Confidence-Guided Local Structure Prediction with HHfrag

Ivan Kalev[1]*, **Michael Habeck**[2]*

[1] Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany, [2] Institute for Mathematical Stochastics, Georg-August-University of Göttingen, Göttingen, Germany

Abstract

We present a method to assess the reliability of local structure prediction from sequence. We introduce a greedy algorithm for filtering and enrichment of dynamic fragment libraries, compiled with remote-homology detection methods such as HHfrag. After filtering false hits at each target position, we reduce the fragment library to a minimal set of representative fragments, which are guaranteed to have correct local structure in regions of detectable conservation. We demonstrate that the location of conserved motifs in a protein sequence can be predicted by examining the recurrence and structural homogeneity of detected fragments. The resulting confidence score correlates with the local RMSD of the representative fragments and allows us to predict torsion angles from sequence with better accuracy compared to existing machine learning methods.

Citation: Kalev I, Habeck M (2013) Confidence-Guided Local Structure Prediction with HHfrag. PLoS ONE 8(10): e76512. doi:10.1371/journal.pone.0076512

Editor: Yang Zhang, University of Michigan, United States of America

Received July 4, 2013; **Accepted** August 28, 2013; **Published** October 16, 2013

Funding: This work has been supported by contract research "Methoden in den Lebenswissenschaften" of the Baden-Württemberg Stiftung, by Deutsche Forschungsgemeinschaft (DFG) grant HA 5918/1-1, and by the Max Planck Society. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ivan.kalev@tuebingen.mpg.de (IK); mhabeck@gwdg.de (MH)

Introduction

Deciphering the protein folding problem remains a fundamental challenge. Although experimental and theoretical studies have improved our understanding of the process, there are still many open problems, one of which is to reliably predict the native structure from sequence only. It has become clear that protein sequences do not adopt unlimited varieties of global and local shapes. Proteins that fold do not explore the complete conformational space [1]. Rather, the local structure of each polypeptide segment is biased by the geometrical and chemical properties of its constituent amino acids [2]. This observation prompted the development of structural alphabets in an attempt to partition known protein structures into a dictionary of discrete motif prototypes [3]. It has been reported that such fragment libraries may be sufficient to describe all protein folds in terms of recurrent building blocks [4,5].

One of the first efforts to systematically study the amino acid preferences of known structural motifs is the I-Sites fragment library [6,7], which was later adapted for use in *ab initio* fragment assembly [8] with Rosetta [9]. This has proven to be a successful strategy for local structure prediction from sequence. However, not all motifs have identifiable sequence preferences. A fixed set of sequence-based prototypes is generally insufficient to detect all structural elements in existing protein structures [10]. Although the sensitivity of sequence-based fragment detection can be pushed to higher levels by dynamic fragment selection [9], our studies show that protein structures are not simple combinations of conserved sequence motifs. Rather, we observe an alternating pattern of easily detectable elements (often matching to known I-Sites), connected by highly variable regions with no detectable sequence conservation (typically flexible coils and linkers) [10]. The ability to discriminate between these high- and low-precision regions is important, since local structure prediction in non-conserved regions is unreliable and should not be trusted. Conversely, protocols that assemble fragments should be able to obtain information about the locations of the conserved motif instances in a given protein sequence and utilize the corresponding torsion angle predictions with higher precedence.

In this study, we introduce a reliable algorithm for blind prediction of local high- and low-precision regions in protein sequences. By analysing the structural consistency and recurrence of motifs in dynamic HHfrag libraries, this method quantifies the quality of fragment assignment at each query position and nominates representative fragments, which are most likely to match the local structure of the query closely. We show that the confidence score of local structure prediction correlates well with the local RMSD and torsion angle error of representative fragments and can be used as a reliable predictor for the presence of high- or low-precision regions. Finally, we illustrate how this property can be used to predict torsion angles from sequence with higher accuracy than existing machine learning methods [11,12].

Methods

To predict the torsion angles of a target protein of unknown structure, we first build a dynamic fragment library using the HHfrag method for fragment detection from sequence [10]. The reliability of fragment detection at each target position is then analyzed by clustering and filtering all fragments, covering a given target residue. If the obtained confidence score for this position is indicative of a local region of high precision, our algorithm

```
procedure ShrinkCluster($G$)
    while $G.D > 1.5$ and $|G.V| > 1$ do
        $outlier = nil$
        $D_{opt} = \infty$

        for each $v \in G.V$ do
            $D'_v = (G.W - v.W)/(|G.E| - |v.Adj|)$
            if $D'_v < D_{opt}$ then
                $outlier = v$
                $D_{opt} = D'_v$
            end if
        end for

        if $D_{opt} < G.D$ then
            RejectVertex($G$, $outlier$)
        else
            Error("Diverging cluster")
        end if
    end while

    return ComputeCentroid($G$)
end procedure
```

Figure 1. The outlier rejection algorithm. The *ShrinkCluster* procedure operates on an undirected graph G. The algorithm keeps track of the total sum of all edge weights in the graph ($G.W$) and their average ($G.D$). $V.W$ denotes the total weight of all edges incident to a given vertex V.
doi:10.1371/journal.pone.0076512.g001

proceeds by selecting a representative fragment (the centroid) and extracts the torsion angles of the centroid. The final list of predicted torsion angles for the entire target sequence is compiled from the set of all nominated centroids.

Dynamic Fragment Selection

For a given target sequence, we compute a dynamic library of variable-length fragments using the standard HHfrag protocol [10]. HHfrag is a sensitive and accurate fragment detection method, which uses internally the HHsearch [13] algorithm for local alignment of pairs of profile hidden Markov models (HMM).

For this purpose, both the query sequence and all template proteins are represented as profile HMMs with secondary structure information incorporated into the HHM files [14]. Sequence profiles are generated with PSI-BLAST [15]. Secondary structure information is calculated with DSSP [16] from the experimental structures of all templates or predicted with PSIPRED [17] for the query. The database of template structures (PDBS25) is a non-redundant subset of PDB [18], derived from the April 2010 build of PDBselect25 [19] (4824 protein chains, filtered at 25% sequence identity).

All detected fragments are directly excised by HHfrag from their corresponding experimental structures. The resulting fragment library is a position-specific, ordered set of structural motifs, ranging from 6 to 21 residues in length [10]. Each fragment is described by its matching query/subject positions, (φ,ψ) torsion angle pairs and backbone coordinate trace.

```
procedure RejectVertex($G$, $outlier$)
    $G.V = G.V - \{outlier\}$
    $G.Adj = G.Adj - \{outlier\}$
    for each $v \in G.V$ do
        if $outlier \in G.Adj[v]$ then
            $G.W = G.W - \omega(v, outlier)$
            $v.W = v.W - \omega(v, outlier)$
            $G.Adj[v] = G.Adj[v] - \{outlier\}$
        end if
    end for
end procedure

procedure ComputeCentroid($G$)
    $centroid = nil$
    $D_{min} = \infty$
    for each $v \in G$ do
        if $|G.Adj[v]|/|G.V| \geq 0.5$ then
            if $v.D < D_{min}$ then
                $centroid = v$
                $D_{min} = v.D$
            end if
        end if
    end for
    return centroid
end procedure
```

Figure 2. Supporting procedures.
doi:10.1371/journal.pone.0076512.g002

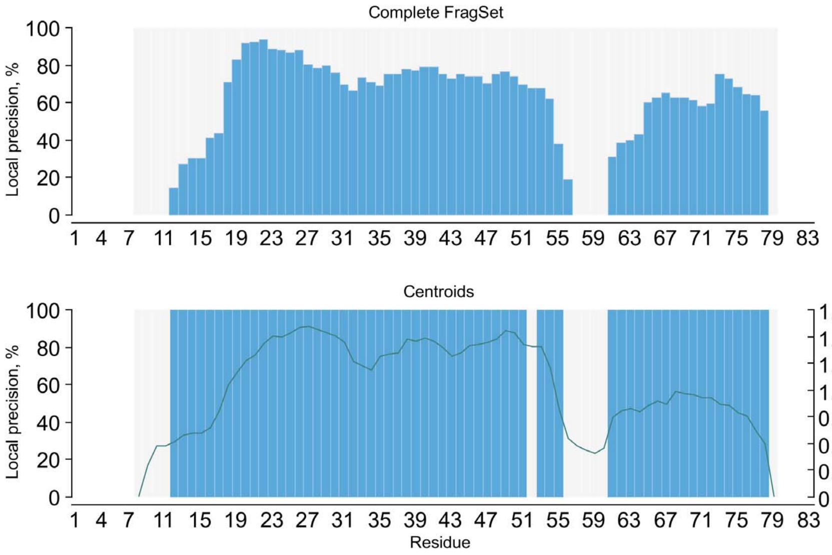

Figure 3. Local precision of filtered fragment libraries. The local precision of the complete HHfrag library for benchmark target 3 nzl was measured by counting the percentage of assigned fragments below 1.5 Å C_α-RMSD to the native structure for each target residue (blue bars; see [10]). The library was then filtered using the outlier rejection algorithm, described in Methods. The precision of the resulting library of representative fragments was measured in a similar way, except that only one fragment per target position (its associated centroid) was taken into account. The green curve shows the corresponding confidence values for each target residue.

doi:10.1371/journal.pone.0076512.g003

Fragment Clusters

To eliminate outliers in the raw dynamic fragment libraries and select representative fragments for each target position, we propose a greedy *outlier rejection* algorithm.

For every position i in the target sequence, we build a cluster of all fragments, covering this position. Each *fragment cluster* is represented by a graph whose nodes are fragments connected by weighted edges. The edge weights are the C_α-RMSDs between each given pair of fragments. Since all HHfrag libraries are composed of motifs of varying length and start positions, some fragments in a given cluster may overlap by less than 6 residues. In such cases, the RMSD cannot be a meaningful indicator for the structural divergence between fragments. Therefore, we do not connect these pairs with an edge.

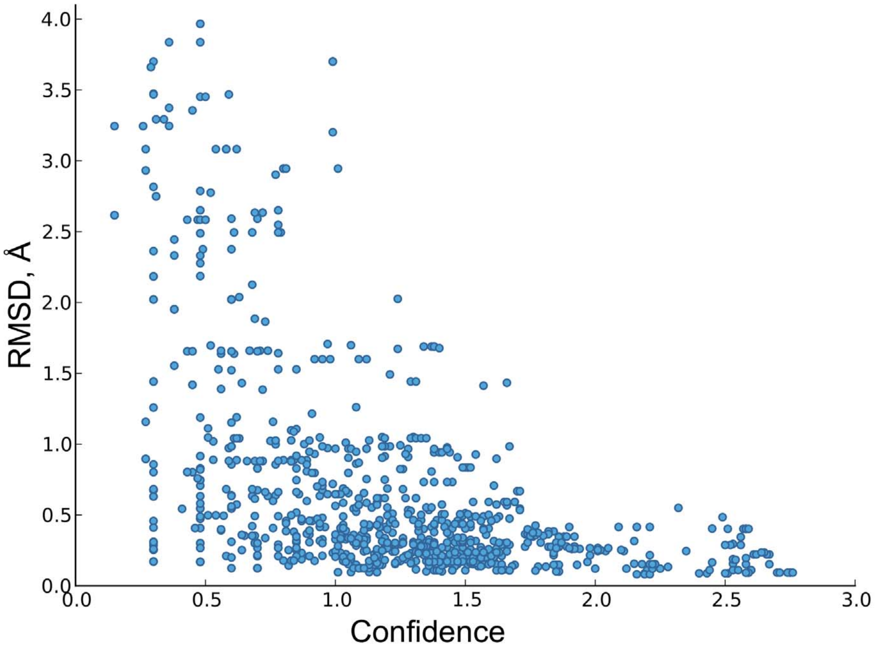

Figure 4. Reliability of the confidence score. Shown is the correlation between the confidence scores of all residue-wise centroids in our CASP9 benchmark and their C_α-RMSD similarity to the corresponding native structures. The RMSD values were calculated over the entire lengths of the representative fragments.

doi:10.1371/journal.pone.0076512.g004

Table 1. Torsion angle prediction performance.

Method	Confidence	MAE (φ)	MAE (ψ)
TANGLE	0.8	31.9±34.9°	90.7±30.6°
ANGLOR	0.8	18.7±25.8°	86.4±43.0°
HHfrag	0.8	18.6±27.0°	22.5±36.2°
TANGLE	0.0	34.2±36.4°	87.4±32.3°
ANGLOR	0.0	23.5±30.0°	84.7±47.6°
HHfrag	0.0	25.4±34.7°	34.9±48.9°

Mean absolute error (MAE) of φ and ψ torsion angle prediction for high-confidence ($C \geq 0.8$) and all residues ($C \geq 0$) in our benchmark.
doi:10.1371/journal.pone.0076512.t001

Each fragment cluster has two key properties:

1. *Recurrence* (r) – refers to the sequence conservation of a structural motif. The recurrence of a given motif is measured by counting the number of its instances in the non-redundant database of profiles (PDBS25). We compute the recurrence of a query position and its associated cluster by simply counting the number of assigned fragments, covering this position, which is the number of vertices in the cluster.
2. *Consistency* (c) – characterizes the structural homogeneity of the fragments. We measure the consistency of a cluster by calculating the subset of structurally similar pairs of fragments. Two vertices are considered similar if the weight of their connecting edge, measured by the C_α-RMSD between the two fragments, does not exceed a critical threshold of 1.5 Å.

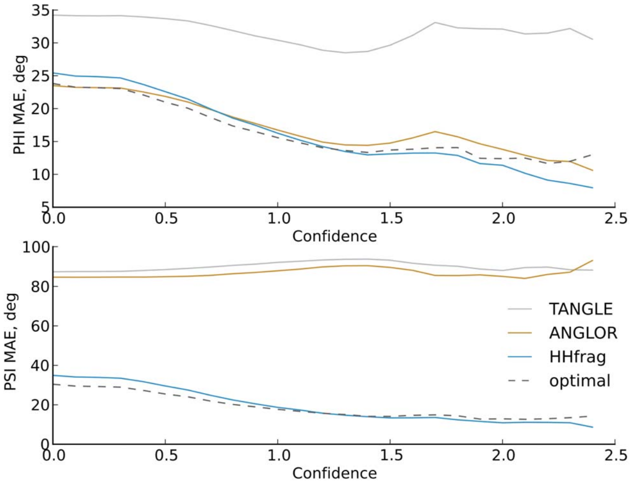

Figure 6. Torsion angle prediction accuracy at increasing confidence cutoffs. We measured the mean absolute error (MAE) of φ and ψ angle prediction at increasing confidence cutoffs. For each cutoff, we computed the φ and ψ MAE for all target residues in our benchmark, having a confidence greater or equal to the cutoff. The optimal curve represents the lowest possible MAE, which our filtering algorithm could achieve by always picking a centroid identical to the best-fitting fragment at each position (the fragment with the lowest C_α-RMSD).
doi:10.1371/journal.pone.0076512.g006

Filtering

The goal of the outlier rejection algorithm is to improve the structural consistency of a given cluster by performing a minimum number of node deletions (Figure 1).

Every cluster (G) keeps track of the total sum of all of its pairwise RMSDs (W). Each cluster vertex (v) also maintains an up-to-date

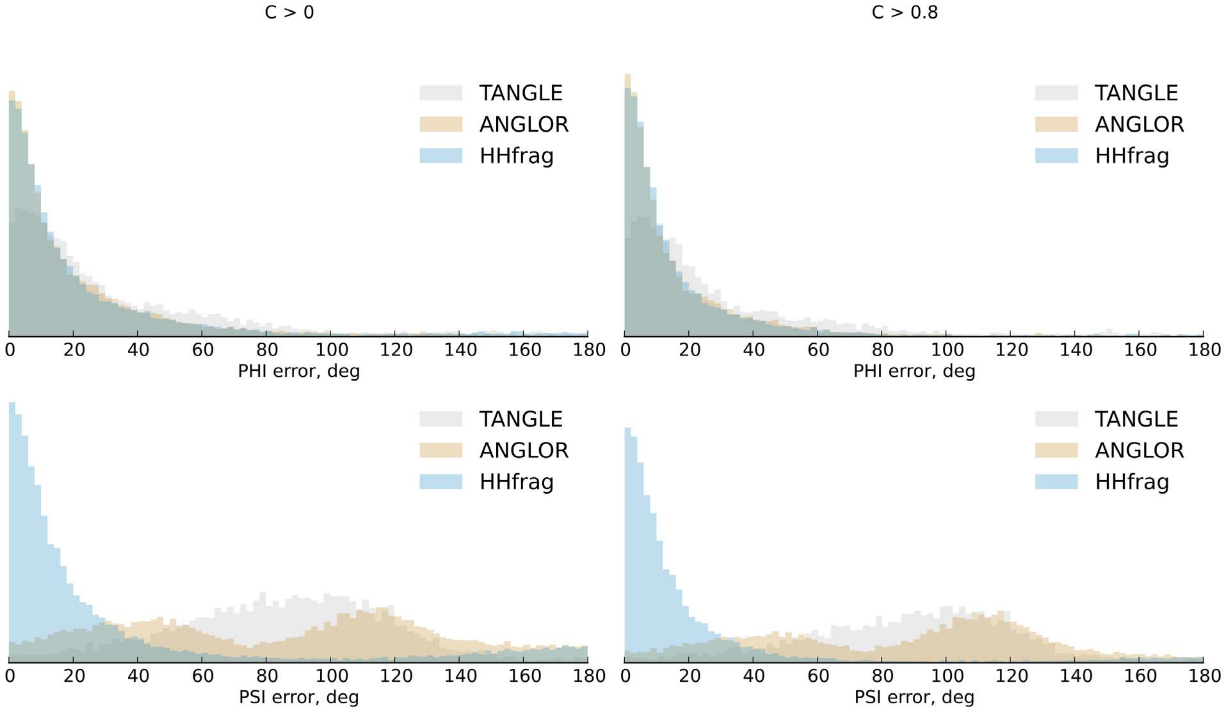

Figure 5. Distributions of the absolute errors of predicted torsion angles. Shown are the distributions of φ and ψ prediction errors for high-confidence (right) and all target residues (left) in our benchmark.
doi:10.1371/journal.pone.0076512.g005

sum of the weights of all edges incident to it W_v. A fragment cluster is said to be *stable*, when the average RMSD D between all adjacent vertices is lower than the threshold of 1.5 Å:

$$D=\frac{1}{|E|}\sum_{(u,v)\in E}\omega(u,v)\leq 1.5\,\text{Å} \tag{1}$$

where E is the set of all edges and $\omega(u,v)$ is the RMSD between a pair of fragments u and v. The algorithm performs iterative rejections, until the cluster stability criterion is satisfied. On each iteration, we probe all nodes by calculating the resulting average RMSD D'_v if vertex v is excluded. This is given by the following greedy criterion:

$$D'_{opt}=\min_{v\in V}\frac{\sum_{e\in E}\omega(e)-\sum_{u\in N(v)}\omega(u,v)}{|E|-|N(v)|} \tag{2}$$

where $N(v)$ is the adjacency set of vertex v. The fragment, whose exclusion from the cluster would produce the most significant drop in D' towards stability ($D\leq 1.5$Å), is selected for deletion and removed.

When the graph is implemented using an adjacency sets data structure, each removal requires linear time of $O(|V|)$ (Figure 2), needed to update all adjacency sets (linear complexity) and recalculate the cached sums of weights W_v of affected nodes (constant time per fragment). If no fragment removal results in decrease of the average RMSD D, this cluster is not able to shrink further. Such clusters are considered *diverging*, which indicates heterogeneous aggregates of false positive fragments. In such cases we terminate the filtering procedure and the corresponding target position remains unassigned, additionally marked to be part of a low-precision region. The same result is also obtained if all cluster nodes are rejected before stability has been reached.

The maximum number of iterations equals the number of fragments $|V|$ in a cluster. The amount of work performed on each iteration k is equal to the current number of nodes $|V_k|$ on iteration k (to identify the candidate for rejection) plus additional $|V_k|$ (to remove the candidate and update all adjacency sets and cached weight sums). The worst-case running time of the filtering algorithm is thus given by:

$$\sum_{k=1}^{|V|}2|V_k|=\Theta(|V|^2+|V|) \tag{3}$$

However, most clusters reach stability much earlier than $k=|V|$ number of iterations, so the typical running time is in practice much better.

Representative Fragments and Confidence

We define the *representative* fragment of a *stable cluster* to be its centroid, which is the node with the minimum average RMSD to its adjacent vertices (Figure 2). Since the number of edges per node may vary, we consider only fragments connected to a significant number of cluster elements ($\geq 50\%$).

The *confidence* score C, assigned to a given target position i and its representative fragment, we derive from the recurrence (r) and structural consistency values (c) of the corresponding filtered (stable) cluster:

$$C=c\,\log_{10}\,r=\frac{e}{|E|}\log_{10}|V| \tag{4}$$

where $|V|$ is the number of fragments in the filtered cluster, $|E|$ is its total number of edges and e is the number of pairwise RMSDs not greater than 1.5 Å.

Prediction of Torsion Angles

We use the filtered fragment libraries and their associated representative fragments for direct prediction of torsion angles from sequence – a strategy, which bears resemblance to earlier approaches [20]. For each position i in a given target protein, we build a fragment cluster and compute the centroid fragment, as outlined above. The pair of torsion angle values (φ_i,ψ_i) of the representative fragment at target position i is extracted from the centroid's experimental structure and directly reported as the final prediction at that position. A confidence value of 0.8 or higher indicates a reliable prediction within a local region of high precision.

Performance Evaluation

We used a set of 106 protein targets from the CASP9 competition [21] to benchmark the accuracy of torsion angle prediction. The PDBS25 database of template structures, used for fragment extraction by HHfrag, contains only older entries and no homologous chains. For each target, we obtain a prediction for its torsion angles with the procedure, described above. The prediction accuracy was measured by the mean absolute error (MAE) between the predicted (P) and experimental (E) torsion angle values:

$$MAE=\frac{1}{\sum_{i=1}^{M}L_i}\sum_{i=1}^{M}\sum_{j=1}^{L_i}|P_{ij}-E_{ij}| \tag{5}$$

where M is the number of proteins and L_i is the number of residues in protein i of confidence greater than a chosen cutoff ($C>x$). All predicted and experimental torsion angles are computed in degrees within the $[-180^{\circ},180^{\circ}]$ range. To keep the error values in that range as well, we apply the following rule when calculating the absolute angular errors $|AE_{ij}|$:

$$|AE_{ij}|=\begin{cases}|AE_{ij}+360| \text{ if } AE_{ij}<-180\\ |AE_{ij}-360| \text{ if } AE_{ij}>+180\\ |AE_{ij}| \text{ otherwise}\end{cases} \tag{6}$$

Predicted torsion angles for the same set of proteins were also obtained with ANGLOR [11] and TANGLE [12] using their web server interfaces. MAE values were calculated using an identical procedure.

The quality of the fragment library is also evaluated in terms of local *precision* and *coverage* (see [10]). The *accuracy* of an assigned fragment is assessed by the local C_α-RMSD of the fragment to the native structure. If this RMSD is below 1.5 Å, the fragment is considered a positive (compatible) hit. The percentage of correctly assigned fragments that cover the same residue is the local precision of the fragment library. The coverage is the percentage of residues that are covered by at least one compatible fragment.

Results and Discussion

In an earlier study, we discussed the precision of popular methods for local structure prediction by sequence-based remote homology detection [10]. We have shown that the precision of fragment detection with this strategy is relatively low (40% on average for Rosetta NNmake [9,22] and 70% for HHfrag). Additionally, the precision of dynamic fragment libraries is never uniform along the target sequences. Local zones of high precision emerge in regions, containing detectable, well-known structural motifs [10]. These motifs were found to exhibit a certain degree of sequence profile conservation and can be observed in a wide range of evolutionary unrelated proteins, thus serving the purpose of structural design patterns [4,6]. The quality of fragment libraries however rapidly decreases as we move away from the high-precision regions and enter areas of very high variability, such as loops and linkers.

In this study, we propose an intuitive model for the prediction of local high- and low-precision zones and demonstrate how this method can be applied to increase the reliability of local structure prediction.

The Confidence Score

A key property of locally conserved motifs is that they usually have good local sequence-structure correlation [6]. Such query sequence regions generate lists of matching fragment instances of higher *structural consistency* (Figure 3). Additionally, the most conserved motifs, such as the *GD box* [23], can be highly ubiquitous and often contain hundreds of detectable instances in a non-redundant structural database. The *recurrence* of a given motif was determined to be an equally strong indicator for reliable local structure prediction and this observation already plays an integral role in the HHfrag fragment detection method [14].

The *confidence* score of fragment assignment (Equation 4) integrates these two properties. The recurrence term in the expression is a weighting factor for the structural consistency of all instances of a given motif. Highly conserved motifs have a recurrence of 50–100 or more, which increases confidence because the structural consistency amplified by a factor greater than one. Clusters of size greater than 10 are up-weighted because 10 is the critical number of HHsearch hits, below which the program switches to a less strict, greedy hit-ranking mode [10,13]. At $r = 10$, the logarithm of the recurrence is 1 and the confidence is determined entirely by the degree of structural consistency. Clusters of size less than 10 are associated with increasing uncertainty and thus severely penalized. We can follow the same intuition to define natural thresholds for the confidence score:

1. $C > 1$: credible local structure prediction, which is guaranteed to be accurate. For example, a confidence value of 1.5 can be obtained for a highly homogeneous cluster (75%) of large size (100 fragment instances).
2. $C = 1$: transitional zone. Confidence value of 1 corresponds to a rare motif (10 instances) with maximum structural conservation or a highly abundant motif (100 instances) at moderate consistency of 50%.
3. $C < 0.8$: uncertainty. This confidence threshold is equivalent to a small fragment cluster (10 instances) at consistency equal to the average precision of HHfrag for ordinary I-Sites (80%) or a highly recurrent motif (100 instances) at low precision (40%).

Filtered Fragment Libraries

To maximise the precision of fragment-based local structure prediction in conserved regions, we propose a filtering algorithm used to compile reduced fragment libraries of low complexity and very high local precision. For each query position, we build a fragment cluster, as outlined in Methods. Inconsistent fragments in each cluster are iteratively rejected until a sufficient level of structural consistency is reached. A single, representative fragment is then selected out of the pool of surviving cluster members. After removing the outliers in all clusters, the entire library is enriched with high-quality fragments. In regions of local conservation, this always results in local centroid precision of 100%, i.e. representative fragments in those regions are guaranteed to have a low RMSD to the native structure (C_α-RMSD ≤ 1.5 Å).

This is illustrated in Figure 3. After filtering the raw HHfrag library for target 3 nzl, we obtain a list of position-specific representative fragments (one fragment per query position). The local precision of the resulting filtered library of centroids is 100% for all high-accuracy regions, observed on the original plot (see Figure 3 for details). The confidence curve correlates well with the observed local precision pattern, dropping rapidly in regions where inaccurate centroids have been selected. Similar results were obtained after filtering all remaining CASP9 targets from the standard HHfrag benchmark [10] (see supplementary material).

Confidence-guided Local Structure Prediction

To study the reliability of the confidence score as an indicator for local motif conservation, we measured the local precision of calculated centroids for all targets in the standard HHfrag benchmark [10].

Figure 4 shows the overall correlation between local accuracy of cluster centroids and confidence in our benchmark. A weak confidence value (0.1–0.6) is a clear signal for the presence of a low-accuracy region. Higher confidence values (0.8–1.0) indicate generally conserved motifs, which sometimes cannot be predicted reliably. The overall centroid precision in this confidence interval is $80 \pm 17\%$ with an average RMSD to native structures of 1.0 ± 0.9 Å. Confidence greater than 1.0 guarantees an accurate and reliable local structure prediction with a very low chance of an error. The overall precision in such regions reaches $92 \pm 13\%$ with an average RMSD to native structures as low as 0.58 ± 0.57 Å. These results confirm the expected confidence thresholds derived in the previous section.

Based on these observations, we propose a context-aware method for prediction of torsion angles from sequence (see Methods). In query regions of high confidence (0.8 or better), we rely on the corresponding centroids as a source of highly accurate torsion angle predictions. This allows client methods to consume predicted torsion angles in order of decreasing confidence. The confidence score brings contextual perspective to the local structure prediction, allowing fragment assembly applications to decide at runtime whether a given prediction should be trusted or rather replaced by exhaustive sampling of a generic structural alphabet.

Benchmark

We examined the performance of our centroid-based torsion angle predictor on 106 protein targets from the CASP9 competition [21]. The mean absolute error (MAE) of predicted φ and ψ angles was compared against the values, obtained with two popular methods for torsion angle prediction from sequence: ANGLOR [11] and TANGLE [12]. The overall precision of HHFrag in comparison with these methods is summarized in Table 1.

When regions of any confidence are considered, our method predicts φ angles with slightly lower accuracy than ANGLOR (2° higher MAE), but better than TANGLE. For ψ angles however, HHfrag is significantly more accurate, improving on both ANGLOR and TANGLE by a 50° lower MAE (Figure 5). The observed MAE of HHFrag is 25.4° for φ and 34.9° for ψ angles on average.

As expected, the quality of torsion angle prediction with HHfrag improves further when the confidence score of each query position is taken into account (Figure 6. In target regions of $C \geq 0.8$, the average MAE drops by 6.8° and 12.4° for φ and ψ angles respectively. Generally, the MAE of HHfrag predictions gradually decreases as we discard regions of lower confidence (Figure 6). Such tendency is less pronounced for φ angle predictions with ANGLOR or TANGLE and completely lacking when these methods are used to predict ψ angles (Figure 6). HHfrag does not always select optimal centroids in low-confidence regions ($C < 0.8$) as the lack of sufficient recurrence and consistency of such clusters hinders the filtering algorithm. However, in transitional zones ($C \geq 0.8$), the deviation from the optimal MAE becomes negligible and for high-confidence regions ($C \geq 1$) our method is guaranteed to extract torsion angles from the best-fitting fragment at each position. These results highlight the importance of taking the local conservation into account and confirm the utility of our confidence-guided prediction strategy.

Availability

The fragment filtering algorithm and the confidence-guided torsion angle predictor are implemented as an HHfrag extension in version 1.2 of the CSB open-source SDK [24]. An HHfrag web server is available at http://toolkit.tuebingen.mpg.de/hhfrag. The standalone executable, source code and Python API are freely available for download at http://csb.codeplex.com/releases.

Conclusion

We discussed the correlation between the quality of local structure prediction from sequence and the degree of local motif conservation. We introduced a greedy algorithm for fragment filtering, which can be used to decrease the complexity of dynamic, sequence-based fragment libraries. This algorithm takes a central part of our confidence-guided framework for prediction of local conservation, which captures the structural homogeneity and recurrence of detected fragments. Protein sequence regions, containing instances of ubiquitous and structurally consistent motifs, generally correspond to zones of very high local accuracy. We showed that this information can be used for reliable prediction of torsion angles from sequence with better accuracy compared to existing machine learning methods.

Supporting Information

Benchmark S1 Local centroid precision for each target in the benchmark set and a breakdown of the torsion angle prediction performance by residue type and secondary structure.
(ZIP)

Author Contributions

Conceived and designed the experiments: IK MH. Performed the experiments: IK. Analyzed the data: IK. Wrote the paper: IK MH.

References

1. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7: 95–99.
2. Bystroff C, Simons KT, Han KF, Baker D (1996) Local sequence-structure correlations in proteins. Curr Opin Biotechnol 7: 417–421.
3. Offmann B, Tyagi M, de Brevern AG (2007) Local protein structures. Current Bioinformatics 2: 165–202(38).
4. Fernandez-Fuentes N, Dybas JM, Fiser A (2010) Structural characteristics of novel protein folds. PLoS Comput Biol 6: e1000750.
5. Fernandez-Fuentes N, Oliva B, Fiser A (2006) A supersecondary structure library and search algorithm for modeling loops in protein structures. Nucleic Acids Res 34: 2085–2097.
6. Bystroff C, Baker D (1998) Prediction of local structure in proteins using a library of sequencestructure motifs. J Mol Biol 281: 565–577.
7. Bystroff C, Webb-Robertson BJ (2008) Pairwise covariance adds little to secondary structure prediction but improves the prediction of non-canonical local structure. BMC Bioinformatics 9: 429.
8. Bystroff C, Shao Y (2002) Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. Bioinformatics 18 Suppl 1: 54–61.
9. Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. Meth Enzymol 383: 66–93.
10. Kalev I, Habeck M (2011) HHfrag: HMM-based fragment detection using HHpred. Bioinformatics 27: 3110–3116.
11. Wu S, Zhang Y (2008) ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction. PLoS ONE 3: e3400.
12. Song J, Tan H, Wang M, Webb GI, Akutsu T (2012) TANGLE: two-level support vector regression approach for protein backbone torsion angle prediction from primary sequences. PLoS ONE 7: e30361.
13. Söding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21: 951–960.
14. Hildebrand A, Remmert M, Biegert A, Söding J (2009) Fast and accurate automatic structure prediction with HHpred. Proteins 77 Suppl 9: 128–132.
15. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSIBLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
16. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.
17. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195–202.
18. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.
19. Griep S, Hobohm U (2010) PDBselect 1992–2009 and PDBfilter-select. Nucleic Acids Res 38: D318–319.
20. Xu D, Zhang Y (2013) Toward optimal fragment generations for ab initio protein structure assembly. Proteins 81: 229–239.
21. Moult J, Fidelis K, Kryshtafovych A, Tramontano A (2011) Critical assessment of methods of protein structure prediction (CASP)-round IX. Proteins 79 Suppl 10: 1–5.
22. Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 32: W526–531.
23. Alva V, Dunin-Horkawicz S, Habeck M, Coles M, Lupas AN (2009) The gd box: A widespread non-contiguous supersecondary structural element. Protein Science 18: 1961–1966.
24. Kalev I, Mechelke M, Kopec KO, Holder T, Carstens S, et al. (2012) CSB: a Python framework for structural bioinformatics. Bioinformatics 28: 2996–2997.