Beyond the Twilight Zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information

Catherine Mooney, Gianluca Pollastri*

Complex and Adaptive Systems Laboratory and School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4

Email: Catherine Mooney - catherine.mooney@ucd.ie; Gianluca Pollastri*- gianluca.pollastri@ucd.ie;

*Corresponding author

Short Title: Protein Structure Prediction with Templates

Key Words: alignments, homology detection, secondary structure, solvent accessibility, machine learning

Corresponding author: Gianluca Pollastri

Address: Complex and Adaptive Systems Laboratory, University College Dublin, Belfield, Dublin 4

Phone: (353) 1 716 5382

E-mail: gianluca.pollastri@ucd.ie

Abstract

The prediction of 1D structural properties of proteins is an important step towards the prediction of protein structure and function, not only in the *ab initio* case but also when homology information to known structures is available. Despite this the vast majority of 1D predictors do not incorporate homology information into the prediction process. We develop a novel structural alignment method, SAMD, which we use to build alignments of putative remote homologues which we compress into templates of structural frequency profiles. We use these templates as additional input to ensembles of recursive neural networks, which we specialise for the prediction of query sequences which show only remote homology to any PDB structure. We predict four 1D structural properties – secondary structure, relative solvent accessibility, backbone structural motifs and contact density. Secondary structure prediction accuracy, tested by 5-fold cross validation on a large set of proteins allowing less that 25% sequence identity between training and test set and query sequences and templates, exceeds 82%, outperforming its *ab initio* counterpart, other state-of-the-art secondary structure predictors (Jpred 3 and PSIPRED) and two other systems based on PSI-BLAST and COMPASS templates. We show that structural information from homologues improves prediction accuracy well beyond the Twilight Zone of sequence similarity, even below 5% sequence identity, for all four structural properties. Significant improvement over the extraction of structural information directly from PDB templates suggests that the combination of sequence and template information is more informative than templates alone.

1 Introduction

The three-dimensional (3D) structure of a protein provides important information about its ligand binding, catalytic sites and protein-protein interactions. Compared to over 7 million known protein sequences, as of December 2008 there are only about 50,000 proteins of known structure deposited in the Protein Data Bank (PDB) [1]. As experimental determination of a protein's structure is difficult, expensive and time consuming the gap between sequence-known and structure-known proteins is growing.
Structural genomic (SG) projects aim to narrow this gap by obtaining useful 3D models of all protein families by a combination of experimental structure determination and comparative model building [2] in a high-throughput manner [3]. This makes prediction methods that incorporate structural information increasingly important. If the SG goal is achieved template structures will be available for most proteins.

Structure prediction methods can be divided into two groups: those that use similarity to proteins of known structure to model all or part of the query protein and *ab initio* prediction methods where no similarity to a protein of known structure can be found. If a close homologue can be found (e.g. a protein of known structure with a sequence identity greater than approximately 30% to the query) then a model can be produced with a high degree of confidence in its accuracy [4]. However many proteins share similar structures even though their sequences may share less than 15% sequence identity [5]. Finding these remote homologues is a much more difficult task with two major issues needing to be addressed: finding the best true homologue and building the correct alignment. Many methods have been developed to address both problems, most of which specialise on different degrees of sequence identity between query and template.

Sequence based methods compare a query sequence with a database of target sequences. If there is a close homologue to the query these methods can generally find it and align it to the query, fast. Two of the most popular of such methods are FASTA [6] and BLAST [7].

Profile-sequence methods, which use a position-specific scoring matrix (PSSM) or profile hidden Markov models (HMM) to search a sequence database are more sensitive at detecting remote homology if the pairwise sequence identity is in the Twilight Zone (between approximately 20% and 30%). Examples of the more popular profile-sequence methods are PSI-BLAST [7], HMMer [8] and SAM [9]. Profile-profile methods such as prof_sim [10], HHSearch [11] and COMPASS [12] use a profile to search a database of profiles, enabling them to identify even more remote evolutionary relationships.

The combination of sequence and structure has been shown to improve fold recognition and alignment quality especially in the case where pairwise sequence identity drops below the Twilight Zone but how best to include this information remains uncertain. Methods that include secondary structure [13,14], solvent accessibility [15] or both [16] have been developed. The addition of other information has also been shown to be of benefit, for example more detailed alphabets of local protein structure [17,18]. The most successful of these methods at present are FUGUE [19], GenThreader [20], ORFeus [13], SAM-T02 [9], SParks2 and SP3 [21]. For a comparison of some of these methods see Livebench [22], EVA [23] and CASP [24] as well as Cheng and Baldi [25] and Elofsson [26].

An alternative approach, based on similarity searching and clustering of homologous sequences, may provide insights into family relationships and has been studied and implemented in various systems such as SYSTERS [27], ProtoNet [28], ProtoMap [29].

Approach

The prediction of 1D structural properties of proteins (i.e. properties which may be represented as a string of the same length as the amino acid sequence) is an intermediate step towards the prediction of protein structure and function. The incorporation of such local structural properties can improve the performance of alignment algorithms, and can lead to the detection of more remote homologues, thus improving the accuracy of structure prediction methods. Until recently predictors of 1D structural properties have generally been *ab initio*. However it is clear that homology information can contribute to more accurate 1D predictions [30,31].

Previously [31] we have described a system (Distill_H) based on bidirectional recursive neural networks (BRNN) [32], which relies on homologues detected by PSI-BLAST to predict a protein's secondary structure (SS) and relative solvent accessibility (RSA). Here we develop a system to deal with situations where no clear homologue can be found i.e. there is no template exceeding 25% sequence identity to the query sequence. The system has two stages: one in which four 1D structural properties are predicted *ab initio* and the predictions are combined into a profile, which is used to search for remote homologues; the second stage, in which a specialised machine learning system harnesses information from remote homologues and the query's sequence to output a refined prediction for the same four 1D properties. The properties we predict are: SS; RSA; an alphabet of 14 backbone structural motifs (SM) [33]; and contact density (CD) [34].

We first compare three different methods for homology detection: PSI-BLAST, COMPASS (a popular profile-profile method) [12] and SAMD, the novel algorithm for remote homology detection we present here. We search for templates by these three methods and investigate the improvements they yield over *ab initio* predictions when used as input to Distill_H. Next we develop a novel system, Distill_RH, trained on SAMD templates, and compare it to its state-of-the-art *ab initio* counterpart, Distill [35], and to a baseline which copies the SS, RSA, SM and CD directly from the best template, for each of the four 1D predictors. Finally we compare Distill_RH SS prediction accuracy with that of PSIPRED [36], Jpred 3 [37] and PROTEUS [30], the only other publicly available SS predictor that we are aware of which exploits similarity to proteins of known structure during the prediction process.

We show that SAMD templates greatly improve prediction accuracy over *ab initio*, well beyond the case in which clear homology is available (essentially, down to any level of sequence identity), and substantially more than PSI-BLAST and COMPASS templates, whilst significantly improving over the best PDB template. We demonstrate that Distill_RH predictions, incorporating these SAMD templates, are significantly more accurate than Distill, Distill_H, PSIPRED, Jpred 3 and PROTEUS.

2 Methods

2.1 Datasets

The data set used to train and test our predictors is extracted from the January 2007 25% pdb_select list [38]. We assign each residue's SS, RSA and ϕ and ψ dihedral angles using DSSP [39]. We remove all sequences for which DSSP does not produce an output. The final set (S3129) contains 3129 proteins and 461,633 amino acids. SS is mapped from the eight DSSP classes into three classes as follows: H, G, I $\rightarrow$ Helix; E, B $\rightarrow$ Strand; S, T, . $\rightarrow$ Coil. RSA is mapped into four roughly equal classes: completely buried (0-4% exposed), partly buried (4-25% exposed), partly exposed (25-50%) and exposed (more than 50%). SM [33] are constructed by mapping tetra-peptides, represented as vectors of ϕ and ψ angles, into 14 conformational clusters determined by Sims *et al* [40]. CD [34] is defined as the principal eigenvector of a protein's residue contact map at $8\mathring{A}$, multiplied by the principal eigenvalue, and is assigned to one of 4 roughly equal classes, corresponding to very low, medium-low, medium-high and very high CD [34]. We use this set to train all systems, in 5-fold cross-validation. A second dataset (for unbiased testing) is extracted from the October 2007 25% pdb_select list. None of the 555 sequences in this dataset have more than 30% sequence identity to any sequence in S3129. S555 is then processed in the same way as S3129.
Multiple sequence alignments (MSA) for S3129, S555 and every sequence in the October 2007 25% pdb_select list are extracted from a redundancy reduced version of the NR database containing 1.05 million sequences. The alignments and PSSM are generated by three runs of PSI-BLAST with parameters $b = 3000$ (maximum number of hits), $e = 10^{-3}$ (expectation of a random hit) and $h = 10^{-10}$ (expectation of a random hit for sequences used to generate the PSSM). MSA for the October 2007 25% pdb_select list become the target database for COMPASS alignments after pre-processing by that method.

2.2 Structure Prediction With Templates

Predictive architectures

To learn the mapping between our input space $\mathcal{I}$ and output space $\mathcal{O}$ we use two-layered architectures composed of BRNN [32] of the same length N as the amino acid sequence. Similarly to Pollastri and McLysaght [41] and Vullo *et al* [34] we use BRNNs with shortcut connections.
These networks take the form:

$$\begin{aligned} o_j &= \mathcal{N}^{(O)}\left(i_j, h_j^{(F)}, h_j^{(B)}\right) \\ h_j^{(F)} &= \mathcal{N}^{(F)}\left(i_j, h_{j-1}^{(F)}, \ldots, h_{j-S}^{(F)}\right) \\ h_j^{(B)} &= \mathcal{N}^{(B)}\left(i_j, h_{j+1}^{(B)}, \ldots, h_{j+S}^{(B)}\right) \end{aligned}$$

$$j = 1, \ldots, N$$

where i_j (resp. o_j) is the input (resp. output) of the network in position j, and $h_j^{(F)}$ and $h_j^{(B)}$ are forward and backward chains of hidden vectors with $h_0^{(F)} = h_{N+1}^{(B)} = 0$. We parametrise the output update, forward update and backward update functions (respectively $\mathcal{N}^{(O)}$, $\mathcal{N}^{(F)}$ and $\mathcal{N}^{(B)}$) using three two-layered feed-forward neural networks.

Input i_j associated with the j-th residue contains primary sequence information and evolutionary information, and direct structural information derived from PDB templates:

$$i_j = (i_{j,1}^{(E)}, \ldots, i_{j,e}^{(E)}, i_{j,1}^{(T)}, \ldots, i_{j,t}^{(T)}) \tag{1}$$

where e units are devoted to sequence and evolutionary information, and t to structural information.

In the case of SS prediction we use $t = 10$ for representing structural information from the templates. The first eight structural input units contain the average 8-class (DSSP style) SS composition in the PDB templates, while the last two encode the average sequence identity of the template column to the query and average coverage of the query by the templates. The averages are weighed by the cubed sequence identity of the query and the PDB template, and by the inverse of the quality of the template, measured as X-ray resolution + R-factor/20, as in Hobohm *et al* [38].

Template information for RSA, SM and CD is encoded similarly to SS, except that four, fourteen and four units are adopted respectively to represent average RSA, SM and CD from PDB-derived templates. The two units encoding the profile quality are the same as in the SS case. When no template is available for a residue, the template section of the input is left blank.

Each sequence-to-structural feature BRNN is cascaded with a structure-to-structure BRNN [41]. Both BRNNs are trained at the same time, but supervised independently. For each prediction task five two-stage BRNN models are trained independently and ensemble averaged to build the final predictors. 1000 epochs of training are performed for each model; the learning rate is halved every time we do not observe a reduction of the error for more than 50 epochs. The number of free parameters per model ranges between 5,800 and 8,000. Template-based models are only on average 7% larger than the corresponding *ab initio* ones.

All 1D predictors are trained and tested using 5-fold cross-validation procedure on S3129. The five folds are of roughly equal sizes, composed of 625 or 626 proteins and ranging between 91,049 and 93,474 residues. We train in identical conditions (training/test set splits, training strategy, network architectures) three different systems: Distill, in which the template-based part of the input is left blank; Distill_H, in which

templates are obtained via PSI-BLAST; Distill_RH, in which templates are obtained by a novel remote homology detection algorithm, SAMD, which we describe below.

2.3 SAMD Template Generation

Building alignments based on structural information

We combine SS (3 classes), RSA (4 classes), SM (14) and CD (4) structural properties into a set of 672 features ($3 \times 4 \times 14 \times 4$), which we call SAMDs (Secondary **S**tructure, Solvent **A**ccessibility, Structural **M**otif and Contact **D**ensity). The SAMD are predicted *ab initio* for the query, and then aligned against the SAMDs in the target set (the PDB, or part thereof) to detect homologues. We use a basic local alignment dynamic programming [42] approach to align query and target sequences. Similarly to Gong and Rose [17] the alignment score between two residues i and j is composed of two parts:

$$P(i,j) = PSSM(i,j) + SAMD(i,j) \tag{2}$$

The sequence penalty for aligning amino acid i against amino acid j is obtained from a position-specific scoring matrix (PSSM) (see Datasets section). The penalty for aligning the structural properties, SAMD i against SAMD j, is obtained from a Structural (SAMD) substitution matrix. We add SAMD and PSSM scores without any relative weighing. The gap penalty for the SAMD matrix is fixed at -15 and the penalty for the PSSM is the minimum penalty value of the matrix. The total gap penalty is the sum of the two.

SAMD substitution matrix

The choice of substitution matrix greatly affects the quality of any pairwise alignment method [43]. The most popular matrices, Pam [44] and BLOSUM [45], have been used successfully to identify homologues when sequence identity exceeds approximately 30% [7]. More specialised matrices have been shown to improve the rate of detection of more remote homologues, for example the environment specific substitution tables of Fugue [19], matrices based on backbone dihedral angles [17], matrices for different sequence-structure contexts [46] or matrices created from structurally aligned protein pairs [47–49].

To create a SAMD substitution matrix we use the formalisation of Henikoff and Henikoff [45] and construct a BLOSUM-style matrix. We start with the structural annotation of the alignments in the BLOCKS database [50] (version 14.2, March 2006). We label each sequence in a block with its SAMDs, extracted from the corresponding PDB file. We count all possible pairs of SAMDs in each column of every block. The result is a frequency table listing the number of times each SAMD pair occurs. This table is used to

calculate a 672 × 672 matrix which represents the log odds ratio between these observed frequencies and those expected to occur randomly.

Estimating statistical significance

To determine if a score does in fact suggest that two sequences share a common ancestor it needs to be compared to a background of random scores and supported by some measure of statistical significance [51]. Much work has been done to develop and improve such methods [52–55]. These methods assume that gapped local alignments scores follow the same extreme value or Gumbel distribution [56] as gap-less local alignments:

$$P(S > x) \approx 1 - exp(-Kmne^{-\lambda x}) \tag{3}$$

In this case, where the substitution matrix is very large (672 × 672), an extreme value distribution is used to approximate the distribution of alignment scores. The e, or expectation, value is a measure of the reliability of the alignment score, or how likely it is for a score equal to or greater than the given score to occur by chance in the database being searched using a given scoring system:

$$E \approx Kmne^{-\lambda x}, \tag{4}$$

where x is the score and m and n are the query and target sequence lengths. K and λ are parameters that depend on the amino acid compositions of the sequences and on the scoring system, and therefore need to be re-estimated for every new scoring system used [53,54]. The key to using these two equations is the accurate calculation of K and λ. For this purpose we use the method of Bailey and Gribskov [54] based on maximum likelihood estimation and the values are recalculated for ever sequence based on the size of the target database.

3 Results

3.1 Template Comparison

We use S3129 to train Distill_H in 5-fold cross-validation. Templates are obtained by running a round of PSI-BLAST against the PDB (available on March 25th, 2008) using a PSSM generated against the NR database (see Datasets section) with an expectation cutoff of 10. To train template-based predictions in marginal similarity conditions we exclude all hits that have a PSI-BLAST hit exceeding 20% sequence

identity to the query sequence. We compare the performance of PSI-BLAST based templates, COMPASS based templates and SAMD templates on Distill_H for SS prediction.

To obtain unbiased results we use S555 as our test set. S555 is a subset of the October 2007 25% pdb_select list [38] and none of the sequences in S555 share more than 30% sequence identity to any sequence in S3129. We use this pdb_select list as the target template database. Since we exclude a protein's own PDB file from its list of templates we do not expect to find any templates showing more than 25% identity to the query, except for short sequences for which the 25% threshold is relaxed.

As the choice of the e value cutoff is critical to the performance of any alignment method we compare the three methods with seven different values of e and test the effect this has on SS prediction accuracy. We start with a very low e value cutoff of $e = 0.0005$. This gives each of the three methods an opportunity to find good templates first. If at least one template is found we stop searching, otherwise we gradually increase the e value six further times ($e = 0.001, 0.01, 0.1, 1, 10, 50$) until at least one template is found or the e value cutoff reaches 50. After these seven rounds COMPASS finds templates for 554 of the 555 sequences, PSI-BLAST finds templates for 545 sequences and SAMD finds templates for all of the sequences. This results in 73.28% of the 81,141 residues in S555 being covered by a COMPASS template, 51.37% are covered by a PSI-BLAST template and 53.15% by a SAMD template. Table 1 shows the accuracy of the best template for each template type compared to the accuracy of Distill_H when these templates are used during the prediction process. Although COMPASS covers more than 20% more residues than PSI-BLAST the template accuracy and Distill_H SS prediction accuracy is similar. In contrast, SAMD templates cover a similar number of residues to PSI-BLAST but the templates are over 10% more accurate and predictions using these templates are more than 4% more accurate than either COMPASS or PSI-BLAST. Figure 1 shows the distribution of SS prediction accuracy for S555 as a function of sequence identity to the best COMPASS, PSI-BLAST and SAMD template for each query sequence. For predictions using COMPASS templates with less than 15% sequence identity between the query sequence and the template sequence *ab initio* predictions are more accurate than template-based predictions. For PSI-BLAST templates *ab initio* predictions are more accurate at just under 10% sequence identity. However, for SAMD templates the Distill_H predictions are always more accurate demonstrating that the SAMD templates carry informative structural information which can be exploited by the Distill_H predictive system even with little or no sequence identity between query and template.

In the 0-10% range of sequence identity SAMD finds nearly five times as many templates as either PSI-BLAST or COMPASS, however these templates are on average less than half of the length of the

	Coverage	Baseline	Distill_H
COMPASS	73.28%	69.12%	81.79%
PSI-BLAST	51.37%	71.33%	82.13%
SAMD	53.15%	82.00%	86.53%

Table 1: Template coverage and accuracy of the best template (Baseline) for each template type compared to the performances of Distill_H SS prediction for residues covered by the best template.

	Baseline	Distill	Distill_RH
SS Q3	80.06%	83.73%	87.26%
RSA Q4	45.57%	56.07%	57.04%
SM Q14	63.19%	68.24%	71.63%
CD Q4	51.39%	51.66%	54.49%

Table 2: Performances of the four 1D predictors for Distill (*ab initio*), and Distill_RH (using SAMD templates) compared with a baseline predictor. Accurcy measured for residues covered by the best template.

COMPASS ones. This suggests that, at least in some cases, we are finding local fragments of templates that have similar 1D features to the query rather than finding a template that is overall correct. As the sequence identity of the templates increases so does template length for both SAMD and PSI-BLAST, however COMPASS remains constant at an average template length of 87.46 residues.

3.2 Distill_RH vs. Distill

We have shown that low sequence identity SAMD templates can be used to improve on *ab initio* SS prediction accuracy and that these gains in accuracy are greater than either PSI-BLAST or COMPASS templates under similar conditions. To investigate whether the SAMD templates can yield similar improvement for other structural property prediction (RSA, SM and CD) as for SS, and to investigate if we could further increase gains over *ab initio* SS predictions with additional specialisation of Distill_H, we re-trained Distill_H with SAMD templates but with a relaxed sequence identity threshold of 25% and a reduced expectation cutoff of $e = 0.1$ as follows.

We re-train Distill_H in 5-fold cross-validation on S3129 with SAMD templates and term it Distill_RH. The target database for template searching is the PDB available on March 25th, 2008. To train template-based predictions in marginal similarity conditions we remove from the PDB all sequences that have a PSI-BLAST hit exceeding 25% sequence identity to the query, prior to searching for SAMD templates, we then search the PDB for templates with an expectation cutoff of $e = 0.1$. We relax the 20% sequence identity threshold used previously for training Distill_H to 25% – given the different way the templates are

obtained (redundancy reduced beforehand, or after the PSI-BLAST search), this results in very similar distributions of template identity in the two cases. We test both predictors with identical sequence identity threshold and expectation cutoff.

We test the performance of Distill_RH on S555 and compare the results to *ab initio* (Distill) predictions. Templates for S555 are created using three rounds of SAMD template generation with increasing e value cutoffs ($e = 0.00005, 0.1, 50$), stopping as soon as we find at least one template using the October 2007 25% pdb_select list as the template database. Figure 2 shows the distribution of the best and the average sequence identity of the SAMD templates for S555. For over a third of these sequences the average sequence identity between the query and the template is 5% or less.

Figure 3 shows the distribution of SS, RSA, SM and CD prediction accuracy for S555 as a function of the sequence identity of the query to its best SAMD template, for both Distill_H and Distill_RH. For SS and SM Distill_RH outperforms Distill even as the sequence identity between the query and the top ranked template approaches zero. RSA and CD are the least responsive to template input, possibly due to their less conserved nature. Predictions do improve by 1.35% and 2.91% on average in the 10-30% range, this represents a real gain in accuracy of 2.42% and 5.29%. If the sequence identity between the query and the best template is less than 10% the *ab initio* predictions are marginally more accurate. For SS in the 0-10% range predictions with SAMD templates exceed *ab initio* predictions on average by 3.12%, and are nearly 4.54% better in the 10-30% range, representing real gains of 4.04% and 5.67%. SM SAMD predictions improve on *ab initio* if the sequence identity between query and template exceeds on average 3%. In the 10-30% range they improve on *ab initio* predictions by 3.79%, a real gain of 6.41%.

When we consider only the residues that are covered by a template (73.55% of 81,141) the Distill_RH improvement over *ab initio* Distill predictions is substantial. Here SS predictions are on average over 4.3% more accurate than *ab initio*, regardless of the sequence identity between that query and the template. Note that sequence identity may be as low as zero and the e value as high as 50. SM and CD show gains of 2.48% and 1.32% respectively. Again RSA shows less improvement at just 0.26%.

We also compare Distill_RH to a baseline which copies the SS, RSA, SM and CD directly from the best template (Table 2). In every case Distill_RH predictions for these same residues outperform the baseline. Distill_RH is 7.2%, 11.47%, 8.44% and 3.1% more accurate than the baseline for SS, RSA, SM and CD respectively, a real improvement in accuracy of 8.99%, 25.16%, 13.36% and 6.04%. The largest gain over the baseline is for RSA - even where the templates are substantially less accurate than the *ab initio* prediction Distill_RH can harness their information to improve on *ab initio* prediction accuracy.

Target Database	Distill_RH	Distill_H	Distill	Jpred 3	PSIPRED	PROTEUS
2007 pdb_select	82.53%	80.63%	79.52%	78.57%	79.82%	77.01%
2008 PDB	87.60%	86.12%	79.52%	78.57%	79.82%	81.86%

Table 3: Secondary structure prediction accuracy (Q3) for S555 using Distill (*ab initio*), Distill_H (PSI-BLAST templates) and Distill_RH (SAMD templates) compared with Jpred 3, PSIPRED and PROTEUS. The October 2007 25% pdb_select list (3,654 sequences) and the full version of the PDB downloaded with PROTEUS in April 2008 (16,623 sequences) are used as the target databases for template searching by Distill_H, Distill_RH and PROTEUS.

3.3 Comparison to Other Methods

We compare SS predictions by Distill, Distill_H and Distill_RH with SS predictions by PROTEUS [30], PSIPRED [36] and Jpred 3 [37] using S555 as our test set (Table 3). We create two sets of templates, PSI-BLAST for Distill_H and SAMD for Distill_RH, as previously described using the October 2007 25% pdb_select list as the target database for the first set and a full version of the PDB supplied with PROTEUS as the target database for the second set. Although Distill_RH incorporates template information, only templates showing less than 25% sequence identity to a query are allowed for the first set. Since this is as strict as the usual separation between training and test set for *ab initio* predictors, we consider the results on the first line of Table 3 to be a fair comparison.

PROTEUS is designed to incorporate structural information into SS prediction, but only when a homologue with more than 25% sequence identity or an e value of $e < 10^{-7}$ is found, hence, when we use the pdb_select as the target database, its predictions are generally *ab initio*. Proteus predictions using the full version of the PDB are more accurate, 81.86%, than our *ab initio* predictions but still less accurate than either Distill_H (86.12%) or Distill_RH (87.60%) when they use this version of the PDB. This again highlights the fact that a soft combination of sequence profiles and structural templates greatly improves prediction accuracy compared to directly extracting structural information from PDB templates as in the 3D-to-2D mapping method of PROTEUS.

We also show that Distill_H and Distill_RH perform well when compared to other methods that do not incorporate structural information, PSIPRED and Jpred 3. PSIPRED is slightly more accurate than the *ab initio* Distill but Distill_RH is 2.71% more accurate using only remote homologues. This increases to 7.78% when we use the full version of the PDB (we exclude the query sequence from its list of templates). Jpred 3 is also slightly more accurate than our *ab initio* Distill (80.12%) using their "easier" DSSP eight to three class conversion: H → Helix; E, B → Strand; G, I, S, T, . → Coil. Again, both our Distill_H and Distill_RH outperform Jpred 3 using either the "easier" or "hard" SS assignment.

Distill_RH	Distill_H	Distill	[57]	PP	SABLE	ACCpro	JNET	PredAcc	NETASA
82.28%	81.6%	79.82%	78.1%	78.1%	77.6%	77.2%	75.0%	70.7%	70.3%

Table 4: Performances of the two-class Distill, Distill_H and Distill_RH for RSA prediction compared with a number of recent methods on the Manesh dataset [60]. Performances of the various methods from [57]. The class threshold is 25% for all methods. Templates up to 90% adopted by Distill_H and Distill_RH. PP [58], SABLE [61], ACCpro [62], JNET [63], PredAcc [64] and NETASA [65].

The overall correct prediction of SS for Distill_RH is 82.53%, 3.01% more accurate than the *ab initio* Distill, an improvement of 3.78%. To estimate the statistical significance of this result, we measure the standard deviation of the error distribution by sampling with replacement N residues from the S555 set M times. In our case $M = 1000$ and $N = 81141$ (the size of the set). Standard deviation can be estimated from N samples x_i and from their average $\bar{x}$. We obtain nearly identical standard deviations of 0.14% and 0.13% for Distill and Distill_RH respectively for the error of both predictors. Given these deviations the observed difference of 3.01% is significant at $p << 0.01$.

In Table 4 we show that Distill_H and Distill_RH both perform well when compared to other well known state-of-the-art RSA prediction methods. Our *ab initio* predictor, Distill, is 1.72% better than the next best methods [57,58], however this increases to 3.5% and 4.18% for Distill_H and Distill_RH respectively. The only RSA prediction method that incorporates structural information from the PDB into the prediction process that we are aware of is ACCpro [59]. In this case BLAST is used to identify homologs with high sequence identity in the PDB to improve ACCpro predictions. When high sequence identity templates are used ACCpro prediction accuracy is $> 77\%$.

4 Conclusion and Future Directions

1D structural properties of protein residues are useful intermediate representations between the primary sequence and the full 3D structure. Traditionally, predictors of 1D structural properties have been *ab initio*. However, as the universe of known folds expands, any detectable degree of similarity to proteins of known structure needs to be fully exploited. In turn, improved predictions may feed into, and help to further improve, comparative modelling and fold recognition systems.

We have developed a high-throughput system for the prediction of 1D structural properties of proteins which takes advantage of similarity to proteins of known structure. We retrieve template information by a novel structure-based algorithm that we have developed in this work, and obtain improved predictions of 1D structural properties well below the Twilight Zone of sequence identity, down to sequence identities of

less than 5% in some cases. This suggests that our approach is very robust with respect to template noise, and may glean information beyond the case where templates represent genuine homologues. Predicted 1D properties are considerably more accurate than those directly derived from the best template, showing that sequence and evolutionary information can help correct errors associated with low sequence identity templates. The methods we presented also deal naturally with predictions in areas not covered by templates.

We are currently looking into a number of directions of further research, namely: under which conditions SAMD templates are genuine homologues and when they are fragments of local structure similar to the query but not from an overall similar fold; whether SAMD-based predictions may be fed back into the loop to refine SAMD templates; whether SAMD-based predictions may be of help in 3D modelling, i.e. whether the refinement of local structure they yield is also captured by fold recognition systems or not.

All the systems are automated and are publicly available at http://distill.ucd.ie/. When appropriate templates are available they are automatically used in the prediction process.

Funding

This work is supported by Science Foundation Ireland grant 05/RFP/CMS0029, grant RP/2005/219 from the Health Research Board of Ireland and a UCD President's Award 2004.

Acknowledgement

Many thanks to Dr Chris Cole of the Barton Group for providing Jpred 3 predictions for S555. We thank Brett Becker for useful suggestions.

References

1. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. The protein data bank. Nucleic Acids Res 28:235–242, 2000.
2. Vitkup, D., Melamud, E., Moult, J., and Sander, C. Completeness in structural genomics. Nat Struct Biol 8:559 – 566, 2001.
3. Chandonia, J. and Brenner, S. The impact of structural genomics: Expectations and outcomes. Science 311:347, 2006.
4. Aloy, P., Pichaud, M., and Russell, B. Protein complexes: structure prediction challenges for the 21st century. Curr Opin Struct Biol 15:15–22, 2005.
5. Chothia, C. and Lesk, A. The relation between the divergence of sequence and structure in proteins. EMBO J 5(4):823–826, 1986.
6. Pearson, W. and Lipman, D. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85:2444–2448, 1988.

7. Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402, 1997.
8. Eddy, S. Profile hidden markov models. Bioinformatics 14(9):755–763, 1998.
9. Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M., and Hughey, R. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins 53:491–496, 2003.
10. Yona, G. and Levitt, M. Within the twilight zone: A sensitive profile-profile comparison tool based on information theory. J Mol Biol 315:1257–1275, 2002.
11. Söding, J. Protein homology detection by HMM-HMM comparison. Bioinformatics 21(7):951–960, 2005.
12. Sadreyev, R. and Grishin, N. COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 326(1):317–336, 2003.
13. Ginalski, K., Pas, J., Wyrwicz, L., Grotthuss, M., Bujnicki, J., and Rychlewski, L. ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res 31(13):3804–3807, 2003.
14. Tang, C., Xie, L., Koh, I., Posy, S., Alexov, E., and Honig, B. On the role of structural information in remote homology detection and sequence alignment: New methods using hybrid sequence profiles. J Mol Biol 334:1043–1062, 2003.
15. Karchin, R., Cline, M., and Karplus, K. Evaluation of local structure alphabets based on residue burial. Proteins 55:508–518, 2004.
16. Przybylski, D. and Rost, B. Improving fold recognition without folds. J Mol Biol 341:255–269, 2004.
17. Gong, H. and Rose, G. Does secondary structure determine tertiary structure in proteins? Proteins 61:338–343, 2005.
18. Karchin, R., Cline, M., Mandel-Gutfreund, Y., and Karplus, K. Hidden markov models that use predicted local structure for fold recognition: Alphabets of backbone geometry. Proteins 51(4):504–514, 2003.
19. Shi, J., Blundell, T., and Mizuguchi, K. FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310:243–257, 2001.
20. McGuffin, L. and Jones, D. Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 19(7):874–881, 2003.
21. Zhou, H. and Zhou, Y. SPARKS 2 and SP3 servers in CASP6. Proteins Suppl 7:152–156, 2005.
22. Rychlewski, L. and Fischer, D. Livebench-8: The large-scale, continuous assessment of automated protein structure prediction. Protein Sci 14:240–245, 2005.
23. Koh, I., Eyrich, V., Marti-Renom, M., Przybylski, D., Madhusudhan, M., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A., and Rost, B. EVA: evaluation of protein structure prediction servers. Nucleic Acids Res 31(13):3311–3315, 2003.
24. Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15:285–289, 2005.
25. Cheng, J. and Baldi, P. A machine learning information retrieval approach to protein fold recognition. Bioinformatics 22(12):1456–1463, 2006.
26. Elofsson, A. A study on protein sequence alignment quality. Proteins 46:330–339, 2002.
27. Krause, A., Stoye, J., and Vingron, M. The SYSTERS protein sequence cluster set. Nucleic Acids Res 28(1):270–272, 2000.
28. Sasson, O., Vaaknin, A., Fleischer, H., Portugaly, E., Bilu, Y., Linial, N., and Linial, M. ProtoNet: hierarchical classification of the protein space. Nucleic Acids Res 31(1):348–352, 2003.
29. Yona, G., Linial, N., and Linial, M. ProtoMap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. Proteins: Structure, Function, and Bioinformatics 37(3):360 – 378, 1999.
30. Montgomerie, S., Sundararaj, S., Gallin, W., and Wishart, D. Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics 7:301, 2006.

31. Pollastri, G., Martin, A., Mooney, C., and Vullo, A. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. BMC Bioinformatics 8:201, 2007.

32. Baldi, P. and Pollastri, G. The principled design of large-scale recursive neural network architectures – DAG-RNNs and the protein structure prediction problem. Journal of Machine Learning Research 4(Sep):575–602, 2003.

33. Mooney, C., Vullo, A., and Pollastri, G. Protein structural motif prediction in multidimensional phi-psi space leads to improved secondary structure prediction. J Comput Biol 13(8):1489–1502, 2006.

34. Vullo, A., Walsh, I., and Pollastri, G. A two-stage approach for improved prediction of residue contact maps. BMC Bioinformatics 7:180, 2006.

35. Baù, D., Martin, A., Mooney, C Vullo, A., Walsh, I., and Pollastri, G. Distill: A suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. BMC Bioinformatics 7:402, 2006.

36. Jones, D. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287:797–815, 1999.

37. Cole, C., Barber, J., and Barton, G. The Jpred 3 secondary structure prediction server. Nucleic Acids Res 36:W197–W201, 2008.

38. Hobohm, U., Schard, M., Schneider, R., and Sander, C. Selection of a representative set of structures from the brookhaven protein data bank. Protein Sci 1:409–417, 1992.

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

40. Sims, G., Choi, I., and Kim, S. Protein conformational space in higher order ψ-ϕ maps. Proc Natl Acad Sci U S A 18:618–621, 2005.

41. Pollastri, G. and McLysaght, A. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21(8):1719–20, 2005.

42. Smith, T. and Waterman, M. Identification of common molecular subsequences. J Mol Biol 147:195–197, 1981.

43. Madhusudhan, M., Marti-Renom, M., Eswar, N., John, B., Pieper, U., Karchin, R., Shen, M.-Y., and Sali, A. Comparative protein structure modeling. In *The Proteomics Protocols Handbook,* Walker, J. M., editor, 831–860. Humana Press Inc., Totowa, NJ, 2005.

44. Dayhoff, M., Schwartz, R., and Orcutt, B. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure,* Dayhoff, M., editor, volume 5, 345–352. National Biomedical Research Foundation, Washington, DC,, 1978.

45. Henikoff, S. and Henikoff, J. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915–10919, 1992.

46. Huang, Y. and Bystroff, C. Improved pairwise alignments of proteins in the twilight zone using local structure predictions. Bioinformatics 22(4):413–422, 2006.

47. Qiu, J. and Elber, R. SSALN: An alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs. Proteins 62:881–891, 2006.

48. Blake, J. and Cohen, F. Pairwise sequence alignment below the twilight zone. J Mol Biol 307:721–735, 2001.

49. Prlic, A., Domingues, F., and Sippl, M. Structure-derived substitution matrices for alignment of distantly related sequences. Protein Eng 13(8):545–50, 2000.

50. Henikoff, S. and Henikoff, J. Automated assembly of protein blocks for database searching. Nucleic Acids Res 19(23):6565–6572, 1991.

51. Pearson, W. and Sierk, M. The limits of protein sequence comparison? Curr Opin Struct Biol 15:254–260, 2005.

52. Altschul, S., Bundschuh, R., Olsen, R., and Hwa, T. The estimation of statistical parameters for local alignment score distributions. Nucleic Acids Res 29(2):351–361, 2001.

53. Bundschuh, R. Rapid significance estimation in local sequence alignment with gaps. J Comput Biol 9(2):243–260, 2002.

54. Bailey, T. and Gribskov, M. Estimating and evaluating the statistics of gapped local-alignment scores. J Comput Biol 9(3):575–593, 2002.

55. Karplus, K., Karchin, R., Shackelford, G., and Hughey, R. Calibrating E-values for hidden markov models using reverse-sequence null models. Bioinformatics 21(22):4107–4115, 2005.

56. Gumbel, E. Statistics of Extremes. Columbia University Press, New York, , 1958.

57. Nguyen, M. and Rajapakse, J. Prediction of protein relative solvent accessibility with a two-stage SVM approach. Proteins 59:30–7, 2005.

58. Gianese, G., Bossa, F., and Pascarella, S. Improvement in prediction of solvent accessibility by probability profiles. Protein Engineering 16(12):987–92, 2003.

59. Cheng, J., Randall, A., Sweredoski, M., and Baldi, P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 33:w72–76, 2005.

60. Naderi-Manesh, H., Sadeghi, M., Araf, S., and Movahedi, A. Prediction of protein surface accessibility with information theory. Proteins 42(4):452–9, 2001.

61. Adamczak, R., Porollo, A., and Meller, J. Accurate prediction of solvent accessibility using neural networks-based regression. Proteins 56(4):753–67, 2004.

62. Pollastri, G., Fariselli, P., Casadio, R., and Baldi, P. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 47:142–235, 2002.

63. Cuff, J. and Barton, G. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 40(3):502–11, 2000.

64. Mucchielli-Giorgi, M., Hazout, S., and Tuffery, P. PredAcc: prediction of solvent accessibility. Bioinformatics 15(2):176–7, 1999.

65. Ahmad, S. and Gromiha, M. NETASA: neural network based prediction of solvent accessibility. Bioinformatics 18(6):819–24, 2002.