Testing homology with Contact Accepted mutatiOn (CAO): a contact-based Markov model of protein evolution

Kuang Lin[1], Jens Kleinjung[2], William R. Taylor[1], Jaap Heringa[2]

22nd January 2003

[1] Division of Mathematical Biology, National Institute for Medical Research, Mill Hill, London NW7 1AA, U.K.; Tel +44-20-8816 2633; Fax +44-20-8816 2460; Email: kxlin@nimr.mrc.ac.uk

[2] Bioinformatics Unit, Faculty of Sciences and Faculty of Earth and Life Sciences, Free University of Amsterdam, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands; Tel +31-20-444-7649; Fax +31-20-444-7653; Email: heringa@cs.vu.nl

Keywords: Sequence alignment, structure alignment, PAM model, CAO model

Abstract

PAM (Point Accepted Mutation) is the Markov model of amino acid replacements in proteins introduced by Dayhoff and co-workers (Dayhoff *et al.*, 1978). The PAM matrices and other matrices based on the PAM model have been widely accepted as the standard scoring system of protein sequence similarity in protein sequence alignment tools. Here we present CAO (Contact Accepted mutatiOn), a Markov model of protein residue contact mutations. The CAO model simulates the interchanging of structurally defined side-chain contacts, and introduces additional structural information into protein sequence alignments. Therefore, similarities between structurally conserved sequences can be detected even without apparent sequence similarity. CAO has been benchmarked on the HOMSTRAD database and a subset of the CATH database, by comparing sequence alignments with reference alignments derived from structural superposition. CAO yields scores that reflect coherently the structural quality of sequence alignments, which has implications particularly for homology modelling and threading techniques.

Introduction

Protein alignments describe similarities between protein sequences and structures, which are the result of pattern conservation in proteins in the evolution of organisms. Amino acids with similar physiochemical properties are more likely to interchange in evolution, because these changes are less affected by the pressure of preserving certain folds or functions of proteins, while substitutions between amino acids with very different physico-chemical character are often prohibited. Different models have been developed to describe these evolutionary changes.

The PAM (Point Accepted Mutation) model of amino acid mutations introduced by Dayhoff and co-workers (Dayhoff *et al.*, 1978) is among the most widely used evolutionary models. As a Markov model it is based on the assumption that replacements of residues are independent of the evolutionary history of proteins. Moreover, potential correlations between mutations are ignored and only alternations of single residues are considered. Given these premises, 20∗20 PAM matrices of transition probabilities at different evolutionary distances can be derived from protein alignments by observing residue pair frequencies. The derived PAM scoring matrices for residue mutations provide a statistical quality measure for the objective function of protein alignment programs. The success of the PAM model is due to the fact that the PAM score is an accurate description of the information content (or the relative entropy) of an alignment (Altschul, 1991).

The PAM matrices have been successively updated on larger protein alignment sets and with more accurate estimation algorithms (Jones *et al.*, 1992; Benner *et al.*, 1994; Müller and Vingron, 2000). However, sequence alignment methods have severe con-

ceptual limitations in detecting evolutionary relationships between proteins. While the similarity between two proteins in terms of structure and biological function indicates an evolutionary relationship, the similarity between their sequences remain frequently undetectable by alignment scores, because protein sequences are less conserved than structures (Chothia and Lesk, 1986; Abagyan and Batalov, 1997). Therefore, the generation of accurate (multiple) alignments using sequence-only tools can still be challenging for protein families with highly divergent sequences (Thompson *et al.*, 1999; Koehl and Levitt, 2002).

Given the difficulties of sequence-only alignments, structural alignments are often taken as the standard of truth in describing the relationships between proteins, if high-resolution structures from crystallography or nuclear magnetic resonance (NMR) experiments are available. With the help of protein structures, similarities in the range from global protein folds to local active sites can be located. Many scoring schemes for protein structure similarity have been developed and assessed (Hubbard, 1999; Zemla *et al.*, 2001; Harrison *et al.*, 2002). The most common distance measure is the RMSD value (Root Mean Squared Distance), which requires protein structures to be superimposed, usually as rigid bodies. Using the RMSD value as objective function, optimised superpositions can be generated by minimising the (weighted) average distances of atoms (or residues) between them. If a favourable superposition can be achieved, the RMSD value may be taken as a rough measure of homology (Chothia and Lesk, 1986). The most obvious limitation of the RMSD value is the total absence of sequence information. Proteins can adopt different conformations, in which case structure alignment programs using the RMSD value will report them as dissimilar or even unrelated proteins, while

sequence alignment programs will recognise their identity. The problem of ambiguous protein shape in structural superpositions arises for example from flexible loops, functional rearrangements, relative domain motions of multi-domain proteins, artifacts from crystal lattice interactions, or variation in ensembles from NMR experiments.

In the CAO model presented here, sequence and structural information are unified in a single matrix. This has previously been done explicitly by combining (adding) a score matrix based on structural geometry with a PAM matrix (Taylor and Orengo, 1989; Orengo *et al.*, 1992). This approach, however, uses geometric features that are specific to the two proteins being compared. With the CAO method developed below, the structural component is generic to all proteins and the method is more equivalent to a "threading" (sequence/structure alignment) method in which the compatibility of a sequence is assessed in the context of a structure (Jones *et al.*, 1992). Threading methods assess the compatibility of a residue type in a particular situation through the summation of its pairwise interactions with neighbouring (or all) residues. However, by contrast to threading methods, the CAO method evaluates the compatibility of the residue **changes** seen at both positions in a pairwise interaction. This consideration of residue changes makes it, in spirit, closer to a simple (Dayhoff) residue substitution matrix, but one that takes account of structural interactions.

The CAO matrices can be used to detect both, sequence and structure similarities, and thus serve as an intermediate between the sequence-only PAM score and the structure-only RMSD value. Here, the CAO matrices have been benchmarked on the HOMSTRAD (Mizuguchi *et al.*, 1998) database and on a large structure set from the CATH database (Orengo *et al.*, 1997). The CAO contact definition is based exclusively

on residue side-chain contacts. An underlying assumption of the CAO model is that within short evolutionary periods (*e.g.* 1 CAO) the pattern of side-chain contacts is preserved. Main-chain/main-chain contacts and main-chain/side-chain contacts are ignored. Main-chain contacts are less discriminative for different residue pairs due to the chemical identity of the backbone fragment for each amino acid. Thus, the characteristic folds of native protein sequences are mostly due to the specific patterns of side chain contacts (Heringa and Argos, 1991; Heringa *et al.*, 1995).

Long-range interactions, *i.e.* interactions between residues that are distant in sequence, are most important for the folding and stability of protein structures. It has been observed within families of homologous protein structures, *e.g.* immunoglobulins, globins and γ-crystallins, that strongly interacting side-chain clusters at structurally corresponding locations can have varying amino acid compositions (Heringa and Argos, 1991; Heringa *et al.*, 1995). CAO is a quantitative measure of the evolution of interacting residue types, and it incorporates information about long-range contacts in addition to local interactions. The results given here show that CAO is a promising approach to improve protein structure prediction (Frishman and Argos, 1996; Baldi *et al.*, 1999). The fact that the method only requires a single structure to evaluate sequence alignments in terms of the above structural features makes it significant for homology modelling applications, which are crucially dependent on the structural quality of the alignment.

Materials and Methods

Programs were written in the 'ANSI C/C++' programming languages, compiled with the GNU 'gcc' compiler and executed on PC III processors running on the Linux 2.4 kernel.

Contact definition The definition of residue contacts is based on high-resolution structures from the Protein Data Bank (Bernstein *et al.*, 1977). Only protein main-chain atoms are used. A pseudo side-chain centre is built for each residue (including glycine) using the main-chain as geometrical reference. A detailed description of the positioning of side-chain centres is given elsewhere (Lin *et al.*, 2002). When the distance between two side-chain centres is less than the sum of radii of both side-chains plus twice the radius of the solvent molecule (e.g. water), the two residues are considered as having a side-chain contact (Figure 1). In the work presented here, side-chain radii were defined as previously described (Lin *et al.*, 2002) and the solvent radius was set to 1.4Å. Over the CATH database set, on average each residue is involved in 8.2 side-chain contacts.

Model estimation For the estimation of the CAO model, 6912 pairs of domains were selected from the CATH (v2.0) database (Orengo *et al.*, 1997). Although the two protein domains of each pair are in non-identical families, they are structurally related according to CATH.

The estimation procedure for the Markov model is based on a method to estimate amino acid substitution frequencies from sequence alignments (Müller and Vingron, 2000). Here, the estimator (called the resolvent method) is applied to estimate the

substitution model of residue contacts from structure alignments. The main difference between the models is the dimension of the matrices, which increases from 20∗20 in PAM to 400∗400 in CAO.

The estimation procedure was performed in two steps:

Firstly, for each pairwise alignment in the training set, a maximum-likelihood estimator is applied to estimate the evolutionary distance between the two proteins. In the first iteration step (when the CAO model was not yet established), the estimated PAM distances were used instead of CAO distances.

Secondly, the rate matrix Q is derived from the obtained set of evolutionary distances (for a detailed derivation see Müller and Vingron, 2000). The estimation is based on the relation

$$Q \;=\; \alpha\, I \;-\; R_\alpha^{-1}\;, \tag{1}$$

where I is the identity matrix, α is a constant parameter and the resolvent R_α is defined by

$$R_\alpha \;=\; \int_0^\infty \; e^{-\alpha t} P(t) dt\;. \tag{2}$$

$P(t)$ are the observed probabilities of contact substitutions at the evolutionary distance t.

After the rate matrix Q is produced, matrices of transition probabilities at distance t can be calculated as

$$C(t) \;=\; F e^{tQ}\;, \tag{3}$$

where F is a diagonal matrix with the contact frequencies as entries.

Then, the log-likelihood of n independent alignments produced by the model can be

expressed as

$$L(\alpha) \;=\; \sum_{t} \sum_{i,j} N_{ij}^{t} \log\, C_{ij}^{t}(\alpha)\;, \tag{4}$$

where N_{ij}^{t} is the number of observed substitutions between contacts i and j at the evolutionary distance t.

The process was iterated until convergence was reached, and the α factor was optimised by maximising the log-likelihood value.

Benchmark Benchmarks were performed on the HOMSTRAD database (Mizuguchi *et al.*, 1998) and a subset of the CATH database (Orengo *et al.*, 1997). The HOMSTRAD database with structures, release Sep2002, was used for benchmarking CAO. The database contains 950 protein families, 572 of which contain two sequences. In the CATH benchmark set there are 2652 pairs of domains; 499 pairs are composed of domains in the same topology family but in different homology families; 1421 pairs are composed of domains in the same homologous family but in different sequence families; 732 pairs are composed of domains from the same sequence family (all classifications according to CATH). All pairs were aligned with the structure alignment program SAP (Taylor and Orengo, 1989; Taylor, 1999). The multiple sequence alignment programs Clustal W (Thompson *et al.*, 1994) and Praline (Heringa, 1999; Heringa, 2002) were used with standard settings and with the Blosum62 matrix (Henikoff and Henikoff, 1993), which is similar to the PAM80 matrix. The PAM score was computed as sum-of-pairs score, *i.e.* the sum of substitution values over all column-wise residue matches (pairs) of a multiple alignment, excluding gaps. The raw PAM score was normalised by the total number of pairs to correct for alignment length or gap distribution deviations. The CAO score was

computed as sum of CAO substitution values over all column-wise contact pair matches (contact pairs) of a multiple alignment, excluding gaps. For this purpose, a contact list was pre-compiled for each structure in the HOMSTRAD and CATH databases. Taking every alignment member as reference in turn, those residue pairs aligned to the residue pairs from the contact list of the reference sequence (*i.e.* pairs that do interact in the reference structure) were scored using the CAO matrix; residue pairs without contact were ignored. All thus compiled CAO scores were summed up and normalised by the total number of contacts to correct for alignment length or gap distribution deviations.

Results

Comparison of CAO with PAM and RMSD The CAO model was compared with the PAM model to illustrate their correlation. For each alignment in the CATH dataset the CAO and PAM distances were estimated according to the maximum likelihood. Figure 2 shows that the CAO and PAM distances of CATH alignments are roughly related. Alignments of short CAO distances tend to have short PAM distances and both distances increase together.

In Figure 3, the sequence identities of CATH alignments are plotted against their estimated distances according to the PAM (A) and CAO (B) models and RMSD values (C). The PAM distances (A) are closely related to the sequence identities, in particular at high sequence identity values. Alignments with high sequence identities have short PAM distances and those with low sequence identities show high PAM distances. The CAO model (B) shows a similar tendency, but the variance of sequence identities is much larger. The CAO distance of an alignment can be 100 or lower while its sequence identity is only about 20%. Owing to the additional structure information, the CAO distance is less affected by the sequence similarities in alignments. Thus, CAO distances of seemingly unrelated sequences can be significantly shorter than the corresponding PAM distances. The RMSD value (C) is (as a measure) unaffected by sequence similarity, and the plot shows a large variation in RMSD distances for all sequence identity values. Although the RMSD value can be taken as a measure of homology, it requires structural information for both domains and is highly sensitive to the quality of structures and the superposition procedure. The CAO measure requires only one structure, from which a contact list can be compiled that is used to assess a (multiple) sequence alignment.

Benchmark Benchmarks were performed on the HOMSTRAD database and a subset of the CATH database. The HOMSTRAD database (Mizuguchi *et al.*, 1998) contains protein families represented by protein structures in superimposed form and the corresponding multiple sequence alignments derived from these structural alignments. The CATH database (Orengo *et al.*, 1997) is a hierarchical classification of proteins, derived by an automated procedure and manual curation from the Protein Data Bank (Bernstein *et al.*, 1977). The selected subset of the CATH databank (see Methods) was structurally superimposed using the program SAP (Taylor and Orengo, 1989; Taylor, 1999), which produced the reference sequence alignments corresponding to the superposition.

In this study, the sequences of each family of HOMSTRAD and each pairwise alignment of CATH were re-aligned using the multiple alignment programs ClustalW (Thompson *et al.*, 1994) and Praline (Heringa, 1999; Heringa, 2002). Both programs yielded similar results. In the following, we will refer to the structural alignments as 'reference alignments' and to the ClustalW or Praline re-alignments as 'test alignments'. The PAM scores and CAO scores of the reference alignments and the test alignments were calculated and plotted in Figure 4. A good scoring function should always yield a higher score for the reference alignments than for the test alignments, because the former represent the sequence alignment that reflects the structural homology, while the latter are often biased by sequence similarity. In practise, however, PAM scores of test alignments (from sequence only) are often higher than reference alignments (from structure) (Heringa, 2002). Ideally, all data points should be on or below the diagonal. It is obvious that the PAM scores are mostly above the diagonal due to the absence of structural information in the PAM model. In contrast, most of the CAO scores are below the diagonal, because

the additional structural information of the CAO matrix leads to the assignment of low scores to those test alignments that deviate from the structurally optimal arrangement.

The HOMSTRAD database contains selected proteins from the same family, and the range of sequence similarity (and scores) is smaller than that of the CATH set. The 'outliers' (see for example arrows in Figure 4) are caused by sequence shifts in the structural superposition procedure leading to errors in the reference alignments. It is not surprising that high-scoring alignments show a very narrow distribution around the diagonal (Figure 4), because alignments with closely related sequences have few ambiguities in the sequence arrangement. In the high-score region, both test alignments and reference alignments are of high quality in terms of structural and sequence matching. The most interesting difference between PAM and CAO are low-scoring alignments for which recognising structural relationships is difficult. The CATH test set contains a large number of these alignments (with PAM scores below -1): nearly all of these test alignments yield higher PAM scores than the corresponding reference alignments, while most of them still have lower CAO scores. We therefore conclude that the CAO model is more effective in discriminating alignments of different structural qualities.

For a more quantitative analysis, Δ-scores (= reference score - test score) were calculated, their distributions are plotted as histograms in Figure 5 and values are summarised in Table 1. The difference between the models is apparent in the negative tail of the PAM Δ-score distributions of CATH in Figure 5, which is absent in the CAO Δ-score distribution.

To assess the differences between PAM and CAO scores of reference alignments directly, CAO scores were plotted against the PAM scores for HOMSTRAD and CATH

in Figure 6. As we mentioned before, HOMSTRAD includes manually crafted families of remotely related proteins, while the CATH set contains alignments of sequences from different sequence families. The apparent difference between the HOMSTRAD and CATH data in Figure 6 is in the very low-scoring and very high-scoring regions. The CATH set has more data points in both extreme regions. With alignments of closely related sequences, both models show similar scoring behaviour, yielding scores with a linear relationship. However, in the CATH set, the curve approximates a horizontal line at low PAM scores. According to the difference between the HOMSTRAD and CATH sets, this region contains mainly pairs of very distantly related proteins that yield sequence and structural alignments of low quality. The interesting feature of CAO is that all low-quality alignments achieve a narrow score distribution around the diagonal.

The relationship between the CAO scores and the PAM scores are not greatly affected by the different CAO distances of scoring matrices. The lower panel in Figure 6 shows the same curve for CAO50 (blue squares), CAO80 (grey circles), and CAO200 (red crosses), revealing that the effect is modulated by changes in the CAO distance, but the relation between PAM and CAO remains essentially the same.

To illustrate the predictive capability of CAO, the structure corresponding to the 'outlier' data point in Figure 6 (arrow) was investigated more closely, because the CAO score (2.2) is extremely low considering the high PAM score (3.9). Other outliers are mainly alignments of short peptides or structures without many internal contacts.

The two proteins of this family (ANATO) are C5 anaphylatoxins from human (PDB code 1kjs) and from pig (PDB code 1c5a). Their superimposed structures are shown in Figure 7a, where the match quality is colour coded, and the sequence alignment is

depicted in Figure 7b, with a conservation score as given by Praline. At a high sequence identity of 67.7% and a an RMSD value of 4.9Å over 65 residues, the structures match well only in the core formed by the two central helices; the two flanking helices and the C-termini are partially distorted. The CAO model is capable to predict the low structural match quality using the residue pair contact list, while the PAM sequence alignment score yields a overly high score. An RMSD value of 4.9 suggests a very distant structural relationship and therefore information from the close sequence relationship would be lost. The CAO model yields a high score only in regions of high sequence conservation and good structural match.

Discussion

The CAO model of residue pair contacts has been designed to capture the evolutionary change in proteins by combining sequence and structure information. Therefore, the CAO score is an intermediate between the purely sequence-orientated PAM score and the purely structure-orientated RMSD value, as we showed in the comparison of the CAO score, PAM score and RMSD values on the HOMSTRAD database and the CATH set. The unification of sequence and structure information in CAO allows for a clear definition of evolutionary similarity. As mentioned before, the lack of structural information in PAM is seriously limiting its utility in describing the relationship between evolutionary distant proteins. On the other hand, structural information alone is insufficient, because rigid-body models of proteins can be misleading, which impairs the statistical interpretation of RMSD values in terms of evolutionary distance.

However, the CAO model itself is based on a side-chain model of two coordinates for

each residue (C^{α}, pseudo-C^{β}) and therefore it ignores many details at the atomic level. Contacts are defined between pseudo-C^{β} spheres, ignoring the main-chain interactions in the protein, and each contact is defined as a binary value, *i.e.* a contact exists below a certain distance threshold. These limitation are mainly implied by the objective of CAO. Full atomic details would interfere with the matching of remotely related structures. The signal from main-chains contacts would include predominantly secondary structure information from local contacts and less information about the fold pattern from long-range contacts (Fariselli *et al.*, 2001; Singer *et al.*, 2002).

The CAO, PAM and RMSD measures are derived respectively from a 400∗400 matrix, a 20∗20 matrix and 6 values defining the superposition of two rigid-body models of proteins (describing a plane in each of the structures). This means that the CAO matrix contains the most information. Moreover, the CAO scores are computed using selected regions of the available reference structures, with a high proportion of evolutionary important long-range interactions, while the PAM and RMSD scores often include the whole protein sequence or structure indiscriminately (although often in a weighted form). Furthermore, CAO contacts are correlated, while mutations of residues in the PAM model are independent of each other. As we observed, some peptides (*e.g.* less than 20 amino acids) and short segments (*e.g.* a single helix) don't have many internal contacts and therefore cannot be accurately assessed by CAO scores. However, because the evolutionary statistical basis for comparing such short peptides is absent, it is useful that our CAO measure will never produce a high score or will abstain from assigning a score at all. If it is deemed important to compare short peptides and segments, for example when aligning loop regions in large protein domains, combined CAO/PAM or

CAO/RMSD scores could yield more reliable results for these cases than PAM or RMSD scores in isolation.

We have shown that the CAO scoring system yields improved results when sequence similarities between structures fall into the so-called 'twilight zone' (Figure 4), where signals of amino acid conservation are disrupted, but the structural information of contact conservation is still preserved (Sander and Schneider, 1991; Abagyan and Batalov, 1997). The CAO model therefore should be useful to generate protein structure/structure and structure/sequence alignments, particularly in cases of low sequence similarity such as in homology modelling efforts of distantly related query sequence/template pairs.

References

Abagyan,R.A., Batalov,S., 1997. Do aligned sequences share the same fold?. *J. Mol. Biol.* **273**, 355–368.

Altschul,S.F., 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* **219**, 555–565.

Baldi,P., Brunak,S., Frasconi,P., Pollastri,G., Soda,G., 1999. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* **15**, 937–946.

Benner,S.A., Cohen,M.A., Gonnet,G.H., 1994. Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng.* **7**, 1323–1332.

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

Chothia,C., Lesk,A., 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.

Dayhoff,M.O., Schwart,R.M., Orcutt,B.C., 1978. A model of evolutionary change in proteins, in: Dayhoff, M. (Ed.), Atlas of Protein Sequence and Structure, vol. 5, National Biomedical Research Foundation Washington D.C., pp. 345–352.

Fariselli,P., Olmea,O., Valencia,A., Casadio,R., 2001. Progress in predicting inter-residue contacts of proteins with neural networks and corelated mutations. *Proteins: Struct. Func. Gen.* **Suppl. 5**, 157–162.

Frishman,D., Argos,P., 1996. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.* **9**, 133–142.

Harrison,A., Pearl,F., Mott,R., Thornton,J., Orengo,C., 2002. Quantifying the similarities within fold space. *J. Mol. Biol.* **323**, 909–926.

Henikoff,S., Henikoff,J.G., 1993. Performance evaluation of amino acid substitution matrices. *Proteins: Struct. Func. Gen.* **17**, 49–61.

Heringa,J., 1999. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Comp. Chem.* **23**, 341–364.

Heringa,J., 2002. Local weighting schemes of protein multiple sequence alignment. *Comp. Chem.* **26**, 459477.

Heringa,J., Argos,P., 1991. Side-chain clusters in protein structures and their role in protein folding. *J. Mol. Biol.* **220**, 151–171.

Heringa,J., Argos,P., Egmond,M.R., de Vlieg,J., 1995. Increasing thermal stability of subtilisin from mutations suggested by strongly interacting side-chain clusters. *Protein Eng.* **8**, 21–30.

Hubbard,T.J., 1999. RMS coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. *Proteins: Struct. Func. Gen.* **Suppl. 3**, 15–21.

Jones,D.T., Taylor,W.R., Thornton,J.M., 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275–282.

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

Koehl,P., Levitt,M., 2002. Sequence variations within protein families are linearly related to structural variations. *J. Mol. Biol.* **323**, 551–562.

Koradi,R., Billeter,M., Wüthrich,K., 1996. MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graphics.* **14**, 51–55.

Kraulis,P.J., 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography* **24**, 946–950.

Lin,K., May,A.C.W., Taylor,W.R., 2002. Threading Using Neural nEtwork (TUNE): the measure of protein sequence-structure compatibility. *Bioinformatics* **18**, 1350–1357.

Mizuguchi,K., Deane,C.M., Blundell,T.L., Overington,J.P., 1998. Homstrad: a database of protein structure alignments for homologous families. *Protein Science* **7**, 2469–2471.

Müller,T., Vingron,M., 2000. Modeling amino acid replacement. *J. Comput. Biol.* **7**, 761–776.

Orengo,C.A., Brown,N.P., Taylor,W.R., 1992. Fast structure alignment for protein databank searching. *Proteins: Struc. Func. Gen.* **14**, 139–167.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B., Thornton,J.M., 1997. CATH - a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108.

Sander,C., Schneider,R., 1991. Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.

Singer,M.S., Vriend,G., Bywater,R.P., 2002. Prediction of protein residue contacts with a PDB-derived likelihood matrix. *Protein Engineering* **15**, 721–725.

Taylor,W.R., 1999. Protein structure comparison using iterated double dynamic programming. *Protein Science* **8**, 654–665.

Taylor,W.R., Orengo,C.A., 1989. Protein-structure alignment. *J. Mol. Biol.* **208**, 1–22.

Thompson,J., Plewniak,F., Poch,O., 1999. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15**, 87–88.

Thompson,J.D., Higgins,D.G., Gibson,T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.

Zemla,A., Venclovas,C., Moult,J., Fidelis,K., 2001. Processing and evaluation of predictions in CASP4. *Proteins: Struc. Funct. Gen.* **Suppl. 5**, 13–21.

Tables

Table 1: Statistics of PAM scores and CAO scores of the HOMSTRAD database

	$N^{+\Delta}$	$N^{-\Delta}$
HOMSTRAD PAM score	421 (44%)	545 (56%)
HOMSTRAD CAO score	589 (61%)	377 (39%)
CATH PAM score	995 (38%)	1657 (62%)
CATH CAO score	1422 (54%)	1230 (46%)

Figure Legends

Figure 1. CAO contact definition. A) Each residue (1 and 2) is represented by a backbone centre (C^{α}) and a side-chain centre (C^{β}). A contact between side-chains $C^{\beta 1}$ and $C^{\beta 2}$ (here of residue types Y and L) is defined as existing if the distance D between the side-chains is smaller than the sum of their radii $R_{C^{\beta}}$ plus twice the radius of the solvent R_{Sol}. B) The above residue contact YL of the top sequence is aligned with the residue pair SN of the bottom sequence, for which no structure is known. The CAO matrix provides a score for the contact pair $YL - SN$ that is proportional to the probability of an evolutionary contact mutation from YL to SN. The total CAO score is the sum of CAO matrix scores over all contacts in the top sequence.

Figure 2. Comparison between PAM distance and CAO distance of proteins in the CATH dataset. The solid line is a polynomial fit to show the central trend of the distribution.

Figure 3. Comparison between sequence identity and the scoring models PAM (sequence-only), CAO (sequence and structure) and RMSD (structure-only). The CAO model is conceptually between PAM and RMSD. Even sequence alignments with identities as low as 15% can be detected by CAO as being closely related.

Figure 4. PAM and CAO scores of protein domains in the HOMSTRAD database and the CATH set. Plotted are the scores of multiple alignments derived from structure superpositioning (reference alignments) over those derived from sequence alignment only (test alignments). The top panels show the comparison of PAM scores, the lower panels show the results for the CAO scores. The solid line is the diagonal y = x. An ideal scoring system would yield only values below the diagonal, because the alignments derived from structure represent the best solution and therefore their scores should be higher than (or at worst equal to) the scores

of alignments derived from sequences only. The plots illustrate that most CAO scores are below the diagonal, whereas most PAM scores are above

Figure 5. Δ-score distribution of HOMSTRAD and CATH. The Δ-score is defined as test score minus reference score. In HOMSTRAD (left), the majority of PAM scores is negative, while the majority of CAO scores is positive. The CATH database reveals the inadequate scoring of PAM at low sequence identities. The tail of negative Δ-scores originates from misleadingly high PAM sequence scores for protein alignments of low structural quality. The CAO score distribution of CATH is symmetric and quite narrow around the midpoint.

Figure 6. CAO scores and PAM scores are correlated, but the extreme values of the CATH data illustrate the difference. At high sequence identities (and scores) both scores are linearly correlated; at low sequence identities CAO yields nearly constant minimal scores whereas PAM scores vary from about 0 to -2. The general curve shape is also obtained with the CAO50 and CAO200 matrices, which shows that the observation is not a function of the different distance of the CAO and PAM matrices.

Figure 7. Structural superposition and sequence alignment of two C5 anaphylatoxins (1kjs and 1c5a) from the ANATO family in HOMSTRAD. The quality of the structural match is colour coded (in decreasing order) red, yellow, grey, and blue. Note the discrepancy between the high sequence identity and the poor structural fit except for the two central helices. Below the sequences are given the sequence segments with α-helical conformation in both structures (HELIX) calculated with DSSP (Kabsch and Sander, 1983). The conservation (CON.) values were calculated with Praline and normalised to the range 1 to 10, where 10 is designated by '*' denoting total conservation. Positional residue conservation was calculated as $10 * S(x,y)/\sqrt{S(x,x) * S(y,y)}$, where S is the matrix score for substituting residues x, y. Av-

eraged over five-residue segments are the RMSD values of C^{α} atoms calculated with MolMol (Koradi *et al.*, 1996) and the CAO scores. High CAO scores are obtained in regions with high sequence conservation and low RMSD values. Low sequence conservation and high RMSD values both decrease the CAO score. The structure plot was generated using Molscript (Kraulis, 1991).

Figures

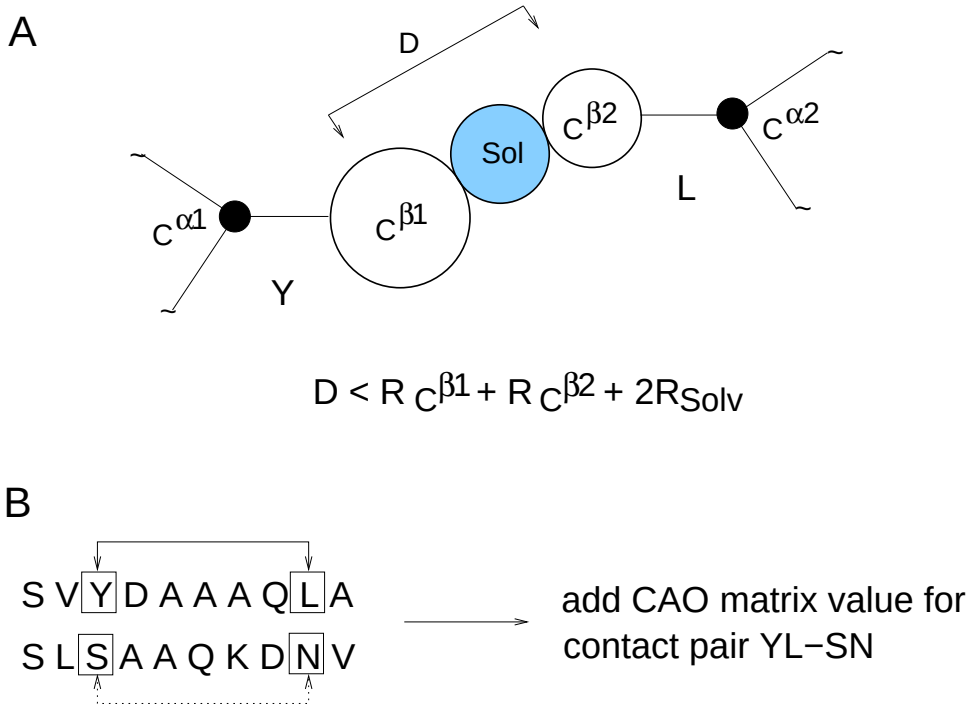

A
D
Sol
$C^{\beta 2}$
$C^{\alpha 2}$
L
$C^{\alpha 1}$
$C^{\beta 1}$
Y
$D < R_{C^{\beta 1}} + R_{C^{\beta 2}} + 2R_{Solv}$
B
S V Y D A A A Q L A
S L S A A Q K D N V
add CAO matrix value for
contact pair YL−SN

Figure 1:

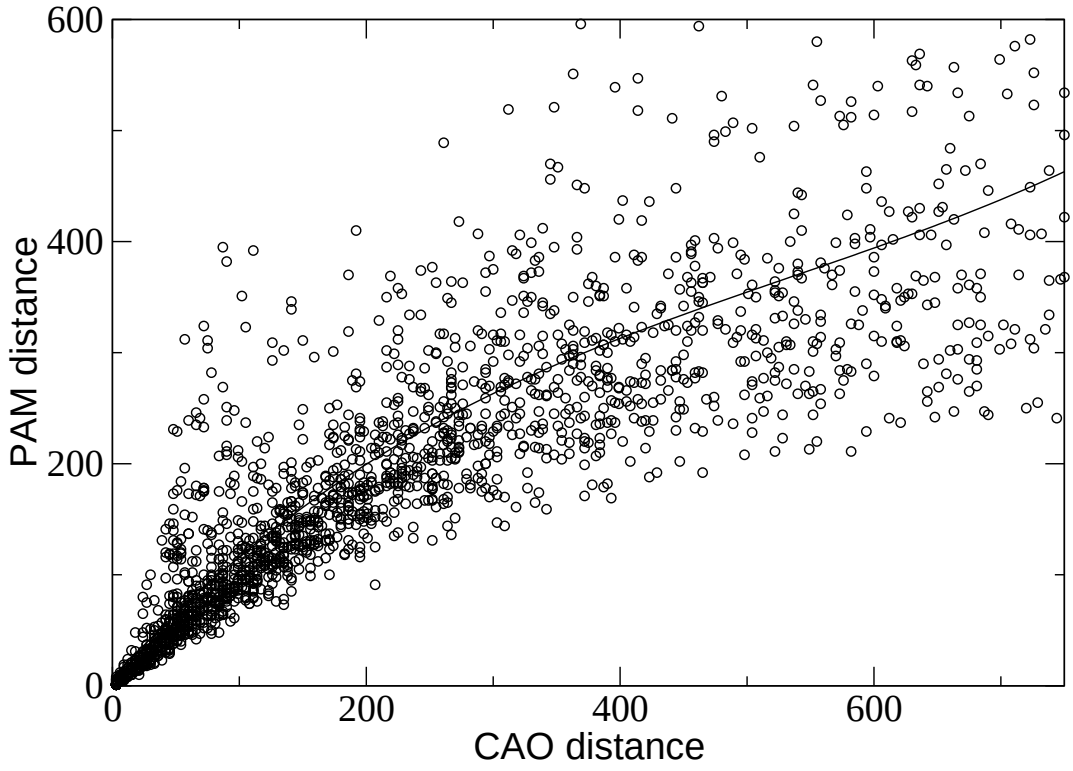


Figure 2:

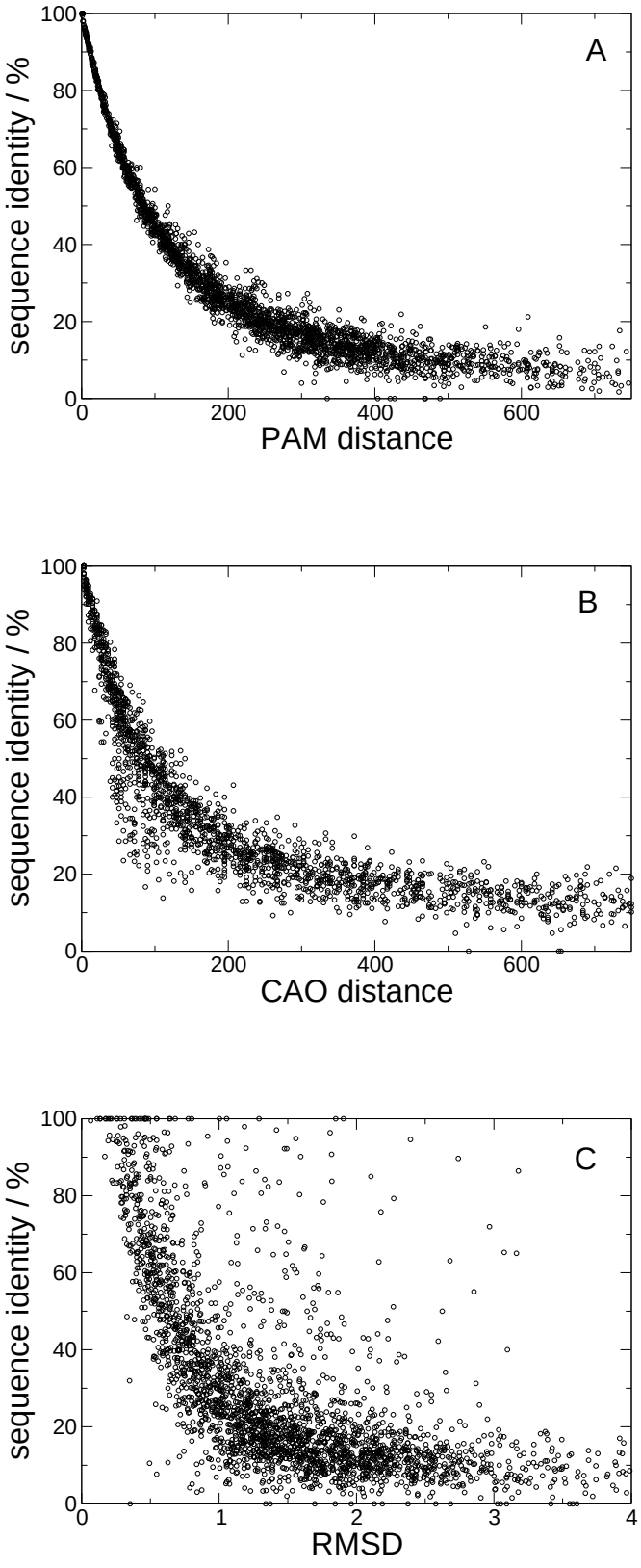

A
B
C
sequence identity / %
PAM distance
CAO distance
RMSD
0
20
40
60
80
100
200
400
600
1
2
3
4

Figure 3:

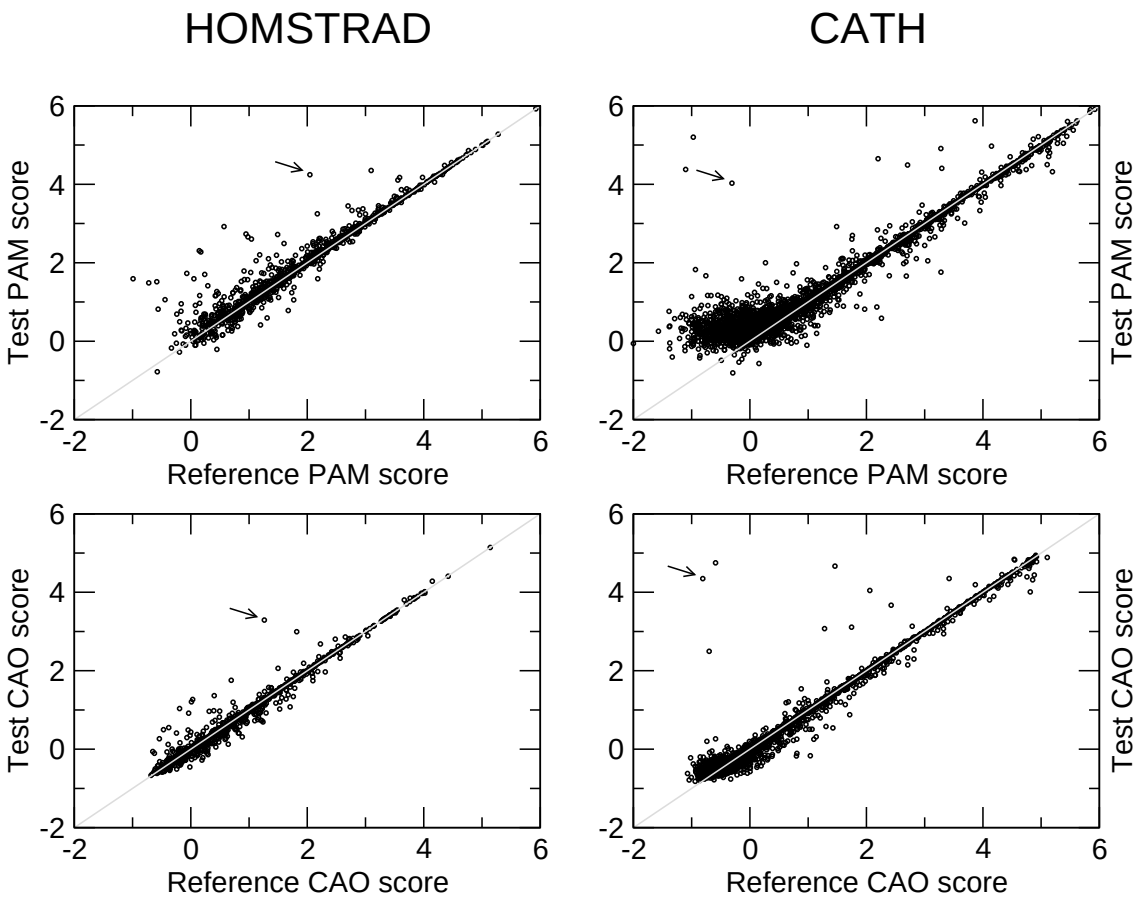

HOMSTRAD
CATH
Test PAM score
Reference PAM score
Test CAO score
Reference CAO score
-2
0
2
4
6

Figure 4:

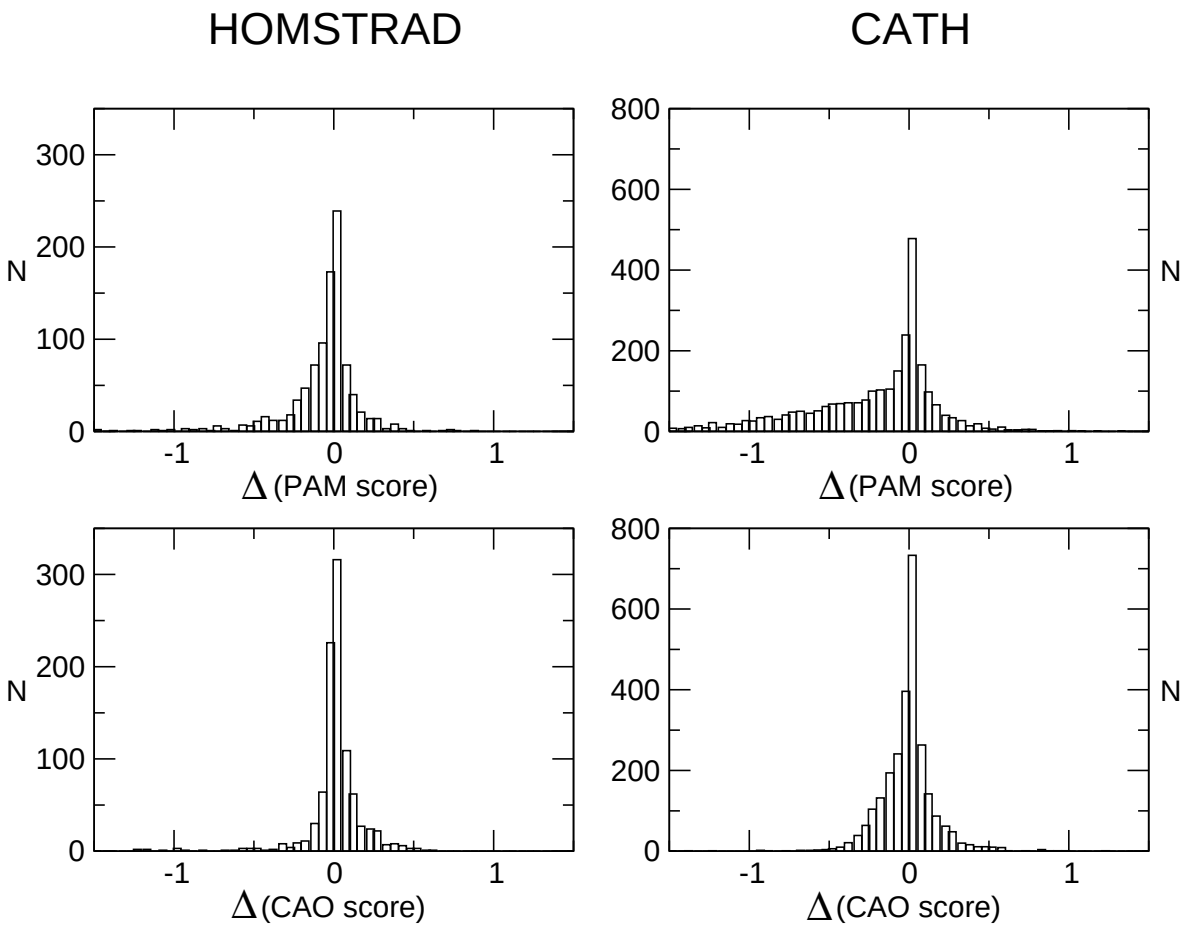

HOMSTRAD
CATH
N
300
200
100
0
800
600
400
-1
0
1
Δ(PAM score)
Δ(CAO score)

Figure 5:

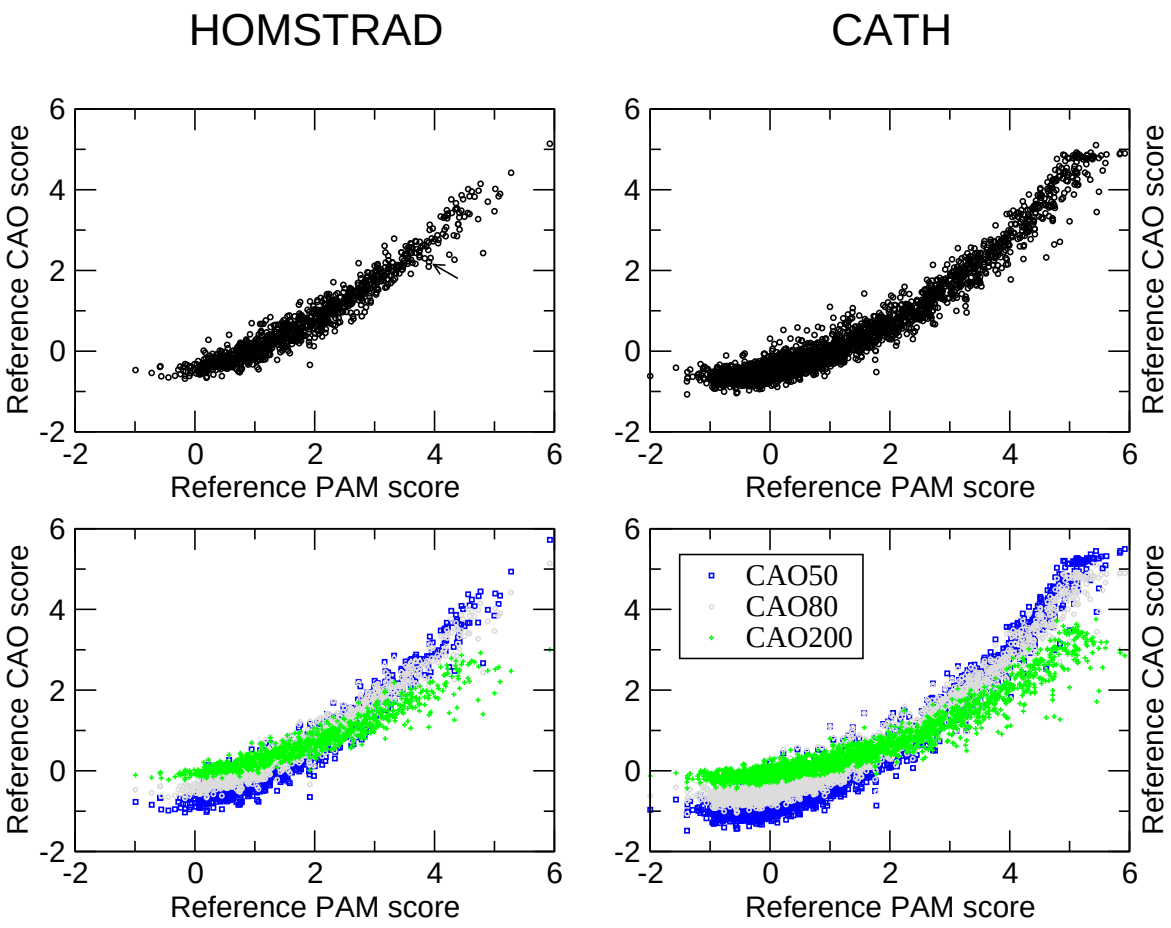
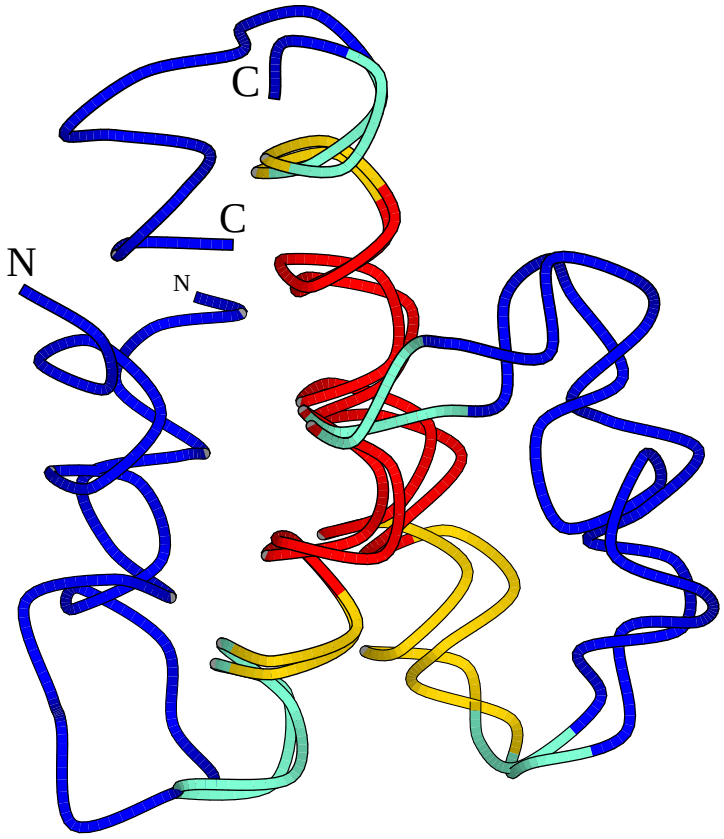

HOMSTRAD
CATH
Reference CAO score
Reference PAM score
CAO50
CAO80
CAO200

Figure 6:




```
1kjs   MLQKKIEEIA AKYKHSVVKK CCYDGACVNN DETCEQRAAR
1c5a   MLQKKIEEEA AKYKYAMLKK CCYDGAYRND DETCEERAAR

HELIX     HHHHHH H      HHH HHHH          HHHHH
CON.   ********1* ****5656** ******11*5 *****6****
RMSD    10.8  4.7   6.2  2.3   1.4  3.2   4.2  5.3
CAO      3.6  2.3   2.2  1.6   4.5  0.8   3.7  3.1


1kjs   ISLGPRCIKA FIECCVVASQ LRANISHKDM QLGR
1c5a   IKIGPKCVKA FKDCCYIANQ VRAEQ----- ----

HELIX     HHHHHH HHHHHHHHHH HH
CON.   *47**6*8** *36**38*5* 6**41

RMSD    2.6  2.0   1.9  3.0   4.3
CAO     1.3  1.7   2.0  0.3   0.2


Sequence Identity: 67.7 %
```

Figure 7: