A High Performing Tool for Residue Solvent Accessibility Prediction

Lorenzo Palmieri [1], Maria Federico[1], Mauro Leoncini[1,2], and Manuela Montangero[1,2]

[1] Dipartimento di Ingegneria dell'Informazione, Università di Modena e Reggio Emilia (Italy)
[2] CNR, Istituto di Informatica e Telematica, Pisa (Italy)
{lorenzo.palmieri,maria.federico,mauro.leoncini,
manuela.montangero}@unimore.it

Abstract. Many efforts were spent in the last years in bridging the gap between the huge number of sequenced proteins and the relatively few solved structures. Relative Solvent Accessibility (RSA) prediction of residues in protein complexes is a key step towards secondary structure and protein-protein interaction sites prediction. With very different approaches, a number of software tools for RSA prediction have been produced throughout the last twenty years. Here, we present a binary classifier which implements a new method mainly based on sequence homology and implemented by means of look-up tables. The tool exploits residue similarity in solvent exposure pattern of neighboring context in similar protein chains, using BLAST search and DSSP structure. A two-state classification with 89.5% accuracy and 0.79 correlation coefficient against the real data is achieved on a widely used dataset.

1 Introduction

Protein folding is the physical process by which a polypeptide chain folds into its characteristic and functional native structure from a simple sequence of amino acids. It's widely believed that this structure is determined as a whole by the residues sequence. Understanding the key mechanisms of protein folding is for the time being one of the major concern in molecular biology and drug design. However, assessment of solvent accessibility is strongly connected to folding because of the high correlation between hydrophobic forces driving core residues towards a buried exposure state, hence determining the folded structure. Also, solvent accessibility is a strong discriminant for residues lying on the surface, thus becoming likely candidates for being Protein-Protein Interaction sites [1].
As the gap between the number of sequenced proteins and three-dimensional solved structures keeps increasing, many investigation efforts are being made to develop methods able to determine solvent accessibility using only primary sequence data [2,3]. Several exposure state interpretations of residue surface area have been proposed. "Classifiers" ideally divide side-chains exposition area of amino acids in a number of discrete intervals, typically two, three or ten. On the other hand, "real value" approaches describe exposure state by a "continuous" range of values in the [0,1] interval, depending on the exposed surface of the residue. The exposure area is usually computed, for each amino acid, as percent of the maximum area of its side chain that can be exposed to

the solvent. Several different threshold values have been used by discrete classifiers described in the literature, with 5%, 10%, 20%, and 25% being the most popular. Depending on this percent value, amino acids are then classified as either buried or exposed on a binary basis, or with discrete exposure levels in case of multiple threshold systems [4]. Several different approaches have been proposed to cope with the solvent accessibility problem: Information Theory [5,6], Bayesian Statistics [7], Probability Profiles [8], Neural Networks [4,9,11,12,13,14,15], Linear Regression [16,17], Support Vector Machines [18,19], Support Vector Regression [20], Look-up Tables [21], meta-methods [22] and many others [23]. However, exploiting sequence similarity to known structures, namely sequence homology, proved to be a substantial improvement strategy for all these methods, both for secondary structure and Solvent Accessibility prediction [9,24]. In many cases sequence homology dramatically improved accuracy of prediction [7,25]. The improvement rate given by this approach is getting more and more tangible with time, by virtue of the thousands new structures solved every year and deposited in the PDB [26].

We developed a software tool for predicting Solvent Accessibility starting from the amino acidic sequence, which exploits the sequence homology information in an efficient and effective way. The underlying algorithm is based on dictionary-like data structures, and takes advantage of information stored in online databases, providing a very high performance on different kinds of datasets, matching the most popular released software tools, and often outperforming them.

2 Tool overview

Our Relative Solvent Accessibility (RSA) prediction tool is a binary classifier which assigns a *buried* or *exposed* state to each residue of the query sequence. The tool works in two phases, as outlined in Figure 1. Given a query sequence, in the first phase the tool: (a) performs a BLAST homology search in order to obtain a list of sequences homologous to the query, rated by similarity [27]; (b) selects a subset of the returned sequences and fetches the corresponding structure information from the Dictionary of Protein Secondary Structure (DSSP) data bank [28]; (c) computes RSA values for residues using these information and appropriately stores them in pattern-based look-up tables. In the second phase the tool makes the predictions by repeatedly accessing the look-up tables for each residue in the query sequence.

The tool is written in Java and shell scripts, runs under Unix/Linux operating systems, and makes use of Protein-Protein BLAST (v2.2.23+).

In the next sections we will describe each phase in more details.

2.1 Fill-up Phase

Homology Search The query sequence is aligned, using the local alignment algorithm BLAST, against the PDB Data Bank to obtain a list of the most similar sequences whose structures have already been solved. This list is parsed by PDB-Id and the first *N* solved structures are fetched from the DSSP Data Bank, where *N* is a tool parameter (in Section 3.1 we will discuss how to choose a proper value for such parameter). In the following, this set of sequences will be addressed as the set of *hit sequences*.

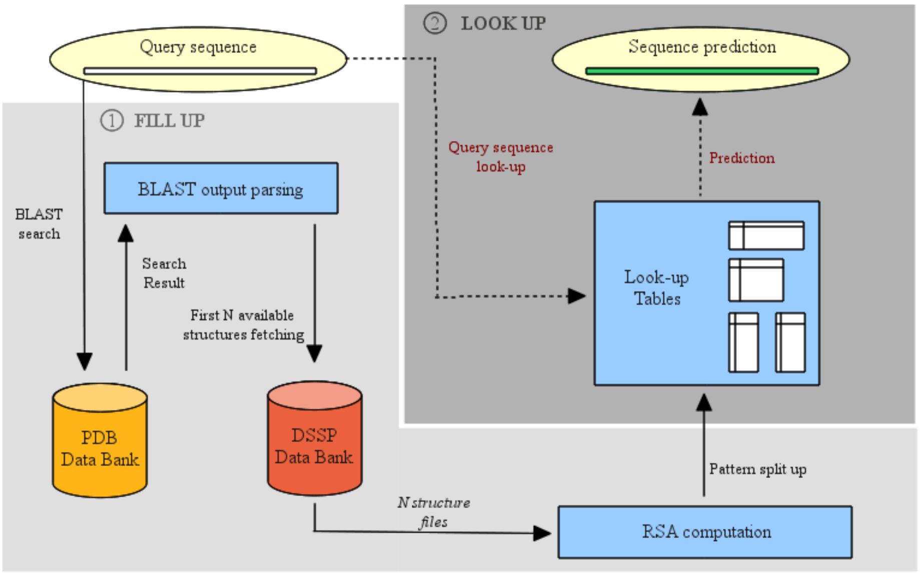


Fig. 1. Two stages prediction workflow. The Fill-up phase includes a BLAST search on PDB known structures, retrieval of structures from DSSP, RSA computation, and Look-up tables creation. The Prediction phase includes looking up the query sequence on the tables, and, finally, residue-by-residue solvent accessibility prediction.

Look-up Tables Creation The DSSP structure files are parsed to obtain Accessible Surface Area (ASA) values of residues in the hit sequences, then these values are used (details in section 3.1) to obtain residue Relative Solvent Accessibility (RSA) values and used to fill-up some specific look-up tables.

For one such table, the entries correspond to k-tuples of residues. We refer to the residue in the middle position of the tuple as the *central* one, and to the other residues as the *context*. The value stored in a given entry is precisely the average RSA value, computed over all the hit sequences, of its central residue in that context. More formally, an entry could be identified by the pair (central,context) = $(r, \langle\alpha, \beta\rangle)$ (where α and β are oligopeptides of total length $k-1$), and the value stored therein as the average of the RSA values of residue r computed over all k-tuples $\alpha r \beta$ appearing in the hit sequence set.

In particular, the tool creates the following four tables:

- 2P2N, standing for "2 Previous 2 Next", is a 21×21^4 table with the 20 standard amino acids on rows (plus a generic *X* amino acid sometimes found in DSSP structures) and 21^4 columns, representing all the possible four residues *context* surrounding the central residue (two residues before, two after, corresponding to the oligos α and β in the notation used above); each entry contains the average RSA value for the central amino acid when surrounded (in the hit sequences) by the specific context represented by the column index.

- 1P1N, standing for "1 Previous 1 Next", is a 21×21^2 table, that stores, analogously to the previous one, the average RSA value when the context consists of only two residues (one before, one after).
- 1P and 1N, standing for "1 Previous" and "1 Next", respectively, are two 21×21 tables. Here the context is composed of only one residue, that can be placed before or after the "central" one.

Explorative experiments (data not given here) showed that larger contexts do not significantly improve the tool performance. This result makes sense if we think that the further we move from one residue, the less probable it is that a residue influences the state of the one under consideration (in this paper, we do not take into consideration the possibility that an independent portion of the sequence, at a large and unpredictable distance, might influence the state of the residue because of the 3D structure of the protein).

Although the amount of space required to store our largest table explicitly would not be a problem for modern PCs (around 30MB, using double precision arithmetic), since the tables are typically sparse we chose an implementation based on hashing.

We observe that a similar approach, based on look-up tables, was used in a previous works by Wang et al. and Carugo [21,5,30]. The crucial differences with our work is that tables there were filled up using information derived from the dataset under study, and not from an independent set of homologous sequences. In particular, we look for sequences showing a high degree of similarity with the one to be predicted, under the hypothesis that sequence similarity implies similarities in protein functions and, hence, also structure similarity. Thus, some of the major differences in the design of our tool are: the introduction of homology search for each sequence, decisions on how to use information coming from homology search has to be taken, look-up tables are computed once for each sequence and not once for the entire dataset.

2.2 Look-up Phase

In the prediction process our tool scans the query sequence residue by residue and, for each residue, accesses the look-up tables in a "hierarchical" fashion, starting from 2P2N down to 1P1N until possibly 1P and 1N.

In details, given a specific query residue, the tool uses its four residue context in the query sequence to access the 2P2N table. In case of a hit (*i.e.*, the value associated to that table entry is non-zero), the RSA value stored in the table is assigned to the analyzed residue of the query sequence as predicted RSA value, and the prediction process moves to the next residue. Otherwise, the two residue context is considered and 1P1N table is examined. In case of another miss (*i.e.*, the value in the appropriate 1P1N entry is zero) a one-residue long context is taken into consideration. We arbitrarily decided to access the 1P table first, and in the case of a miss, the 1N table (by further experiments - data not shown - this assumption turned out not to appreciably influence the prediction performance).

After the look-up phase is completed, the tool assigns a state to the residue that might be *buried* or *exposed*. The decision is made according to the so called *exposure* *threshold* (given as input parameter) on the RSA value associated to query residues: if the RSA value is under the threshold, the residue is classified as buried, otherwise it is labeled as exposed.

In the rare cases of four misses (i.e., a miss in each of the four tables), the exposed or buried state is assigned to the query residue by means of a default value obtained by a Principal Component Analysis (PCA) study of amino acids physiochemical properties [31]. This study suggests to predict standard amino acids exposure state in the following way: buried for A (Ala), C (Cys), F (Phe), I (Ile), L (Leu), M (Met), V (Val), Y (Tyr), W (Trp) and exposed for the others.

3 Experiments

To evaluate our tool, we worked on already solved protein structures. To avoid overfitting and allow fair comparisons with other tools, the experiments were carried using a minor variant of the algorithm described in the previous section: given the result of the homology search, we discard sequences that show an exact match with the query sequence PDB identifier; *i.e.*, we discard chains strictly related with the query sequence (sometimes the query sequence itself). Note, however, that the number of selected hit sequences is always equal to the parameter N.

3.1 Datasets and Experimental setup

We ran experiments using the two datasets described below, which are among those most studied.

Dataset 1 (NM215) This dataset consists of 215 non-homologous protein chains (50878 residues) with no more than 25% pairwise-sequence identity and crystallographic resolution < 2.5Å [6].

Dataset 2 (RS126) This dataset contains 126 non-homologous protein chains (23360 residues), again with no more than 25% pairwise-sequence identity [10].

Prediction Evaluation Indicators To evaluate the performance of our tool, we used two performance indicators: *accuracy* and *correlation*.

At sequence level, accuracy is simply the percentage of correctly predicted residues over the total number of residues in the sequence. At data set level, accuracy is the average sequence accuracy of the sequences in the dataset. At residue level, accuracy is meant as the percentage of correctly predicted residue occurrences in the dataset, over the total number of occurrences of that particular residue in the dataset.

Correlation is computed by means of Pearson's Correlation Coefficient (PCC). Given an R-residue long protein chain, let o_i and p_i denote the observed o_i and predicted p_i solvent exposure states of residue i, for $i = 1, \ldots, R$. Then the correlation c of the chain is given by:

$$c = \frac{R \sum_i o_i p_i - \sum_i o_i \sum_i p_i}{\sqrt{R \sum_i o_i^2 - (\sum_i o_i)^2} \sqrt{R \sum_i p_i^2 - (\sum_i p_i)^2}}$$

with $o_i, p_i \in \{0, 1\}$. The correlation coefficient for a whole dataset is then computed by averaging over the chains in that dataset. PCC values lie in the $[-1, 1]$ continuous interval, with 0 denoting complete uncorrelation, and ± 1 indicating direct and inverse perfect correlation, respectively. In our case, correlation values as closest to 1 as possible are desirable, meaning high similarity between observed and predicted data.

Relative Solvent Accessibility (RSA) Intuitively, the RSA value is an indicator of the percentage of the residue surface area that is exposed. Our tool computes RSA values of the residues in the hit sequences as follows. First, residue absolute Accessible Surface Area (ASA) values are retrieved form DSSP, then these values are normalized using Chothia method [29], i.e., ASA of a given residue X is divided by the maximum exposure area. We recall that the latter quantity is given by the ASA value of the same residue type in a Gly-X-Gly oligopeptide, with the side chain in a fully extended and standard conformation.

Exposure threshold Our tool classifies residues into buried or exposed by means of the exposure threshold on RSA values associated to query residues. The threshold value is an input parameter.

In this work we set the default value of the exposure threshold to 20%. This default value has been used in [32] for the first time as the value that allows an even distribution of residues, with respect to solvent accessibility value, of the sequences in the considered dataset. This threshold value has often been considered as a reference in later works [4,5,7,8,12,13,14,16,18,19,20,22,23].

Similarity Depth (SD) In the first phase, our tool searches for sequences that are homologous to the query sequence using BLAST, and keeps the first hits to be processed later. We call *Similarity Depth* (SD) the number of hits selected by the tool at this stage, which is an input parameter of the tool.

The influence of SD on the overall performance has been studied empirically (with the exposure threshold at the default value). The observed results are shown in Figure 2.

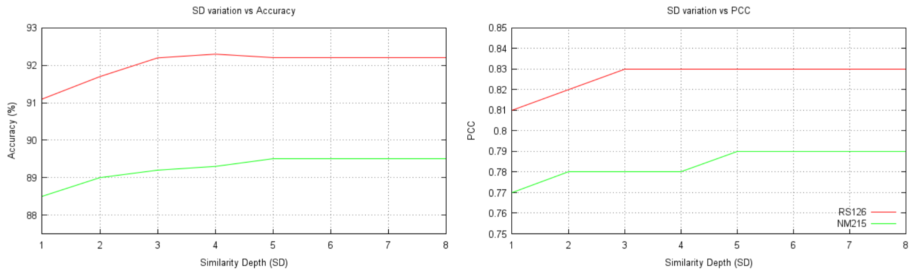


Fig. 2. Accuracy and PCC VS Similarity Depth value for the two datasets. The tool was run in default configuration (20% exposure threshold).

Observe that even with just one hit sequence (the most similar one), remarkable values of accuracy and correlation are obtained: 91.1% (resp. 88.5%) for accuracy and 0.81 (0.77) for correlation for the RS126 (resp. NM215) dataset. Increasing SD leads to a gradual refinement of prediction accuracy, apparently tending to a limit value for both accuracy and correlation when SD is greater than or equal to 5.

The above behavior could be explained by the following observation. Initially, the use of more hit sequences indeed helps to fill up the "higher-level" tables (2P2N being the highest), so that there is a clear improvement in the subsequent look-up phase. As the number of sequences considered further increases, either the pairwise-sequence identity keeps high, and thus the average values stored in the higher tables do not vary much, or otherwise the new RSA values go into lower-level tables, whose entries are likely not to be looked-up because of more probable higher-level hits.

These considerations led us to set SD to a default value of 5.

3.2 Results and Discussion

We compared our tool with some of the most representative and best performing RSA prediction tools available in literature. Competitors use very different approaches to predict the exposure state of residues. For each such approach we selected the best performing tool: [6] for the Information Theory (IT) approach, [8] for Probability Profiles (PP), SARpred [14] for Neural Networks (NN), RSA-PRP [20] for Support Vector Regression (SVR), [23] for a combination of Linear Regression and Support Vector Regression (LR+SVR), and SABLE [13] for a combination of Neural Networks and Linear Regression (NN+LR). We did not compare our tool against those that used a Real Values approach [33,21,15] (including the look-up table approach by Carugo *et al.* [5]), as these are not binary classifiers, which makes output comparison not straightforward. One might post-process Carugo *et al.* tool output using the same threshold used by our tool to produce a binary result, but the comparison might not result fair, as the tools were designed to address different problems.

Results are shown in Table 1: when a tool could not be downloaded or run properly, the reported results are taken from published papers. Missing entries are due to missing results in the original papers. Remember that the majority of tools do not take the exposure threshold as an input parameters, hence for some tests results are not available and we could not test all tools using all exposure threshold values. On the contrary, we ran our tool with any threshold value that has been used by other tool, always providing direct comparison.

The results show that our tool performs very well (in the considered datasets) both in terms of Correlation Coefficient and Accuracy, always outperforming the other tools where comparisons were possible. The obtained results are likely very close to the theoretical limit to solvent accessibility prediction, due to the intrinsic nature of variability for residues of proteins in their native state. In fact, RSA can reach about 10% of variability overall in protein chains with 100% of sequence identity [24].

Our tool is particularly reliable when the query sequence shows high similarity with known sequences, and less reliable otherwise. Nevertheless, our tool is positively affected by the continuous update of the PDB Data Bank: when new solved structures are

Tool/Approach (YEAR)	*Exposure threshold*			
NM215 dataset	**5%**	**10%**	**20%**	**25%**
IT (2001) [6]	75.1% (0.49)^a	75.9% (0.51)^a	-	74.4% (0.47)
PP (2003) [8]	75.7% (0.34)	73.4% (0.40)	-	71.6% (0.43)
SABLE/NN+LR (2004) [13]	76.8% (—)	77.5% (—)	77.9% (—)	77.6% (—)
SARpred/NN (2005) [14]	74.9% (0.31)^b	77.2% (0.50)^b	77.7% (0.56)^b	-
SVR+LR (2008) [23]	81.1% (0.68)	79.7% (0.68)	78.8% (0.68)	-
RSA-PRP/SVR (2010) [20]	77.1% (—)	77.0% (—)	77.5% (—)	77.4% (—)
Our Tool	**91.7% (0.78)**	**90.7% (0.79)**	**89.5% (0.79)**	**89.1% (0.78)**
Our Tool on RS126^c	**94.4% (0.78)**	**93.7% (0.80)**	**92.2% (0.83)**	**91.9% (0.83)**
RS126 dataset	**9%**	**16%**	**23%**	
IT (2001) [6]	78.2% (–)	77.5% (–)	77.4% (–)	
PP (2003) [8]	72.8% (0.39)	71.5% (0.42)	71.4% (0.43)	
Our Tool	**93.4% (0.80)**	**92.3% (0.81)**	**91.7% (0.82)**	
Our Tool on NM215^c	**90.9% (0.79)**	**90% (0.79)**	**89% (0.78)**	

Table 1. Accuracy (and PCC) comparison with other methods and different threshold values on the NM215 and RS126 datasets. ^{c}Accuracy (and PCC) obtained by our tool on RS126 (resp. NM215) with exposure thresholds not used by other tools predicting RS126 (resp. NM215). Our tool is set to default configuration with $SD = 5$. For blank entries see the discussion in the text. ^{a} Only results for 4% and 9% threshold values are available, respectively. ^{b} Mattews' Correlation Coefficient (MCC) used, instead of PCC.

added to the data bank, low performing query sequences might get better predictions if similar enough to the newly added ones.

The following examples clearly show how powerful our tool might be: our prediction for the protein chain 119LA [35] in the NM215 dataset (with default parameters) reaches accuracy 93% and correlation 0.86, compared to accuracy and correlation results, respectively, of 77% and 0.58 for SABLE [13], 83% and 0.62 for RSA-PRP [20], 80% and 0.56 for SARPRED [14]. Even better, our prediction for the protein chain 1bmv 1 [36] in the RS126 dataset (with default parameters) reaches accuracy 97% and correlation 0.95, compared to accuracy and correlation results, respectively, of 74% and 0.52 for SABLE, 70% and 0.42 for RSA-PRP, 69% and 0.37 for SARPRED.

With the aim of making a finer investigation of the good results obtained at the dataset level, we analyzed results also at single sequence level. Figure 3 and Figure 4 show the distributions of correlation values, respectively accuracy values, on the sequences composing the dataset. It can be seen that both values are clustered around the average: accuracy 92.2% and correlation 0.83 for RS126, accuracy 89.5% and correlation 0.79 for NM215.

The worst performing sequences bring accuracy down to 59% for RS126 and 57% for NM215, and correlation down to 0.19 and to 0.13, respectively. We deeply investigated the prediction process for low performing predictions and we found out that this

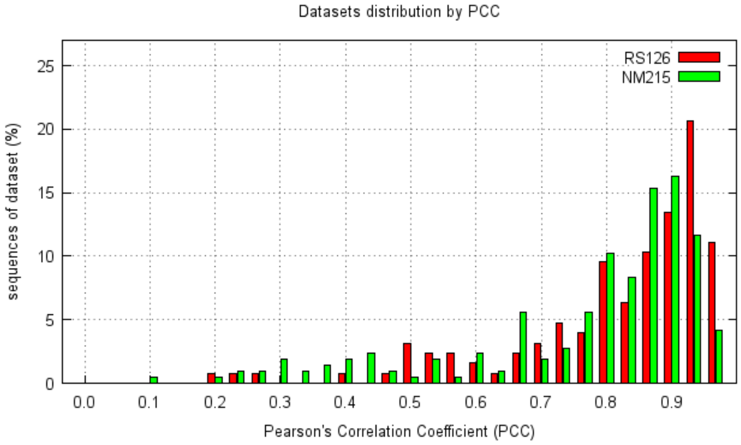


Fig. 3. PCC values reached in prediction, related with the percent number of sequences obtaining specific PCC value. Our tool was run in default configuration.

happens mainly for one (or both) of the following reasons: (1) in the set of hit sequences there are sequences showing less than 30% of identity with the query and sharing local identity of at most three consecutive residues. This implies that the most reliable data for prediction, those in the 2P2N table, are completely absent, and hence that the prediction relies only on shorter contexts. (2) The set of hit sequences contains short sequences that do not cover the entire length of the query sequence; in this way, the prediction of uncovered portions of the query sequence is done according to data that refer to unaligned portions of the sequence.

The former problem is deeply connected with the approach adopted by our tool: if there is no solved structure similar enough to the sequence we wish to predict, then there is a small chance to return a reliable prediction. Nevertheless, the user might be advised of such a situation. On the other hand, the second problem can be somehow worked out (see Section 4 for some intuitions), but we leave this for the “work still to be done” agenda.

We also investigated the obtained results at residue level. Figure 5 shows frequency and prediction accuracy distributions among standard amino acids. Observe that frequency distribution of residues is quite conserved among the two datasets, allowing us to make comparisons between them.

The first and probably most important observation is that the range of accuracy prediction distribution is reasonably small, being about 7%. A finer look reveals that the worst and best predicted residues in both datasets are M (Met) and K (Lys), respectively. Note, however, that even M exhibits sufficiently high accuracies (namely 87% and 89% in NM215 and RS126, respectively), while K reaches such very good figures as 93.5% and 96.5% in NM215 and RS126, respectively.

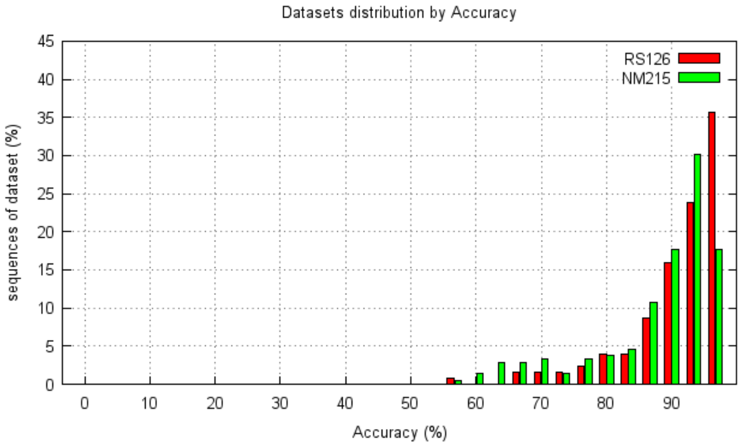


Fig. 4. Accuracy values reached in prediction, related with the percent number of sequences obtaining specific accuracy value. Our tool was run in default configuration.

To improve the tool's overall performance, we should address our attention to those low-performing amino acids that appear with high frequency. In this respect, M itself is not very interesting, since it only appears approximately twice every 100 residues. One such candidate is instead Alanine, whose frequency is among the highest (around 8%). The problem with A (like with other low-performing amino acids) is that it does not have a strong hydrophobic nor hydrophilic preference, and its exposure state floats between buried and exposed depending on the surrounding local environment. It is thus clear that for A's accuracy to improve more context information is desirable. Note that this behavior of A is in agreement with the already mentioned PCA study [31].

Other effects that can be noticed in some residue behavior are probably due to the mixing influence of the two problems (i.e., low local sequence-identity in the hit sequences in the neighborhoods of the considered residue, and low query sequence coverage) that we mentioned when discussing the results at the sequence level. In particular, we may notice that for some amino acid the ranked performance is completely reversed in the two datasets. This is the case of T (Thr) and S (Ser): these are among the best predicted for the RS126 dataset (92.5% and 95% accuracy, respectively), and among the worst ones for the NM215 dataset (around 87%). On the other hand, amino acid C (Cys) is one of the best predicted for the NM215 dataset (with an accuracy of 92%) and one of the worst for the RS126 dataset (accuracy 90%).

Our last investigation deals with look-up tables statistics. As it might be expected, 2P2N is generally a very sparse matrix (on the average, no more than 0.1% of the cells contain a non zero value, for both datasets), nevertheless hits do occur frequently during the prediction process: 84.3% of the times the tool finds a hit in the 2P2N table, for the NM215 dataset, and 93.9% of the times for RS126. Table 1P1N is clearly less sparse than 2P2N, with 3.9% (resp. $2,5\%$) of non zero entries for the NM215 (resp.

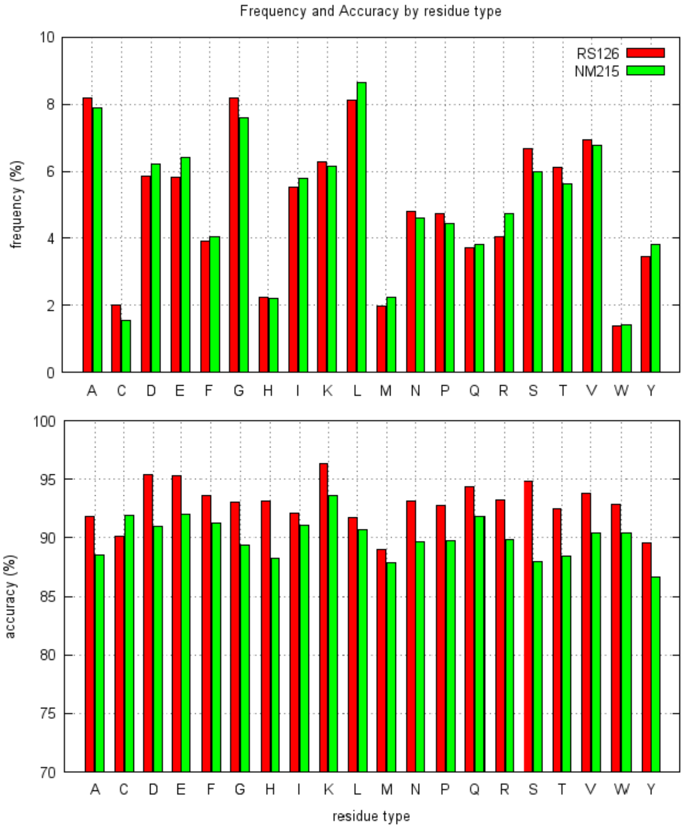


Fig. 5. Residue frequency (top) and accuracy (bottom). Frequency gives the number of times each type of residue appears in the dataset. Accuracy values refer to percent of correctly predicted exposure state for type of residue. Our tool was run in default configuration.

RS126) dataset, but its contents are used only around twice to four times, depending on the dataset, for each 100 table look-ups. Finally, the most populated tables are 1P and 1N, which together have 43.3% and 32.8% of non zero entries in NM215 and RS126, respectively, and are accessed from 3 to 10 times every 100 look-ups, depending on the dataset. For the sake of completeness, we also mention that 1.4% of the times the exposure states ate recovered from the default PCA values.

This data clearly show how 2P2N function is truly relevant, since it stores information for the largest context (of the central residue), and the most similar replication of the environment surrounding each amino acid. Should a stretch of the query sequence match a 5-residues pattern in the tables, this would be a very close replica of the former one, hence representing a very similar peptide environment and providing a reliable prediction. Indeed, we observed that predictions done with very few hits in 2P2N are not very reliable predictions, and vice-versa. This also suggests that the main avenue for a further improvement is not an increase of the context size, but rather an increase in the number of hits in 2P2N.

4 Conclusions and Future work

In this paper we described a tool that is able to produce very reliable predictions on the exposed/buried state of protein amino acids. The tool bases its predictions mainly on sequence homology, by using information of already solved protein structures that show some degree of similarity with the sequence under prediction. Results obtained on consolidated benchmarks show that our tool clearly outperforms existing tools adopting alternative strategies.

Although the results obtained up to now are extremely encouraging, there still is enough room for further analysis and possible improvements. First of all, the tools high performances should be confirmed (or re-assessed) on larger datasets containing chains that have been solved more recently. Secondly, additional work must be done to address some of the problems discussed in the previous section and get possibly even better results.

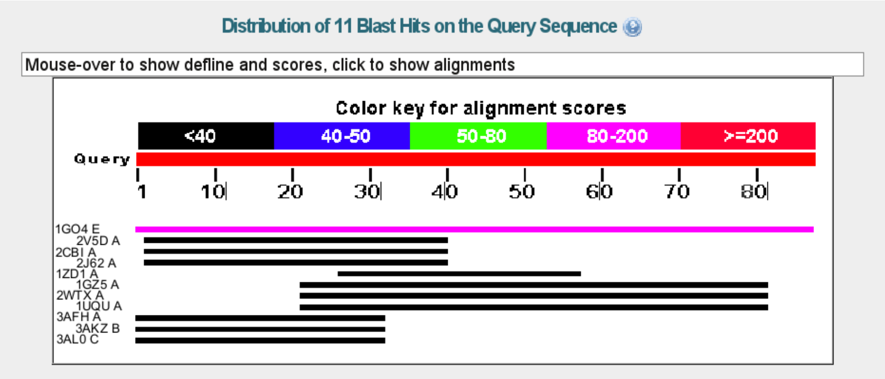
As for the latter point, we plan to address at least the following two issues.

(1) As we pointed out in section 3.2, our predictions are less reliable when the set of hit sequences does not cover the entire query sequence; *i.e.*, there are large enough portions of the query sequence that are not aligned with any portions of the hit sequences. A major optimization of the tool would be to select, in the output returned by BLAST, a set of sequences that covers the entire length of the query sequence, while maintaining a high similarity level.

Figure 6 shows an example where the sequence with PDB identifier 1GO4 E [37] is only partially covered by the best scoring sequences found by BLAST. If we run the current version of the tool, considering only the the two best scoring similar sequences obtained by a BLAST search (namely 2V5D A and 2CBI A), we achieve 62% accuracy and 0.18 correlation in prediction, while a better result of 70% accuracy and 0.40 correlation is achieved if the prediction is done by using the two most similar sequences that span the whole length of the query sequence (namely 2V5D A and 1GZ5 A) [38,39,40]. Although the use of 1GZ5 A in place of 2CBI A gives only a moderate improvement in accuracy, this example does suggest that it is possible to achieve better predictions by improving the query coverage (yet simultaneously storing statistically relevant values in the 2P2N table).

Conversely, neglecting “outlying” stretches of hit sequences much longer than the query, which only share a high local similarity with it, might also facilitate data reliability and noise reduction in prediction, because RSA values relative to the residues in those portions of the sequences will not affect look-up table entries.

(2) In its current version, the tool makes predictions by accessing look-up tables by means of the exact context surrounding the residue under consideration. It might be interesting to investigate the possibility of accessing tables using similar, but possibly not equal, contexts. Here “similar” means that we allow the substitution of some (one or two) context residues with others that do not significantly alter its neighbor exposure



Distribution of 11 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

Fig. 6. BLAST output showing the coverage of the most similar sequences to the 1GO4 E query. Prediction using 2V5D A and 1GZ5 A, instead of 2V5D A and 2CBI A, leads to a significant improvement, as the former sequences span the whole length of the query sequence. Our tool was run with SD = 2, and exposure threshold of 20%.

state. Clearly, the choice of appropriate substitution matrices is crucial here, but the payoff could be an increase in the number of hits in the highest 2P2N look-up table, with the already pointed out benefits on the performance.

Acknowledgments. Authors wish to thank Prof. M.C. Menziani for the useful discussions during the research process that led to the present work.

References

1. Jones, S., Thornton, J. M.: Analysis of Protein-Protein Interaction Sites Using Surface Patches. J. Mol. Biol. 272, 132-143 (1997)
2. Wako, H., Blundell, T. L.: Use of Amino Acid Environment-Dependent Substitution Tables and Conformational Propensities in Structure Prediction from Aligned Sequences of Homologous Proteins. I. Solvent accessibility classes. J. Mol. Biol. 238, 682-692 (1994)
3. Chakrabarti, P., Janin, J.: Dissecting Protein-Protein Recognition Sites. Proteins 47, 334-343 (2002)
4. Rost, B., Sander, C.: Conservation and Prediction of Solvent Accessibility in Protein Families. Proteins 20, 216-226 (1994)
5. Carugo, O.: Predicting Residue Solvent Accessibility From Protein Sequence by Considering the Sequence Environment. Protein Eng. 13, 607-609 (2000)
6. Naderi-Manesh, H., Sadeghi, M., Arab, S., Moosavi Movahedi, A.A.: Prediction of Protein Surface Accessibility with Information Theory. Proteins 42,452-459 (2001)
7. Thompson, M.J., Goldstein, R.A.: Predicting Solvent Accessibility: Higher Accuracy Using Bayesian Statistics and Optimized Residue Substitution Classes. Proteins 25, 38-47 (1996)
8. Gianese, G., Bossa, F., Pascarella, S.: Improvement in Prediction of Solvent Accessibility by Probability Profiles. Protein Eng. 16, 987-992 (2003)

9. Holbrook, S.R., Muskal, S.M., Kim, S.H.: Predicting Surface Exposure of Amino Acids from Protein Sequences. Protein Eng. 3, 659-665 (1990)
10. Rost, B., Sander, C.: Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure. Proteins 19, 55-72 (1994)
11. Ahmad, S., Gromiha, M.M.: NETASA: Neural Network Based Prediction of Solvent Accessibility. Bioinformatics 18, 819-824 (2002)
12. Pollastri, G., Baldi, P., Fariselli, P., Casadio, R.: Prediction of Coordination Number and Relative Solvent Accessibility in Proteins. Proteins 47, 142-153 (2002)
13. Adamczak, R., Porollo, A., Meller, J.: Accurate Prediction of Solvent Accessibility Using Neural Networks Based Regression. Proteins 56, 753-767 (2004)
14. Garg, A., Kaur, H., Raghava, G.P.S.: Real Value Prediction of Solvent Accessibility in Proteins Using Multiple Sequence Alignment and Secondary Structure. Proteins 61, 318-324 (2005)
15. Dor, O., Zhou. Y.: Real-SPINE: An Integrated System of Neural Networks for Real-value Prediction of Protein Structural Properties. Proteins 68, 76-81 (2007)
16. Li, X., Pan, X.M.: New Method for Accurate Prediction of Aolvent Accessibility from Protein Sequence. Proteins 42, 1-5 (2001)
17. Wang, J., Lee, H., Ahmad, S.: Prediction and Evolutionary Information Analysis of Protein Solvent Accessibility Using Multiple Linear Regression. Proteins 61, 481-491 (2005)
18. Yuan, Z., Burrage, K., Mattick, J.S.: Prediction of Protein Solvent Accessibility Using Support Vector Machines. Proteins 48, 566-570 (2002)
19. Nguyen, M., Rajapakse, J.: Prediction of Protein Relative Solvent Accessibility with a two-stage SVM Approach. Proteins 59, 30-37 (2005)
20. Meshkin, A., Ghafuri, H.: Prediction of Relative Solvent Accesibility by Support Vector Regression and Best-First Method. EXCLI Journal 9, 29-38 (2010)
21. Wang, J-Y., Ahmad, S., Gromiha, M.M., Sarai, A.: Look-up Tables for Protein Solvent Accessibility Prediction and Nearest Neighbor Effect Analysis. Biopolymers 75, 209-216 (2004)
22. Chen, H., Zhou, H.X.: Prediction of Solvent Accessibility and Sites of Deleterious Mutations from Protein Sequence. Nucleic Acids Res. 33, 3193-3199 (2005)
23. Chen, K., Kurgan, M., Kurgan, L.: Sequence Based Prediction of Relative Solvent Accessibility Using two-stage Support Vector Regression with Confidence Values. J. Biomed. Sci. Eng. 1, 1-9 (2008)
24. Flores, T.P., Orengo, C.A., Moss, D.S., Thornton, J.M.: Comparison of Conformational Characteristics in Structurally Similar Protein Pairs. Protein Sci. 2, 1811-1826 (1993)
25. Cuff, J.A., Barton, G.J.: Application of Multiple Sequence Alignments Profiles to Improve Protein Secondary Structure Prediction. Proteins 40, 502-511 (2000)
26. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res. 28, 235-242 (2000)
27. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic Local Alignment Search Tool. J. Mol. Biol. 215, 403-410 (1990)
28. Kabsch, W., Sander, C.: Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. Biopolymers 22, 2577-2637 (1983)
29. Chothia, C.: The Nature of the Accessible and Buried Surfaces in Proteins. J. Mol. Biol. 105, 1-12 (1976)
30. Carugo, O.: Prediction of Polypeptide Fragments Exposed to the Solvent. In Silico Biology 3, 35 (2003)
31. Palmieri, L., Federico, M., Leoncini, M., Montangero, M., : Sequence-Based Prediction of Solvent Accessibility in Proteins. University of Modena and Reggio Emilia, M2CSC doctoral research school, internal report 2009.

32. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H.: Hydrophobicity of Amino Acid Residues in Globular Proteins. Science 229, 834-838 (1985)
33. Ahmad, S., Gromiha, M.M., Sarai, A.: Real Value Prediction of Solvent Accessibility from Amino Acid Sequence. Proteins 50, 629-635 (2003)
34. Brenner, S.E., Chothia, C., Hubbard, T.J.P.: PNAS 95, 6073-6078 (1998)
35. Blaber, M., Lindstrom, J.D., Gassner, N., Xu, J., Heinz, D.W., Matthews, B.W.: Energetic Cost and Structural Consequences of Burying a Hydroxyl Group within the Core of a Protein Determined from Ala–¿Ser and Val–¿Thr Substitutions in T4 lysozyme. Biochemistry 32, 11363-11373 (1993)
36. Chen, Z.G., Stauffacher, C., Li, Y., Schmidt, T., Bomu, W., Kamer, G., Shanks, M., Lomonossoff, G., Johnson, J.E.: Protein-RNA Interactions in an Icosahedral Virus at 3.0 A Resolution. Science 245, 154-159 (1998)
37. Sironi, L., Mapelli, M., Knapp, S., Antoni, A., Jeang, K.T., Musacchio, A.: Crystal Structure of the Tetrameric Mad1-Mad2 Core Complex: Implications of a 'Safety Belt' Binding Mechanism for the Spindle Checkpoint. Embo J. 21, 2496 (2002)
38. Ficko-Blean, E., Gregg, K.J., Adams, J.J., Hehemann, J.H., Smith, S.J., Czjzek, M., Boraston, A.B.: Portrait of an Enzyme, a Complete Structural Analysis of a Multimodular beta-N-acetylglucosaminidase from Clostridium Perfringens. J. Biol. Chem. 284, 9876-9884 (2009)
39. Rao, F.V., Dorfmueller, H.C., Villa, F., Allwood, M., Eggleston, I.M., Van Aalten, D.M.F.: Structural Insights into the Mechanism and Inhibition of Eukaryotic O-GlcNAc Hydrolysis. Embo J. 25, 1569 (2006)
40. Gibson, R.P., Turkenburg, J.P., Charnock, S.J., Lloyd, R., Davies, G.J.: Insights into Trehalose Synthesis Provided by the Structure of the Retaining Glucosyltransferase OtsA. Chem. Biol. 9, 1337 (2002)