A modular learning environment for protein modeling

Jérôme Gracy*, Laurent Chiche, Jean Sallantin***

*Laboratoire d'Informatique, de Robotique et de Micro-électronique de Montpellier
161 rue Ada, 34392 Montpellier Cedex 5, France.

**Centre de Biochimie Structurale
Faculté de Pharmacie
15 avenue Charles Flahault, 34060 Montpellier.

jg@lirmm.fr

Abstract

We propose in this paper a modular learning environment for protein modeling. In this system, the protein modeling problem is tackled in two successive phases. First, partial structural informations are determined via numerical learning techniques. Then, in the second phase, the multiple available informations are combined in pattern matching searches via dynamic programming. It is shown on real problems that various protein structure predictions can be improved in this way, such as secondary structure prediction, alignment of weakly homologous protein sequences or protein model evaluations.

Introduction

Knowledge of the three-dimensional structure of proteins is a necessary condition to fully understand their mechanism of action. However the experimental determination of protein structures is still a long and difficult task and the number of known structures remains much lower than the number of known sequences. This situation led to development of various modeling techniques aimed to predict protein 3D structures from their primary sequences, either *ab initio* or with the help of other data (low resolution experimental model, structure of homologous protein, probable presence of a given structural motif inferred from the protein function, etc.).

Unfortunately, the most popular methods display weaknesses that preclude protein structure prediction from the sequence alone. Molecular mechanics and molecular dynamic are based on empirical potential energy functions known to have difficulties in accounting for electrostatic and solvation effects (Wiener et al. 1986; Van Gunsteren & Berendsen 1990). This, together with the need of very large cpu time, prevents simulation of the folding process. On the other hand, secondary structure prediction methods based on statistical analysis (Chou & Fasman 1974; Garnier, Osguthorpe & Robson 1978), on primary pattern matching (Cohen et al. 1986; Rooman & Wodak 1988; Villareal 1989) or on neural networks (Qian & Sejnowski 1988; Holley & Karplus 1989; Mc Gregor, Flores & Sternberg 1989) show an accuracy limited to 60-70%.

So far, the most efficient prediction method uses the homology between the protein to be modeled and those with known three-dimensional structures (Blundell et al. 1987; Lee 1992). In this case, the known structures can be used as templates for modeling the newly sequenced protein. This approach assumes a structural conservation within the protein family under study during the evolution process (Hubbard & Blundell 1987).

But the most promising recent advances came from the use of multiple sequences to increase the prediction accuracy (Benner & Gerloff 1990) and from methods to solve the "inverse problem", i.e. evaluating the compatibility between a sequence and a given 3D fold (Bowie, Lüthy, Eisenberg 1991; Sippl 1990; Finkelstein & Reva 1991), or the correctness of a predicted model (Chiche et al. 1990; Lüthy, Bowie & Eisenberg 1992; Holm & Sander 1992).

Need for a modular modeling system

Indeed, the emergence of new tools based on quite different approaches prompted us to develop a modular system able to efficiently combine these tools to afford automatic improved predictions that had previously to be done by hand (see for example, Rippmann et al. 1991). The proposed system is based on various learning units that extract knowledge about sequence-structure relationships and offers methods to combine the multiple learned informations with external constraints (multiple homologous sequences and 3D structures) in order to solve specific problems.

There is indeed a large variety of structure related questions that can be raised about proteins according to various experimental materials or knowledge that are available in each particular case. Several typical problems are described below and will serve to illustrate the various cooperative strategies and information synthesis methods provided by the system.

Problem 1. Several possible structural models are sometimes proposed for new proteins as a result of modeling studies (ab initio predictions, homology modeling, energy calculations ...). In this case, the models can be ranked according to sequence-structure compatibility scores. These scores are best computed

from learned functions optimized to correctly evaluate the compatibility between sequences and given structural features.

Problem 2. When the studied protein has no evolutionary relationships with any protein with known structure, prediction rules can be used to obtain crude structural informations as, for instance, the protein secondary structure. But if several homologous sequences are available, the alignment of their prediction profiles will significantly improve the accuracy of the prediction.

Problem 3. When a protein is homologous to others proteins with known structures, their probable 3D similarity can be reliably used to predict the folding of the new protein. Such homology modeling actually consists in finding the best fit (optimal alignment) of the new sequence onto the target 3D templates given by the known structure.

Problem 4. When weakly homologous proteins are compared, the optimal alignment relative to any scoring method does not often exactly correspond to the true structural alignment. This correct alignment will more probably belong to a set of near-optimal alignments that can be generated by modification of the standard alignment algorithm. Then, this limited listing of alignments can be further filtered using scoring functions computed with the learning units. Differences observed among the near-optimal alignments can also help the user to focus on ambiguous sequential positions.

Overview of the system

The protein structure prediction problem can be easily expressed using a pattern recognition framework : the learning environment constructs a global function able to translate the source description of proteins - their primary sequence - into a target description - their 3D structure -. The effective construction of such a sequence-structure associative function involves two main successive steps. During the knowledge acquisition step, prediction functions that deal with multiple protein structure organization levels are learned, then, during the prediction and evaluation step, structural informations are deduced and combined to map the source sequence onto the target structure.

Knowledge acquisition step. The complete system is organized in units that deal with complementary aspects of the protein structure organization levels (primary structure - i.e. chemical sequence -, secondary and tertiary structure - i.e. short-range and long-range internal interactions -, solvent exposure - i.e. external interactions -). Within each unit, a selected set of protein structures is used as learning database to optimize the parameters of local prediction rules that value the propensity of polypeptides to adopt particular folds.

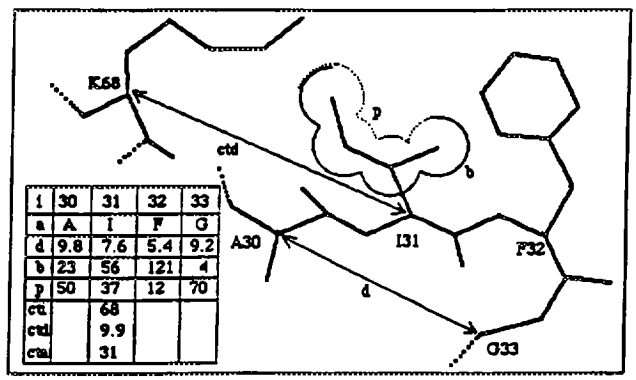
Prediction and evaluation step. The protein sequence under study is compared, by using the learned rules within each unit, to the proteins with known 3D structures or to the predicted structural profiles computed with their sequences. The resulting local compatibility scores help us to select ranked matches between residues along the sequence and 3D environments extracted from the structures. Depending on the problem, the new sequence can be aligned onto the target 3D structures by a dynamic programming algorithm that optimally satisfies the constraints expressed by the complementary scoring profiles.

Knowledge processing methods

Description of protein structures

The Protein Data Bank (Bernstein et al. 1977) provides three-dimensional coordinates of protein atoms. In order to be efficiently processed, proteins must preferably be described as linear structured objects. The chosen descriptive parameters are (Figure 1):

- The *sequence a* expressed by an alphanumeric chain of amino acids.
- The *secondary structure d* which describes the local folding of the main chain (helix, sheet, loop). We retained as a characteristic the distance between two alpha-carbons separated by three positions along the sequence (d_i=Distance($C\alpha_{i-2}$,$C\alpha_{i+1}$)).
- The *buried area b* of a side-chain which corresponds to the part of its Van der Waals surface that is inaccessible to the solvent. This area was calculated with a local method that implement the Kabsch and Sander algorithm (1983).
- The *percentage of polarization p* which corresponds to the fraction of the total area covered by polar atoms (O, N) or by the solvent (H_2O) (Lüthy, McLachlan & Eisenberg 1991).



i	30	31	32	33
a	A	I	F	G
d	9.8	7.6	5.4	9.2
b	23	56	121	4
p	50	37	12	70
cti		68		
ctd		9.9		
cta		31		

Figure 1: A polypeptide and its descriptive code.

Finally, two matrices were set up to describe the long range interactions in the protein:

- The *distance matrix ctd* gives, for every pair of residues, the distance between their alpha carbons.
- The *orientation matrix cta* gives the relative orientation of their side-chains. This orientation parameter is indeed the angle between vectors directed from their alpha-carbons to the center of gravity of their side-chains.

Induction units

Let us describe three units which illustrate different inductive techniques used to afford knowledge about different features of protein structures.

Unit e_{2D}: local structural environment. If there is at least one known 3D structure, it can be used as a target geometrical template. For each position i along the sequence, a structural context is then defined by its buried area b_i, its polarized surface p_i and two successive inter Cα distances (d_i,d_{i+1}). The local compatibility between a residue and a structural context can be evaluated via empirical scores (Figure 2).

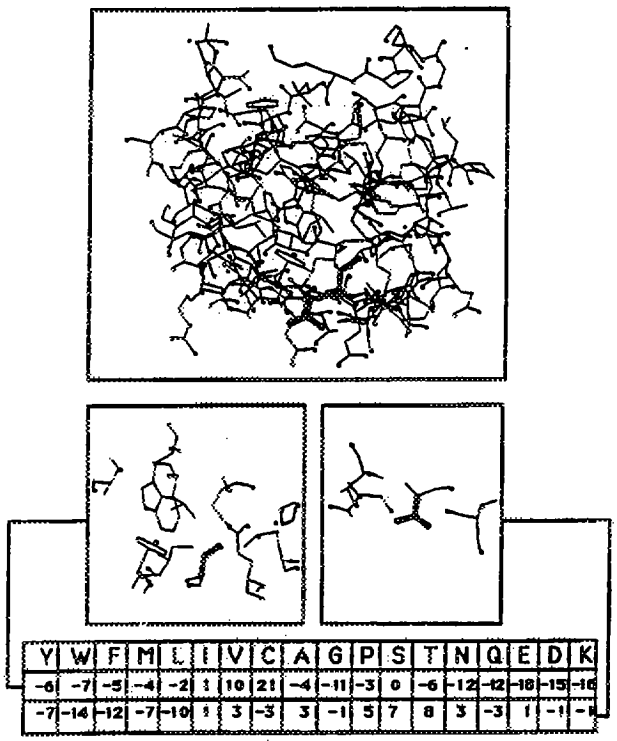


Figure 2: Evaluation of the sequence-structure compatibility. The local structural environments of a buried cysteine and of an exposed threonine were extracted from the VL domain of an immunoglobulin (1REI) and their compatibility against every type of residue was quantified using the potentials. In the first case, hydrophobic residues are positively evaluated. In the second case, polar amino acids are preferred.

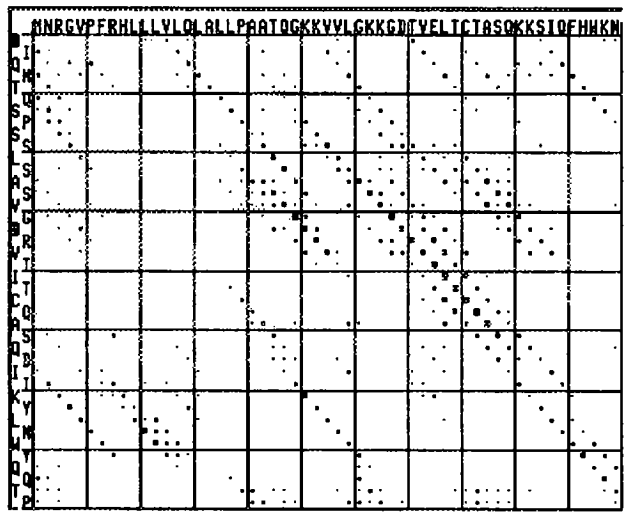


Figure 3 : Comparaison of the beginnings of CD4 sequence and 1REI immunoglobulin structure through the e_{2D} score (averaged over 5 consecutive positions). Squares indicate compatible sequence-structure matches.

Assuming a Boltzmann distribution, compatibility scores between a residue and a structural context are calculated from the frequencies with which each type of amino acid is found in a particular structural environment in a selected set of protein structures. Let a be a given amino acid and c be a structural context expressed by the vector (b,p,d_1,d_2). The compatibility between a and c is then calculated with the formula:

$$\pi_{2D}(a,c)=\ln[P(a|c)/P(a)],$$

where $P(a/c)$ is the conditional probability of a knowing c and $P(a)$ is the frequency of a.

When a sequence $(a_1...a_J)$ is compared to a structural profile $(c_1...c_J)$, local scores are evaluated by (Figure 3):

$$e_{2D}(i,j)=\pi_{2D}(a_i,c_j).$$

Unit e_{p2D}: secondary structure prediction. If there is no known structure in the protein family under study, partial structural informations can be obtained via secondary structure predictions on fragments of the sequence. The prediction of a particular secondary structure s is computed on a local window of *a priori* fixed length L around position i :

$$p_{2D}(s,i)=\Sigma_k\, \pi_{p2D}(s,k,a_{i+k}),$$

where the weight $\pi_{p2D}(s,k,a_{i+k})$ is expressing the propensity of residue a at position $i+k$ in the window to cause a local fold of type s at position i.

Knowledge about local folding expressed by these weights is extracted from a set of protein structures during a preliminary learning phase. This learning phase is performed in two successive steps. During an unsupervised learning step, a dictionary of representative secondary patterns is created using a clustering method. Then, for each structural pattern, a prediction rule is iteratively optimized by a supervised error minimization algorithm. Let us describe both steps.

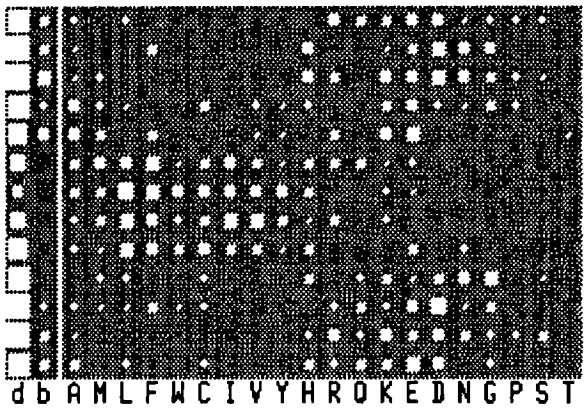


Figure 4 : A learned prediction rule. Each line corresponds to one of the 13 positions in the segment. Columns d and b display for each position the average distance and buried area of the predicted prototypical pattern. White squares respectively correspond to d_i>7Å and b_i<50%. The related matrix shows the weight assigned to each residue at each position of the window. Black squares correspond to $\pi_{p2D}(s,k,a_k)>0$. The larger the square, the larger the absolute value.

Initially, a set of proteins with known structures is chosen by the user and each protein is split into overlapping polypeptides of fixed length L. The structure of each segment is described by two vectors of geometrical variables $(d_1...d_L)$ and $(b_1...b_L)$. For each residue a_i at position i on the segment, d_i is the distance between alpha carbons of residue i-2 and residue i+1 and b_i is the portion of the side chain buried into the core of the protein.

The resulting listing of structural fragments $(d_1...d_L,b_1...b_L)$ is then clustered into N classes according to their euclidian distance (Diday 80). Initially, the fragments are arbitrarly clustered. At each iteration, the distance d and the buried area b found at each position of the fragments assigned to a particular class are averaged in order to compute a prototypical description of the cluster. Then, the partition is updated by assigning each fragment to the class of the nearest prototype and the iterative process is repeated until convergence. This adaptive clustering step leads to a partition of the learning set into classes of polypeptides $(a_1...a_L)$ found in the database with similar conformations.

During the supervised learning step, the weights $\pi_{p2D}(s,k,a)$ depending on the structural class, the relative position in the window and the residue type are tuned using the back-propagation algorithm in order to minimize the global error of prediction computed on the whole learning base (Rumelhart, Hinton & Williams 1986; Qian & Sejnowski 1988; Holley & Karplus 1989; McGregor, Flores & Sternberg 1989).

It is worth noting that the learning set of 3D protein structures will preferably be chosen to learn prediction rules specific to the protein class to which the new sequence belongs if it is known (all-α, all-β, α/β or α+β). The window length L and the number of classes N must be empirically determined by trials and errors in order to find a good compromise between the accuracy of the learned rules and their generality measured on a test set external to the learning database.

Unit e_{3D} : tertiary interactions. Contrarily to the structural environment that describes local fold, the long range residue-residue interactions describe the global folding of the protein. Interaction scores between amino acids are computed from a set of 3D protein structures as described above for the e_{2D} term.

Let $P(a,a'/ct)$ be the conditional distribution of the pair of residues (a,a') relative to a contact ct described by its distance ctd and its orientation cta. The compatibility between the pair (a,a') and the interaction ct was then calculated as :

$$\pi_{3D}(a,a',ct)= \ln[P(a,a'|ct)/(P(a).P(a'))].$$

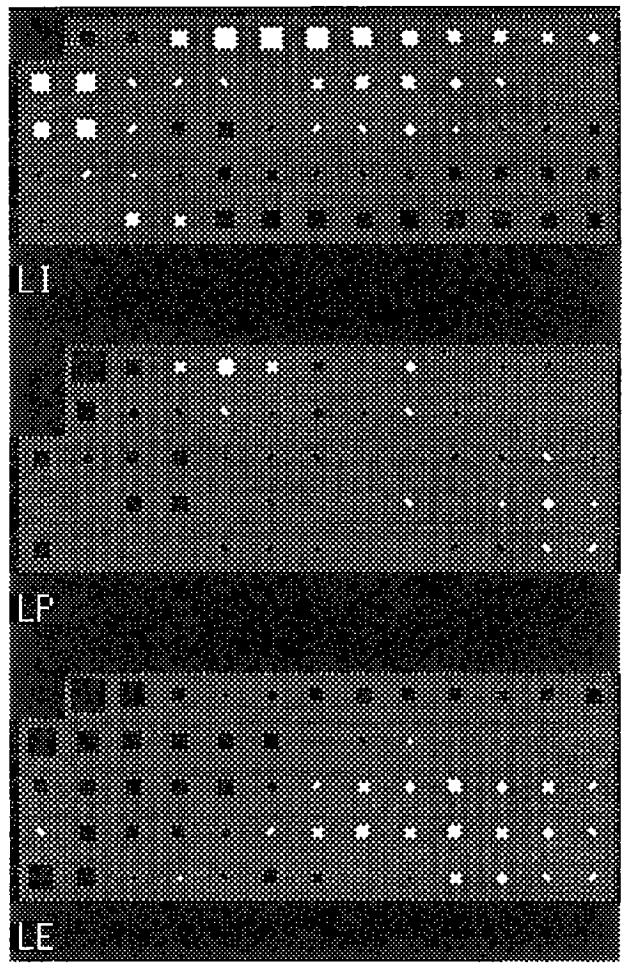


Figure 5: Long range residue-residue interaction scoring tables for few residue pairs. The inter-residue alpha-carbon distance varies along the horizontal axis from 3 to 16 Å and the angle between side chains varies along the vertical one from -180° to 180°. White squares indicate positive (favorable) scores and black squares negative ones.

Figure 5 displays the interaction scoring tables of leucine with isoleucine, proline and glutamine. It can be seen that hydrophobic residues *Leu* and *Ile* are frequent spatial neighbors (optimum from 7 to 10 Å for direct contacts or from 4 to 5 Å for parallel side chain directions - this last case includes interactions between neighbor strands in a beta sheet -) while glutamine is statistically located further from leucine and with side chains in opposite directions. On the other hand the *(Leu,Pro)* pair does not display any clear strong preference. Of course steric clashes are strongly penalized by the 3D score (large black squares).

After the initial statistical induction yielding the scoring table, homologous proteins with known structures can be used to define a set of tertiary contacts that must be verified by the new protein. The tertiary contacts are described by an interaction matrix which gives the relative conformations $(ct_{kl})=(ctd_{kl}, cta_{kl})$ for every pair of residues. The fit between two polypeptides of fixed lengths $(a_i...a_{i+K})$ et $(a_j...a_{j+L})$ and a geometrical pattern defined by the interaction matrix (ct_{kl}) is then evaluated by the tertiary score (figure 6):

$$e_{3D}(i,j,ct) = \Sigma_k \Sigma_l \pi_{3D}(a_{i+k}, a_{j+l}, ct_{kl}).$$

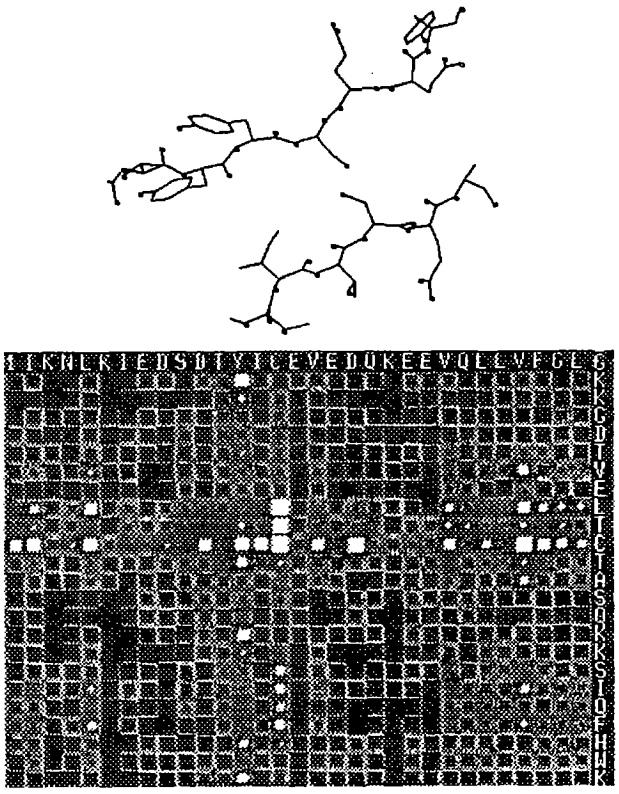


Figure 6: The interaction between the two strands extracted from an immunoglobulin structure has been evaluated using the e_{3D} score against the two fragments of CD4 sequence represented along each axis of the matrix A white square of coordinates (i,j) on the plot indicates a good fit between the (i-4...i+4), (j-1...j+4) polypeptides of each primary fragment and the 3D pattern.

Knowledge synthesis

The different problems initially described in the introduction will now be solved using various combination methods of the information sources provided by the learning units.

Structural hypothesis evaluation (problem1). The system is organized in units that evaluate the compatibility between residues and various structural features. The quality of correspondence between positions i and j of two proteins computed by unit k is given by $e_k(i,j)$. These partial evaluations are then linearly combined into a global score $E(i,j)$:

$$E(i,j) = \Sigma_k \lambda_k e_k(i,j),$$

where parameters λ_k are used to weight the relative influence of each unit.

The particular evaluation terms $E(i,i)$ can be averaged over a protein sequence to score the self compatibility between this sequence and a particular structural model. Application of this strategy to tropomyosin is illustrated on figure 7.

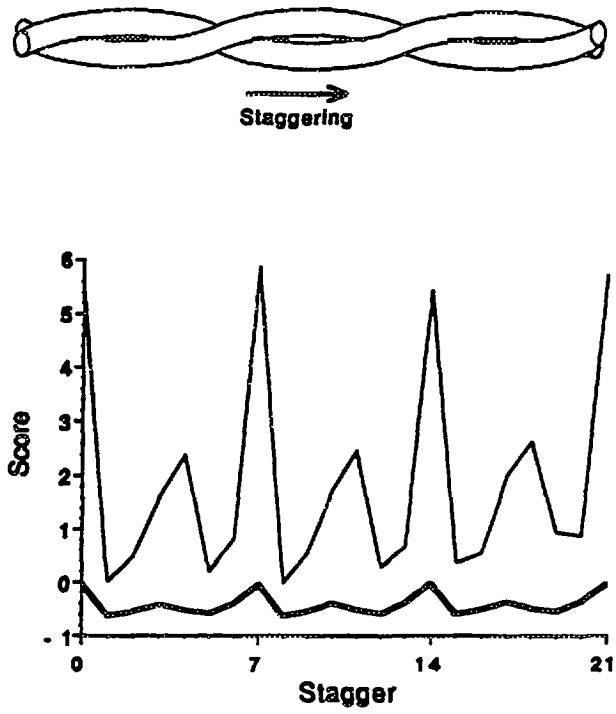


Figure 7. Tropomyosin is made of two identical entirely α-helical chains associated in a parallel coiled-coil structure. Structural models corresponding to various staggering between the helices were obtained from a fast building procedure. The e_{2D} (thin line) and e_{3D} (thick line) global scores of these staggered models show clear differences that agree with molecular dynamics energy calculation (Gracy, Cregut & Chiche 1993).

Homology modeling (problem3). As previously described, homology modeling consists in finding the best fit of the new sequence onto target 3D templates given by known structures. A structural scoring profile can be computed using the above evaluation function. The main difference with probem 1 is that insertions along the sequence are now allowed. The alignment of the new sequence onto the target structural profile is then performed by dynamic programming (Needleman & Wunsch 1970). The cumulated score $Sc(i,j)$ at position (i,j) is calculated from its neighborhood by the recursive formula :

$$Sc(i,j) = \max \begin{vmatrix} Sc(i-1,j-1)+E(i,j) \\ \max_k (Sc(i,j-k)-G(k)) \\ \max_k (Sc(i-k,j)-G(k)) \end{vmatrix}$$

where $E(i,j)$ is the score of the comparison between positions i and j, and $G(k)$ evaluates the cost of k consecutive insertions.

The ability of the 2D and 3D scores to discriminate between homologous and non homologous pairs of proteins were tested on 66 proteins distributed over 17 families These proteins were exhaustively compared by pair. For each pair, the sequence of the first protein was aligned onto the combined sequence and structural profile of the second protein. According to this alignment, the compatibility between sequence 1 and structure 2 was evaluated using 2D and 3D scores.

Figure 8 shows that the e_{2D} and the e_{3D} scores clearly differentiate, for a given fold, the correct sequence, homologous sequences and unrelated sequences (Gracy, Chiche & Sallantin 1992a).

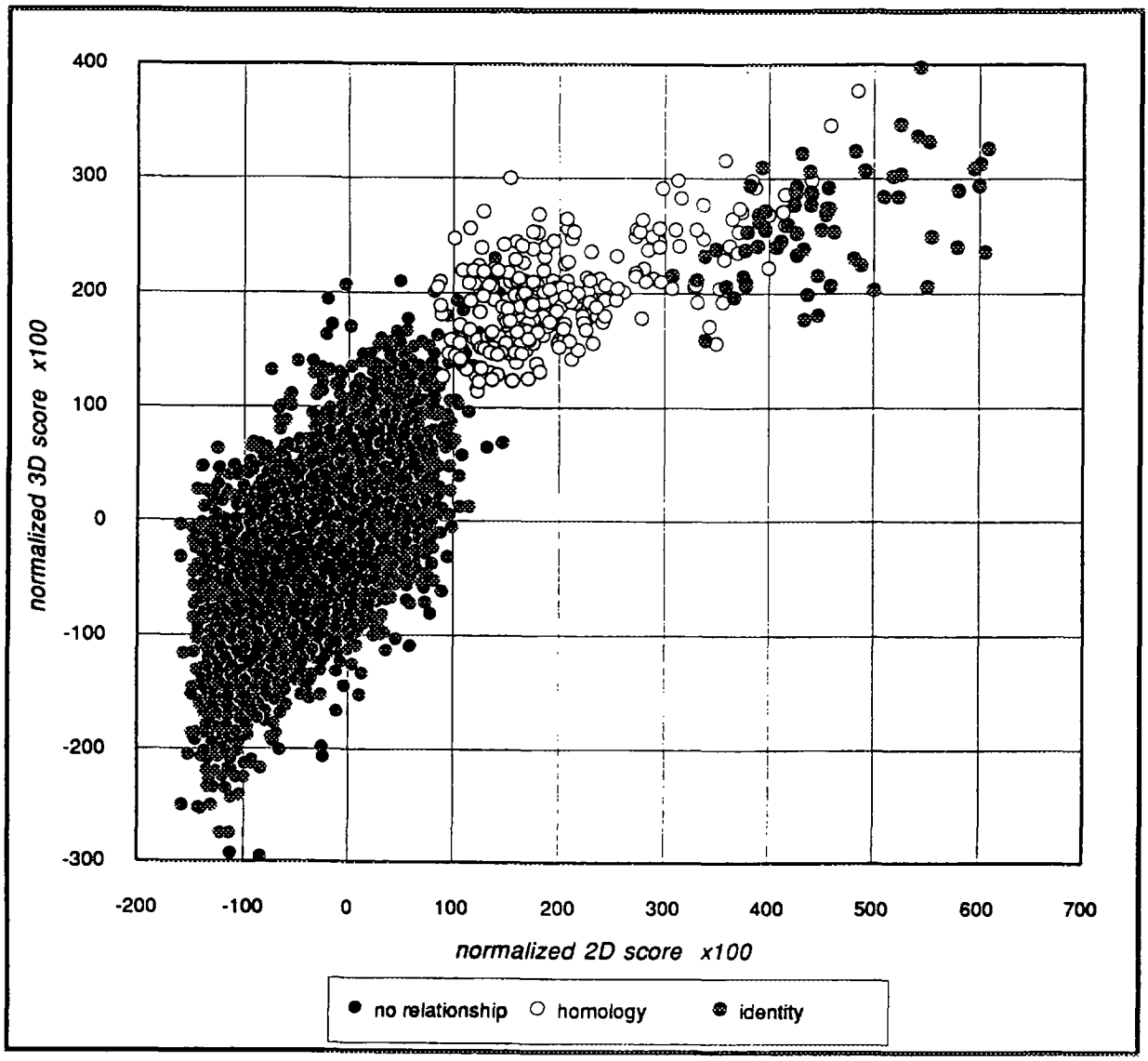


Figure 8: Distribution of normalized 2D and 3D scores for pairwise alignments of proteins obtained from the Brookhaven structural database. Non homologous proteins are indicated by black dots, homologous proteins by white dots and self comparisons by grey dots.

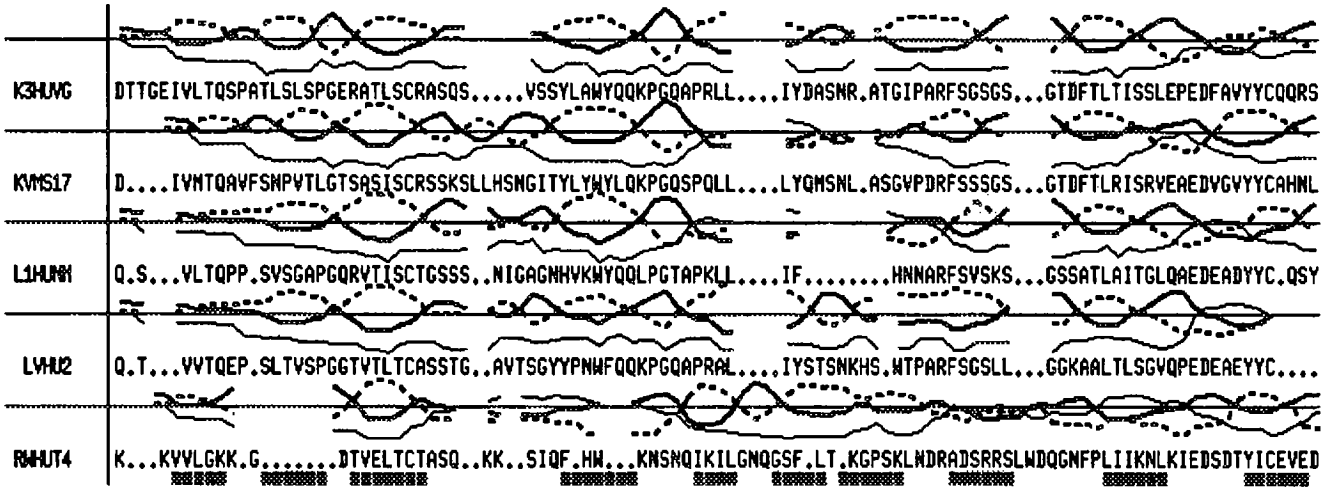


Figure 9: Alignment of the CD4 N-terminal domain (RWHUT4) onto five immunoglobulin VL domains obtained with the combined e_{1D} and e_{p2D} evaluation terms. The prediction profile on three states is shown above its associated sequence (helix = thin line, sheet = dashed line, coil = thick line). Black squares below the CD4 sequence indicate the beta sheet segments assigned according to the average prediction profile.

Secondary structure prediction (problem 2). If no 3D structure is available, then secondary structure prediction rules (e_{p2D} unit) can be used and the e_{1D} primary mutation term may be complemented by a similarity measure between the prediction profiles calculated on both sequences :

$$e_{p2D}(i,j) = -\Sigma_s |p^1{}_{2D}(s,i) - p^2{}_{2D}(s,j)|.$$

If there are more than two proteins to align, a preliminary alignment of every pair of proteins is done. Estimations of the pairwise similarities are then used to order a hierarchical clustering process : at each step, the two most similar proteins are aligned and replaced by a "multiple" profile formed by their association. Two profiles formed by clusters of already aligned sequences, can be compared by averaging the mutation scores between all pairs of residues found at a particular positions along the sequence. The process is repeated until the global alignment corresponding to the root of the clustering tree is obtained. The average prediction profile common to the aligned sequences can then be computed.

An example is shown in Figure 9 for the alignment of the CD4 N-terminal domain with immunoglobulins. The beta strands are correctly predicted in CD4 on the basis of the average prediction profile while structural assignments performed using the CD4 sequence alone only achieve 47% correct predictions (Gracy, Chiche & Sallantin 1992b).

Tertiary evaluation of suboptimal alignments (problem 4). The e_{3D} scores describe long range interactions that cannot be processed by the dynamic programming algorithm which is based on a principle of local optimality. On the other hand, the exponential number of possible alignments prevents their exhaustive evaluation. However, the dynamic programming algorithm can be modified in order to obtain the scores of all the suboptimal alignments (Zuker 1991). This complete information is constructed by executing the algorithm in both forward and backward directions onto the sequences.

Let $FSc(i,j)$ be the partial score of the optimal path relative to the $(1D,2D)$ criterion joining the start of the sequences to position (i,j), let $BSc(i,j)$ be the optimal score calculated from the end of the sequences to the same position (i,j), the score of the optimal path constrained to align positions i and j is :

$$Sc(i,j)=FSc(i,j)+BSc(i,j)-E(i,j).$$

Using this matrix Sc, suboptimal alignments can be extracted by iterating the following procedure :

- Dynamic programming is performed with the (1D,2D) score.
- The scores of matrix Sc along the resulting alignment path are set to negative constant in order to prevent the repetition of alignments.
- The pair (i_0,j_0) achieving the maximum value of the new Sc matrix is selected.
- Positions i_0 and j_0 are artificially constrained to be aligned by over-estimating the term $e_{1D}(i_0,j_0)$.

The $3D$ scores can then be used to evaluate the quality of all the generated suboptimal alignments and to classify them in a rational way. A similar strategy was recently described (Saqui, Bates & Sternberg 1992). However, their approach differs from ours in both the structural criteria used to evaluate the alignments and the method used to generate the suboptimal alignments.

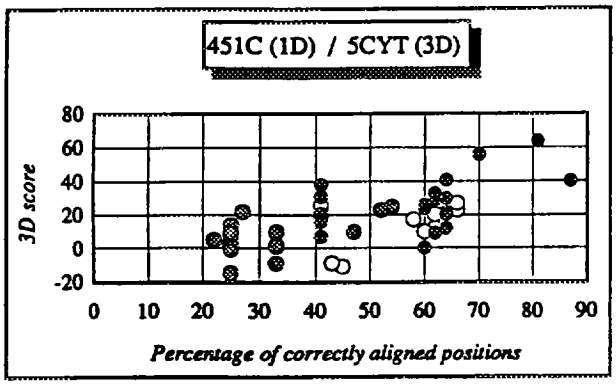


Figure 10: Evaluation of suboptimal pairwise alignments of weakly homologous cytochromes (PDB codes : 451C and 5CYT). 3D scores versus percentage of correctly aligned positions (relative to the true structural alignment) are displayed for the 20 best alignments obtained with the 1D score (white dots), the 2D score (grey dots) and the combined 1D+2D score (black dots).

A careful analysis of the discrepencies observed between the obtained alignments can help the biologist to delimit critical positions along the sequence.This information could be used to suggest mutagenesis experiments that could ascertain the validity of an underlying structural model.

An example is shown in Figure 10 for the alignment of weakly homologous cytochromes. From the data shown, it is clear that: (i) the combination of 1D and 2D informations in the alignment process provides much better results than when the primary information is used alone and (ii) the e3D score quite correctly recognizes the best alignments. The progressive convergence into the neighborhood of the true structural alignment directed by two complementary biases (*1D+2D* score for alignment generation then *3D* score for selection) clearly illustrates the selective power of multi-criteria prediction methods. By this way, it was possible to improve the correctness of the alignment of a sequence on a 3D target by 6.5% on average (Gracy, Chiche & Sallantin 1992a).

Conclusion

We have described a learning environment for protein modeling.

Structural modeling involves a synthesis of informations obtained from multiples sources. In this environment, structural knowledge based on sequence-structure evaluation functions is learned with data-driven inductive methods. In complement, the expert can add its own source of knowledge by selecting suitable interactions between the units according to the nature of the available data (3D models, homologous proteins, functional hypothesis ...).

The various aspects of cooperative strategies between the learning units implemented by the system are illustrated on four typical modeling problems. In each case, improvements due to appropriate evaluation procedures clearly show that inductive units used as complementary information sources allow to increase the quality of the protein structure prediction.

Acknowledgements

This work has been supported by the Centre National de la Recherche Scientifique and Framentec/Cognitech.

References

Benner, S.A., and Gerloff, D. 1990. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Adv. Enz. Reg.* 31: 121-180.

Bernstein, F.C., Koeztle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, Y. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112, 535-542.

Bowie, J. U., Lüthy, R., and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253: 164-170.

Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. and Thornton, J.M. 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 326: 347-352.

Chiche, L., Gregoret, L.M., Cohen, F.E. and Kollman, P.A. 1990. Protein model structure evaluation using the solvatio energy of folding. *Proc. Natl. Acad. Sci. USA* 87: 3240-3243.

Chou, P. and Fasman, G. 1978. Conformational parameters for amino acids in helical, beta-sheets, and random coil regions calculated from proteins. *Biochemistry* 13: 22.

Cohen, F.E., Abarbanel, R.M., Kuntz, I.D. and Fletterick 1983. Secondary structure assignment for alpha/beta proteins by a combinatorial approach. *Biochemistry* 22: 4894-4904.

Diday, E. 1980. *Optimisation en classification automatique.* INRIA, France.

Finkelstein, A.V., and Reva, B. 1991. A search for the most stable folds in proteins. *Nature* 351: 497-499.

Garnier, J., Osguthorpe, D.J. and Robson, B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 120: 97-120

Gracy, J., Chiche, L. and Sallantin, J., 1992a. Improved alignment of weakly homologous protein sequences using structural information. Rapport LIRMM No 92-086.

Gracy, J., Chiche, L. and Sallantin, J., 1992b. Learning and alignment methods applied to protein structure prediction. Rapport LIRMM No 92-085.

Gracy, J., Cregut, D. and Chiche, L. 1993. Statistical potentials and learning methods to evaluate protein models. Application to the coiled-coil Tropomyosin. *THEOCHEM, Computational advances in biomolecular sciences*. Forthcoming.
Holley, L.H. and Karplus, M. 1989. Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. U.S.A.* 86: 152-156.
Holm, L. and Sander, C. 1992. Evaluation of proteins model by atomic solvation preference. *J. Mol. Biol.* 225: 93-105.
Hubbard, T.P.J. and Blundell, T.L. 1987. Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling.*Protein Engng.* 1: 159-171.
Kabsch, W. and Sander, C. 1983. Dictionary of protein structure: pattern recognition of hydrogen-bonded patterns and geometrical features. *Biopolymers*, 22, 2577-2637.
Lee, R.H. 1992. Protein model building using structural homology. *Nature* 356: 543-544.
Lüthy, R., Bowie, J.U. and Eisenberg, D. 1992. Assessment of protein models with three-dimensional profiles. *Nature* 356: 83-85.
Lüthy R., McLachlan A.D. and Eisenberg D. 1991. Secondaray structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins: Struc., Fonc., Gen.* 10: 229-239.
McGregor, M.J., Flores, T.P. and Sternberg, M.J. 1989. Prediction of beta-turns in proteins using neural networks. *Prot. Eng.* 7: 521-526.
Needleman, S.B. and Wunsch, C.D. 1970. General method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48: 443-453.
Qian, N. and Sejnowski, T.J. 1988. Predicting the secondary structure of globular proteins using neural networks models. *J. Mol. Biol.* 202: 865-884.
Rippmann, F., Taylor, W.R., Rothbard, J.B. and Green, N.M. 1991. A hypothetical model for the peptide binding domain of hsp70 based on the peptide binding domain of HLA. *The EMBO J.* 10: 1053-1059.
Rooman, M.J. and Wodak, S.J. 1988. Identification of predictive sequence motifs limited by protein structure database size. *Nature* 335: 45-49.
Rumelhart, D.E., Hinton, G., and Williams, R.J., 1986. Learning internal representations by back-propagating errors. *Nature* 323: 533-536.
Saqui, M.A.S., Bates, P.A, and Sternberg, M.J.E. 1992. Towards an automatic method for predicting protein structure by homology: an evaluation of suboptimal sequence alignments. *Protein Engng.* 5: 305-311.
Sippl, M.J. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* 213: 859-883.
Van Gunsteren, W.F. and Berendsen, H.J.C. 1990. Computer simulations of molecular dynamics: methodology, applications and perspectives in chemistry. *Angew. Chem. Int. ED. Engl.* 29 992-1023.
Villareal Fernandez, M. 1989. Construction par apprentissage de modeles d'objets complexes. Thèse de 3ème cycle, LIRMM, Université de Montpellier II, France.
Weiner, S., Kollman, P.A., Case, D.A., Shingh, U.C., Ghio, C., Alagona, G., Profeta, S. and Weiner P. 1986. An all atom force field for simulations of proteins and nucleic acids. *J. Comp. Chem.* 7: 230-252.
Zuker, M. 1991. Suboptimal sequence alignment in molecular biology. Alignment with error analysis. *J. Mol. Biol.* 221: 403-42.