From local structure to a global framework: recognition of protein folds

Agnel Praveen Joseph1,2,3,4,† and Alexandre G. de Brevern3,4,5,6,†

1 Science and Technology Facilities Council, Rutherford Appleton Laboratory, Harwell Oxford, Didcot OX11 0QX, UK
2 National Centre for Biological Sciences, GKV, Bellary Road, Bangalore 560065, India
3 INSERM, U1134, DSIMB, 75739 Paris, France
4 Université Paris Diderot, Sorbonne Paris Cité, UMR S 1134, 75739 Paris, France
5 Institut National de la Transfusion Sanguine (INTS), 75739 Paris, France
6 Laboratoire d'Excellence, GR-Ex, 75739 Paris, France

Protein folding has been a major area of research for many years. Nonetheless, the mechanisms leading to the formation of an active biological fold are still not fully apprehended. The huge amount of available sequence and structural information provides hints to identify the putative fold for a given sequence. Indeed, protein structures prefer a limited number of local backbone conformations, some being characterized by preferences for certain amino acids. These preferences largely depend on the local structural environment. The prediction of local backbone conformations has become an important factor to correctly identifying the global protein fold. Here, we review the developments in the field of local structure prediction and especially their implication in protein fold recognition.

1. Introduction

A detailed understanding of the function of a protein can be achieved by studying its three-dimensional structure. Insights into the molecular details of protein interactions and enzyme activities are obtained from the three-dimensional structure and the impact of protein structure-based drug design in pharmaceutical development has been impressive [1–3]. Although significant advances have been made in the field of experimental structure determination [4,5], the difficulty and cost associated with this limit the rate at which protein structures are solved. The sequence information, on the other hand, has grown immensely with the help of high-throughput sequencing techniques [6–11]. The huge gap between the sequence and structure space is formed by sequences whose three-dimensional structures are unknown.

Computational approaches provide effective ways for filling this gap. The methods used for structural annotation assign a probable three-dimensional fold for a given sequence. Applications of computational structure modelling have been enormous [12]. The modelled structures are useful in studying protein–protein [13] and protein–ligand interactions [12–14]. They are also used as search models in experimental structure determination [15] and for characterizing structures of huge assemblies [16–18]. Computational model-based drug design [19,20] and studies for understanding protein dynamics based on modelled structures [12,21,22] have been successful. Large-scale proteome-wide structural annotation studies have also been designed upon computational approaches for fold identification [23,24].

Modelling protein structures based on a template fold (comparative/homology modelling) is so far the most reliable technique for generating three-dimensional structure for a given sequence [25]. The most critical step in comparative modelling is the identification of the correct template fold. The possible number of protein folds are estimated to be limited in number and the currently available structures can cover most of it [26,27]. The ability to relate sequences to the correct folds forms the basis for protein fold recognition. The simplest means for deriving structural information is by direct comparison with other related sequences with known structures [28–30] (Figure 1a). The tertiary...
structure of a protein is highly conserved across different species, when compared with its sequence [32]. Figure 1b shows the direct relationship existing between structural similarity in terms of global distance test (GDT) [31] and sequence identity. This analysis has been performed on a randomly chosen subset of about 500 SCOP domain, which were structurally aligned [33,34] with iPBA superimposition method [35,36]. Thus, if the sequence similarity (usually quantified as the percentage amino acid identity) is high, the structures are likely to be similar. Comparative modelling can generate reliable structural models (1–3 Å root mean squared deviation (RMSD) with the real structure), especially when the sequence identity is above 30% [25,37] (figure 1a(i)).

Nonetheless, the sequences of homologous protein structures can diverge beyond the point where pairwise sequence comparison methods fail. Figure 1a highlights the success in detecting the correct fold declines with decrease in sequence identity. The match between the sequence of interest (target) and the known folds is obtained as an alignment of sequences. Below 30% sequence identity (twilight zone, figure 1a(ii)), the accuracy of alignments falls significantly which results in a low sensitivity in fold recognition. At very low sequence identities (below 15%, figure 1a(iii)), the sequence alignments are not very informative. The values provided in the figure are just indicative of the sequence identity ranges, and no fixed thresholds have been characterized [38]. The lectin superfamily is a classic example of this case where the sequence identities fall below 10% but they are known to adopt a jelly-roll-sugar binding fold [39], and a random alignment is expected to have about 12% identity.

Figure 1. ‘Homology’ detection with variation in sequence identity. (a) Schematic demonstrating the use of sequence comparison for detecting structural homology. The sequence alignments are indicated with ‘X’ representing any amino acid. Same amino acids at equivalent positions are highlighted in red, similar ones are in green. At sequence identity levels above 30% (i), simple sequence alignments are largely sufficient to detect similar folds. Below this similarity threshold, the alignments are less accurate and thus less efficient in detecting genuine relationships. (ii) Between 20 and 30%, the correct fold is not often detected as the top hit. (iii) At very low sequence identities, simple sequence alignments are not very useful. (b) Variation of structural similarity (quantified in terms of GDT_TS score) [31] with change in sequence identity. Even at low sequence identities (less than 30%, highlighted in grey background) significant structural similarity could be observed. (Online version in colour.)
In this review, we focus mainly on the scenario where the target has only remote homology with a protein of known fold (template). In such cases, a clear understanding of the association of amino acid sequences with the three-dimensional structure (sequence–structure relationship) is the key factor in determining the success to recognize correct folds. Firstly, a general outline of the methods for fold recognition based on the detection of remote homology is presented. Then the role of local structure prediction in fold recognition is emphasized, followed by a discussion on the importance of careful extraction of sequence–structure relationships.

2. Fold recognition by remote similarity detection

In the absence of closely related proteins of known structure, it is difficult to assign the correct fold using simple sequence alignments (figure 1a). Additional information is required to enhance the sensitivity to detect more distant relationships. The knowledge on the evolutionary and structural constraints associated with the target sequence forms the major contributing factor in detecting distant relationships.

2.1. Evolutionary information: profile–profile alignments

The protein sequence databases nowadays contain homologous sequences from different species. A multiple sequence alignment (MSA) can be used to trace the extent of evolutionary divergence among the related sequences. As against a single sequence, MSA consists of information about amino acid propensities at each position in the sequence which characterizes a sequence ‘profile’. This information is highly valuable in differentiating conserved and variable positions, and these position-specific amino acid propensities are more effective in identifying a protein that is evolutionarily related to this group. Thus, the use of MSAs as sequence profiles has a huge impact on homology detection (figure 2a,b) [40,41].

A sequence profile is often described as a position-specific scoring matrix (PSSM), which holds information on amino acid propensities for each position in the sequence, derived based on an MSA. PSI-BLAST and RPS-BLAST [42,43] are widely used sensitive PSSM-based search algorithms for remote homology detection. The search for related sequences is usually carried out iteratively where the profile (PSSM) gets updated at each step with new sequence information. These new sequences, which can be termed as ‘intermediates’, could also be used as new targets to trace the linkage to the distant homologue [44,45]. Recently, a method named HangOut has been proposed to improve the sensitivity of PSI-BLAST searches especially for domains with long insertions [46].

Several approaches have been proposed to improve the search for related proteins based on sequence profiles. An MSA can be used to generate hidden Markov models (HMMs), which can also be used as efficient search tools [47–50]. HMM-based profile comparison tools are reported to have a higher accuracy than PSSM-based methods [45,51]. To improve the accuracy of profiles generated, efficient methods that can weigh conserved sequence regions have also been developed [52–56]. Amino acid substitution probabilities that drive the alignment scores are influenced by the local structural [57] or sequence neighbourhood [58,59] in the protein. CSI-BLAST [60] uses mutation probabilities from context-specific profiles based on 13-residue windows while DELTA-BLAST [61] incorporates frequencies from longer profiles corresponding to conserved domain sequences [62]. Two iterations of CSI-BLAST [60] are more sensitive than five iterations of PSI-BLAST, whereas DELTA-BLAST [61] is more sensitive than both CS-BLAST and PSI-BLAST in detecting relatively similar and remote homologues. Profile–profile comparison methods that optimize local alignments are also reported to be more effective in remote homology detection [55,63,64].

Fast and efficient HMM profile-based iterative search approaches similar to PSI-BLAST marked further advancements in this field. JACKHMMER [65] relies on a series of database filtering steps to reduce search time, whereas HHBLITS implements HMM–HMM comparisons with approximation of profile columns using 219 extended alphabets [66]. The sensitivity of HHBLITS in detecting remote homologues was about 50–100% more than that of PSI-BLAST.

Increasing amount of evidence of sequence permutations, duplications or fusion events [67–69] raises the need for developing methods that can detect homologues independent of the sequence order. A classic example is the methyl transferase family where gene fusion and permutation events are observed in the evolutionary process resulting in differences in substrate recognition and catalytic activity [70,71]. Interesting developments have also been made in this direction. A few alignment-free techniques were developed for this purpose [72,73], while the majority of methods that employ sequence-order-independent comparisons are optimized for multi-domain proteins [74–77].

However, at a sequence identity well below 20%, the chances of detecting true templates even with profile-based approaches remain low [78]. It is possible that the fold which the sequence adopts is already known but the sequence-based searches fail to trace such relationship. Such cases could be addressed if one can verify whether this protein sequence is compatible with any of the available protein folds. The question is how to derive this compatibility with known folds, given a sequence target.

2.2. Structure information: fold recognition by threading

Two proteins sharing the same fold could have diverged significantly in terms of sequence and it is often difficult to detect the relationship based on sequence- or profile-based methods. Some well-known examples of these are the oligonucleotide/oligosaccharide binding fold, cupins, TIM barrels, serine proteases, etc. Hence it becomes necessary to relate target sequence with the known folds. To assess the fit of a sequence on a template protein fold, it is necessary to quantify the preference for amino acids to occur in the structural environment of the template. This measure can enable us to generate an alignment between the sequence and the template structure (threading). Two major strategies have been used to score the sequence–structure compatibility.

2.2.1. Threading based on global potentials

The more direct approach would be to fit each residue from the query sequence onto each position in the template backbone and check whether it is energetically compatible...
The energy could be quantified as pairwise interaction potentials \[79,80\] or solvation energy \[81\] which were used by many of the earlier methods for fold identification \[82–84\]. The non-local regions of the target residue could have significant differences when compared with that of the template. This means that the equivalent regions (or residues) have to be known a priori to calculate the compatibility score. Hence the target–template alignment has to be generated in advance and this makes the problem extremely hard. An easier solution is the ‘frozen approximation’ \[81\], where each residue in the target is scored against the structural environment formed by the neighboring template residues. However, as mentioned earlier, the target and the template often have different environments involving variations in the amino acid residues. Hence this method is frequently prone to accumulate errors that affect the final estimation of compatibility. Approximate solutions to identify residue neighbors could be derived using Monte Carlo simulations \[85\] or double-dynamic programming \[83\]. The divide-and-conquer approach used in Prospect \[86\] aims to solve several sub-structure alignments and integrate them in an optimal way. The complexity associated with fold identification purely based on global interaction potentials was alleviated by combining local structural information \[87\] and the local structural context is observed to have the major role in obtaining an optimal template alignment \[88\].

2.2.2. Threading based on local scoring functions
The second category of methods adopts a more localized approach for quantifying the compatibility of the target sequence on a template. They score the preference for each residue to occur in a local environment of the template structure. The preference for an amino acid to occur in a local structural

---

**Figure 2.** Different strategies for protein fold recognition. The fold space is highlighted by the blue background and the lengths of the black arrows joining the target sequence (space) and fold space give an idea of the distance of relationship. (a) Close relationships are often detected by simple sequence alignment techniques. (b) Addition of evolutionary information using sequence profiles derived from MSAs helps in detecting more distantly related folds. When the sequence-based alignments are not informative, sequence–structure matching needs to be carried out. (c) The target sequence can be threaded on to the known folds to check the compatibility. The compatibility is usually quantified based on the global interaction or energy potential. Obtaining an optimal alignment between the sequence and a fold is however difficult and computationally quite expensive. (d) The other alternative is to carry out prediction of different structural features like local backbone conformation, solvent accessibility or contact order and then matching the predicted features with that found in the known fold. (Online version in colour.)
environment can be referred to as the sequence--structure relationship. As mentioned earlier, several structural features can be used to describe the local structural environment of a protein. The most commonly used features are secondary structure, solvent accessibility, residue depth, backbone conformation, potential for hydrogen bonding or hydrophobic interactions [57,89]. A combination of some or all of these features can define one local environment. The issue is to choose the right set of features and weigh each feature to derive the combination.

Two different ways in which the preference of an amino acid for a local structural environment can be calculated are: (i) the amino acid preferences associated with a local structural environment can be extracted as a one- to three-dimensional substitution matrix which gives the score for assigning each amino acid in this environment [57,90–92]. These methods are based on the premise that the amino acid substitution patterns vary based on their structural environments [57], (ii) The other strategy adopted by most of the recent methods is to predict the structural features separately based on the target sequence and then combine the predictions carefully to search for related structures (figure 2f) [93]. The use of predicted local structural features as strings can bring down the problem of sequence--structure comparison to that of a sequence--sequence comparison. As in the case of classical sequence alignment procedures, dynamic programming algorithm can be used to achieve this task [94]. Recognition of protein folds based on predicted local structural features is competitive with methods that consider non-local structural details [95–98].

The efficiency of different profile comparison methods is coupled with the predicted structural information to improve the quality of profile-based fold recognition [54,95,99–106]. The successful approaches for fold recognition use profile comparisons to obtain a limited set of alignments which are then evaluated and refined using the energy-based potentials [101,102,107,108]. An interesting recent observation is that the accuracy of fold recognition improves with the use of artificially evolved sequences compatible with the template folds [109]. These sequences are optimized to stabilize a given structure by simulated annealing based on a variety of potentials, similar to those commonly used in protein threading.

In the following sections, we discuss the developments in the field of local structure prediction which has been the most essential component of prediction-based threading methods.

2.2.2.1. Secondary structure prediction

The protein backbone prefers mainly a limited number of stable conformations constrained by the backbone dihedral angles and hydrogen bonds. The two important regularities seen in the local conformation of the backbone constitute α-helices and β-strands [110]. The rest of the backbone is usually considered as coils which represent that not assigned to one of these two conformations. These three states (helix, strand and coil) constitute the classical definition of secondary structures. This description has been used either to complement sequence-based searches or to identify the best structural relative from the results of sequence-based searches. It has been a strategic step in most of the fold recognition approaches [89,95,108,111,112]. The fold recognition methodologies like PHYRE [95], MUSTER [113], SPARKS [114] and SP5 [105] incorporate methods for predicting secondary structure of the target sequence and then comparing that with the secondary structure assigned for the templates.

Efficient machine learning methods like artificial neural networks (ANNs) [115] or support vector machines (SVMs) [116] are being developed to predict secondary structure from a protein sequence. They are trained to learn the amino acid preferences associated with different secondary structures and make predictions for new sequences. Most of the successful and recent approaches use the information on amino acid propensities derived from sequence profiles [115,117–122]. The use of multiple sequences (as a profile) adds details on amino acid variability at each position in the sequence. Studying the association of relatively large fragments of about 15–20 residues with a secondary structural state helped to incorporate the effect of long-range interactions to a certain extent [123].

It must be noted that ‘the machine learning algorithms predict what they learn’. The secondary structure assignments for known structures are made using a non-redundant dataset of protein structures and then the sequence association with each secondary structural unit is learnt. There are several methods available for assigning secondary structures, based on local backbone dihedral angles or hydrogen bonds. It is interesting to note that the assignments made by different methods do not agree to a significant extent [124–126]. Figure 3 presents the comparison of secondary structure assigned based on the three-dimensional structure of methyglyoxal synthase [127] with that of predictions made by different methods. The ambiguities are generally seen in predicting the short repetitive structures and especially boundaries of secondary structures. The underlying cause is that the capping regions do not necessarily follow the same rules set for defining the backbone conformation of helices and strands [132].

Taking into account the variation of secondary structure assignments among different conformations of the same structure (inherent flexibility) and between close structural homologues, a theoretical limit of 88% was proposed for secondary structure prediction accuracy [123]. The most popular and not very recent methods like PSIPRED [115], PROF [122,123] and SSPro [130] could already achieve a prediction accuracy close to 80%. PHD [129] and PSIPRED [115] use two levels of neural network predictions where the initial predictions are refined at the second step. SSPro [130] implements a recursive single neural network instead of multiple layers. The use of dual layer SVMs also gave performance comparable to these methods [116,119]. A few improvements or alternative methods have been proposed in the last few years [116–118,120,121,131,133–136] but the increase in prediction accuracy is minimal. Use of consensus approaches based on the results of multiple predictions [133,136] gave accuracy above 80%. The use of a dictionary of words of amino acid stretches associated with a secondary structure state also resulted in predictions comparable to the best machine learning methods available [137].

A recent assessment of available secondary structure prediction methods suggests that a prediction accuracy of about 82% could be reached [138]. The exposed coils were predicted higher than helices and strands. The accuracies of prediction for residues in 3_10 helices and β-bridges were less than 50% and 48%, respectively [138]. Complementing profile–profile comparisons with predicted secondary structure resulted in a significant improvement in the efficiency of fold recognition [95,108,139,140]. The prediction of
non-local structural details mainly based on the predicted local features has also reported some success. The prediction of protein contact order which reflects the average length of sequence separating contacting residues, with the aid of predicted secondary structure information is reported to achieve a correlation of about 0.85 [141].

2.2.2. Beyond secondary structures

In the absence of reasonable sequence similarity, the secondary structure prediction makes a huge contribution to remote homology detection. However, proteins with similar secondary structure topology need not necessarily have the same tertiary fold. Figure 4 gives an example of one such case where a wrong fold is chosen purely based on the similarity in secondary structure. This is often true for proteins with repeating secondary structural elements, i.e. helix or strand repeats. The three-state secondary structural information (helix, sheet and coil) does not give precise details of the backbone conformation for the complete protein backbone [111]. More than 50% of the residues in the available protein structures are assigned to the coil state which by definition corresponds to irregular backbone conformations. The loss of information caused by considering such a large majority of residues as coils is a main contributor to the recognition of wrong folds [112,144].

On the contrary, the coil state is not strictly irregular. Next to helices and strands, turns characterize another frequent regularity in the protein backbone [145]. β-Turns account for about 25% of the residues in proteins [146]. Other favourable local structures involve PolyProline II helices [147], hairpin loops, corner motifs, β-bulges, etc. [148]. Most of these repeating units are associated with strong sequence–structure relationships. Thus, a major portion of protein backbone can be associated with repeating local structural elements.

Attempts have been made to generate a minimum set of local backbone conformations, a combination of which can be used to represent most or all of the protein backbone conformation. A set of prototypes of local structures that can be used to approximately define the conformation of a protein backbone is called as a structural alphabet (SA) [149]. The number and size of these fragments vary depending on the approach used. Apart from their use in modelling, structural alignment and functional analysis, these libraries also help to extract precise information on the amino acid preferences associated with local structures.

Several approaches have been developed for the prediction of local structures based on fragment libraries [150–157]. The underlying methods used for prediction are based on both probabilistic [150,153,158] and machine learning algorithms [152,159–162]. Bystroff and Baker generated a

![Figure 3. Comparison of secondary structure prediction methods. For a recently solved structure of methylglyoxal synthase (PDB ID 2X8W) [127], the assigned secondary structure by (a) DSSP [128] and predicted ones are shown. The α-helices are shown in red, β-strands in yellow and coils in green. Different secondary structure prediction methods are shown: (b) PSIPRED [115], (c) PROF [129], (d) SSPRO [130] and (e) YASPIN [131]. The predictions are also shown as sequence alignment in (f). Helices, strands and coils are indicated by H, E and C, respectively. (Online version in colour.)](http://rsif.royalsocietypublishing.org/doi/abs/10.1098/rsif.2013.1147)
library by clustering frequently observed short sequence patterns (I-sites) in protein structures. An efficient local structure prediction method was developed based on this library [150, 163]. As the number of prototypes increase, the prediction accuracy was found to be lower. However, the regions predicted with high confidence add fine details on structural motifs and local backbone conformation. The use of secondary structure information is proved to improve the accuracy of local structure prediction. However, the repetitive structures are sometimes over-predicted in place of other local conformations [164].

Protein blocks (PBs) are a widely used SA composed of 16 pentapeptide conformations characterized by the series of backbone dihedral angles [153]. These conformations are characterized by strong sequence–structure relationships determined by the high predictability. Figure 4d gives an example of two different protein folds having the same secondary structure topology. It highlights that the PB description is more informative when compared to the three-state secondary structure, especially for the coil regions.

It also underlines the significance of precise local structural information in fold recognition. A series of methods have been tested for the prediction of PBs from sequence data. Bayesian prediction models gave an accuracy ranging from 34 to 48% [153, 164–166]. Dual layer ANNs and SVMs improved the prediction rate to about 58.5% [159], 61% [167] and 67% [160]. Combining information on secondary structure and solvent accessibility further enhanced the prediction rate of PBs [168]. Longer fragments based on PBs were predicted with better accuracy [166,169]. Frequently observed 11 residue fragments described based on the dihedrals were predicted with a significant accuracy of about 63% using SVMs [170]. Prediction of structural flexibility based on local structure preferences has also been successful [171]. Table 1 gives a list of popular and more recent methods for local structure prediction, based on fragment libraries.

In recent years, a few attempts have been carried out to predict the $\phi/\psi$ backbone dihedral angles associated with each residue, based on the preferences observed in the
Table 1. Fragment-based local structure prediction methods. The table gives the list of methods for predicting local backbone conformation based on a library of fragments (prototypes). The length, number of prototypes and the distance measure used to generate the library are also mentioned. The reported prediction rates from the original publication are also listed. MDA, maximum deviation in torsion angles; ANN, artificial neural networks; SVM, support vector machines.

<table>
<thead>
<tr>
<th>research team</th>
<th>fragment length</th>
<th>distance measure</th>
<th>prediction method</th>
<th>prototype number</th>
<th>prediction rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rooman et al.</td>
<td>4, 5, 6, 7</td>
<td>Cα RMSD</td>
<td>statistical mechanics</td>
<td>4</td>
<td>40–47</td>
</tr>
<tr>
<td>最小化代价函数 (mean force potential)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bystroff &amp; Baker</td>
<td>3–19</td>
<td>sequence profiles, dihedral angles</td>
<td>profile–profile matching</td>
<td>13 (later updated to 16)</td>
<td>50</td>
</tr>
<tr>
<td>de Brevem et al.</td>
<td>5</td>
<td>dihedral angles</td>
<td>Bayesian</td>
<td>16</td>
<td>34.4</td>
</tr>
<tr>
<td>Hunter &amp; Subramaniam</td>
<td>4</td>
<td>hypercosine Cα</td>
<td>Bayesian</td>
<td>28–16336</td>
<td>40</td>
</tr>
<tr>
<td>Yang &amp; Wang</td>
<td>9</td>
<td>dihedral angles</td>
<td>sequence profile matching</td>
<td>138604</td>
<td>79</td>
</tr>
<tr>
<td>Etchebest et al.</td>
<td>5</td>
<td>dihedral angles</td>
<td>Bayesian</td>
<td>16 (153)</td>
<td>49</td>
</tr>
<tr>
<td>(simulated annealing)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benros et al.</td>
<td>11</td>
<td>Cα RMSD, PB based</td>
<td>hybrid protein model</td>
<td>120</td>
<td>51.2</td>
</tr>
<tr>
<td>Sander et al.</td>
<td>7</td>
<td>Cα distance</td>
<td>decision trees, SVM, random forest</td>
<td>28</td>
<td>23–36</td>
</tr>
<tr>
<td>Dong et al.</td>
<td>7</td>
<td>Cα distance</td>
<td>ANN</td>
<td>28</td>
<td>45.6</td>
</tr>
<tr>
<td>Dong et al.</td>
<td>5</td>
<td>dihedral angles</td>
<td>ANN</td>
<td>16 (153)</td>
<td>58.5</td>
</tr>
<tr>
<td>Zimmermann et al.</td>
<td>5</td>
<td>dihedral angles</td>
<td>SVM</td>
<td>16 (153)</td>
<td>61</td>
</tr>
<tr>
<td>Chen &amp; Johnson</td>
<td>9</td>
<td>Cα distance</td>
<td>SVM</td>
<td>800</td>
<td>72</td>
</tr>
<tr>
<td>Bornot et al.</td>
<td>11</td>
<td>Cα RMSD, PB based</td>
<td>hybrid protein model</td>
<td>120</td>
<td>63.1</td>
</tr>
<tr>
<td>Rangwala et al.</td>
<td>5</td>
<td>dihedral angles</td>
<td>SVM</td>
<td>16 (153)</td>
<td>67</td>
</tr>
<tr>
<td>Yu et al.</td>
<td>7–19</td>
<td>dihedral angles</td>
<td>Bayesian</td>
<td>82</td>
<td>62</td>
</tr>
</tbody>
</table>

Ramachandran map [172]. The highly preferred regions of dihedral angle pairs define the predicted classes. Both ANN- and SVM-based predictions [173,174] have been carried out, with a special focus on the coil regions. Prediction accuracy of above 80% was reported. Another approach predicting the probable dihedral angle associated with the coil residues has been developed [175]. The favoured dihedral region in the Ramachandran plot was divided into bins (30 × 30) and the probability of occurrence in these bins is predicted. Significant improvement could be achieved over random prediction and about 80% accuracy is reported for the prediction of the top 20 populated bins. Prediction of dihedral angles also helped in improving the accuracy of secondary structure prediction [135]. In this approach based on SVMs, an accuracy of prediction of about 54–57% was obtained by clustering the ϕ/ψ space into seven regions based on the population distribution.

Absolute values of ϕ and ψ angles are also predicted by several methods mainly based on ANNs [176–178]. Predicted secondary structure and/or solvent accessibility were combined with the sequence or profile information, as inputs for these learning methods. The ANGLOR method [177] was assessed to provide an error of 28° and 46° for the ϕ and ψ dihedral angles, respectively. These values were reduced to 22° and 36° with the improved Real-SPINE method [179].

Local structure prediction beyond the three-state secondary structures has been shown to improve recognition of remote homologues [180]. Prediction of absolute values of the dihedral angles has been incorporated in many recent fold recognition tools [105,113]. More than 5% increase in efficiency of detection of homologous folds was reported, with the addition of dihedral angle predictions [105]. A recent study also demonstrated that the use of local structure preference information significantly improves the quality of fold recognition [181].

2.2.3. Fragment-based fold recognition

The local regions of strong sequence–structure relationships are extracted as fragments and are assembled to generate the fold. The recognition of related folds by I-TASSER is mainly based on a profile–profile alignment method [98]. The efficiency of the profile search (see §2.1) is enhanced by the addition of secondary structure information. Various strategies involving HMMs [182], PSI-BLAST profiles [42], global [29] and local [30] dynamic programming algorithms are employed in the profile-based search of I-TASSER. The continuous fragments aligned with regions of multiple templates are reassembled into complete models [183]. The unaligned regions without structurally similar segments in the template are built by ab initio modelling, i.e. generating structural models based on physico-chemical and mechanical properties, without the use of a template [184].

In the absence of a template of similar fold, the target protein could still have local structural motifs observed in the available protein structures [185]. Though a homologous fold cannot be identified reliably, a probable model could be
 protocols by complementing the local conformation predictions. Figure 5 highlights the high confidence predictions of secondary structures, flexible and disordered regions on glutamate mutase [218]. Structural dynamic profiles may also be used as another feature that can be used for fold detection.

### 2.2.5. In search of specificities

The knowledge of sequence–sequence and sequence–structure relationships is gained on the premise that it is applicable to all proteins. Even though this is often true, some specificities have also been identified among certain sets of proteins.

The amino acid composition varies between different proteomes. Depending on the organism, the variation could be sometimes quite significant such that the standard substitution matrices may fail to find the right homologues. Based on this perspective, substitution matrices adjusted for the background composition are derived [221]. These matrices are reported to be quite efficient in detecting homologues from the species. Species-specific variations are also observed with respect to amino acid propensities for local protein structures [222]. This observation is quite interesting and the specificities could be exploited in improving the accuracy of approaches for local structure prediction. A study conducted in this direction showed promising results with *Plasmodium falciparum*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae* [223]. Considering the compositions corresponding to both query and template improves profile comparisons [224]. The amino acid preferences for local structures also show variations among different structural classes of proteins [225]. These specific preferences could be carefully exploited to enhance the accuracy of local structure prediction. The information on the specific amino acid preferences has been used to develop fold recognition approaches dedicated for specific classes of proteins [226,227].

### 2.3. Some successful methods

In order to assess the performance of different methods for protein structure prediction, Moult and co-workers introduced a public experiment called critical assessment of techniques for protein structure prediction (CASP). The participants are provided with protein sequences for which the structures have not been released. The I-TASSER server [108] gave the best performance in the recent CASP assessments (CASP 8, 9 and 10) of fold prediction servers [228]. The other servers in the top category were Robetta [186], Quark [229], HHpred [140], pmodeller [230], RAPTOR [87], pro-sp3 TASSER [231] and FALCON [232].

HHpred uses a novel HMM-based profile comparison method with enhanced accuracy by incorporating information on predicted secondary structure [51]. FALCON employs a fragment-HMM approach [232] where both the preferred dihedral angles of each residue and sequence–structure relationships from fragment libraries are integrated to predict structural models for a sequence. It has become a common strategy to combine different predicted structural features along with amino acid sequences to improve the accuracy of profile-based fold detection [55,106,177,181,233]. Weighted matching of multiple profiles (sequence and structural features) has also been employed to enhance the accuracy of recognizing homologous folds [105,234].
2.3.1. Combining different approaches

One of the main limitations of the available fold recognition procedures is that they often fail to pick the correct template as the top hit. However in most cases, the right fold occurs among the top 10 or 20 hits [235]. The specificity can be improved by assessing the results of several different methods. This idea has led to the development of many meta-servers which run different fold recognition approaches and report the probable folds based on a consensus approach [98,183,236–241]. They also integrate methods for prediction of local structural features and scoring functions to assess and refine the results.

Hybrid methods generate models by combining structures of multiple templates which are either a set of initial predicted models or those obtained from different fold recognition methods [89,198,231,242–245]. Pro-sp3-TASSSER [231] is one such method that identifies template fold based on PROSPECTOR [102] and SP3 [106]. They also involve realignment of target–template in the uncertain regions [197,246,247]. Pcons [230,248] generates structural models based on the target–template alignments generated by different approaches and has been quite competitive in the recent CASP experiments. Nevertheless, a consensus approach may also suppress the true result, if only one of the underlying methods is successful [249].

The recent IntFOLD-TS approach [250] is one of these successful consensus methods; it integrates multiple sequence–structure alignments from methods such as SP3 [106], SPARKS [114], HHsearch [140] and COMA [251] and generates many models with multiple and single templates and scores them. The combination of these methods resulted in a better accuracy in generating models for 117 CASP targets when compared with the component methods individually. eThread [252] is another meta-threading approach that uses a combination of 10 methods and generates consensus alignments using machine learning techniques. This approach is reported to be more sensitive than any of the component methods and nearly 50% of the models were of high quality (TMscore > 0.5) [253].

3. Conclusion

The favourable local backbone conformations and the associated amino acid preferences can be carefully extracted to predict the conformation of a new sequence. Several methods for the efficient prediction of local backbone conformations have been proposed in recent years. It is becoming increasingly clear that these methods can contribute significantly to improve the accuracy of identifying related folds. Careful extraction of sequence–structure relationships is important for reducing the extent of false predictions. The knowledge of preferred local conformations is also used in generating structural models by the assembly of fragments. Strong signals of sequence–structure preferences need to discriminate from weak associations like chameleon sequences, flexible and disordered regions. The use of predicted flexibility profiles of a protein structure may also add to the accuracy of recognizing correct folds. The sequence–structure relationships learnt can be further refined based on specific proteomes or based on the structural class of proteins. Preliminary studies with available data point towards an improvement in prediction accuracy with such dataset-specific learning. Hence an efficient integration of local structural preferences with the global features helps significantly in our efforts for recognizing the fold that an amino acid sequence will adopt.

Acknowledgements. We thank Dr Adam Wilkins (Frontiers in Life Sciences doctorate school (FDV Paris, France), and Biochemistry, Biotherapy, Molecular Biology, Infectiology—B3MI doctorate school (Univ Paris Diderot and Univ Paris Descartes, Paris, France)) for useful discussions and suggestions. The authors also thank Jean-Christophe Gelly, Catherine Etchebest, N. Srinivasan, R. Sowdhamini, R.M. Bhaskara, S. Mahajan and G. Agarwal for fruitful discussions.

Funding statement. This work was supported by grants from the Ministry of Research (France), Univ Paris Diderot (France), Sorbonne Paris Cité (France), the National Institute for Blood Transfusion (INTS, France) and the National Institute for Health and Medical Research (INSMR, France). A.P.J. received a grant from CEFIPRA number 3903-E, and acknowledges NCBS for the post-doctoral fellowship. A.B. acknowledges CEFIPRA for a collaborative grant (no. 3903-E).


213. Pailla U, Kondam R, Ranjan A. 2008 Genome bias influences amino acid choices: analysis of

223. Xu D, Zhang Y. 2012


