**PredAcc: prediction of solvent accessibility**

M. H. Mucchielli-Giorgi, S. Hazout and P. Tuffery

Equipe de Bioinformatique Moléculaire, INSERM U155, Université Paris 7, case 7113, 2, place Jussieu, 75251 Paris cedex 05, France

Received on November 11, 1998; accepted on December 3, 1998

---

**Abstract**

**Summary:** PredAcc is a tool for predicting the solvent accessibility of protein residues from the sequence at different relative accessibility levels (0–55%). The prediction rate varies between 70.7% (for 25% relative accessibility) and 85.7% (for 0% relative accessibility). Amino acids are predicted in four categories: almost certainly hidden and almost certainly exposed with a given a posteriori prediction error, probably hidden and probably exposed otherwise.

**Availability:** http://condor.urbb.jussieu.fr/PredAccCfg.html

**Contact:** tuffery@urbb.jussieu.fr

Several methods for predicting the solvent accessibility in two states (hidden or exposed) have been proposed, issued from learning based on Bayesian statistics (Thompson and Goldstein, 1996) or neural networks and starting from a single sequence (Holbrook et al., 1990) or from multiple aligned sequences of homologous proteins (Rost and Sander, 1994). Here, we propose a prediction of the solvent accessibility in four states with learning based on an improved logistic function and starting from a single sequence.

The main principle of the method is the use of a logistic function. The inputs of this function are: the whole sequence length, the distance to the C- and N-terminal extremities, and the amino acid frequencies in the sequence relative to the database [for details, see Mucchielli-Giorgi et al. (1998)]. This function is improved by a prediction-correction procedure and by a thresholding dependent on each amino acid. It returns for a given residue defined by the data {X} the probability P(X) to be hidden. When this probability for an amino acid i is more than a value P5(i) (i = 1, …, 20), the residue is then predicted hidden else it is predicted exposed. The thresholds P5(i) adapted to each amino acid i were determined by maximizing the percentage of correct prediction on a set of monomeric, single-domain proteins of a non-redundant database (Hobohm and Sander, 1996), subset of the PDB (Bernstein et al., 1977). This set was clearly divided into two-thirds for learning, one-third for testing. The solvent accessibility were taken as the output of the DSSP program (Kabsch and Sander, 1983). We have introduced two reliability thresholds for each decision (hidden or exposed): Pinf(i) and Psup(i) determined for an a posteriori error gamma (i.e. the proportion of hidden residues among the predicted exposed residues, and conversely) fixed to 5 or 10%. In fact, when an amino acid i is predicted hidden [i.e. P(X) > P5(i)], we check whether this probability P(X) exceeds the reliability threshold Psup(i); if yes, then the residue is ‘almost certainly hidden’ with an error a posteriori of 5 or 10% and denoted by H. Otherwise, it is only ‘probably hidden’ and denoted by p. The same control is carried out for the predicted exposed residues which are shared in two classes: ‘almost certainly exposed’ denoted by E [i.e. when P(X) le; Pinf(i)] and ‘probably exposed’ denoted by e [i.e. when Pinf(i) < P(X) < P5(i)]. For each threshold of the relative accessibility (values from 0 to 55% by 5% steps), the set parameters of the logistic function are estimated and the values of thresholds P5(i), Pinf(i) and Psup(i) for each amino acid i are determined for an a posteriori error fixed to 5 or 10%.

The percentages of correct predictions are >70.7% (minimal value obtained for 25% relative accessibility) and <85.7% (maximal value obtained for 0% relative accessibility). For example, for 10 and 15% relative accessibility, the accuracies are 77.0 and 73.4%, respectively. The system is more accurate for predicting cyclic residues F, W and Y and the charged residue K with a percentage of correct prediction >80% (values for 25% relative accessibility). For more information, we give an example of the proportions of correctly or not correctly predicted residues, for three relative accessibilities (0%, 25%, 55%). (i) Among the hidden residues: 13.5%, 14.4% and 1% are not correctly predicted, 1.5%, 30.6% and 27.2% are predicted ‘probably hidden’, 0.1%, 5.3% and 55% are predicted ‘almost certainly hidden’. (ii) Among the exposed residues: 1%, 14.9% and 15.5% are not correctly predicted, 21.4%, 27.2% and 1.3% are predicted ‘probably exposed’, and 62.5%, 7.6% and 0.1% are predicted ‘almost certainly exposed’.

The PredAcc program is called by a WWW interface that is simple to use. The user can input its query protein sequence on the submission under the FASTA format. A lot of protein sequences, aligned or not, can be input in a single request. Two buttons allow the user to select relative accessibility between 0 and 55% (by default 25%), and the value of the a posteriori error labelled gamma (5 or 10%, by default 10%). Input sequence length is not limited and an error message is returned when at least one amino acid of the sequence is not recognized. Results are returned in the form of a series of
Fig. 1. Example of output of the PredAcc WWW server for the protein 1aak [relative accessibility of 20% and an a posteriori error (gamma risk) of 10%]. The prediction thresholds of the 20 amino acids: $P_{sup}(i)$ (line with label H), $P_{S}(i)$ (line with label h/e) and $P_{inf}(i)$ (line with label E) are given. Each residue of the sequence is predicted as: ‘almost certainly’ hidden (H) or exposed (E), ‘probably’ hidden (h) or exposed (e).

Characters e, E, h and H. The user can choose between three output formats. The ‘short’ format gives only the character series, the ‘embedded’ format the results completed by the input amino acid sequences, and the ‘long’ format gives in addition the series of probabilities to be hidden $P(X)$ along the sequence as well as the thresholds $P_{S}(i)$, $P_{inf}(i)$ and $P_{sup}(i)$ for the whole amino acids. The user can study the closeness of the values of $P(X)$ relative to the thresholds. The result can be returned either by e-mail directly or on the fly (the e-mail option can be only used when the request provides an identified ip address international). On-line help is provided. Figure 1 gives the results of accessibility prediction for the protein 1aak (a ubiquitin conjugating enzyme).

PredAcc is then an accurate tool for the prediction of solvent accessibility. It is able to predict the accessibility of each amino acid for 20 different thresholds of the relative accessibility (from 0 to 55%) and identifies the residues which are predicted with a given a posteriori prediction error.

References


