A graphical interface for correlated mutations and other protein structure prediction methods

Florencio Pazos, Osvaldo Olmea[1] and Alfonso Valencia[2]

Introduction

We present here a graphical interface for the representation of different types of structural predictions derived from sequence information.

The interface (Figure 1) is able to represent: (i) secondary structure; (ii) accessibility predictions, both of them derived from the Predict-Protein server (Rost *et al.*, 1994); (iii) sequence conserved residues, as calculated by Sander and Schneider (1993) also obtained from the Predict-Protein server; (iv) prediction of contacts by correlated mutations. The degree of conservation, level of correlation and confidence of the secondary structure predictions can be set interactively by the user. This is especially useful since the quality of the different predictions depends on their reliability values.

The need to integrate all this information in a single representation became apparent in a recent protein structure prediction exercise (Hubbard *et al.*, 1996) and in the context of the recent contest on structure prediction (CASP2 http://www.mrc-cpe.cam.ac.uk/casp2/). We have found the combined representation useful. Its use will certainly contribute to new improvements and hopefully will facilitate the access to otherwise unconnected prediction methods.

The newest information displayed is the contact prediction based on correlated mutations. Correlated mutations are those changes that occurred during evolution to compensate for structural drift and/or to maintain stability. They can be found in multiple sequence alignments as groups of positions with similar patterns of variation. Our initial approach to the calculation of correlated mutations was described in Göbel *et al.* (1994). The main difference with other published methods [see Rost *et al.* (1994) for a review] is that in our method the positions in the alignments are not compared directly. Instead, the relative variation of the sequences at each position is compared. It is important to notice that our approach is also completely different from mutual information calculated as conditional probabilities for each type of

residue pair at two different positions (Clarke, 1995). Our procedure can be viewed as the comparison of two grey-colour matrices by the relative distribution of the tones in each rather than by comparing the actual intensities between corresponding pixels.

We have recently made some technical improvements to the published method (Göbel, *et al.*, 1994). First, we use the McLachlan scoring matrix directly, avoiding a previous normalization step that was found to be responsible for numerical rounding errors. Also, we treat gaps as dummy observations, so that gaps are set to an arbitrary value of 0 in the scoring matrix. This value has no effect on the calculation as assessed independently. In this way, the number of sequences at each position (N) is kept constant along the whole alignment, regardless of the presence of gaps. This new treatment of gaps only affects positions with <10% gaps, since positions with more gaps as well as completely conserved positions are excluded from the calculation.

These rather technical changes have a quite dramatic effect on the predictive power of the method, with a 2-fold improvement in the prediction of three-dimensional contacts over our previous implementation. In Figure 2, the accuracy of the current implementation is represented for a large set of non-homologous proteins of known three-dimensional structure. accuracy, also called precision by others, is defined as the fraction of correct predictions over the total number of predictions. A typical example could be gamma-b crystallin (170 residues) for which nine contacts are predicted correctly out of 71 predictions (accuracy of 0.12 when contacts between sequence neighbours closer than six residues are excluded). Accuracy depends on protein size, since it is easier to predict contacts in small proteins that have higher densities in their contact maps.

The graphical interface itself is independent of the program used to predict the contacts. Any method can be used provided that its output fits the simple format described in our Web page (http://gredos.cnb.uam.es/pazos/plotcorr.html). Indeed, a more comprehensive approach that integrates correlated mutations with other sources of information derived from multiple sequence alignments is described elsewhere (O.Olmea *et al.*, submitted). The three main sources of information are: (i) sequence conservation; (ii) contact density per residue type and environment; (iii) stability of correlated mutations to changes in the input

Protein Design Group, CNB-CSIC, Campus U, Autonoma, Cantoblanco, Madrid 28049, Spain
[1]*Permanent address: Physical Chemistry Division, CIGB, Havana, Cuba*
[2]*To whom correspondence should be addressed*

E-mail: valencia@samba.cnb.uam.es
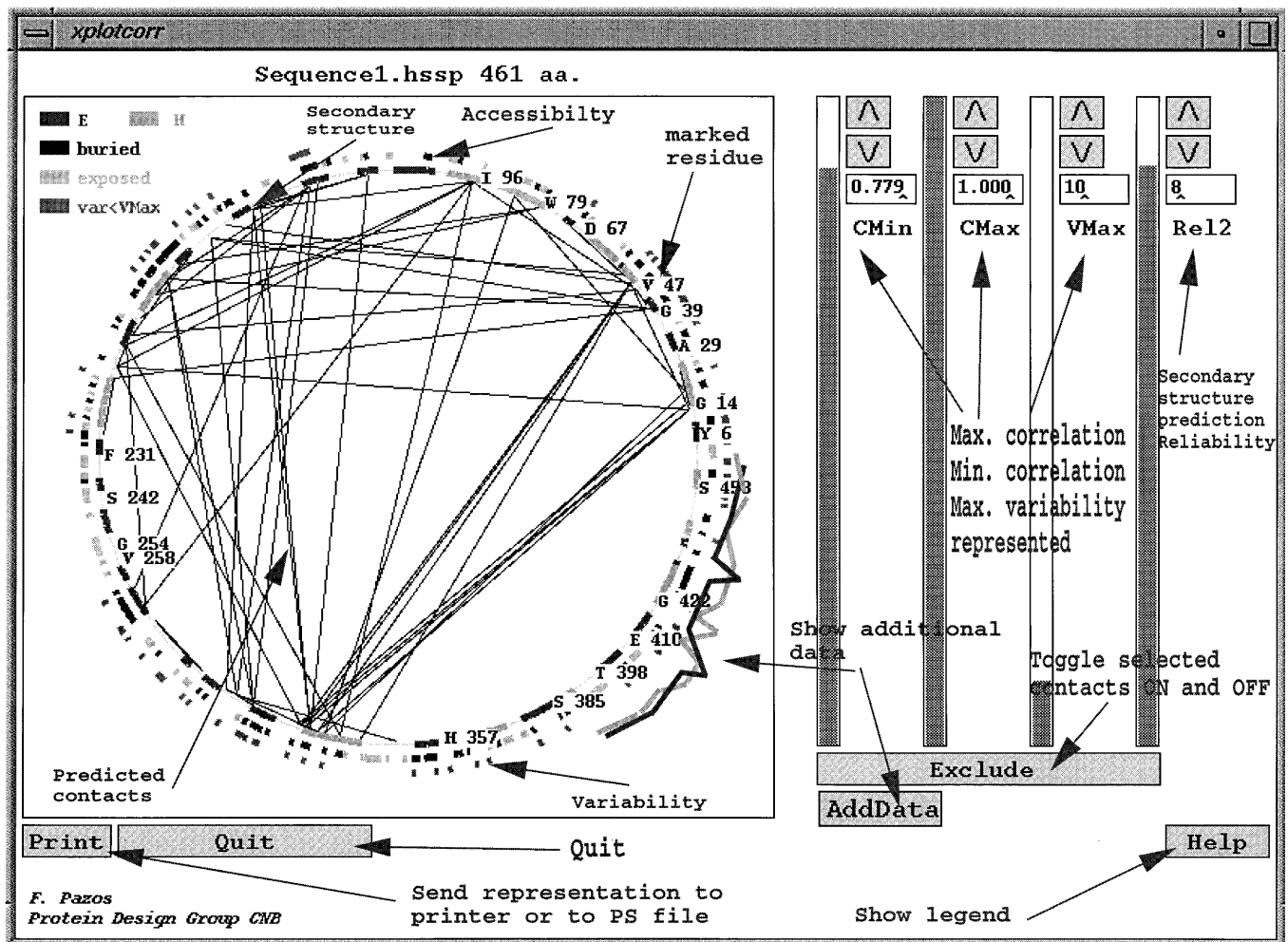*url http://gredos.cnb.uam.es*

Fig. 1. Graphical interface for the representation of correlated positions. In the X-window, the protein sequence is represented as a circle with the N- and C-termini on the right of the figure and the sequence running counterclockwise. Starting from the outside, the following information is displayed in circles. (a) Position of the conserved residues. Conservation is taken from the Predict-Protein server (Rost *et al.*, 1994) as calculated by the program MAXHOM (Sander and Schneider, 1991). The level of variability, starting with 0 (conserved), can be chosen from the dials. (b) Predicted accessibility and (c) secondary structure schema are given, as obtained from the Predict-Protein server (Rost *et al.*, 1994). One of the dial buttons controls the reliability level of the predictions. A colour code indicates the different structural elements. (d) Selected residues: it is possible to pick up residues from the screen to get residue type and number labelled. (e) Correlated pairs of residues are represented by lines connecting the corresponding residues. The level of correlation, between −1 and 1, can be set by the user. In many cases, it is interesting to visualize the results excluding some residues. Residues to be excluded can be read from an external file. This option is particularly useful for excluding tree-determinant residues as calculated by the program SEQUENCESPACE (Casari *et al.*, 1994).

sequence alignment. The new program is already available at the same site and the results can be examined with the graphical interface. The accuracy of this combined approach is better: in the example of gamma-b crystallin mentioned above, 15% of contact predictions are correct.

Availability

The graphical interface is written in C using X graphical libraries. The program that calculates correlated mutations is coded in C++. The typical execution time for the whole procedure is 20 s on a R4400 MIPS processor for a 170 residue long alignment with 50 sequences. User instructions, detailed description of the methods, and the binary code for SGI IRIX 5.2 and Linux operating systems are available at http://gredos.cnb.uam.es/pazos/plotcorr.html. The programs are free for academic use. Commercial users are requested to contact valencia@samba.cnb.uam.es.
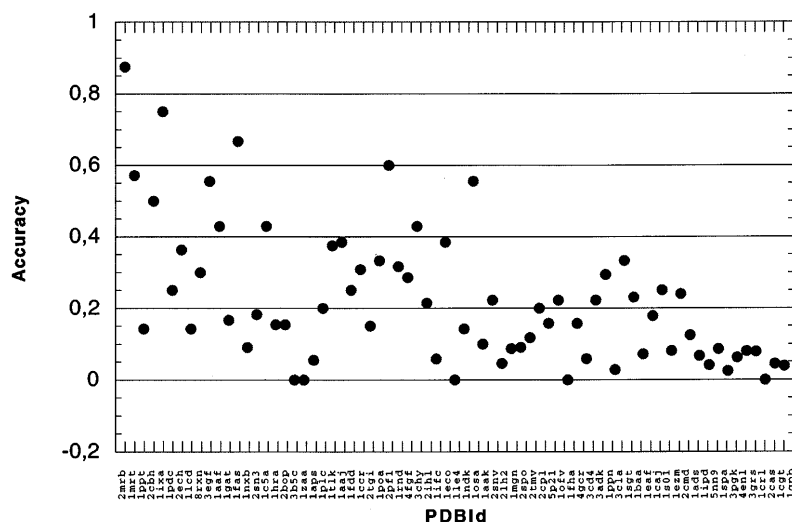
Acknowledgements

Fig. 2. Performances of correlated mutations in contact prediction. *x*-axis: protein families sorted by number of aligned positions. Proteins are labelled by their PDB code. The list includes all proteins in the PDB-select list of non sequence-redundant protein structures (Hobohm and Sander, 1994) for which there were more than alignment of 15 sequences and at least 40 valid positions. in the HSSP database (Sander and Schneider, 1993). Valid positions are those with <10% gaps and not completely conserved. *y*-axis: scale of accuracy (number of true predictions over true plus false predictions). True predictions are those that correspond to pairs of residues with their beta carbons closer than 8 Å (alpha carbons are used for Gly). The number of predicted pairs for each protein is half of protein length. These pairs are chosen from the most correlated positions in the sorted list of correlation values.

References

Casari,G., Sander,C. and Valencia,A. (1994) Functional residues in protein sequence space. *Nature Struct. Biol.*, **2**, 171–178.

Clarke,N. (1995) Covariation of residues in the homeodomain sequence family. *Protein Sci.*, **4**, 2269–2278.

Göbel,U., Sander,C., Schneider,R. and Valencia,A. (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.

Hubbard,T., Tramontano,A. and IRBM Workshop Team (1996) Update on protein structure prediction: results of the 1995 IRBM Workshop. Review. *Folding Design*, **1**, R55–R63.

Rost,B., Sander,C. and Schneider,R. (1994) PHD—A mail server for protein secondary structure prediction. *Comput. Applic. Biosci.*, **10**, 53–60.

Sander,C. and Schneider,R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Sander,C. and Schneider,R. (1993) The HSSP data base of protein structure-sequence alignments. *Nucleic Acids Res.*, **21**, 3105–3109.