Journal of Theoretical Biology 328 (2013) 77–88

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

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

The utility of artificially evolved sequences in protein threading and fold recognition

Michal Brylinski [a,b,*]

[a] Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA
[b] Center for Computation & Technology, Louisiana State University, Baton Rouge, LA 70803, USA

HIGHLIGHTS

- A novel modeling stratagem to improve fold recognition is introduced.
- We developed a new method for the optimization of amino acid sequences.
- Artificial sequences have significant capabilities to recognize correct structures.
- Fold recognition systematically improves the detection of structural analogs.
- More sensitive threading methods to target midnight zone templates are suggested.

ARTICLE INFO

Article history:
Received 19 October 2012
Received in revised form
24 January 2013
Accepted 18 March 2013
Available online 27 March 2013

Keywords:
Artificial sequences
Evolved sequences
Protein threading
Protein structure modeling
Template-based modeling

ABSTRACT

Template-based protein structure prediction plays an important role in Functional Genomics by providing structural models of gene products, which can be utilized by structure-based approaches to function inference. From a systems level perspective, the high structural coverage of gene products in a given organism is critical. Despite continuous efforts towards the development of more sensitive threading approaches, confident structural models cannot be constructed for a considerable fraction of proteins due to difficulties in recognizing low-sequence identity templates with a similar fold to the target. Here we introduce a new modeling stratagem, which employs a library of synthetic sequences to improve template ranking in fold recognition by sequence profile-based methods. We developed a new method for the optimization of generic protein-like amino acid sequences to stabilize the respective structures using a combined empirical scoring function, which is compatible with these commonly used in protein threading and fold recognition. We show that the artificially evolved sequences, whose average sequence identity to the wild-type sequences is as low as 13.8%, have significant capabilities to recognize the correct structures. Importantly, the quality of the corresponding threading alignments is comparable to these constructed using conventional wild-type approaches (the average TM-score is 0.48 and 0.54, respectively). Fold recognition that uses data fusion to combine ranks calculated for both wild-type and synthetic template libraries systematically improves the detection of structural analogs. Depending on the threading algorithm used, it yields on average 4–16% higher recognition rates than using the wild-type template library alone. Synthetic sequences artificially evolved for the template structures provide an orthogonal source of signal that could be exploited to detect these templates unrecognized by standard modeling techniques. It opens up new directions in the development of more sensitive threading methods with the enhanced capabilities of targeting difficult, midnight zone templates.

1. Introduction

Systems biology has emerged to help understand how the components of complex living systems interact and how their malfunction causes disease (Kitano, 2002). To perform a systems-level analysis of molecular interactions, one requires a complete list of functionally annotated genes and proteins specified by these genes (or briefly gene products) within an organism. Numerous genome sequencing projects have already provided the community with a vast amount of sequence information; however, due to the lack of functionally annotated close homologs, the molecular functions of many of these gene products remain unknown (Mi et al., 2003). To carry out functional inference in the low sequence

* Correspondence address: Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA.
Tel.: +1 225 578 2791.
E-mail address: michal@brylinski.org

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.03.018

identity regime, a number of structure-based methods have been developed (Brylinski et al., 2007; Elcock, 2001; Hetenyi and van der Spoel, 2006; Huang and Schroeder, 2006). Early methods for function inference from protein structure were extremely sensitive to the quality of the target structures and typically required these solved experimentally by X-ray crystallography or NMR. More recent approaches, many of which employ global structure alignments (Wass et al., 2010), evolutionary restraints (Brylinski and Skolnick, 2008) or a low-resolution description of protein structures (Liu and Vakser, 2011), are devoid of these limitations and show a considerable promise for proteome-scale functional annotation (Skolnick and Brylinski, 2009). On that account, protein structure modeling plays an important role in Functional Genomics by providing structural information on gene products that is subsequently utilized by powerful structure-based approaches to protein function inference (Drew et al., 2011; McGuffin et al., 2006; Shah et al., 2003).

From a point of view of a system-wide modeling, the high coverage of protein sequences in a given organism with structural information is critical. Experimental structure determination methods provide high-resolution structure information only for a small subset of proteins, whilst the computational structure prediction supplies valuable information for a large number of sequences, whose structures will not be determined experimentally. The most successful algorithms for protein structure modeling build on homology to proteins with known structures. These approaches first identify template proteins, whose structures are presumably similar to that adopted by the target sequence. Next, the alignment of the target sequence to the template structure is generated and a three-dimensional model of the target is constructed. A successful modeling requires structurally related templates in the available databases, such as the Protein Data Bank (PDB) (Berman et al., 2000). A systematic analysis has demonstrated that the PDB is likely complete at the level of compact, single domain protein structures, i.e. for almost every target, a significantly similar structure is present in the PDB (Zhang et al., 2006); however, finding them still renders significant challenges for a subset of targets.

Protein threading, also known as fold recognition, has been developed to search for high- as well as low-sequence identity templates that can be used to construct the structural model of a target protein (Jones et al., 1992). Currently, state-of-the-art threading techniques can identify templates with a structurally significant alignment for about 70% of the targets. In proteome-scale structure and function prediction projects, these target proteins for which threading fails to identify structural templates are typically excluded from the modeling process (Brylinski and Skolnick, 2011), which reduces the structural and functional coverage of gene products.

Currently, considerable efforts are directed toward increasing the sensitivity of protein threading and fold recognition. Many recently developed approaches employ suboptimal alignments (Chen and Kihara, 2011), low-ranked templates (Pandit and Skolnick, 2010), multiple sequence/template alignments (Peng and Xu, 2011), the ensembles of "fragment alignments" (Kuziemko et al., 2011) or new scoring functions for low-homology template detection (Peng and Xu, 2010). In this study, instead of improving scoring functions or alignment construction algorithms, we investigate whether the imperfections of existing approaches can be turned into their advantages by applying an effective modeling stratagem. The motivation for this project has its origin in an interesting observation made during the development of sequence profile hidden Markov models (HMMs) for protein homology detection (Soding, 2005). One of the important components of a scoring function is the secondary structure match between the target and the template. Here, the secondary structure predicted for a target sequence can be scored either against known secondary structure assigned based on the experimental structure of a template (Frishman and Argos, 1995) or against that predicted from the template sequence (Jones, 1999). It turned out that, paradoxically, the overall sensitivity of HMMs increases when predicted instead of known secondary structure is used. One possible explanation is that the scoring of the predicted vs. predicted secondary structure is better optimized for remote homologies, which have diverged more during the course of evolution. Also, this type of scoring may account for systematic errors in secondary structure prediction. In other words, for two sequence fragments that adopt a similar secondary structure, a prediction algorithm can make a systematic mistake, which would still result in a good matching score, despite the incorrect predictions. Of course, a mispredicted fragment of the target would be scored poorly against the known secondary structure assigned based on the experimental structure of the template. Here, we introduce a similar strategy, but in a more general context, *viz.* we explore the possibility of using synthetic sequences artificially evolved for the template structures rather than (or in addition to) wild-type sequences in threading and fold recognition. These artificial sequences, also referred to as evolved or synthetic sequences, are optimized to stabilize the respective structures by a variety of potentials, which are compatible with those already commonly used in protein threading. We demonstrate that such synthetic sequences may provide an orthogonal source of signal that could be advantageously exploited in protein structure modeling.

2. Methods

2.1. Dataset

The development and optimization of the scoring function and simulation protocols as well as benchmarking calculations are carried out for a non-redundant at the 35% sequence identity level set of 10,558 proteins 50–600 residues in length selected from the Protein Structure Classification (CATH) database (Orengo et al., 1997). The pairwise sequence identity threshold automatically excludes close homologs from benchmarks. Pairwise structure alignments for the dataset proteins are constructed by fr-TM-align (Pandit and Skolnick, 2008) and evaluated by a TM-score, which is a length-independent measure that ranges from 0 to 1. A value of 0.4 indicates a statistically significant similarity with a p-value of 3.4×10^{-5} (Zhang and Skolnick, 2004). We note that a TM-score of 0.4 is an appropriate fold similarity assignment threshold; template structures above a TM-score of 0.4 contain sufficient information to enable the full-length reconstruction of the target structure (Skolnick et al., 2012).

2.2. Components of the scoring function

The scoring function consists of a linear combination of the following pseudo-energy terms:

2.2.1. Burial potential

The burial potential considers a 7-state alphabet, BURIAL-Cβ-14-7, based on the count of Cβ (Cα for glycine) atoms within a 14 Å radius sphere centered on the Cβ of a residue of interest. The count ranges for states A–G are <27, 27–33, 34–39, 40–46, 47–54, 55–65 and >66, correspondingly. This classification arranges protein residues according to their exposure to solvent and neighboring atoms and was previously found to be highly effective in fold recognition (Karchin et al., 2004). First, we calculated the composition of the amino acid of type A within a state B in the

non-redundant CATH dataset as follows:

$$C_{A,B}^{\mathrm{bur}} = N_{A,B}^{\mathrm{bur}} / \sum_{i=1}^{20} N_{i,B}^{\mathrm{bur}} \quad (1)$$

where $N_{i,B}^{\mathrm{bur}}$ is the number of amino acids of type i within the state B.

The burial score for a given amino acid A is defined as its composition in the particular state B (calculated from the structure) normalized by its frequency of occurrence, f_A, in the dataset:

$$S_{A,B}^{\mathrm{bur}} = \frac{C_{A,B}^{\mathrm{bur}}}{f_A} \quad (2)$$

For a given amino acid sequence of length n mounted in a structure, the burial contribution to the pseudo-energy, E^{bur}, is calculated as the burial score averaged over all residues:

$$E^{\mathrm{bur}} = \frac{1}{n} \sum_{i=1}^{n} S_{i,B}^{\mathrm{bur}} \quad (3)$$

2.2.2. Secondary structure preferences

These preferences are calculated according to a 7-state classification by STRIDE (Frishman and Argos, 1995): H—α-helix, G—3–10 helix, I—π-helix, E—extended conformation, B—isolated bridge, T—turn, and C—coil. Similarly to the burial potential, we calculated the composition of the amino acid of type A within a secondary structure D in the non-redundant CATH dataset, $C_{A,D}^{\mathrm{sec}}$. Secondary structure pseudo-energy, E^{sec}, is calculated as the normalized secondary structure composition, $S_{A,D}^{\mathrm{sec}}$, averaged over all residues:

$$E^{\mathrm{sec}} = \frac{1}{n} \sum_{i=1}^{n} S_{i,D}^{\mathrm{sec}} \quad (4)$$

Note that despite the similar notation (each state is assigned a capital letter), the 7-state secondary structure classification is neither equivalent nor compatible with the 7-state burial alphabet described in the previous section.

2.2.3. Sequence profile score

For each target structure, we derive sequence profiles from structure alignments constructed by fr-TM-align (Pandit and Skolnick, 2008) against the CATH database. Only these structures that have a TM-score to the target of ≥0.4 are used. Moreover, since we operate in a low ($<35\%$) sequence identity regime, we use the following 7-state amino acid classification (Guharoy and Chakrabarti, 2005): (1)—A, V, L, I, M, C; (2)—G, S, T; (3)—D, E; (4)—N, Q; (5)—R, K; (6)—P, F, Y, W; and (7)—H. For each residue position in a target structure, a 7-class profile is derived including pseudo-counts:

$$P_j^{\mathrm{seq}} = \frac{c_j + F_j\sqrt{m}}{m + \sqrt{m}} \quad (5)$$

where P_j^{seq} is the probability of a residue class j to be found in this position, c_j is the number of templates that have a class j residue in the equivalent position, m is the total number of similar structures, and F_j is the frequency of occurrence of residue class j in the CATH database. Residue equivalences are calculated from structure alignments generated by fr-TM-align (Pandit and Skolnick, 2008).

For a given sequence, the sequence profile score, E^{seq}, is calculated as the P^{seq} probabilities averaged over all residues:

$$E^{\mathrm{seq}} = \frac{1}{n} \sum_{i=1}^{n} P_i^{\mathrm{seq}} \quad (6)$$

2.2.4. Statistical potential

As a distant-dependent statistical potential, we use the protein conformation free energy score by dFire (Zhang et al., 2004), separately for Cα atoms (*dFire-Cα*) and the side chain centers of mass (*dFire-SC*). The scores *dFire-Cα* and *dFire-SC* are subject to the following transformation to calculate the final pseudo-energies, $E^{\mathrm{dF-C\alpha}}$ and $E^{\mathrm{dF-SC}}$, respectively, which are independent of protein length:

$$E^{\mathrm{dF-C\alpha}} = \frac{dFire - C\alpha}{1.8031 \times n - 18.669} \quad (7)$$

$$E^{\mathrm{dF-SC}} = \frac{dFire - SC}{1.5766 \times n - 44.863} \quad (8)$$

where n is the protein length. The regression parameters were derived from the CATH database; Fig. 1 shows the $E^{\mathrm{dF-C\alpha}}$ and $E^{\mathrm{dF-SC}}$ pseudo-energies plotted as a function of protein length for wild-type proteins from CATH.

2.2.5. Anti-bunching restraints

Grouping artifacts are suppressed by Helmut Schmidt's test of force-like runs, also known as the Pot statistics, which measures the bunching relative to the spacing of a single state, e.g. a particular residue type, within a series of other states (Schmidt, 2000). First, we calculated a standard score of the Pot statistics for each amino acid types using wild-type sequences from the CATH database. For an arbitrary sequence, the mean value, $\overline{\mathrm{Pot}_j}$, and the corresponding standard deviation, σ_j, is then used to calculate C_j^{pot},

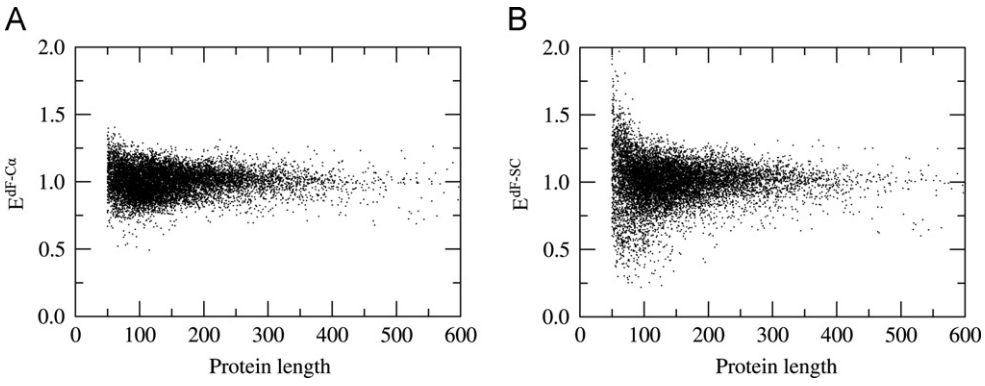

Fig. 1. *Normalized dFire pseudo-energy scores*. Scores are calculated for CATH proteins using (A) Cα atoms and (B) side chain centers of mass and plotted as a function of protein length.

a single Gaussian restraint for an amino acid type j:

$$C_j^{\mathrm{pot}} = 0.5\left(\frac{\mathrm{Pot}_j - \overline{\mathrm{Pot}_j}}{\sigma_j}\right)^2 - ln\ \frac{1}{\sigma_j\sqrt{2\pi}} \quad (9)$$

where Pot_j is a standard score of the Pot statistics for amino acid j.

These restraints are finally averaged over 20 residue types to give anti-bunching pseudo-energy score, E^{pot}:

$$E^{\mathrm{pot}} = \frac{1}{n}\sum_{i=1}^{20} C_i^{\mathrm{pot}} \quad (10)$$

E^{pot} penalizes artificial short-range bunching of particular amino acid types. Without this term, α-helices would likely be populated by e.g. bunched alanine residues and β-structures by isoleucine and valine residues.

2.3. Scoring function optimization

The total pseudo-energy score E for an arbitrary sequence mounted in a given structure is calculated using a linear combination of individual terms

$$E = w_1E^{\mathrm{bur}} + w_2E^{\mathrm{sec}} + w_3E^{\mathrm{seq}} + w_4E^{\mathrm{dF-C\alpha}} + w_5E^{\mathrm{dF-SC}} + w_6E^{\mathrm{pot}} \quad (11)$$

The weight factors were optimized on the CATH dataset. First, for each structure, we generated a large number of sequences, by iteratively shuffling the native one. Then, we selected 20 native-like sequences, whose identity to the wild type sequence was $>50\%$ (on average every second residue is identical) as well as 30 decoy sequences with the identity of $<10\%$ (evidently dissimilar sequences). Native-like sequences should exhibit similar behavior in profile-based modeling while decoy sequences will have very different properties. This procedure resulted in the total number of 527,900 sequences. We employed an algorithm that belongs to Evolution Strategies, which imitate the principles of natural evolution as a method to solve parameter optimization problems (Back and Schwefel, 1993; Back et al., 1992), to select a set weight factors that maximize the *Z-score* (the dimensionless ratio of the first and second moments of the pseudo-energy distribution within the native-like pool and the decoy pool)

$$Z\text{–}score = \frac{\overline{E}_{nat} - \overline{E}_{dec}}{\sigma} \quad (12)$$

Furthermore, to ensure that the optimal set of weight factors does not depend on the selection of training proteins, the CATH dataset was randomly divided into two equal subsets, for which weight factors were optimized independently. In both cases, the optimal values for weights w_1–w_6 were consistently estimated as 0.10, 0.49, 1.00, 0.10, 0.16 and 0.66, correspondingly.

2.4. Sequence evolution engine

The optimization of an amino acid sequence to stabilize a given structure is carried out by Simulated Annealing (SA). Here we use the C++ implementation from GNU Scientific Library (http://www.gnu.org/software/gsl), with the following SA parameters: N_TRIES=200, ITERS_FIXED_T=2000, K=1.0, T_INITIAL=5000, MU_T=1.002 and T_MIN=0.005. A single Monte Carlo step swaps a pair of randomly selected residues in the evolving sequence. The starting sequences are always random with a generic protein-like composition according to amino acid frequencies provided by UniProtKB/Swiss-Prot (Boutet et al., 2007).

2.5. Threading and fold recognition

For template selection and the construction of target-to-template alignments, we use two popular algorithms: HHpred (Soding, 2005) and CSI-BLAST (Biegert and Soding, 2009). HHpred detects distant homologous relationships between proteins based on the pairwise alignments of profile hidden Markov models and was shown to outperform other profile-profile comparison methods. CSI-BLAST is a modified version of PSI-BLAST (Altschul et al., 1997) that derives context-specific amino acid similarities from short windows centered on each query sequence residue, which results in a significant increase of sensitivity as well as alignment quality, particularly for difficult cases. For each program, we constructed three template libraries using wild-type sequences from CATH, these artificially evolved from random to stabilize CATH structures as well as random protein-like sequences.

2.6. Data fusion

Template rankings obtained by threading a wild-type target sequence against wild-type and evolved template libraries were merged using data fusion and a SUM rule. For a given template t, a combined score CS is defined as

$$CS^t = r_t^{\mathrm{WT}} + r_t^{\mathrm{EV}} \quad (13)$$

where r_t^{WT} and r_t^{EV} are the template rank in the wild-type and evolved library, respectively.

2.7. Assessment measures

The ability to select structurally similar templates from the library is assessed by the area under the accumulation curve (*AUAC*), where positives are defined as these structures that have a TM-score to the target of ≥0.4. The remaining templates are considered negatives. Furthermore, we calculate the number of "good" templates (TM-score ≥0.4) found among the top 10 identified templates. Alignment accuracy is evaluated by Matthew's correlation coefficient (*MCC*) against reference structure alignments constructed by fr-TM-align (Pandit and Skolnick, 2008):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (14)$$

where *TP*, *FP* and *FN* are the number of aligned residue positions correctly predicted, overpredicted and missed, respectively. *TN* is the number of residue pairs correctly predicted not to align to each other.

In addition to *MCC*, we also assess the quality of the target-to-template alignments by a TM-score (Zhang and Skolnick, 2004) calculated over the aligned residue positions reported by threading.

3. Results

All benchmarking calculations reported here are carried out using a non-redundant and representative CATH library (Orengo et al., 1997). First, we developed a combined scoring function that artificially evolves a protein-like amino acid sequence to stabilize a given structure. Next, such evolved synthetic sequences were generated for the entire CATH library. Using two popular threading/fold recognition algorithms, HHpred (Soding, 2005) and CSI-BLAST (Biegert and Soding, 2009), we demonstrate that the artificial sequences have similar capabilities to recognize correct structures as the wild-type sequences. Finally, we explore the possibility of using the evolved sequences in addition to the standard template library to improve the performance of

threading in template selection, maintaining the high quality of constructed target-to-template alignments.

3.1. Scoring function discriminates between wild-type and random sequences

A critical element is the scoring function used to optimize amino acid sequences mounted in the respective structures. It needs to be effective, consistent with these used in threading and fold recognition and devoid of potential modeling artifacts, such as the bunching of a particular type of residues. The scoring function developed in this study includes several energy terms: a burial potential, secondary structure preferences, a distant-dependent contact potential, sequence profiles and anti-bunching restraints. First, for each structure present in the CATH dataset, we generated a random sequence with the protein-like amino acid composition and assessed the values returned by individual scoring terms. Fig. 2 shows the distribution of scores for four pseudo-energy terms that are the most effective in discriminating between wild-type and random sequences. For E^{sec} (secondary structure preferences), E^{bur} (burial potential), $E^{dF\text{-}SC}$ (dFire for side chain centers of mass) and E^{seq} (sequence profiles), the wild type sequence was scored higher than the random one in 98.7%, 98.0%, 99.9% and 99.2% of the cases, respectively. Next, using a large dataset of 527,900 generated sequences, we derived the optimal set of weights that maximize the gap between native-like and random decoy sequences; see Section 2.3 for details.

3.2. Evolved sequences share low sequence identity with the wild-type ones

For each structure in the CATH dataset, we evolved a synthetic sequence using the optimized scoring function and a Simulated Annealing protocol. We note that these artificial sequences were evolved from entirely random sequences with a protein-like composition and are designed only to stabilize the respective structures in our force field. No information on the wild-type sequence or native amino acid composition is used in these simulations. Interestingly, as shown in Fig. 3A, the sequence identity to wild-type for the evolved sequences is significantly higher than for the random sequences; the median value is 13.8% and 5.8%, respectively. However, this similarity is still way below commonly accepted thresholds for a safe zone of sequence alignments and the evolved sequences clearly fall into the midnight zone (Rost, 1999).

Despite the low sequence identity between wild-type and evolved sequences, the latter carry sufficient amount of information to build non-degenerate sequence profiles using PSI-BLAST (Altschul et al., 1997). Here, sequence profile degeneracy is defined as the distribution of probabilities of amino acids at numerous positions close to the background frequencies of occurrence of amino acids in proteins. It is measured by Pearson's chi-squared test (χ^2) using amino acid frequencies provided by UniProtKB/Swiss-Prot (Boutet et al., 2007). Fig. 3B shows the distribution of χ^2-statistic values calculated for wild-type, evolved and random sequences; note that higher χ^2-statistic values indicate lower levels of degeneracy. Sequence profiles constructed for artificially evolved sequences are more degenerate than these obtained from wild-type sequences; the median χ^2-statistic (mean $\pm$ standard deviation) is 4.1 (6.7 ± 8.5) and 4.8 (7.7 ± 9.8), respectively. Nevertheless, this level of profile degeneracy is lower than that calculated for random protein sequences, which give the median χ^2-statistic (mean $\pm$ standard deviation) of 3.4 (4.9 ± 5.6). These results suggest that information carried by synthetic sequences can be utilized by sequence profile-based threading and fold recognition methods; this is explored further in the following sections.

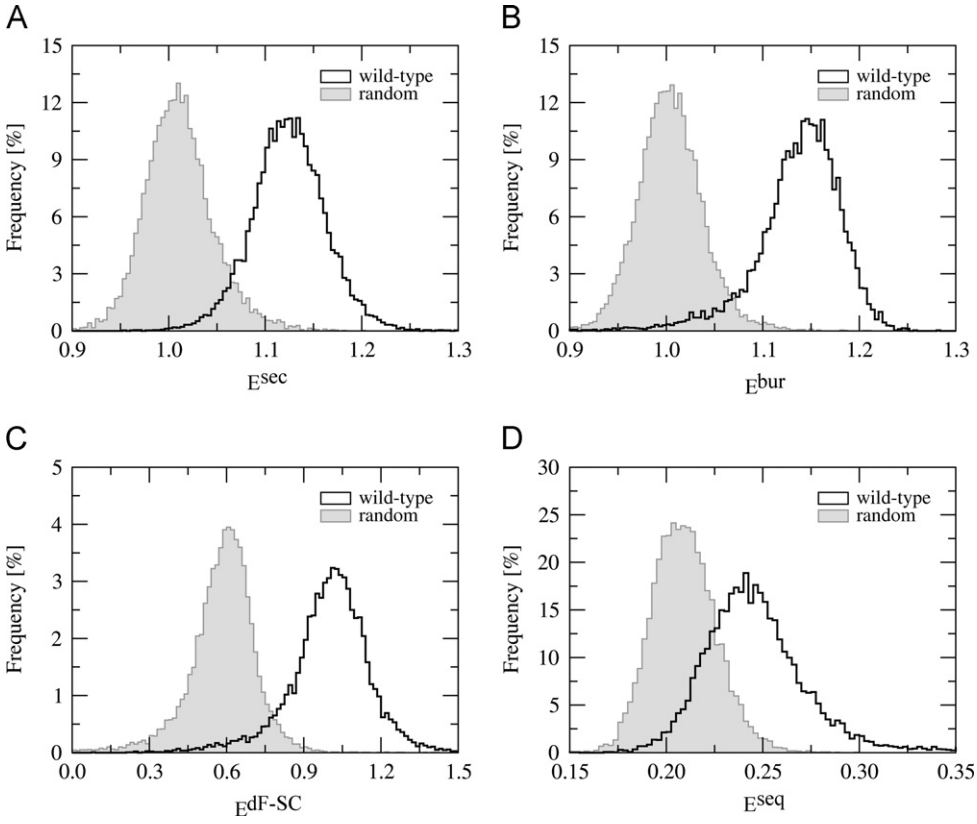

Fig. 2. *Most effective scoring terms in discriminating between wild-type and random sequences*. Distribution of selected pseudo-energy scores assigned to wild-type as well as random protein-like sequences across the CATH dataset: (A) secondary structure score, (B) burial score, (C) dFire-SC score and (D) sequence profile score.

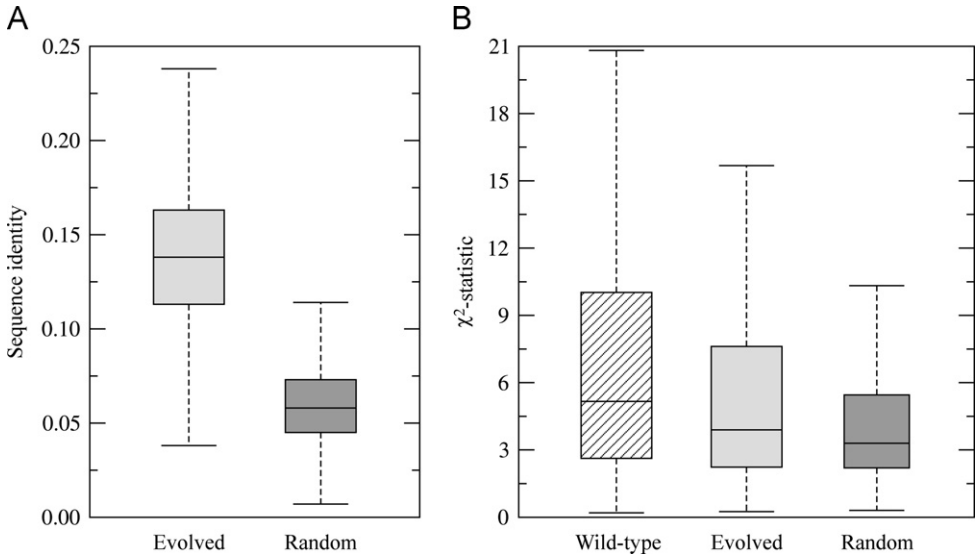

Fig. 3. *Characteristics* of *evolved sequences*. (A) Sequence identity to wild-type calculated for evolved and random sequences. (B) Degeneracy of sequence profiles constructed for wild-type, evolved and random sequences measured by Pearson's chi-squared test (χ^2-statistic). Boxes end at the quartiles Q_1 and Q_3; a horizontal line in a box is the median. Whiskers point at the farthest points that are within 3/2 times the interquartile range.

3.3. Evolved sequences recognize the native-like fold

Initial benchmarks consider threading of the artificially evolved sequences against a template library to ascertain whether they carry a sufficient amount of information to recover the native-like fold. Here, we use two fold recognition algorithms: HHpred (Soding, 2005) and CSI-BLAST (Biegert and Soding, 2009). For both programs, we constructed three different template libraries using protein structures from the CATH database. First library contains wild-type sequences, second was built from the artificially evolved sequences and the third one contains protein structures with mounted random protein-like sequences.

Fig. 4A shows that when the evolved target sequence is threaded against the wild-type library using HHpred and CSI-BLAST, for 85% (18%) and 79% (11%) of proteins the area under the accumulation curve (*AUAC*) is >0.5 (>0.6), respectively. This performance is significantly better than when a random library is used; here only 50% (0.8%) proteins have the *AUAC* of >0.5 (>0.6). More interestingly, the artificially evolved target sequences give the accuracy not only better than random, but also fairly close to the performance of wild-type target sequences. Standard benchmarks of HHpred and CSI-BLAST on the CATH dataset considering threading wild-type sequences against the wild-type library result in 86% (22%) and 82% (11%) of target proteins having the *AUAC* of >0.5 (>0.6), respectively (see Fig. 5A). This is a surprising result since the evolved sequences, which share on average only 13.8% sequence identity with their wild-type counterparts, reside in the midnight zone of sequence similarity (Rost, 1999). We note that all benchmarks reported here are carried out below a 35% pairwise sequence identity level. The accuracy further increases when the synthetic sequences are threaded against the synthetic library; here the *AUAC* of >0.5 (>0.6) for HHpred and CSI-BLAST is found for 94% (26%) and 86% (30%) of the targets, respectively. This can be easily explained. In a traditional scenario, threading frequently fails detecting templates in the midnight zone, which is populated by protein structure pairs that may have become similar by convergent or divergent evolution (Doolittle, 1994; Rost, 1997). In our computer experiment, there is no convergent evolution, thus protein sequences artificially evolved to stabilize a pair of similar structures will have on average high capabilities to recognize each other.

In addition to the *AUAC* analysis, we also assess the results in terms of the number of "good" (structurally similar) templates detected within the top 10 ranked templates. This seems more practical from a point of view of structure modeling, where typically only a few top ranked templates are used to build a model for the target sequence (Ginalski, 2006; Marti-Renom et al., 2000; Zhang, 2009). Fig. 4B demonstrates that for evolved sequences threaded against the wild-type and artificial library, HHpred (CSI-BLAST) recovers at least 5 good templates within the top 10 ranks in 52% (39%) and 71% (52%) of the cases, respectively. Again, it confirms that the ability of artificially evolved sequences to recognize each other is comparable to that of wild-type sequences (Fig. 5B); here the corresponding fraction of targets with at least 5 "good" templates for HHpred (CSI-BLAST) is 72% (53%).

3.4. Artificially evolved template sequences can be used in threading

In the previous section, we demonstrated that the evolved sequences have fairly high capabilities to recognize native-like folds. However, an important question is whether a wild-type target sequence can recognize these templates, whose amino acid sequences were artificially optimized to stabilize a fold, which is similar to that adopted by the target. That could have an immediate practical value in selecting templates from the midnight zone of sequence similarity, where traditional threading often fail due to the absence of detectable signal (Doolittle, 1994; Rost, 1997). By replacing wild-type template sequences, many of which are unrelated or related only remotely to the target sequence, by these artificially evolved, we provide an orthogonal source of signal that could be exploited in threading and fold recognition.

Fig. 5 shows the *AUAC* as well as the number of good templates identified by HHpred and CSI-BLAST using wild-type target sequences and wild-type, evolved and random CATH libraries. Here, the performance of HHpred using the evolved template library is only slightly worse than for the wild-type library (Fig. 5A); 82% (16%) proteins have the *AUAC* of >0.5 (>0.6). Interestingly, for CSI-BLAST, which is less sensitive than HHpred, the fraction of targets that have the *AUAC* of >0.5 (>0.6) significantly increased to 81% (21%) when the evolved library was used instead of the wild-type one. However, for the

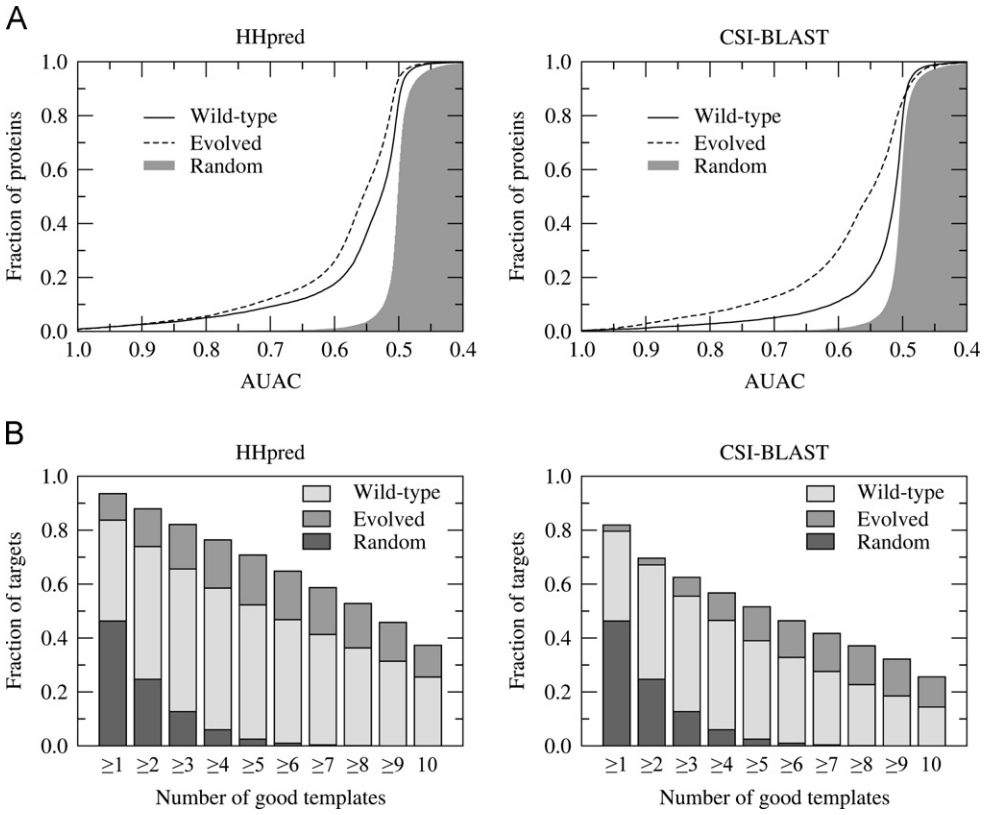

Fig. 4. *Ability of evolved sequences to recognize the native-like fold.* Accuracy of template selection by HHpred and CSI-BLAST from the wild-type, evolved as well as random CATH libraries using evolved target sequences: (A) area under the accumulation curve (AUAC), (B) number of structurally similar proteins within the top 10 identified templates.

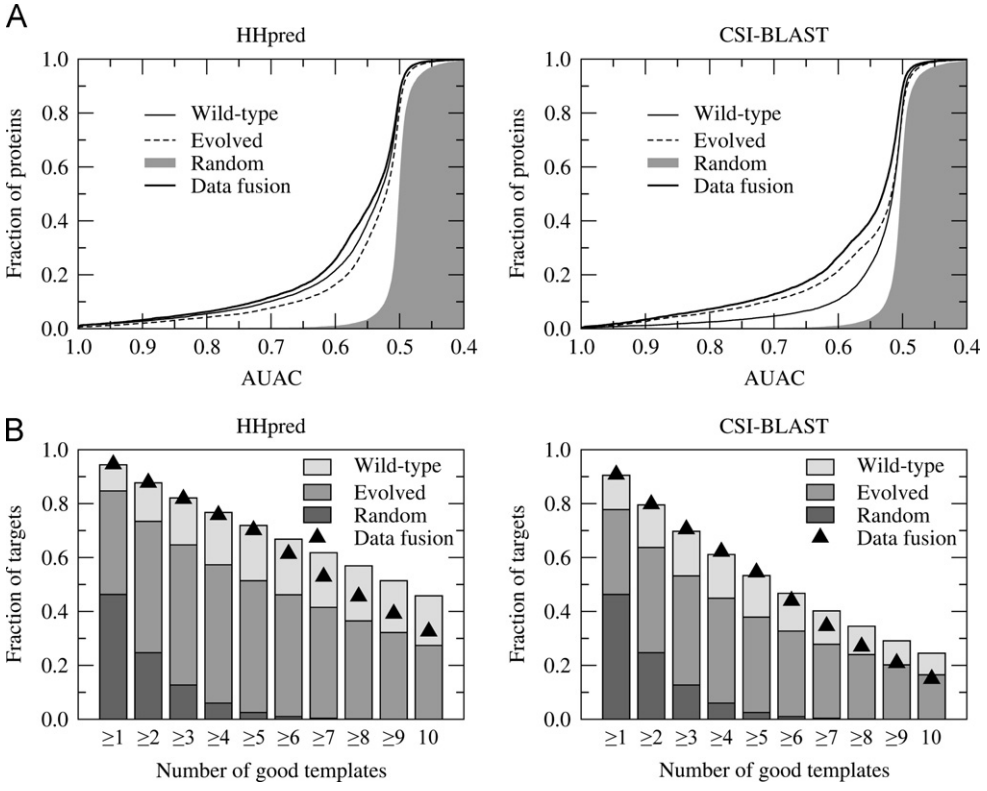

Fig. 5. *Performance of artificially evolved templates in fold recognition.* Accuracy of template selection by HHpred and CSI-BLAST from the wild-type, evolved as well as random CATH libraries using wild-type target sequences: (A) area under the accumulation curve (AUAC), (B) number of structurally similar proteins within the top 10 identified templates. Data fusion ranking is obtained by merging ranks assigned using wild-type and evolved libraries.

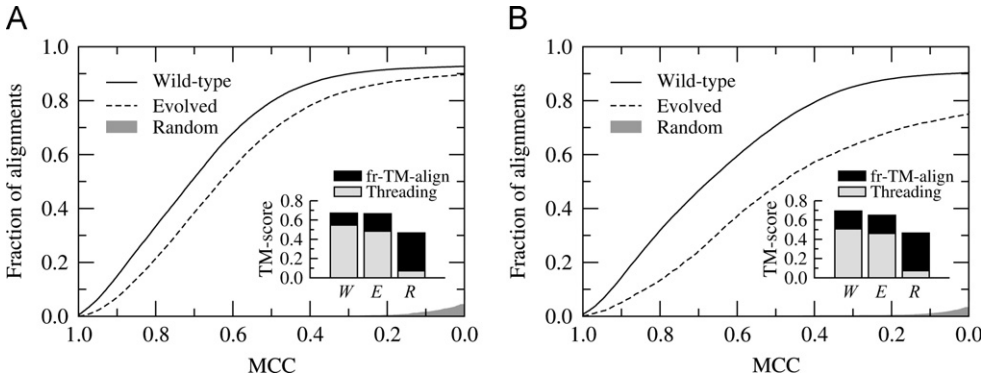

Fig. 6. *Accuracy of target-to-template threading alignments*. Alignments are constructed for wild-type target sequences by (A) HHpred and (B) CSI-BLAST using wild-type, evolved as well as random CATH libraries and assessed by Matthew's correlation coefficient (MCC) against structure alignments. Inset plots show the average TM-score calculated for threading alignments compared to the corresponding structure alignments for (W) wild-type, (E) evolved and (R) random CATH libraries.

wild-type library, both threading algorithms still recover ~15% more good templates on average at the top 10 ranks (Fig. 5B).

3.5. Target-to-template threading alignments are accurate

Effective template selection is important, but not sufficient for practical applications, such as protein structure modeling. In addition, target-to-template alignments should also be accurate to build a correct model. In Fig. 6, using reference structure alignments constructed by fr-TM-align, we evaluate the quality of threading alignments generated by HHpred and CSI-BLAST for wild-type target sequences. The accuracy is assessed by Matthew's correlation coefficient (*MCC*). For HHpred (Fig. 6A), the overall quality of alignments constructed using evolved template library is fairly high and almost comparable to the standard, wild-type library; the fraction of alignments with an *MCC* of >0.5 is 68.7% and 79.6%, respectively. Alignments by CSI-BLAST contain more errors; here the fraction is 48.1% and 70.7% for the evolved and wild-type library, respectively. However, in both cases the errors are often caused by only small shifts by 1–2 residues. This is shown as the inset plots in Fig. 6, which report the average TM-score calculated directly from the constructed alignments; TM-score is a geometrical measure (Zhang and Skolnick, 2004), less sensitive to small local errors in the aligned positions. For HHpred, the average TM-score for structure/threading alignments generated using wild-type and evolved template library is 0.67/0.54 and 0.66/0.48, respectively. For CSI-BLAST, the average TM-scores are comparable: 0.69/0.51 and 0.65/0.46, respectively. These results indicate that target-to-template alignments constructed using the artificially evolved library would be as useful in protein structure modeling as these constructed using the wild-type template sequences. This is again quite a surprising result, since both libraries share on average only 13.8% sequence identity. The high quality of threading alignments arises from the optimized scoring function used to design synthetic sequences.

3.6. Template ranking by data fusion improves recognition rates

Next, we explore the possibility of combining threading results obtained for wild-type and artificially evolved libraries to improve the overall performance in template recognition. For this purpose, we apply data fusion (Hall and Llinas, 1997) to merge template ranks calculated using the wild-type and evolved libraries. Here, we use the SUM rule that is expected to be less sensitive to a rugged input than MAX and MIN rules (Ginn et al., 2000) and is generally preferred when fusion is by rank (Hert et al., 2004). As shown in Fig. 5A, data fusion improves the overall performance in template ranking; using HHpred and CSI-BLAST, for 87% (26%) and 89% (27%) of the target proteins the *AUAC* is >0.5 (>0.6), respectively. These results suggest that particularly these templates, for which the signal at the wild-type sequence level cannot be detected, are systematically assigned better ranks when artificially optimized sequences are used. Nevertheless, the signal is still not strong enough to enrich the very top fraction of the ranked library with good templates (Fig. 5B). It may suggest that more sophisticated protocols are needed to fully exploit the additional information provided by the artificial template sequences.

Data fusion results and the correlation plots shown in Fig. 7 indicate that, in principle, such an improvement should be possible. The Pearson correlation coefficient for the *AUAC* values calculated for template ranking using the wild-type and artificially evolved libraries is 0.89 and 0.59 for HHpred and CSI-BLAST, respectively. Particularly for CSI-BLAST, a significant improvement in template ranking could be achieved by taking advantage of the evolved sequences (dots above the diagonal in Fig. 7B).

3.7. Confidence estimates are accurate

Each individual threading/fold recognition algorithm assesses structures present in the template library using some scoring system, e.g. CSI-BLAST employs a scoring system based on analytically estimated E-values and HHpred uses calibrated probabilities for true relationships between proteins. Despite the fact that these confidence scores were optimized for wild-type sequence libraries, we find that they are applicable to artificially evolved template sequences as well. Fig. 8 shows the distribution of *AUAC* for template selection for targets assigned different confidence. Here, the confidence corresponds to the mean value of scores returned by each algorithm for the top 10 ranked templates. The confidence estimates are not only well correlated with the overall performance, but they also are independent on the library used (either wild-type or evolved).

3.8. Case studies

To illustrate the improved performance of threading and fold recognition by using evolved template sequences, we selected several representative examples. In the first case study, we focus on the improved fold recognition rate. Using a wild-type sequence of the calcium-binding C-terminal domain of BM-40 osteonectin (CATH domain code: 1sraA00) and the wild-type template library, the following templates are assigned ranks lower than 10 by

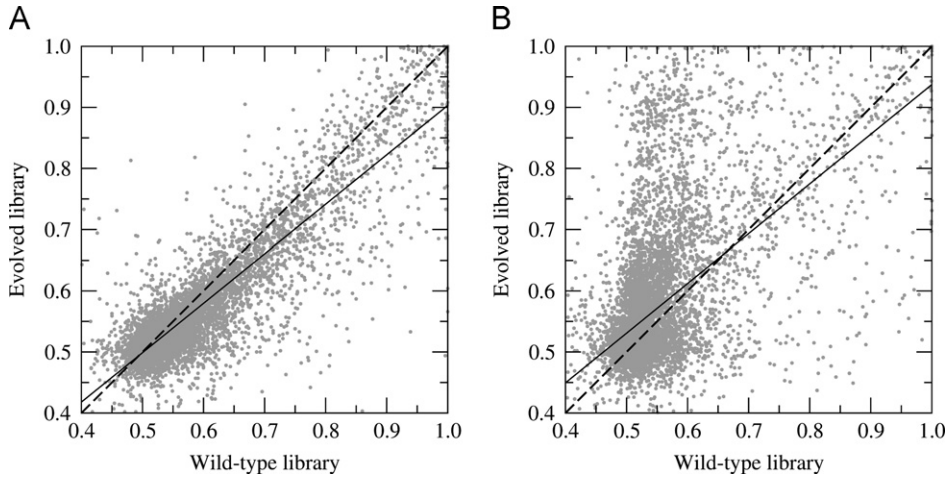

Fig. 7. *Template ranking using wild-type and artificially evolved libraries*. Area under the accumulation curve for template selection from wild-type and evolved libraries using wild-type target sequences. (A) HHpred and (B) CSI-BLAST. Each dot represents one CATH target, the regression line and the diagonal is solid and dashed, respectively.

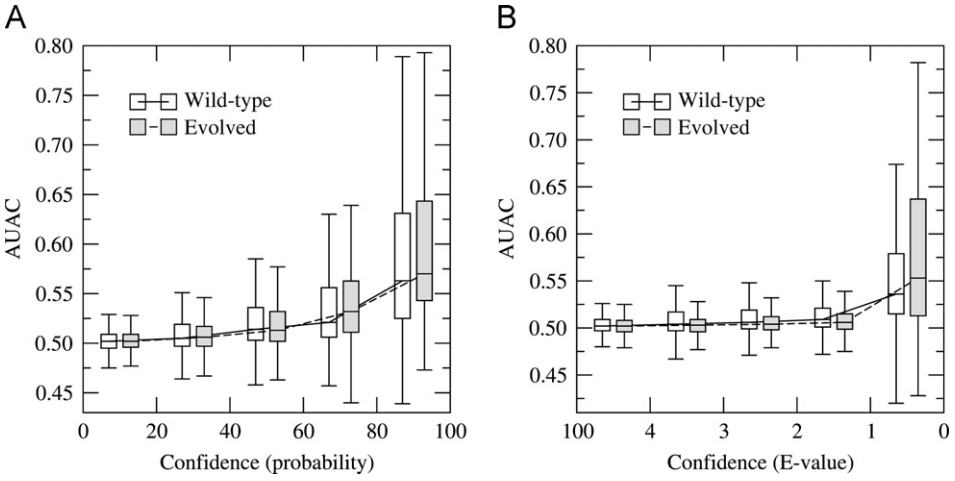

Fig. 8. *Confidence estimates for template selection*. Distribution of the area under the accumulation curve (AUAC) for template selection from wild-type and evolved libraries using wild-type target sequences. Targets are grouped into 5 confidence bins based on the (A) probability values estimated by HHpred and (B) E-values by CSI-BLAST.

HHpred: 2hpkA00 (rank 22), 1qv0A00 (rank 41), 2zfdA00 (rank 12) and 1yr5A00 (rank 24). Despite their low sequence identity to the target (25.2%, 17.6%, 20.8% and 22.3%, respectively), all these protein domains are structurally related with a TM-score of 0.49, 0.48, 0.43 and 0.43, respectively (see Fig. 9B and D). By using HHpred and the evolved template library, these templates are found within the top 10 ranks; 2hpkA00, 1qv0A00, 2zfdA00 and 1yr5A00 are now ranked 2, 3, 4 and 8, respectively. Note that the similarity of evolved sequences to the target sequence remains at a comparable level: 22.9%, 18.8%, 23.6% and 17.8%, respectively. Another interesting example is Gram-negative porin from *Rhodobacter capsulatus* (CATH domain code: 2porA00) and its four weakly homologous templates: 3dwnA00, 2gskA02, 2iahA03 and 1kmoA02. Their TM-score/wild-type/evolved sequence identity to the target is 0.50/23.8%/18.9%, 0.57/23.6%/22.1%, 0.58/23.2%/22.3% and 0.59/24.5%/19.8%, respectively. Structural alignments of these templates to the target are shown in Fig. 9E–H. By using evolved template sequences rather than the wild-type ones, the ranking of these templates systematically improves from 13 to 1, from 22 to 3, from >100 to 3, and from >100 to 6, respectively.

Our second case study shows that even when threading of wild-type target sequences using both wild-type and evolved template libraries correctly identifies structurally similar proteins, the latter can still provide more accurate target-to-template alignments. Here, our first example is a cupin domain from oxalate decarboxylase (CATH domain code: 1j58A02) and its weakly homologous (22.1% sequence identity) template protein 1gqgC02, correctly identified by CSI-BLAST. Fig. 10A shows the template-to-target structure alignment and the corresponding Cα–RMSD per residue (orange circles). The structure alignment covers 89% of the target sequence with most of the template residues well aligned to the target within a distance of 3 Å. CSI-BLAST using the wild type template sequence only partially recovers the structure alignment (residues 50–90, blue squares in Fig. 10A). However, when the evolved template sequence, whose similarity to the target sequence is 20.2%, is used, the threading alignment significantly improves, particularly over residues 1–50 and 90–120 (green triangles in Fig. 10A). Another example is the subunit C of urease domain of hydantoinase (CATH domain code: 1gkpA02) and its low sequence identity (29.9%) template protein 1rk6A01, see Fig. 10B. Here, the target-to-template alignment constructed by HHpred is significantly improved when the evolved template sequence (27.3% identity to the target) is used instead of the wild-type one and closely follows the optimal structure alignment (most green triangles in the right panel of Fig. 10B are only slightly higher than orange circles).

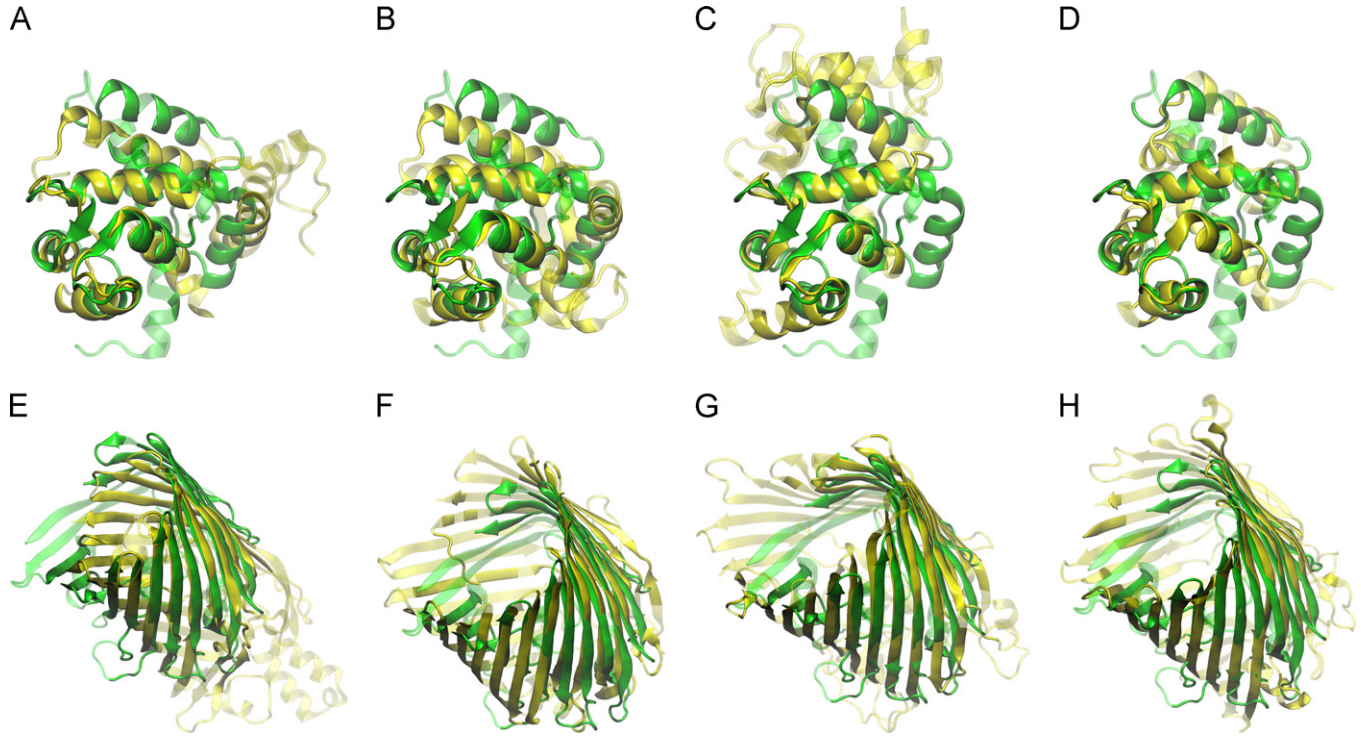

Fig. 9. *Examples of improved template recognition using evolved template library*. Top panel: template-to-target structural alignments of 1sraA00 (target) and (A) 2hpkA00, (B) 1qv0A00, (C) 2zfdA00, and (D) 1yr5A00 (templates). Bottom panel: template-to-target structural alignments of 2porA00 (target) and (E) 3dwnA00, (F) 2gskA02, (G) 2iahA03, and (H) 1kmoA02 (templates). Target and template structures are colored in green and yellow, respectively; the aligned region is solid. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

Many existing approaches to protein threading and fold recognition exhibit extremely high false negative rates due to the fact that the vast majority of pairs of proteins with similar structures populate the midnight zone of sequence identity (Rost, 1999). Most of the scoring schemes use as a basis a strong sequence profile component, which is designed to detect evolutionary relationships at the sequence level rather than structure similarities (Biegert and Soding, 2009; Eddy, 1998; Hughey and Krogh, 1996; Sadreyev and Grishin, 2003; Soding, 2005). Since most proteins with similar structures in the midnight zone are likely the products of convergent or divergent evolution (Doolittle, 1994; Rost, 1997), they remain undetectable even for very sensitive homology-based methods. It has been demonstrated that for an arbitrary protein, structural analogs are likely present in the PDB (Zhang and Skolnick, 2005; Zhang et al., 2006), yet a considerable fraction of protein sequences fall into the "hard target" category (Peng and Xu, 2010). One possible solution to this problem would be to develop a new class of purely structure-based algorithms; however, extensive efforts in this direction brought about rather limited success (Kryshtafovych et al., 2005; Moult et al., 2011, 2009). In this study we explore the possibility of using synthetic amino acid sequences instead of the wild-type ones to enhance the detection of structural analogs.

We developed a method for the optimization of generic protein-like amino acid sequences to stabilize the respective structures using several statistical potentials, which are compatible with these used in protein threading and fold recognition. In extensive benchmarks, we show that the artificially evolved sequences, despite their low sequence identity to the wild-type counterparts, have significant capabilities to recognize the correct structures. Furthermore, when state-of-the-art threading is applied to both wild-type and artificially evolved template libraries, even as simple technique as data fusion systematically improves template ranking. Also, we demonstrate that the quality of the corresponding alignments generated for synthetic sequences, which would have an impact on the accuracy of subsequent protein structure modeling, is fairly high and comparable to these constructed using a standard threading approach.

Notwithstanding these encouraging results, the proposed method still has important limitations and further developments are needed. Before it will have a practical value, more sophisticated algorithms are required to provide better enrichment of the very top fraction of the threading library with structurally similar templates. This could be achieved by e.g. developing a custom threading approach with the parameters specifically tailored to the synthetic library, designing a better and more sensitive scoring function for the sequence optimization or developing a meta-threading pipeline (Kurowski and Bujnicki, 2003; Lundstrom et al., 2001; Wu and Zhang, 2007) with an advanced machine learning-based system for template selection. Moreover, the subsequent structure modeling may require a template pre-clustering approach, which on average improves the accuracy of the final models constructed from multiple low-ranked templates (Pandit and Skolnick, 2010). One can also imagine enriching the library of experimental structures with these constructed *in silico* (Dai and Zhou, 2011; Skolnick et al., 2009; Taylor et al., 2009). These new directions will be investigated in future research.

Finally, the presented work opens up additional areas for further exploration, which mostly relate to protein evolution, engineering and design as well as to current studies on the completeness of protein structure space and the origin of folds and protein universe. The effective procedure for the design of a quasi-stable sequence for an arbitrary structure provides a desired linkage between protein structure and function in computer experiments, thus can facilitate *au courant* studies on the origin of biochemical function.

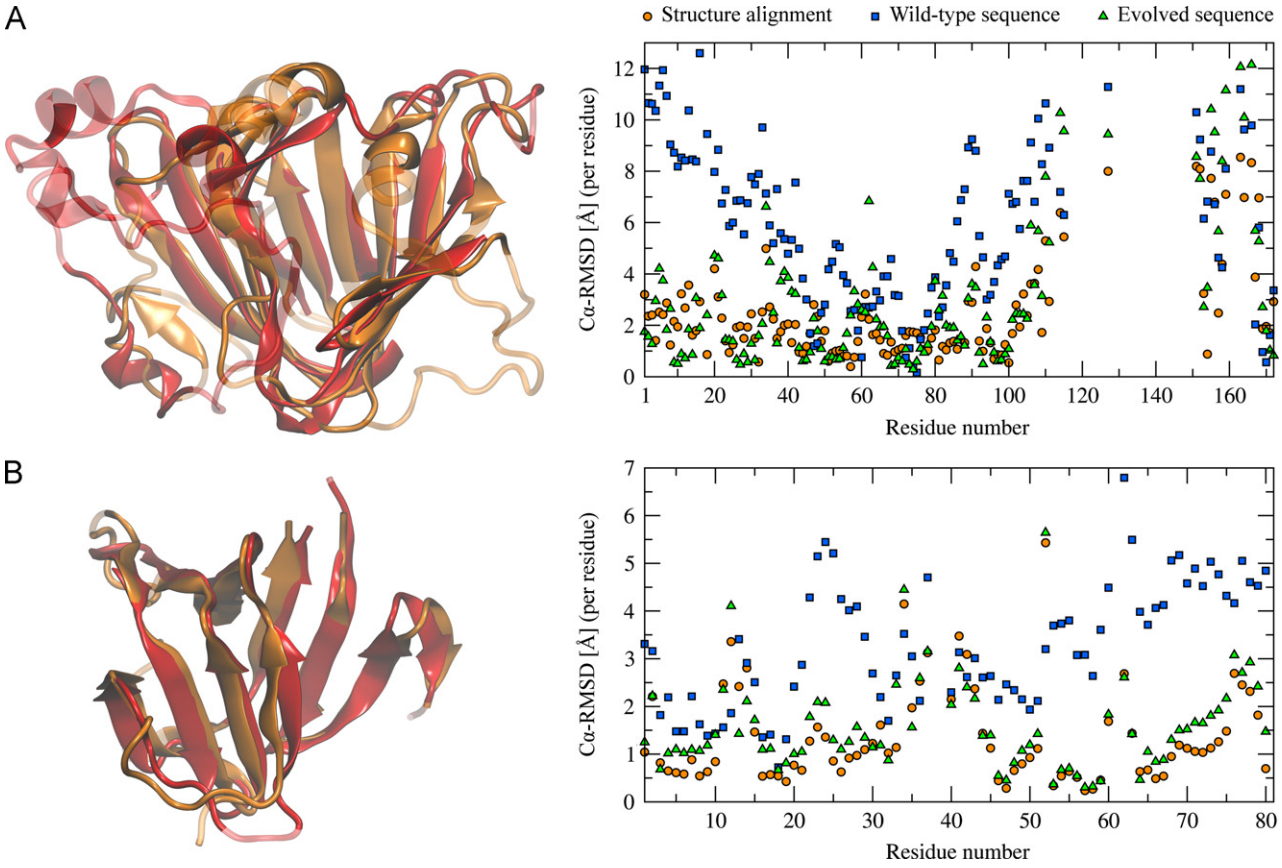

Fig. 10. *Examples of more accurate target-to-template alignments constructed using evolved template sequences*. (A) Alignments constructed by CSI-BLAST for 1j58A02 (target) and 1gqgC02 (template); (B) alignments constructed by HHpred for 1gkpA01 (target) and 1rk6A01 (template). Left panel shows template-to-target structural alignment by fr-TM-align. Target and template structures are colored in red and orange, respectively; the aligned region is solid. Plots on the right show Cα-RMSD per residue calculated over residue positions aligned by fr-TM-align; reference structural alignments are compared to threading alignments obtained using wild-type and evolved template sequences. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Conclusions

We introduce a novel modeling stratagem, which employs a library of synthetic sequences to improve template ranking in fold recognition by sequence profile-based methods. Regardless of its current limitations, it represents a new direction in the development of more sensitive threading approaches with the enhanced capabilities of targeting difficult, midnight zone templates.

Availability

Datasets and modeling results are available free of charge at ⟨http://www.brylinski.org/evolver⟩.

Acknowledgments

This study was supported by LSU Council on Research through the 2012 Summer Stipend Program. Portions of this research were conducted with high performance computational resources provided by Louisiana State University (http://www.hpc.lsu.edu).

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25, 3389–3402.

Back, T., Schwefel, H.P., 1993. An overview of evolutionary algorithms for parameter optimization. Evol. Comput. 1, 1–23.

Back, T., Hoffmeister, F., Schwefel, H.P., 1992. A Survey of Evolution Strategies. In: Proceedings of the Fourth International Conference on Genetic Algorithms, San Mateo, CA, pp. 2–9.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucl. Acids Res. 28, 235–242.

Biegert, A., Soding, J., 2009. Sequence context-specific profiles for homology searching. Proc. Natl. Acad. Sci. USA 106, 3770–3775, http://dx.doi.org/10.1073/pnas.0810767106.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bairoch, A., 2007. UniProtKB/Swiss-Prot. Methods Mol. Biol. 406, 89–112.

Brylinski, M., Skolnick, J., 2008. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc. Natl. Acad. Sci. USA 105, 129–134.

Brylinski, M., Skolnick, J., 2011. FINDSITE-metal: integrating evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level. Proteins 79, 735–751, http://dx.doi.org/10.1002/prot.22913. [doi].

Brylinski, M., Prymula, K., Jurkowski, W., Kochanczyk, M., Stawowczyk, E., Konieczny, L., Roterman, I., 2007. Prediction of functional sites based on the fuzzy oil drop model. PLoS Comput. Biol. 3, e94.

Chen, H., Kihara, D., 2011. Effect of using suboptimal alignments in template-based protein structure prediction. Proteins 79, 315–334, http://dx.doi.org/10.1002/prot.22885.

Dai, L., Zhou, Y., 2011. Characterizing the existing and potential structural space of proteins by large-scale multiple loop permutations. J. Mol. Biol. 408, 585–595, http://dx.doi.org/10.1016/j.jmb.2011.02.056.

Doolittle, R.F., 1994. Convergent evolution: the need to be explicit. Trends Biochem. Sci. 19, 15–18.

Drew, K., Winters, P., Butterfoss, G.L., Berstis, V., Uplinger, K., Armstrong, J., Riffle, M., Schweighofer, E., Bovermann, B., Goodlett, D.R., Davis, T.N., Shasha, D., Malmstrom, L., Bonneau, R., 2011. The Proteome Folding Project: proteome-scale prediction of structure and function. Genome Res. 21, 1981–1994 http://dx.doi.org/10.1101/gr.121475.111.

Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14, 755–763.

Elcock, A.H., 2001. Prediction of functionally important residues based solely on the computed energetics of protein structure. J. Mol. Biol. 312, 885–896, http://dx.doi.org/10.1006/jmbi.2001.5009.

Frishman, D., Argos, P., 1995. Knowledge-based protein secondary structure assignment. Proteins 23, 566–579, http://dx.doi.org/10.1002/prot.340230412.
Ginalski, K., 2006. Comparative modeling for protein structure prediction. Curr. Opin. Struct. Biol. 16, 172–177, http://dx.doi.org/10.1016/j.sbi.2006.02.003.
Ginn, C.M.R., Willett, P., Bradshaw, J., 2000. Combination of molecular similarity measures using data fusion. Perspect. Drug Discov. Des. 20, 1–16.
Guharoy, M., Chakrabarti, P., 2005. Conservation and relative importance of residues across protein-protein interfaces. Proc. Natl. Acad. Sci. USA 102, 15447–15452, http://dx.doi.org/10.1073/pnas.0505425102.
Hall, D.L., Llinas, J., 1997. An introduction to multisensor data fusion. Proc. IEEE 85, 6–23.
Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A., 2004. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. J. Chem. Inf. Comput. Sci. 44, 1177–1185, http://dx.doi.org/10.1021/ci034231b.
Hetenyi, C., van der Spoel, D., 2006. Blind docking of drug-sized compounds to proteins with up to a thousand residues. FEBS Lett. 580, 1447–1450, http://dx.doi.org/10.1016/j.febslet.2006.01.074.
Huang, B., Schroeder, M., 2006. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct. Biol. 6, 19, http://dx.doi.org/10.1186/1472-6807-6-19.
Hughey, R., Krogh, A., 1996. Hidden Markov models for sequence analysis: extension and analysis of the basic method. Comput. Appl. Biosci. 12, 95–107.
Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202, http://dx.doi.org/10.1006/jmbi.1999.3091.
Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. A new approach to protein fold recognition. Nature 358, 86–89, http://dx.doi.org/10.1038/358086a0.
Karchin, R., Cline, M., Karplus, K., 2004. Evaluation of local structure alphabets based on residue burial. Proteins 55, 508–518, http://dx.doi.org/10.1002/prot.20008.
Kitano, H., 2002. Systems biology: a brief overview. Science 295, 1662–1664, http://dx.doi.org/10.1126/science.1069492.
Kryshtafovych, A., Venclovas, C., Fidelis, K., Molt, J., 2005. Progress over the first decade of CASP experiments. Proteins 61 (7), 225–236, http://dx.doi.org/10.1002/prot.20740.
Kurowski, M.A., Bujnicki, J.M., 2003. GeneSilico protein structure prediction meta-server. Nucl. Acids Res. 31, 3305–3307.
Kuziemko, A., Honig, B., Petrey, D., 2011. Using structure to explore the sequence alignment space of remote homologs. PLoS Comput. Biol. 7, e1002175, http://dx.doi.org/10.1371/journal.pcbi.1002175.
Liu, S., Vakser, I.A., 2011. DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. BMC Bioinform. 12, 280, http://dx.doi.org/10.1186/1471-2105-12-280.
Lundstrom, J., Rychlewski, L., Bujnicki, J., Elofsson, A., 2001. Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci. 10, 2354–2362.
Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., Sali, A., 2000. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29, 291–325, http://dx.doi.org/10.1146/annurev.biophys.29.1.291.
McGuffin, L.J., Smith, R.T., Bryson, K., Sorensen, S.A., Jones, D.T., 2006. High throughput profile-profile based fold recognition for the entire human proteome. BMC Bioinform. 7, 288, http://dx.doi.org/10.1186/1471-2105-7-288.
Mi, H., Vandergriff, J., Campbell, M., Narechania, A., Majoros, W., Lewis, S., Thomas, P.D., Ashburner, M., 2003. Assessment of genome-wide protein function classification for *Drosophila melanogaster*. Genome Res. 13, 2118–2128, http://dx.doi.org/10.1101/gr.771603.
Moult, J., Fidelis, K., Kryshtafovych, A., Tramontano, A., 2011. Critical assessment of methods of protein structure prediction (CASP)—round IX. Proteins 79 (10), 1–5, http://dx.doi.org/10.1002/prot.23200.
Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Tramontano, A., 2009. Critical assessment of methods of protein structure prediction—Round VIII. Proteins 77 (9), 1–4, http://dx.doi.org/10.1002/prot.22589.
Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M., 1997. CATH—a hierarchic classification of protein domain structures. Structure 5, 1093–1108.
Pandit, S.B., Skolnick, J., 2008. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. BMC Bioinform. 9, 531, http://dx.doi.org/10.1186/1471-2105-9-531.
Pandit, S.B., Skolnick, J., 2010. TASSER_low-zsc: an approach to improve structure prediction using low z-score-ranked templates. Proteins 78, 2769–2780, http://dx.doi.org/10.1002/prot.22791.
Peng, J., Xu, J., 2010. Low-homology protein threading. Bioinformatics 26, i294––i300, http://dx.doi.org/10.1093/bioinformatics/btq192.
Peng, J., Xu, J., 2011. A multiple-template approach to protein threading. Proteins 79, 1930–1939, http://dx.doi.org/10.1002/prot.23016.
Rost, B., 1997. Protein structures sustain evolutionary drift. Fold Des. 2, S19–S24.
Rost, B., 1999. Twilight zone of protein sequence alignments. Protein Eng. 12, 85–94.
Sadreyev, R., Grishin, N., 2003. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J. Mol. Biol. 326, 317–336.
Schmidt, H., 2000. A proposed measure for psi-induced bunching of randomly spaced events. J. Parapsychol. 64, 301–316.
Shah, M., Passovets, S., Kim, D., Ellrott, K., Wang, L., Vokler, I., LoCascio, P., Xu, D., Xu, Y., 2003. A computational pipeline for protein structure prediction and analysis at genome scale. Bioinformatics 19, 1985–1996.
Skolnick, J., Brylinski, M., 2009. FINDSITE: a combined evolution/structure-based approach to protein function prediction. Brief Bioinform. 10, 378–391, doi: bbp017 [pii]10.1093/bib/bbp017 [doi].
Skolnick, J., Zhou, H., Brylinski, M., 2012. Further evidence for the likely completeness of the library of solved single domain protein structures. J. Phys. Chem. B 116, 6654–6664, http://dx.doi.org/10.1021/jp211052j.
Skolnick, J., Arakaki, A.K., Lee, S.Y., Brylinski, M., 2009. The continuity of protein structure space is an intrinsic property of proteins. Proc. Natl. Acad. Sci. USA 106, 15690–15695, http://dx.doi.org/10.1073/pnas.0907683106.
Soding, J., 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951–960, http://dx.doi.org/10.1093/bioinformatics/bti125.
Taylor, W.R., Chelliah, V., Hollup, S.M., MacDonald, J.T., Jonassen, I., 2009. Probing the "dark matter" of protein fold space. Structure 17, 1244–1252, http://dx.doi.org/10.1016/j.str.2009.07.012.
Wass, M.N., Kelley, L.A., Sternberg, M.J., 2010. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res 38, W469–W473, http://dx.doi.org/10.1093/nar/gkq406.
Wu, S., Zhang, Y., 2007. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res. 35, 3375–3382, http://dx.doi.org/10.1093/nar/gkm251.
Zhang, C., Liu, S., Zhou, H., Zhou, Y., 2004. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Sci. 13, 400–411, http://dx.doi.org/10.1110/ps.03348304.
Zhang, Y., 2009. I-TASSER: fully automated protein structure prediction in CASP8. Proteins 77 (9), 100–113, http://dx.doi.org/10.1002/prot.22588.
Zhang, Y., Skolnick, J., 2004. Scoring function for automated assessment of protein structure template quality. Proteins 57, 702–710, http://dx.doi.org/10.1002/prot.20264.
Zhang, Y., Skolnick, J., 2005. The protein structure prediction problem could be solved using the current PDB library. Proc. Natl. Acad. Sci. USA 102, 1029–1034, http://dx.doi.org/10.1073/pnas.0407152101.
Zhang, Y., Hubner, I.A., Arakaki, A.K., Shakhnovich, E., Skolnick, J., 2006. On the origin and highly likely completeness of single-domain protein structures. Proc. Natl. Acad. Sci. USA 103, 2605–2610, http://dx.doi.org/10.1073/pnas.0509379103.