	58.1	43.2	54.5	0.0
PRED-A	73.3	47.2	79.1	58.5	71.4	3.6
PRED-B	74.3	53.8	82.8	79.8	73.5	2.8
TALOS-A	78.3	46.7	78.2	57.7	75.6	2.0
TALOS-B	87.9	79.7	86.6	58.8	86.4	27.7
TPLUS-A	79.6	51.4	83.2	63.8	77.5	3.3
TPLUS-B	84.4	60.5	86.2	64.4	82.6	14.7
DANG-A	76.4	56.4	79.1	63.6	74.8	0.0
DANG-B	82.8	76.5	84.4	69.5	82.1	17.4

[a] Comparison performed using the test set of 29 proteins.
[b] Percentage of predictions for which both ϕ and ψ deviate from their reference values by less than 30°.

3.7. Comparison with other secondary structure prediction methods

DANGLE also provides information at lower resolution, assigning each residue of the query protein to one of three secondary structure states: helix (H), strand (E) or coil (C). Classification of this sort is a useful step when a new protein system is being characterised: it can guide manual resonance assignment procedures, as well as facilitating visualisation, fold categorisation, homology modelling and sequence alignment [57]. A self-assessment test using the 186 proteins from the fragment database returns the same secondary structure identification as DSSP [35] in 86% of cases, rising slightly to 87% when residues with multi-island GLE diagrams are filtered out. These results fall close to the limit of 88% identified as the maximum achievable value of A_3 for *ab initio* three-state prediction methods, reflecting the level of discrepancy encountered when benchmarking different secondary structure detection algorithms, comparing NMR and X-ray structures of

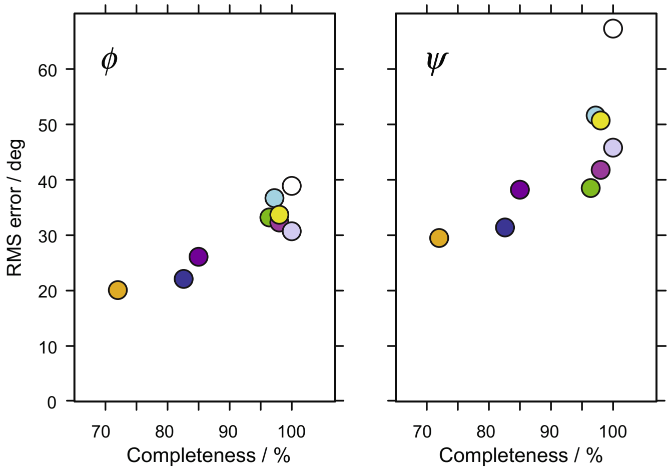

Fig. 7. Scatter plots showing of root mean square (RMS) errors against the degree of completeness for various dihedral angle prediction methods for (A) ϕ and (B) ψ. White data points indicate results from RSPN-3; cyan: PRED-A; green: PRED-B; yellow: TALOS-A; orange: TALOS-B; magenta: TPLUS-A; purple: TPLUS-B: light blue: DANG-A; blue: DANG-B. Predictions were made for all residue types from the 29 protein test set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

identical or homologous proteins, or analysing the variation within an ensemble of solution structures [57,58].

When alternative shift-based procedures are compared using the 29 protein test set, most are inferior to state-of-the-art sequence-only approaches, represented here by results from the Jpred3 server [42], which gives an A_3 index of 79% (see Table 2 and Fig. 8). Taking chemical shift data into account confers a clear advantage over Jpred3 for only two methods: PsiCSI (84%); and DANGLE, which returns 83% when all sites are considered and 85% when ambiguous predictions are discarded. Interestingly, PsiCSI [6] and TALOS+ [14] both use neural network approaches to make deductions about secondary structure class, but in our hands PsiCSI is more successful.

The majority of disagreements between DANGLE and DSSP occur near protein N- or C-termini (where the backbone is likely to be flexible), in linker regions that connect elements of regular secondary structure (where polypeptide chains often sample multiple conformations), or concern exactly where helices and strands start or finish (which may be a genuine reflection of protein dynamics, such as “fuzzy” helix capping) [58].

Table 2
Accuracy comparison for secondary structure prediction methods.[a]

	A_3/%[b]				
	Generic	Glycine	Proline	Pre-pro	All
Jpred3	78.4	85.1	87.8	86.9	79.3
ΔCOB	69.5	–[c]	72.2	68.0	69.6
ΔCAB	71.7	–[c]	75.5	79.2	72.0
cCSI	76.4	72.9	72.2	77.8	76.0
PECAN	78.7	74.0	67.0	72.7	77.7
PSSI	80.0	85.1	77.4	71.7	79.9
PsiCSI	83.6	87.8	81.7	87.8	84.2
TPLUS-A	75.6	77.0	80.7	72.0	75.8
TPLUS-B	76.8	78.6	78.8	72.7	76.8
DANG-A	83.2	83.4	87.0	86.9	83.4
DANG-B	84.8	91.3	87.5	84.1	85.2

[a] Comparison performed using the test set of 29 proteins.
[b] Percentage of three-state secondary structure predictions that are identical to the states in the reference structure.
[c] The ΔCOB and ΔCAB methods are not applicable to glycines, which lack C^{β} nuclei.

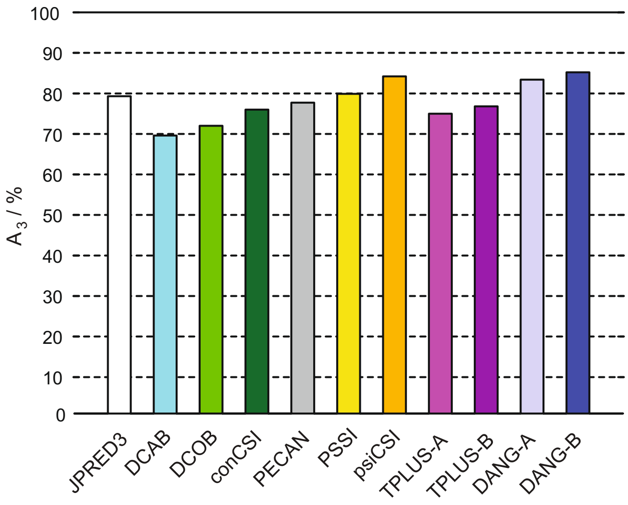

Fig. 8. Histogram showing A_3, the percentage of three-state predictions of secondary structure that agree with the output from DSSP, comparing various prediction methods. Results are for all residue types in the 29 protein test set. Method names are described in the text.

4. Conclusions

This article has introduced DANGLE, a new algorithm that uses protein chemical shift measurements and a database of high-quality structures with known shifts to predict backbone ϕ and ψ dihedral angles and secondary structure states. It employs Bayesian inferential logic to generate a global likelihood estimate map for each residue in the query protein, from which robust estimates of ϕ and ψ and realistic prediction errors can be derived, along with an indication of the degeneracy in the relationship between shift measurements and structure at that site. In addition, each residue is assigned a secondary structure state by assessing the conformations of database fragments that possess similar chemical shifts and amino acid sequence.

At some sites, the measured chemical shifts are not consistent with a single protein conformation, as might occur in regions that undergo regular structural rearrangements. Imposing inappropriate shift-based dihedral angle constraints on such regions during structure calculations would result in an over-restrained ensemble that did not reflect the true dynamic properties of the polypeptide chain. We therefore recommend that dihedral angle predictions for residues with ambiguous, multi-island GLE diagrams should be deleted from the restraint list. Used in this mode, DANGLE offers an ideal compromise between accuracy and coverage when compared with the competing packages TALOS [11], PREDITOR [12,13] and TALOS+ [14]. In contrast to these methods, DANGLE outputs a list of dihedral angle estimates that are defined within realistic boundary ranges, which can therefore be used in conventional structure calculation protocols without further interpretation. Alternatively, DANGLE offers GLE diagrams that could be utilised as restraining probability distributions in inferential structure determination procedures [52].

In addition to providing a screening method for identifying reliable dihedral angle predictions, GLE diagram degeneracy provides a new tool for investigating the properties of unfolded protein states, with a sensitivity to residual structure that is different from the shift-based RCI approach [49]. This property could be investigated further by deriving an expression for the Shannon entropy [39] from the matrix of posterior probability scores.

Finally, we have shown that DANGLE performs predictions of secondary structure class in a highly reliable manner, yielding results that are as accurate as psiCSI [6], the best alternative technique. As excluding ambiguous sites has only a small effect on overall accuracy, we advise that three-state secondary structure predictions should be performed for every residue in the query protein.

5. Authors' contributions

MSC wrote the code, constructed the databases, tested the algorithm and integrated the software into the CCPN package. MLM provided preliminary work. TJS provided insights throughout the project. RWB conceived the project, gave direct supervision and wrote the paper.

Acknowledgments

We acknowledge valuable insights provided by a reviewer, which have opened up several new avenues of research. M.S.C acknowledges the Cambridge Commonwealth Trust for the award of a Prince Philip Graduate Exhibition PhD scholarship. M.L.M. acknowledges BBSRC for the award of a PhD studentship. We also thank: Dr Simon Lovell for providing a copy of his structural database; Dr Vim Vranken for assistance with the RECOORD database; and Professor Ernest Laue of the Collaborative Computational Project for NMR (CCPN), funded by BBSRC, for including dihedral angle predictions in their data model.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmr.2009.11.008.

References

[1] C.C. McDonald, W.D. Phillips, Manifestations of the tertiary structures of proteins in high-frequency nuclear magnetic resonance, J. Am. Chem. Soc. 89 (1967) 6332–6341.

[2] L. Szilágyi, Chemical shifts in proteins come of age, Prog. NMR Spectrosc. 27 (1995) 325–443.

[3] D.S. Wishart, D.A. Case, Use of chemical shifts in macromolecular structure determination, Methods Enzymol. 338 (2001) 3–34.

[4] D.S. Wishart, B.D. Sykes, Chemical shifts as a tool for structure determination, Methods Enzymol. 239 (1994) 363–392.

[5] Y. Wang, O. Jardetzky, Probability-based protein secondary structure identification using combined NMR chemical-shift data, Protein Sci. 11 (2002) 852–861.

[6] L.H. Hung, R. Samudrala, Accurate and automated classification of protein secondary structure with PsiCSI, Protein Sci. 12 (2003) 288–295.

[7] H.R. Eghbalnia, L. Wang, A. Bahrami, A. Assadi, J.L. Markley, Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structure elements, J. Biomol. NMR 32 (2005) 71–81.

[8] W.J. Metzler, K.L. Constantine, M.S. Friedrichs, A.J. Bell, E.G. Ernst, T.B. Lavoie, L. Mueller, Characterization of the three-dimensional solution structure of human profilin: ^{1}H, ^{13}C, and ^{15}N NMR assignments and global folding pattern, Biochemistry 32 (1993) 13818–13829.

[9] R.P. Barnwal, K.V.R. Chary, An efficient method for secondary structure determination in polypeptides by NMR, Curr. Sci. 94 (2008) 1302–1306.

[10] S.P. Mielke, V.V. Krishnan, Characterization of protein secondary structure from NMR chemical shifts, Prog. NMR Spectrosc. 54 (2009) 141–165.

[11] G. Cornilescu, F. Delaglio, A. Bax, Protein backbone angle restraints from searching a database for chemical shift and sequence homology, J. Biomol. NMR 13 (1999) 289–302.

[12] S. Neal, M. Berjanskii, H. Zhang, D.S. Wishart, Accurate prediction of protein torsion angles using chemical shifts and sequence homology, Magn. Reson. Chem. 44 (2006) S158–S167.

[13] M.V. Berjanskii, S. Neal, D.S. Wishart, PREDITOR: a web server for predicting protein torsion angle restraints, Nucleic Acids Res. 34 (2006) W63–W69.

[14] Y. Shen, F. Delaglio, G. Cornilescu, A. Bax, TALOS+: a hybrid method for predicting protein backbone angles from NMR chemical shifts, J. Biomol. NMR 44 (2009) 213–223.

[15] A. Cavalli, X. Salvatella, C.M. Dobson, M. Vendruscolo, Protein structure determination from NMR chemical shifts, Proc. Natl. Acad. Sci. USA 104 (2007) 9615–9620.

[16] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J.M. Aramini, G. Liu, A. Eletsky, Y. Wu, K.K. Singarapu, A. Lemak, A. Ignatchenko, C.H. Arrowsmith, T. Szyperski, G.T. Montelione, D. Baker, A. Bax, Consistent blind protein structure generation from NMR chemical shift data, Proc. Natl. Acad. Sci. USA 105 (2008) 4685–4690.

[17] D.S. Wishart, D. Arndt, M. Berjanskii, P. Tang, J. Zhou, G. Lin, CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data, Nucleic Acids Res. 36 (2008) W496–W502.

[18] P. Robustelli, A. Cavalli, M. Vendruscolo, Determination of protein structures in the solid state from NMR chemical shifts, Structure 16 (2008) 1764–1769.

[19] P. Robustelli, A. Cavalli, C.M. Dobson, M. Vendruscolo, X. Salvatella, Folding of small proteins by Monte Carlo simulations with chemical shift restraints without the use of molecular fragment replacement or structural homology, J. Phys. Chem. B 113 (2009) 7890–7896.

[20] Y. Shen, R. Vernon, D. Baker, A. Bax, De novo protein structure generation from incomplete chemical shift assignments, J. Biomol. NMR 43 (2009) 63–78.

[21] S. Neal, A.M. Nip, H. Zhang, D.S. Wishart, Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts, J. Biomol. NMR 26 (2003) 215–240.

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

[23] S.B. Nabuurs, C.A. Spronk, E. Krieger, H. Maassen, G. Vriend, G.W. Vuister, Quantitative evaluation of experimental NMR restraints, J. Am. Chem. Soc. 125 (2003) 12026–12034.

[24] A. Loquet, B. Bardiaux, C. Gardiennet, C. Blanchet, M. Baldus, M. Nilges, T. Malliavin, A. Böckmann, 3D structure determination of the Crh protein from highly ambiguous solid-state NMR restraints, J. Am. Chem. Soc. 130 (2008) 3579–3589.

[25] J.F. Doreleijers, S. Mading, D. Maziuk, K. Sojourner, L. Yin, J. Zhu, J.L. Markley, E.L. Ulrich, BioMagResBank database with sets of experimental NMR contraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank, J. Biomol. NMR 26 (2003) 139–146.

[26] H. Zhang, S. Neal, D.S. Wishart, RefDB: a database of uniformly referenced protein chemical shifts, J. Biomol. NMR 25 (2003) 173–195.

[27] 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 (2000) 235–242.

[28] A.J. Nederveen, J.F. Doreleijers, W. Vranken, Z. Miller, C.A. Spronk, S.B. Nabuurs, P. Güntert, M. Livny, J.L. Markley, M. Nilges, E.L. Ulrich, R. Kaptein, A.M. Bonvin, RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank, Proteins 59 (2005) 662–672.

[29] G.N. Ramachandran, C. Ramakrishnan, V. Sasisekharan, Stereochemistry of polypeptide chain configurations, J. Mol. Biol. 7 (1963) 95–99.

[30] B.K. Ho, A. Thomas, R. Brasseur, Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the α-helix, Protein Sci. 12 (2003) 2508–2522.

[31] S.C. Lovell, I.W. Davis, W.B. Arendall, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson, D.C. Richardson, Structure validation by C$^{\alpha}$ geometry: ϕ, ψ and C$^{\beta}$ deviation, Proteins 50 (2003) 437–450.

[32] B.K. Ho, R. Brasseur, The Ramachandran plots of glycine and pre-proline, BMC Struct. Biol. 5 (2005) 14–24.

[33] S. Schwarzinger, G.J. Kroon, T.R. Foss, P.E. Wright, H.J. Dyson, Random coil shifts in acidic 8 M urea: implementation of random coil shift data in NMRView, J. Biomol. NMR 18 (2000) 43–48.

[34] S. Schwarzinger, G.J. Kroon, T.R. Foss, J. Chung, P.E. Wright, H.J. Dyson, Sequence-dependent correction of random coil NMR chemical shifts, J. Am. Chem. Soc. 123 (2001) 2970–2978.

[35] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.

[36] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inform. Theory 37 (1991) 145–151.

[37] D.B. Dahl, Z. Bohannan, Q. Mo, M. Vannucci, J. Tsai, Assessing side-chain perturbations of the protein backbone: a knowledge-based classification of residue Ramachandran space, J. Mol. Biol. 378 (2008) 749–758.

[38] E.T. Jaynes, Probability Theory: The Logic of Science, Cambridge University Press, 2003.

[39] T.M. Cover, J.A. Thomas, Elements of Information Theory, second ed., Wiley, 2006.

[40] W.F. Vranken, W. Boucher, T.J. Stevens, R.H. Fogh, A. Pajon, M. Llinas, E.L. Ulrich, J.L. Markley, J. Ionides, E.D. Laue, The CCPN data model for NMR spectroscopy: development of a software pipeline, Proteins 59 (2005) 687–696.

[41] E. Faraggi, B. Xue, Y. Zhou, Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two layer neural network, Proteins 74 (2009) 847–856.

[42] C. Cole, J.D. Barber, G.J. Barton, The Jpred 3 secondary structure prediction server, Nucleic Acids Res. 36 (2008) W197–W201.

[43] G. Meinke, P.A. Bullock, A. Bohm, Crystal structure of the simian virus 40 large T-antigen origin-binding domain, J.Virol. 80 (2006) 4304–4312.

[44] X. Luo, D.G. Sanford, P.A. Bullock, W.W. Bachovchin, Solution structure of the origin DNA-binding domain of SV40 T-antigen, Nat. Struct. Biol. 3 (1996) 1034–1039.

[45] 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 (2001) 3561–3571.

[46] M.V. Berjanskii, D.S. Wishart, A simple method to predict protein flexibility using secondary chemical shifts, J. Am. Chem. Soc. 127 (2005) 14970–14971.

[47] M.V. Berjanski, D.S. Wishart, NMR: prediction of protein flexibility, Nat. Protoc. 1 (2006) 683–688.

[48] M.V. Berjanskii, D.S. Wishart, The RCI server: rapid and accurate calculation of protein flexibility using chemical shifts, Nucleic Acids Res. 35 (2007) W531–537.

[49] M.V. Berjanskii, D.S. Wishart, Application of the random coil index to studying protein flexibility, J. Biomol. NMR 40 (2008) 31–48.

[50] T. Tanaka, S. Yokoyama, Y. Kuroda, Improvement of domain linker prediction by incorporating loop-length-dependent characteristics, Biopolymers 84 (2006) 161–168.

[51] V.N. Uversky, What does it mean to be natively unfolded?, Eur J. Biochem. 269 (2002) 2–12.

[52] W. Rieping, M. Nilges, M. Habeck, ISD: a software package for Bayesian NMR structure calculation, Bioinformatics 24 (2008) 1104–1105.

[53] R.A. Laskowski, M.W. MacArthur, D.S. Moss, J.M. Thornton, PROCHECK: a program to check the stereochemical quality of protein structures, J. Appl. Cryst. 26 (1993) 283–291.

[54] M. Iwadate, T. Asakura, M.P. Williamson, C^{α} and C^{β} ^{13}C chemical shifts in proteins from an empirical database, J. Biomol. NMR 13 (1999) 199–211.

[55] R.M. Jarret, N. Sin, M. Dintzner, Deuterium isotope effects on ^{13}C NMR chemical shifts of amides, Microchem. J. 56 (1997) 19–21.

[56] N.M. Sergeev, N.D. Sergeeva, L.F. Kobets, Negative linear deuterium isotope shift for ^{13}C in iodoform, J. Struct. Chem. 28 (1987) 934.

[57] B. Rost, Review: protein secondary structure prediction continues to rise, J. Struct. Biol. 134 (2001) 204–218.

[58] C.A.F. Andersen, A.G. Palmer, S. Brunak, B. Rost, Continuum secondary structure captures protein flexibility, Structure 10 (2002) 175–184.