Artificial Neural Networks and Hidden Markov Models for Predicting the Protein Structures: The Secondary Structure Prediction in Caspases

Thimmappa S. Anekonda

Computational Molecular Biology
(Biochemistry 218-BioMedical Informatics 231)

Acknowledgments

I am grateful to Dr. Doug Brutlag for his timely advice on the development of this project

Artificial Neural Networks and Hidden Markov Models for Predicting the Protein Structures: The Secondary Structure Prediction in Caspases

The protein structure prediction has been an active research area for the last 40 years or so. The technological progress in computational molecular biology during the last decade has contributed significantly to the progress we see today. The major goal of predicting protein structures underpins the correct assumption that three-dimensional structures confer protein function. The linear amino acid sequences must transform to non-linear secondary structures and then to tertiary and quaternary structures that are responsible for biological functions. Biological functions may remain similar or change in the related organisms through the evolutionary process. The theory of natural selection predicts that evolution occurs by tinkering but not by inventing (Jacob 1977). This means, deletions, mutations, or any other changes occurring in the linear ancestral protein sequence can cause changes in biological structure and function in the descendants. The problem is, structures and biological functions are highly redundant and they are conserved across the evolutionarily diverged organisms. It is not uncommon to have similar 3-D structures for proteins that show less than <30% sequence identity in the pair-wise comparisons, adding further complications to predictions (Rost 1999). Computationally, amino acid sequence is similar to a text string and that string algorithms rooted in computer science can handle structure prediction problem easily (Durbin et al. 2002). The most prominent break through in structure prediction research of the last decade can be attributed to refined architectures provided by two important machine learning methods: artificial neural networks (ANN) and hidden Markov models (HMM) (Rost 2002; Karplus et al. 1998, 1999; Krogh et al. 2001).

In this report, first I briefly review ANN- and HMM-based structure prediction methods covering literature from 1990 through 2002. Then, I present the results of applying four secondary structure prediction methods to predicting structures of three caspases that are responsible for apoptotic cell death in mammals. The Medline MESH search path for "protein conformation" along with "NN" or "HMM" is presented in **Table 1**. In this report, I will cover only the important ideas discussed in these and other relevant references.

Artificial Neural Networks (ANN)

Artificial neural networks were originally developed to model human brain function. ANNs are parameterized graphical models consisting of networks with three prime architectures: recurrent, feed-forward and layered (Baldi & Brunak 2001). The recurrent architecture is more complex and it contains directed loops. The feed-forward architecture does not contain the directed loops. The layered architecture usually contains visible input and output layers and non-visible hidden layers. Feed-forward, layered architecture is more

commonly used in computational molecular biology. Each layer may contain one or many units. The units in the input layer are connected to units in the hidden layers, which in turn are connected to units in the output layer. The connections are associated with weights. The number of units in layers is determined by the problem at hand. A good rule of thumb is that number of units in the hidden layers is equal to half the sum of the number of input and output units (Haykin 1994), but variations to this rule are very common. ANNs are powerful because they are capable of modeling extremely complex biological functions yet they are relatively easy to use, as they can learn from examples. Well-trained ANNs can predict complex biological patterns, structures, or functions of newly discovered sequences. Depending on the type of data, the structure prediction problem can be divided into two main categories: classification and regression. For example, observed secondary structure data (Helix, Strand, Coil) can provide discrete information for secondary structure classification. The continuous hydrophobicity data can be used to fit the regression equation.

ANNs used during the early days were called "black boxes," because the network architecture was neither trained properly nor evaluated on representative sets of sequences. Accumulated knowledge on network implementation rules, specific knowledge-based training networks, and feeding neural networks with evolutionary information for more than a decade (1990-2002) dramatically increased the predictive ability of secondary structures (Rost 2002). For example, today's best secondary structure predictive methods have attained >78% accuracy.

Hidden Markov Models

Hidden Markov models are special cases of neural networks, stochastic grammars and Bayesian networks (Baldi & Brunak 2001). They can convert multiple sequence alignments into position-specific scoring matrices (PSSM), taking into account all matches, mismatches, and gaps in the alignment. The PSSMs in turn can be used for searching distance homologues of the query sequence or for predicting protein structures (Eddy 1998; Karplus et al. 1998). A set of 20-100 sequences is needed to train the HMMs (Mount 2001). The most general HMM takes into account all insertions, deletions, and matches that appear in the related sequences and the associated transition probabilities to generate PSSMs. These matrices can be used for predicting secondary structures (helix, strand, coil) or for modeling 3-D structures of proteins. The most important limitations of HMMs are that they need to be trained on a larger set of sequences (say hundreds) to correctly identify distant homologues. HMMs are unable to efficiently identify long-distance correlations between the amino acid residues of a sequence (Eddy 1998). Limitations of HMMs can be overcome by using them in conjunction with ANNs in hybrid architectures.

Three main methods for predicting protein structures

Functional native protein structures are usually three-dimensional (3-D). Original 1-D protein sequence must fold into 3-D structure. Prediction of 3-D structure of a given 1-D sequence whose structure is not known is a non-trivial task. Prediction of 3-D structure is essential for designing rational drugs and proteins to meet the needs of human health care (Sew & Fischer 2001). Three most important approaches for protein structure design are: (a) homology modeling, (b) threading or fold recognition, and (c) *ab initio*.

Homology modeling is built on the principle that evolutionarily related 1-D amino acid sequences would fold into similar 3-D structures. So, if evolutionary relationship of a new sequence (target) can be established with another sequence (template) whose structure is known, a 3-D model for the target sequence can be readily built. When sequence similarity between the target and template exceeds 30% or so, the homology modeling is usually successful. The homology modeling consists of aligning the target sequence with the template proteins and copying coordinates of the matching residues of the template to target. The next step is building side chains and the regions with loops. Building side chains is generally more difficult than building loops. Finally, the model should be optimized and adjusted to minimize steric clashes between atoms.

When similarity between the target and template sequences falls below 30%, protein structures may still be similar but homology modeling would be ineffective. Under such situations, fold recognition or threading is the method of choice. In threading, a library of known folds is used to establish a compatibility function between 1-D sequence and 3-D fold from the library, taking into account the preferences of amino acids to different 3-D environments. Finally, this compatibility function will be used to thread the target sequence into an appropriate 3-D fold. Threading method would work when the fold library can provide suitable 3-D fold to the target sequence.

Frequently, the fold library may not have a suitable fold for a newly discovered sequence. When this happens, threading does not work any more. For cases when homology modeling or threading method becomes ineffective, *ab initio* method is used. *Ab initio* methods are computationally demanding. They usually search the energy surface using Monte Carlo simulations, genetic algorithms, or molecular dynamics.

Many computational methods that implement the above three structure prediction approaches usually employ ANN or HMM-based architectures. In some cases, hybrid architecture of both ANN and HMM may be used.

Blind and manual evaluation of structure prediction methods

A plethora of protein prediction methods developed by different research groups all over the world are now available on the Internet. The challenge is to identify suitable methods that are truly superior in accurately predicting the protein structures of interest. Evaluation of these methods for performance

accuracy is not a trivial task. In the recent years, such an effort has lead to a development of a whole new area of research. The evaluation effort has developed into two interrelated philosophies: Critical Assessment of Structure Prediction (CASP) initiated in 1994 and Critical Assessment of Fully Automated Structure prediction (CAFASP) initiated in 1998. Currently, several human research groups assess the accuracy of prediction methods in the CASP group and automated web servers assess the accuracy of prediction methods in CAFASP. As of October 2002, participants of CASP have met five times and CAFASP thrice. The most recent results of CAFASP3 and CASP5 are not yet published. As per previous assessments (CAFASP2 and CASP4), there was an agreement between these two groups such that they both assess the accuracy of the same prediction methods and compare their results with each other. The main goal of CAFASP2 were to assess the performance of fully automated servers for structure prediction, provide the assessment results to the community of users, allow human groups to participate in the CASP for non-automated predictions, and compare the results of CAFASP with CASP (Fischer et al 2001). According to CAFASP2 and CASP4 results, there is a considerable agreement between these two assessments in terms of rank orders awarded to the prediction methods. Over 100 CASP human groups and >36 automatic web servers participated in evaluating five main categories: fold recognition, secondary structure prediction, contacts prediction, *ab initio*, and homology modeling.

Dozens of Meta servers have been recently added to CAFASP3 and CASP5 assessments. These servers essentially evaluate many servers and extract the best prediction results. For example a Meta server named ShotGun on 5 (3DS5) is a consensus predictor uses the results of other servers such as FFAS, 3D-PSSM, SHGU, FUGE and nGenTHREADER. It compiles one model from models produced by the 5 servers using parsing of partial structures (the complete list of all servers can be found in http://bioinfo.pl/meta/servers.html). 3DS5 and other consensus servers have ranked very high in most of the structure prediction categories, essentially surpassing all individual servers. Another ANN-based consensus server, Pcons, out performs any single server by producing about 8-10% more correct predictions of folds (Lundstr_m et al. 2001). The fundamental theme of the consensus servers follows the basic assumption of the probability theory that all the relevant evidence must be used in predictions.

The secondary structure prediction in caspases

Introduction

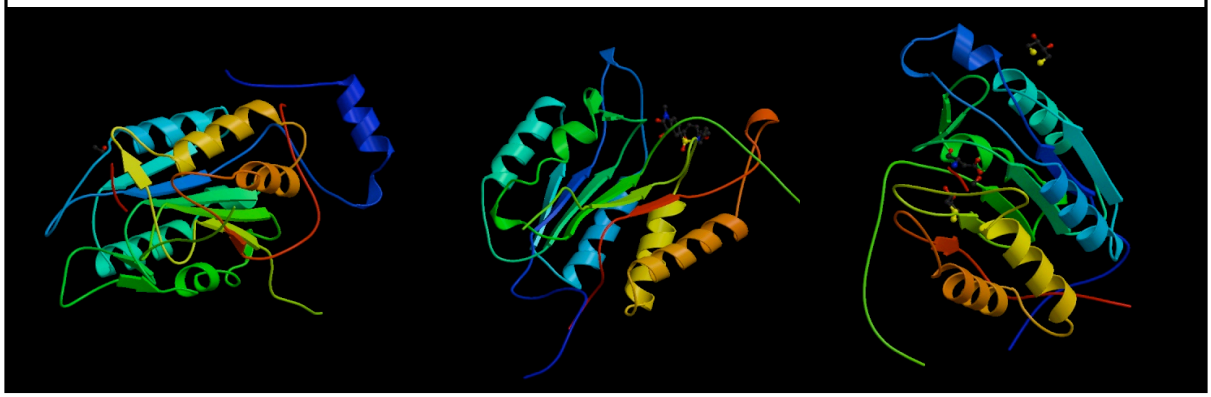
The secondary structure prediction could be useful for predicting some aspects of protein function: identifying putative structural switching regions not known before (Kirshenbaum et al. 1999), or for establishing correlations between local secondary structure and global conditions (Young et al. 1999). In addition, secondary structure prediction could be useful for identifying membrane

proteins, coiled-coil regions and domains, classifying through secondary structure content, and finding binding sites (Rost 2000 and references therein).

In this section I briefly introduce caspases, describe four secondary structure prediction methods, and apply them to predicting secondary structures in caspases.

Caspases are a family of intracellular cysteine endopeptidases. They play a key role in inflammation and mammalian apoptosis or programmed cell death. Procaspses in most cells exist in latent forms and contain three distinct regions in their structures: N-terminal pro-domain, middle large subunit, and C-terminal small sub-unit (Lee et al. 1997; Rano et al. 2000 ; Watt et al. 1999; Z_rnig et al. 2001). The maturation of caspases includes homodimerization and proteolytic processing when pro-domain is cleaved off, and the small and large subunits are cut at internal aspertases by autoproteolysis or by other caspases. The active caspases consist of two large (~20 kDa) and two small (~10 kDa) sub units. Caspases can be grouped into four subfamilies (Wolf & Green 1999): (i) cytokinin processors (caspases-1, 4, 5, 11, 12, 13, 14), (ii) apoptic initiators (caspases-2, 8, 9, 10), (iii) apoptic executioners (caspases-3, 6, 7), and (iv) invertebrate caspases. All caspases contain an active site pentapeptide sequence (**QACXG**), where X could be R, Q, or G. The cysteine within the active site is directly involved in substrate catalysis. In this study I analyze two chains (large, A and small B) each from casp-1, casp-3 and casp-8 (**Figure 1**).

Figure 1. Three-state (alpha-helix, beta-sheet, and coil) Protein Data Bank structures of casp-1 (1IBC; 2.73 Å), casp-3 (1GFW; 2.80 Å), and casp-8 (1QTN; 1.2 Å). All three structures belong to alpha and beta class (a/b), mainly with parallel beta sheets (beta-alpha-beta units), of the SCOP structural classification. Casp-3 is in complex with an isatin sulfonamide inhibitor and casp-8 is in complex with the tetrapeptide inhibitor ace-ietd-aldehyde. In all 3 proteins, the large chain A begins with blue coil and end with green coil. The small chain B begins with red coil and end with yellowish-green coil. In all 3 cases, the pentapeptide active site (QACXG) is located at the junction of the last beta-sheet near its arrow and the beginning of the last coil in chain A.



I applied four secondary structure prediction methods (PHDsec, PSIPRED, SAM-T02, PROF King) to analyze two chains (A and B) each from casp-1, casp-3 and casp-8 proteins. A brief description of these methods is presented in **Table 2**. Architectures of all four methods use ANNs. In addition to ANN, PROF King utilizes seven GOR classifiers and SAM-T02 server employs HMMs in hybrid architecture. These four methods represent a period of more than a decade of technological advances in the area of protein secondary structure prediction. The prediction accuracy was stalled at 60% level for decades prior to the introduction of PHDsec method in 1993 (Rost and Saders 1993). The PHDsec method for the first time utilized evolutionary information systematically from multiple sequence alignments and increased the prediction levels to 70%. Six years later the prediction levels jumped up to 76% thanks to Jones' (1999) PSIPRED method that uses iterated PSI-BLAST profiles as input data set instead of multiple sequence alignments. SAM-T02 method takes advantage of both ANN and HMM principles (Karplus & Hughey 1999). PROF King method can be viewed as a consensus prediction method, because it uses seven GOR-based predictions along with ANNs (Ouali & King 2000). Here I try to answer the following four questions related to the secondary structure prediction in caspases:

1) Which prediction method is most suitable for predicting secondary structures in caspases and why?
2) Does prediction accuracy for alpha helices is better or worse than for beta sheets?
3) What is the influence of conserved blocks on secondary structure prediction?
4) Is there any relationship between the size of the individual secondary structure (residue content) and the prediction accuracy?

Also, I will briefly discuss the strengths and limitations of the four prediction methods.

Materials and Methods

First, I obtained the amino acid sequence data for chains A and B of casp-1, casp-3 and casp-8 from NCBI website (http://www.ncbi.nlm.nih.gov/) (**Box 1**). The EMotif search of the Blocks+ database retrieved four conserved blocks in chain A and one in chain B (IPB001309 A-E). Secondly, I applied four prediction methods (**Table 2**) to each of the 6 sequences individually and generated their 3-state predicted secondary structures. Then, I obtained the observed sequence data for these chains from the PDB (http://www.rcsb.org/pdb/index.html). PDB gives an 8-state secondary structure data. I converted the 8-state structure data into 3-state data ([HGI]>H for alpha helix; [EB]>E for beta strand; [.ST]>C for coil, loop or other type of structure). I also designated unresolved PDB structure to be 'C'. Finally, the sequence data, 3-state observed structure data and 3-state predicted structure data were organized into three columns matching

structures to residues in each row. A total of 24 data sets (3 proteins x 2 chains x 4 methods) were created. In addition, I obtained data on the number of helices and strands, and their residue content for each of the three proteins from PDB.

Statistical Analysis

I used the SOV server (http://predictioncenter.llnl.gov/local/sov/sov.html) to analyze each of 24 data sets. The analysis gave both Q3% and SOV% values. The Q3 index (Qhelix, Qstrand, Qcoil) gives percentage of residues predicted correctly as helix, strand, or coil for all three conformational states. It is the fraction of number of residues correctly predicted from the number of all residues. Q3% measure of overall number of residues predicted correctly can be misleading. It shows how well individual residues are predicted but not how well secondary structure elements are predicted. To make evaluation of secondary structure prediction more structurally meaningful, in this study I used a segment overlap measure (SOV%) that was first proposed by Rost et al (1994) and further described and evaluated by Zemla et al (1999). I also estimated averages and standard deviations for these and other variables when appropriate.

Results and Discussion

Overall and individual secondary structure prediction accuracies for chain 'A' in casp-1, casp-3 and casp-8 from four prediction methods are presented in **Table 3**. The PSIPRED method gave the highest (84.7%) overall average Q3 value and the PROF King method gave the lowest value (77.8%), with other two methods being intermediate. PSIPRED predicted helices and strands with nearly equal Q3 accuracy, but PHDsec, SAM-T02 and PROF King predicted strands better than helices. Overall accuracies for SOV were 6-7% greater for PSIPRED than for other methods. Depending on the prediction method, SOV values for strands were 3-9% greater than the SOV values for helices. These results suggest that beta strands were predicted with slightly greater accuracy than alpha helices in 'A' chains of the three caspases.

The Q3 and SOV values for the secondary structures in chain 'B' of caspases-1, 3 and 8 are presented in **Table 4**. Overall Q3 value was highest for PROF King (73.8%) and lowest for PHDsec (68.5%). Q3 values for helices were nearly 2- 4 times greater than Q3 values for strands, while Q3 values for coils stayed intermediate between these values. Overall average SOV value was 75.6% for PSIPRED method and 58.1% for PHDsec, with values for other two methods being intermediate. SOV values for helices were 96.8% for PHDsec, PSIPRED and SAM-T02 and 87.2% for PROF King method. SOVs for helices were nearly 2 to 5 times greater than SOVs for strands. Both Q3 and SOV values clearly suggest that helices were predicted with much greater accuracy than strands in chain 'B' of the caspases.

Of the four methods, PSIPRED is slightly better than all other methods. Among the remaining three, SAM-T02 and PROF King are slightly better than PHDsec. The superior accuracies for PSIPRED may be because PSI-BLAST alignments are based on pair-wise local alignments that created reliable local alignments, and iterated profiles may have increased the sensitivity of PSI-BLAST (Jones 1999). The estimated overall Q3 and SOV values for caspases from different methods either overestimated or under estimated relative to those shown in **Table 2.** This difference is mainly due to small sample size (6 chains per method) of this study.

Overall prediction accuracy for chain 'A' is expected to be better than prediction accuracy for chain 'B', because on average, chain 'A' is 97-residues (or 2.36 times) longer than chain 'B' (**Table 5**). Overall Q3 and SOV values are respectively 9.5% and 5.8% greater for chain 'A' relative to values for chain 'B'. While strands were predicted with slightly better accuracies in chain 'A', prediction accuracies for chain 'B' were 2-3 times greater for helices than for strands.

The difference in total chain lengths alone can't explain the slightly biased prediction accuracy for strands in chain 'A' or highly biased prediction accuracy for helices in chain 'B'. The prediction bias towards beta strands in chain 'A' can be explained based on the distribution of secondary structural elements within the conserved blocks. The eMotif search identified 4 conserved blocks in 'A' chains of caspases (**Box A**). All except one strand reside within these blocks, suggesting that conserved secondary structures were better predicted than relatively less conserved structural elements in chain 'A'. Highly biased prediction accuracy for helices, however, cannot be explained either by total chain length or by the distribution of helices with the conserved block in chain 'B'. In fact only one-half of one of the two helices is present within the conserved block, yet helices predicted with superior accuracy than the strands. The number of helices and the number of strands, and their percent residue content in chains 'A' and 'B' of caspases-1, 3, and 8 are presented in **Table 6**. The beta strands are twice as many as alpha helices in all chains. The residue content of alpha helices and beta- strands of chain 'A' are about the same. The residue content of alpha helices is 2.22 times greater than residue content of beta strands in chain 'B'. Simply, alpha helices are much longer than beta strands in chain 'B'. The longer alpha helices in chain 'B' are the cause for their greater prediction accuracy by all prediction methods.

The total chain length, presence of conserved blocks, distribution of secondary structures with the blocks and the length of secondary structure itself influenced the prediction accuracy of the methods used.

Strengths and limitation of the prediction methods

PHDsec was the first method to systematically use multiple sequence alignments for training ANNs. This scientific breakthrough pushed the

prediction accuracy levels for secondary structures from 60% to 70% (Rost & Sanders 1993, 1994). Besides using new knowledge from multiple sequences, it can also handle long-distance interactions that usually increase the prediction accuracy. The major limitation of PHDsec is that multiple sequence alignment is time consuming and it is difficult to move the PHD server to another site. The alignments usually correlate to sequence diversity. The diversity of the training set influences the prediction accuracy. This method uses only one line of evidence from the alignments to train the networks.

PSIPRED was the first method to take advantage of position-specific scoring matrices generated by the powerful PSI-BLAST search algorithm for predicting protein secondary structures. The PSI-BLAST search generates greater local alignment reliability. This radical idea helped increasing the prediction accuracies from 70 to over 76% level. This method can also handle long distance interactions. The major limitation is that, like PHDsec, it uses only one line of evidence coming from the profiles for training networks. The greater accuracy of this method could be attributed to homologous tertiary structures rather than more efficient use of all available data.

SAM-T02 makes use of theories from both ANN and multi-track HMM. Although HMM can't handle long distance correlations, the ANN part of architecture helps to handle this problem. The complicated hybrid architecture of this method may mask the simple biological insights.

PROF King uses different background theories and different lines of evidence relevant to prediction (for example, 7 GOR methods and ANN). It even assumes that long-distance interactions may not be important and their importance may have been overstated in the past. PROF King strictly uses local information. The complicated non-linear rule-based statistics (too many rules) of PROF King may mask the biological meaning. Combining different architectures usually adds to complexity and may adversely affect clarity of biological conclusions.

Conclusions

The secondary structure prediction methods have now reached close to 80% accuracy level, which is as good as accuracies coming from actual structure determinations using traditional methods. Further improvement is somewhat difficult but not impossible. Technological advances in many areas are expected to enhance the prediction accuracy even further. Refining the background evolutionary knowledge used for learning and improving the learning techniques, ever increasing database size, developing organism specific networks, and striking effective balance between the multitudes of methods used and preserving biological sense in the consensus methods should all help advancing this area even further. Because consensus predictions use different background theories and different lines of evidence relevant to prediction, they are expected to be superior over individual prediction methods. The next wave of revolution for increased prediction accuracy is expected to come from superior consensus methods of today and tomorrow.

References

Baker, D. (2000). A surprising simplicity to protein folding. Nature 405, 39-42.

Baldi, P., & Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach*. Cambridge: The MIT Press.

de la Cruz, X., et al. (2002) Toward predicting protein topology: An approach to identifying _ hairpins. PNAS 99, 11157-11162.

Dosztanyi, Z., et al. (1997). Stabilization centers in proteins: Identification, characterization and predictions. J Mol. Biol. 272, 597-612.

Durbin, R., et al. (2002). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics 14, 755-763.

Fischer, D. et al. (2001). CAFASP2: The second critical assessment of fully automated structure prediction methods. PROTEINS: Structure, Function & Genetics Suppl 5, 171-183.

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. New York: Macmillan Publishing.

Jacob, F. (1977). Evolution and tinkering. Science 196: 1161-1166.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.

Karchin, R. et al. SAM_T02 Protein Structure Prediction Webserver (Abstract). (http://www.soe.ucsc.edu/research/compbio/sam.html).

Karplus, K., et al. (1998). Hidden Markov models for detecting remote protein homologies. Bioinformatics 14, 846-856.

Karplus, K., et al. (1999). Predicting protein structure using only sequence information. Proteins. Suppl 3, 121-125.

Kirshenbaum, K. et al. (1999). Predicting allosteric switches in myosins. Protein Sci. 8, 1806-1815.

Krogh, et al. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol. Biol. 305, 567-580.

Lee, D., et al. (2000). Potent and Selective Nonpeptide Inhibitors of Caspases 3 and 7 which Inhibit Apoptosis and Maintain Cell Functionality. *J.Biol.Chem*. 275, 16007.

Lundstr_m, J., et al. (2001). Pcons: A neural-network-based consensus predictor that improves fold recognition. Prot. Sci. 10, 2354-2362.

Mount, D. W. (2001). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbour, New York: Cold Spring Harbour Laboratory Press.

Ouali, M., & King, R.D. (2000). Cascaded multiple classifiers for secondary structure prediction. Prot. Sci. 9, 1162-1176

Przybylski, D & Rost, B. (2002) Alignments grow, secondary structure prediction improves. Proteins 46, 197-205.

Rano, T. A., et al. (1997). A combinatorial approach for determining protease specificities: application to interleukin-1beta converting enzyme (ICE). *Chem Biol*. 4, 149.

Rice, D. & Eisenberg, D. (1997). A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol. Biol. 267, 1026-1038.
Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232, 584-599.
Rost, B. & Sander, C. (1994). Combining evolutionary information neural networks to predict protein secondary structure. PROTEINS: Structure, Function & Genetics 19, 55-72.
Rost, B. et al. (1994). Redefining the goals of protein secondary structure prediction . J Mol. Biol. 235, 13-26.
Rost, B. (1999). Twilight zone of protein sequence alignments. Prot. Engin. 12, 85-94.
Rost, B. (2001). Review: Protein secondary structure prediction continues to riase. Structural Bioinformatics. In press.
Rost, B. (2002). Neural networks predict protein structure: hype or hit? In 'Artificial intelligence and heuristic models for bioinformatics', Paolo Frasconi (ed), ISO Press, in press.
Salvesen, G.S. (2002). Caspases: opening the boxes and interpreting tha arrows. Cell Death & Differentiation 9, 3-5.
Sew, N. & Fischer, N. (2001). Convergent evolution of protein structure prediction and computer chess tournaments: CASP, Kasparov, and CAFASP. IBM Systems J. 40, 410-425.
Watt, W., et al. (1999). The Atomic-Resolution Structure of Human Caspase-8, a Key Activator of Apoptosis *Structure (London)* 7 *pp.* 1135.
Wolf, B.B. & Green, D.R. (1999). Suicidal tendencies: Apoptotic cell death by caspase family proteinases. J Biol. Chem. 274, 20049-20052.
Young, M. et al. (1999). Predicting conformational switches in proteins. Protein Sci. 8, 1752-1764.
Zemla, A., et al. (1999). A modified definition of SOV, a segement based measure for protein secondary structure prediction assessment. PROTEINS: Structure, Function & Genetics 34, 220-223.
Z_rnig, M., et al. (2001). Apoptosis regulators and their role in tumorigenesis. Biochim. Biophys. Acta 1551, F1-F37.

Table 1. Search history for the PubMed MeSH terms.

#6	Search **#2 AND #4** Limits: **Publication Date from 1990 to 2002, English**	8
#5	Search **#2 AND #3** Limits: **Publication Date from 1990 to 2002, English**	42
#4	Search **Hidden Markov Model** Limits: **Publication Date from 1990 to 2002, English**	242
#3	Search **Neural Network** Limits: **Publication Date from 1990 to 2002, English**	5913
#2	Search **"protein conformation" [MESH] AND predict** Limits: **Publication Date from 1990 to 2002, English**	1031
#1	Search **"protein conformation" [MESH]** Field: **All Fields,** Limits: **Publication Date from 1990 to 2002, English**	89039

Table 2. Description of the secondary structure prediction methods used in this study.

Attribute	**Secondary Structure Prediction Methods**			
	PHDsec[1]	**PSIPRED**[2]	**SAM-T02**[3]	**PROF**[4] **King**
Architecture	3-layered, feed-forward ANN	2-layered, feed-forward ANN	ANN and multi-track HMM	ANN and linear discrimination classifiers
Input data	Multiple sequence alignments in place of earlier single sequences	Iterated PSI-BLAST profiles rather than sequences	Multiple alignments of sequences	From a cascade of multiple classifiers (GORs) and PSI-BLAST
Non-homologous test data set	130 protein sequences	187 unique protein folds	??	496 protein sequences
Evolutionary information?	Yes	Yes	Yes	Yes (Also tried without)
Overall accuracy (%)	70	76	??	77
Literature	Rost & Sander (1993)	Jones (1999)	Karplus et al. (1999)	Ouali & King (2000)

Websites for the prediction methods:
[1] http://cubic.bioc.columbia.edu/predictprotein
[2] http://bioinf.cs.ucl.ac.uk/psiform.html
[3] http://www.cse.ucsc.edu/research/compbio/HMM-apps/T02-query.html
[4] http://www.aber.ac.uk/~phiwww/prof/

Table 3. Prediction accuracies (Q3%, SOV%) of the secondary structures in chain 'A' of Casp-1, Casp-3, Casp-8 from four prediction methods (PHDsec, PSIPRED, SAM-T02, PROF King).

Method	Protein	N, aa	Q3%				SOV%			
			All	Helix	Strand	Coil	All	Helix	Strand	Coil
PHDsec	Casp-1	194	82.0	69.1	81.8	88.7	76.6	81.8	87.9	71.6
	Casp-3	147	79.6	73.7	72.2	86.3	73.3	82.6	78.9	67.0
	Casp-8	164	79.9	66.0	82.9	86.6	64.5	74.5	94.3	52.7
		AVG	**80.5**	**69.6**	**79.0**	**87.2**	**71.5**	**79.6**	**87.0**	**63.8**
		±STD	**1.3**	**3.9**	**5.9**	**1.3**	**6.3**	**4.5**	**7.7**	**9.9**
PSIPRED	Casp-1	194	84.5	78.2	78.8	89.6	82.3	84.2	85.6	80.5
	Casp-3	147	83.0	81.6	72.2	89.0	74.2	80.3	80.6	68.8
	Casp-8	164	86.6	83.0	82.9	90.2	81.8	87.2	94.3	75.3
		AVG	**84.7**	**80.9**	**78.0**	**89.6**	**79.4**	**83.9**	**86.8**	**74.9**
		±STD	**1.8**	**2.5**	**5.4**	**0.6**	**4.5**	**3.5**	**6.9**	**5.9**
SAM-T02	Casp-1	194	84.0	81.8	81.8	85.8	80.9	82.0	86.1	78.9
	Casp-3	147	76.9	78.9	80.6	74.0	65.9	81.3	87.2	52.5
	Casp-8	164	79.3	78.7	88.6	75.6	72.1	76.5	92.6	62.9
		AVG	**80.1**	**79.8**	**83.7**	**78.5**	**73.0**	**79.9**	**88.6**	**64.8**
		±STD	**3.6**	**1.7**	**4.3**	**6.4**	**7.5**	**3.0**	**3.5**	**13.3**
PROF	Casp-1	194	74.2	54.5	78.8	83.0	63.7	63.6	92.3	57.6
	Casp-3	147	78.2	76.3	63.9	86.3	74.2	90.5	72.2	67.9
	Casp-8	164	81.1	70.2	74.3	90.2	77.1	78.7	85.7	73.4
		AVG	**77.8**	**67.0**	**72.3**	**86.5**	**71.7**	**77.6**	**83.4**	**66.3**
		±STD	**3.5**	**11.2**	**7.6**	**3.6**	**7.1**	**13.5**	**10.2**	**8.0**

Table 4. Prediction accuracies (Q3%, SOV%) of the secondary structures in the chain 'B' of casp-1, casp-3, and casp-8 from four prediction methods (PHDsec, PSIPRED, SAM-T02, PROF King).

Method	Protein	N, aa	Q3%				SOV%			
			All	Helix	Strand	Coil	All	Helix	Strand	Coil
PHDsec	Casp-1	88	67.0	100.0	30.0	59.3	52.2	100.0	17.1	43.2
	Casp-3	97	68.0	90.3	14.3	69.2	62.3	90.3	17.9	58.2
	Casp-8	95	70.5	96.3	35.7	66.7	59.9	100.0	26.2	50.5
		AVG	**68.5**	**95.5**	**26.7**	**65.1**	**58.1**	**96.8**	**20.4**	**50.6**
		±STD	**1.8**	**4.9**	**11.1**	**5.1**	**5.3**	**5.6**	**5.0**	**7.5**
PSIPRED	Casp-1	88	69.0	91.7	33.3	64.8	77.2	100.0	23.0	76.1
	Casp-3	97	78.4	90.3	64.3	75.0	74.1	90.3	54.8	69.7
	Casp-8	95	69.5	96.3	28.6	66.7	75.5	100.0	14.7	79.0
		AVG	**72.3**	**92.8**	**42.1**	**68.8**	**75.6**	**96.8**	**30.8**	**74.9**
		±STD	**5.3**	**3.1**	**19.4**	**5.4**	**1.6**	**5.6**	**21.2**	**4.8**
SAM-T02	Casp-1	88	69.3	100.0	70.0	55.6	66.8	100.0	37.2	59.1
	Casp-3	97	70.1	83.9	42.9	69.2	69.7	90.3	41.4	65.6
	Casp-8	95	72.6	100.0	35.7	68.5	73.6	100.0	34.3	70.5
		AVG	**70.7**	**94.6**	**49.5**	**64.4**	**70.0**	**96.8**	**37.6**	**65.1**
		±STD	**1.7**	**9.3**	**18.1**	**7.7**	**3.4**	**5.6**	**3.6**	**5.7**
PROF	Casp-1	88	69.3	83.3	50.0	66.7	71.2	71.4	34.6	77.8
	Casp-3	97	78.4	80.6	42.9	86.5	65.5	90.3	42.9	59.0
	Casp-8	95	73.7	85.2	35.7	77.8	68.9	100.0	31.6	64.0
		AVG	**73.8**	**83.0**	**42.9**	**77.0**	**68.5**	**87.2**	**36.4**	**66.9**
		±STD	**4.6**	**2.3**	**7.2**	**9.9**	**2.9**	**14.5**	**5.9**	**9.7**

Table 5. Estimated overall averages and standard deviations of accuracy measures across all 4 methods for each chain of caspases.

Caspase	N, aa	Average Q3% (±STD)				Average SOV% (±STD)			
		All	Helix	Strand	Coil	All	Helix	Strand	Coil
Chain A	168.3 ±23.8	80.8 ±3.5	74.3 ±8.3	78.2 ±6.6	85.4 ±5.4	73.9 ±6.5	80.3 ±6.8	86.5 ±6.7	67.4 ±9.4
Chain B	71.3 ±3.8	71.3 ±3.8	91.5 ±7.1	40.3 ±15.4	68.8 ±8.1	68.1 ±7.3	94.4 ±8.6	31.3 ±12.0	64.4 ±11.0

Table 6. The number of helices and the number of strands, and their percent residue content in chains 'A' and 'B' of caspases-1, 3, and 8.

Caspase	Chain	Helix		Strand	
		Number	Content (%)	Number	Content (%)
Casp-1	A	4	23.20	6	16.40
Casp-3	A	3	21.77	7	23.13
Casp-8	A	4	23.17	7	21.34
Casp-1	B	2	27.27	4	11.36
Casp-3	B	2	28.87	4	14.43
Casp-8	B	2	28.42	4	12.63

Box 1. Amino acid sequences of the two chains (A & B) of caspases-1, 3, & 8. Their observed secondary structures are shown in red. The EMotif search of the Blocks+ retrieved four conserved blocks in chain A and one in chain B (IPB001309 A-E), which are shown in blue. These five conserved blocks are underlined in red in the observed secondary structure. The pentapeptide catalytic site in chain 'A' is shown in highlighted and underlined **blue**.

Casp-1 (Chain A)

SQGVLSSFPAPQAVQDNPAMPTSSGSEGNVKLCSLEEAQRIWKQKSAEIYPIMDKSSRTRLALIICNEEFD
SIPRRTGAEVDITGMTMLLQNLGYSVDVKKNLTASDMTTELEAFAHRPEHKTSDSTFLVFMSHGIREGICG
KKHSEQVPDILQLNAIFNMLNTKNCPSLKDKPKVIII**QACRG**DSPGVVWFKD
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHCHHHCCCCCCCCCCCCEEEEEECCCCC
CCCCCCCHHHHHHHHHHHHHHHCCEEEEECCCHHHHHHHHHHHHHCCHHHHCCEEEEEECCECCCEEEEC
CCCCCCCCCEEEHHHHHHHCCCCCCHHHCCCEEEEEECCCCCCCCCCCCCC

Casp-1 (Chain B)

AIKKAHIEKDFIAFCSSTPDNVSWRHPTMGSVFIGRLIEHMQEYACSCDVEEIFRKVRFSFEQPAGRAQMP
TTERVTLTRCFYLFPGH
CCCCCCCCCCCCEEECCCCCCCCCCEECCCEEHHHHHHHHHHHHHCCCCCHHHHHHHHHHCCCCCCCCCCCC
EEECCCCCCCCCCCCCC

Casp-3 (Chain A)

SGISLDNSYKMDYPEMGLCIIINNKNFHKSTGMTSRSGTDVDAANLRETFRNLKYEVRNKNDLTREEIVEL
MRDVSKEDHSKRSSFVCVLLSHGEEGIIFGTNGPVDLKKITNFFRGDRCRSLTGKPKLFII**QACRG**TELDC
GIETD
CCCCCCCCCCCCCCCECCEEEEEECCCCCHHHCCCCCCCHHHHHHHHHHHHHHCCCEEEEEECCCHHHHHHH
HHHHHHCCCCCECEEEEEEECCEECCEEECCCCEEEHHHHHCCCCCCCCHHHCCCEEEEEECCCCCCCCCC
CCCCC

Casp-3 (Chain B)

DMACHKIPVDADFLYAYSTAPGYYSWRNSKDGSWFIQSLCAMLKQYADKLEFMHILTRVNRKVATEFESFS
FDATFHAKKQIPCIVSMLTKELYFYH
CCCCCCCCCCCCEEEEECCCCCCCCCEECCCEEHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHCCCCC
CCHHHCCCCCCCEEEEECCCCCCCCC

Casp-8 (Chain A)

SPREQDSESQTLDKVYQMKSKPRGYCLIINNHNFAKAREKVPKLHSIRDRNGTHLDAGALTTTFEELHFEI
KPHDDCTVEQIYEILKIYQLMDHSNMDCFICCILSHGDKGIIYGTDGQEAPIYELTSQFTGLKCPSLAGKP
KVFFI**QACQG**DNYQKGIPVETD
CCCCCCCCCCCCCCCCCCCCCCCCEEEEEEECCCCHHHHHHCHHHCCCCCCCHHHHHHHHHHHHHHCCCEE
EEEECCCHHHHHHHHHHHHHCCCCCCSEEEEEEECCEECCEEECCCCCEEEHHHHHHHHCCCCCHHHCCCE
EEEEEECCCCCCCCCCCCCCC

Casp-8 (Chain B)

LSSPQTRYIPDEADFLLGMATVNNCVSYRNPAEGTWYIQSLCQSLRERCPRGDDILTILTEVNYEVSNKDD
KKNMGKQMPQPTFTLRKKLVFPSD
CCCCCCCCCCCCCCCCEEEECCCCCCCCCCEECCCEEHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHHCCE
CCCCECCCCEEEECCCCCCCCCCC