LA-UR- 98-2180

Approved for public release;
distribution is unlimited.

Title: Covariation of Mutations: A
Computational Approach for
Determination of Function and
Structure from Sequence

Author(s): Alan S. Lapedes, T-DO

(see attached sheet for additional
authors)

Submitted to: DOE OFFICE OF SCIENTIFIC AND TECHNICAL
INFORMATION (OSTI)

2AT MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. The Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Covariation of Mutations: A Computational Approach for Determination of Function and Structure from Sequence

David Haussler* (Professor, Computer Science, UCSC)

Alan S. Lapedes* (Staff member, Theoretical Division, LANL)

Richard Hughey (Assistant Professor, Computer Engineering, UCSC)

Kevin Karplus (Associate Professor, Computer Engineering, UCSC)

Melissa Cline (Graduate Student, Computer Science, UCSC)

Leslie Grate (Graduate Student, Computer Engineering, UCSC)

David Kulp (Graduate Student, Computer Engineering, UCSC)

Kimmen Sjolander (Ph.D. Alumna, Computer Science, UCSC)

Covariation of Mutations: A Computational Approach for Determination of Function and Structure from Sequence

David Haussler* (Professor, Computer Science, UCSC)
Alan S. Lapedes* (Staff member, Theoretical Division, LANL)
Richard Hughey (Assistant Professor, Computer Engineering, UCSC)
Kevin Karplus (Associate Professor, Computer Engineering, UCSC)
Melissa Cline (Graduate Student, Computer Science, UCSC)
Leslie Grate (Graduate Student, Computer Engineering, UCSC)
David Kulp (Graduate Student, Computer Engineering, UCSC)
Kimmen Sjolander (Ph.D. Alumna, Computer Science, UCSC)

May 1, 1998

Project #95688

Abstract

Classical fold recognition methods are based on the premise of distinct pairwise preferences between two given amino acids. We have studied these preferences extensively and found that in the general case, this information is limited. Yet by modeling pairwise interactions in context of phylogenetic relationships and by modeling one specific type of contact, the contact between interacting beta strand residues, we have recovered significant information for prediction and analysis of protein structure.

1 Background and Research Objectives

At this time, when the structure of a protein is determined, its structure is more likely to belong to some existing fold than to represent a new fold [Mur96]. With this insight, and a number of good fold taxonomies available [HMBC97, PMA96, HS93], fold recognition is one of the more promising sub-problems in protein structure prediction.

We have developed an enhanced a set of tools for fold recognition with hidden Markov models (HMMs), and used these tools effectively in the CASP2 protein structure prediction contest [KKB+97]. HMMs have limitations, and one limitation is that they do not model the long-range pairwise interactions that define the shape of a protein. As such, we are working on modeling pairwise interactions to incorporate them into our HMM-based framework.

Amino acid potentials, the prediction of when two amino acids will be in contact, has a wide body of users as described in recent reviews [Sip95, WR93]. Yet we found the information in these potentials to be limited. As such, we have turned our attention to variants of general potentials: beta strand contacts, and contacts in the context of phylogenetic relationships. Here, we have shown, is significant and useful information.

2 Importance to LANL's Science and Technology Base and National R & D Needs

Adequate pairwise and higher order interaction models for score functions would supply a key component in the protein fold recognition problem. The need for better score functions has been identified as the single most important obstacle to the improvement of inverse protein folding. Without adequate models and score functions, it is not possible to identify the fold of a newly discovered protein sequence in any

reliable way. Currently, the amount of data in the protein databases is growing at a rapid rate. Ultimately these endeavors will have a profound impact on basic science and health care. But this cannot occur until the existing predictive methods are refined.

3 Scientific Approach and Accomplishments

3.1 HMMs

In the last few years, researchers have also begun to make use of hidden Markov models (HMMs) to search for both protein motifs, and larger remote homologies between proteins [KBM94, BCHM94, SWS93, Edd95, EMD95, AHO93] (see recent reviews in [Edd96, Tau96]). HMMs have been used in a variety of fields to do discrete time series analysis, most notably in speech recognition. A general introduction to them can be found in [Rab89]. In biosequence analysis they can be used not only in protein analysis, but also in genefinding [KMH94, JH96] and other areas [LG87, Chu89]. When used as statistical models of protein motifs or domains, they combine the best aspects of weight matrices and Smith-Waterman methods.

The structure of the HMMs that are most commonly used in protein modeling is similar to that of a weight matrix or *profile* [GME87, GLE90, BLE91, HH91], except that it has specific states and transitions at each position to model insertions and deletions, and the probability parameters for these are different in each position. One advantage of an HMM is that it defines a formal statistical model for sequences in the given protein family, so one can calculate the likelihood of a sequence and find the most probable locations for the insertions and deletions, i.e. the most probable alignment of the sequence to the "consensus model" for the family. The likelihood is calculated by the *forward* algorithm and the most probable alignment by the *Viterbi algorithm*. Each is a dynamic programming method similar to the Smith-Waterman method used to align two sequences. The forward algorithm can be used to search a database for homologs of the protein family represented by the HMM, and the Viterbi algorithm can be used to create a multiple alignment of all family members. The parameters of the HMM can be estimated from a set of unaligned family members using an expectation-maximization method known as the *forward-backward* algorithm [Rab89], which is related to the Gibbs Sampling methods used in [LR90, CS92, LAB93].

HMMs for protein families have found several successful applications. They are now used in the analysis of nematode and human DNA sequencing efforts at the genome center at Washington University and at the Sanger Centre [Edd96]. They led to the discovery of fibronectin type III domains in yeast [Bat96] and members of the immunoglobulin superfamily in bacteria [BEC96], were used in the analysis of lectins [B.96], and other discoveries [Bat97, NJD96, DMHM97]. Finally, our group at UCSC collaborated with Lisa Holm and Chris Sander at EBI to test HMM methods in the CASP2 experiment in protein structure prediction [Pen96]. A library of over 1500 HMMs for protein families with representatives of known structure was built specifically by our group for this experiment. In all, 947 predictions were made by 76 international prediction teams on 42 different target sequences. After the solved structures were made available, the predictions were evaluated and presented at a well-attended meeting in Asilomar in December, 1996. While it was clear from the meeting that the problem of predicting protein structure from sequence is still far from solved, HMMs were shown to perform quite well in comparison to other more sophisticated threading methods in the fold recognition and alignment category of the contest. The predictions of our group were ranked 4th overall according to alignment accuracy, a key quality measure employed by the prediction assessor, Michael Levitt.

Yet HMMs do have their limitations, notably that they do not model long-range interactions. In effort to incorporate this long-range contact information into our HMM framework, we have studied contact potentials.

3.2 Analysis of general contact information

Given a vector of environmental features, for each amino acid, using neural nets we estimate the probability that that amino acid will occur in that environment. Table 1 summarizes the contents of the environmental feature vector $\vec{e}$ for each feature set studied.

To represent pairs (a_1, a_2) of interacting amino acids, first we estimated $\hat{P}(a_1|\vec{e})$, the (estimated) probability of the first amino acid given the environmental vector $\vec{e}$. Then we estimated $\hat{P}(a_2|a_1, \vec{e})$, the probability of the second amino acid given the first and given the environmental vector. The product of these is $\hat{P}(a_1, a_2|\vec{e})$, the probability of the interacting pair of amino acids (a_1, a_2) given the environmental

Dataset	Features
Bryant-Lawrence [BL93]	Peptide Midpoint Distance
Crippen [MC94]	O - N distance, N - O distance, C - N distance, N - C distance, O - Cbeta distance, Cbeta - O distance, N - Cbeta distance, Cbeta - N distance, Cbeta - Cbeta distance
Jernigan [JC90]	Side-chain Midpoint Distance
Sippl [Sip90]	Alpha Carbon Distance
True-mrf [WMS94]	Beta Carbon Distance, Omega angle Amino Acid 1 Phi Angle, Amino Acid 2 Phi Angle, Solvent Exposure at Amino Acid 1, Solvent Exposure at Amino Acid 2, Secondary Structure at Amino Acid 1, Secondary Structure at Amino Acid 2
Tetrahedron [CS97]	Window through which Amino Acid 1 looks at Amino Acid 2, Window through which Amino Acid 2 looks at Amino Acid 1, Secondary structure at Amino Acid 1, Secondary structure at Amino Acid 2, Fraction of Volume Accessible via the first window specified, Fraction of Volume Accessible via the second window specified, Solvent Exposure at Amino Acid 1, Solvent Exposure at Amino Acid 2, Beta Carbon Distance

Table 1: Contents of the feature sets. The windows of the Tetrahedron feature set are based on a tetrahedron centered around each residue with one vertex at each of the non-carbon backbone atoms and one vertex situated near the beta carbon [CS97].

vector $\vec{e}$. Table 2 summarizes these results for each feature set. The results are reported in terms of log likelihood ratios $\log_2(\frac{\hat{P}(aa|\vec{x})}{P(aa)})$ where $\hat{P}(aa|\vec{x})$ represents the probability of amino acid aa given input features $\vec{x}$ as estimated by the neural net, and $P(aa)$ represents the background frequency of amino acid aa . The log likelihood ratio of these quantities reflects the amount of information on amino acid aa available in feature vector $\vec{x}$. Because we are using log base 2, we refer to this amount of information as *bits*. Table 2 reports these results for the training set, independent test set, and independent verify set. Given further environmental features, the predictability of amino acids improves, but the amount of information gained is still less than one half of one bit — even for the best feature set.

For each feature set, Table 3 lists $I(a_1, a_2)$, the Shannon mutual information between the contacting residues, expressed in units of bits of information. The other column, $\log_2(\frac{\hat{P}(a_2|a_1, \vec{e})}{\hat{P}(a_2)}}$, represents an approximation of how much information the first amino acid provides on the second, as estimated by the neural net. This quantity is expected to be less than $I(a_1, a_2)$, which represents the amount of

Feature Set	$\log_2(\frac{\hat{P}(a_1 \vec{e})}{\hat{P}(a_1)})$			$\log_2(\frac{\hat{P}(a_2 a_1, \vec{e})}{\hat{P}(a_2)}}$		
	Training Set	Test Set	Verify Set	Training Set	Test Set	Verify Set
Bryant-Lawrence [BL93]	0.0110	0.0062	0.0049	0.0303	0.0071	0.0004
Crippen [MC94]	0.1347	0.1323	0.1212	0.1738	0.1560	0.1526
Jernigan [JC90]	0.0205	0.0186	0.0052	0.1055	0.0483	0.0488
Sippl [Sip90]	0.0003	-0.0014	-0.0010	0.0059	0.0013	0.0017
Tetrahedron [CS97]	0.3227	0.2464	0.2190	0.3565	0.2483	0.2417
True-mrf [WMS94]	0.2647	0.2241	0.2127	0.2917	0.2408	0.2264

Table 2: Results on the Feature Sets. The quantities shown were computed for all examples in the training, test, and verify sets; their average values are shown here.

Feature Set	$I(a_1, a_2)$	$\log_2(\frac{P(a_2 a_1, \mathcal{E})}{P(a_2 \mathcal{E})})$
Bryant-Lawrence [BL93]	0.0239	0.0130
Crippen [MC94]	0.0622	0.0193
Jernigan [JC90]	0.0876	0.0277
Sippl [Sip90]	0.0131	0.0013
True-mrf [WMS94]	0.0394	0.0250
Tetrahedron [CS97]	0.0805	0.0010

Table 3: Mutual Information and Log Likelihood Ratios on the Contact Pairs

information that might be learned by a perfect neural net. In short, our results showed that knowing the identity of either residue yields less than one tenth of one bit of information on the other residue. This suggests that there will be severe limitations to protein threading models based on modeling pairwise interactions alone.

To further investigate general contact information, we addressed the question of how much of the apparent contact preferences are actually side-effects of hydrophobic forces. In collaboration with Rick Lathrop [Lat97], we investigated the effect of solvent exposure on the pairwise mutual information. Hydrophobic forces are considered the most dominant force in protein folding leaving researchers to wonder how much of the apparent pairwise preferences are merely side-effects of hydrophobic forces.

We studied this effect by partitioning the contact pairs into subsets based on their solvent exposure and a threshold value, computing the mutual information of each subset, and computing a weighted average to provide an overall result. Since mutual information exhibits small sample size effects [WW95, WW93], these results are compared to the expected mutual information given the sample size. This expected mutual information was calculated based on the background frequencies of the interacting residues, and was not calculated in the context of solvent exposure. If pairwise contact preferences are largely a side-effect of hydrophobic forces, we would expect to see the observed mutual information drop significantly lower than its expected value.

In Figure 1, we see that this drop does occur, but a more interesting effect is seen in Figure 2, where we investigate the nature of this drop. In Figure 2, the observed mutual information is shown separately for the buried pairs and for the exposed pairs. In the case of exposed pairs, the observed mutual information tracks the expected value fairly closely. However in the buried pairs, the mutual information drops to about half of its expected value. This indicates that the limitations in pairwise mutual information are most pronounced in the hydrophobic core of the protein, the region most critical to structure prediction.

To address the apparent limitations of pairwise interaction models, we have explored two models that incorporate higher order interactions involving more than just two amino acids. One is a general model in which the neighborhood of an amino acid is modeled with abstract geometry such as a sphere or tetrahedron [GFL95] [CS97], and all amino acids within the neighborhood are taken into account in the calculations. The contents of these feature sets are described in Table 4. We measured the mutual information gain over background, as described above. For comparison, we performed the same measures of the singleton potential models of other major groups [BL93] [Sip90]; the results are summarized in Table 5.

3.3 Modeling of beta strand contact information

From a biologist's perspective, knowing the identity of the residues in two beta strands is very informative on whether or not these strands will be in contact. Such factors as the charge and hydrophobicity of the side chains affect how stable the strands will be in proximity to each other. This suggests that the residues bonding the potential contact pair might provide enough context to permit estimating the likelihood of the pair. Furthermore, a beta strand contact is usually comprised of more than one pair, so the contact information accumulates along the beta strand. Contrast this to random coil, where two residues might be close enough to interact, but their sequence neighbors might not be close enough to have any effect on each other. Indeed, in earlier work we found significant mutual information in the hydrogen bond residues of the alpha helices and in the paired beta strands of a beta sheet [Lap94]. This result suggests a strong tertiary effect that could be playing a key role in protein folding. Tim Hubbard has been successful in this area with simple maximum-likelihood predictors [Hub94]; his work further inspired us to address the problem with more involved predictors.

Figure 1: Pairwise mutual information conditioned on exposure

Figure 2: Pairwise mutual information evaluated with an exposure threshold

Dataset	Features
Bryant-Lawrence [BL93]	Number peptide midpoints within 5 angstroms, Number peptide midpoints between 5 and 6 angstroms, Number peptide midpoints between 6 and 7 angstroms, Number peptide midpoints between 7 and 8 angstroms, Number peptide midpoints between 8 and 9 angstroms, Number peptide midpoints greater than 9 angstroms
Sippl [Sip90]	Number alpha carbons within 5 angstroms, Number alpha carbons between 5 and 6 angstroms away, Number alpha carbons between 6 and 7 angstroms away, Number alpha carbons between 7 and 8 angstroms away, Number alpha carbons between 8 and 9 angstroms away, Number alpha carbons between than 9 angstroms away
Neighborhood Spheres [GFL95]	Identity of all residues in neighborhoods 1 through 8 Secondary structure Solvent accessibility Neighborhood and identity of both backbone neighbors
Tetrahedron [CS97]	Number of backbones visible through window 2 Number of backbones visible through window 3 Number of backbones visible through window 4 Number of beta carbons visible through window 2 Number of beta carbons visible through window 3 Number of beta carbons visible through window 4 Fraction of volume accessible via window 2 Fraction of volume accessible via window 3 Fraction of volume accessible via window 4

Table 4: Contents of the feature sets. The neighborhoods of the Neighborhood Spheres feature set are defined by centering two spheres at the beta carbon, with diameters of 5 angstroms and 6 angstroms respectively, and dividing the region defined by each of the spheres into quadrants [GFL95]. The windows of the Tetrahedron feature set are based on a tetrahedron centered around each residue with one vertex at each of the non-carbon backbone atoms and one vertex situated near the beta carbon [CS97].

Feature Set	$\log_2\left(\frac{P(a_1 \bar{c})}{P(a_1)}\right)$		
	Training Set	Test Set	Verify Set
Bryant-Lawrence [BL93]	0.2266	0.1743	0.1532
Neighborhood Spheres [GFL95]	0.3493	0.0961	0.0905
Sippl Singleton [Sip90]	0.0333	0.0111	0.0081
Tetrahedron [CS97]	0.3506	0.2722	0.2054

Table 5: Results on the Singleton Feature Sets

Figure 3: Illustration of the beta strand windowing

To develop the model for beta sheet interactions, we enlisted the additional collaboration of Dr. Lydia Gregoret in the biochemistry department at UCSC [Gre97]. Working with her, we have developed a new neural net method for modeling higher order interactions solely in beta sheet structure. This net models interactions among 10 amino acids: one block of 5 amino acids on each side of a pair of adjacent beta strands within a beta sheet, as depicted in Figure 3.

To build the potential function, a neural net is trained to distinguish actual contacts from decoys based on the input. The decoys consist of pairs of strands that are in the same protein but are not in contact with each other. We have built potential functions for anti-parallel strand pairs with a variety of inputs, and tested their performance on selecting the actual beta strand contacts over decoys. These results are summarized in Table 6. We are currently in the process of applying these potential functions in scoring the likelihood of protein alignments, based on the anti-parallel beta strand contacts.

Properties	Accessibility	Strand Membership	Chain Separation	Number of inputs	Accuracy in True-False Discrimination (%)
X	X	X	X	151	87.45
X	X	X		150	81.44
X	X			140	62.11
X		X	X	131	86.45
X		X		130	80.71
X			X	121	80.74
X				120	63.32
		X	X	11	86.73
		X		10	76.10
			X	1	80.53

Table 6: Percent accuracy in actual/decoy contact discrimination Each row represents a separate experiment, and each X indicates features used in that experiment

3.4 Analyzing pairwise interactions in the context of phylogenetic relationships

Covariation analysis of sets of aligned sequences for RNA molecules is relatively successful in elucidating RNA secondary structure, as well as some aspects of tertiary structure. Covariation analysis of sets of aligned sequences for protein molecules is successful in certain instances in elucidating certain structural and functional links, but in general, pairs of sites displaying highly covarying mutations in protein sequences do not necessarily correspond to sites that are spatially close in the protein structure. We have identified two reasons why naive use of covariation analysis for protein sequences fails to reliably indicate sequence positions that are spatially proximate, and have developed two solutions to address these problems.

Problem 1: Sequences related by a phylogenetic tree do not constitute independent samples. Hence estimation of pairwise probabilities by a frequency counting approximation, resulting from a maximum likelihood analysis assuming independence of the sequence samples, can be biased. We have developed a null-model approach to handle phylogenetic bias in estimation of covariation and validated it in simulation. Given a phylogenetic tree, and a model for independent evolution of sites to be described below, we evolve sequences down the given tree numerous times using the independence model for sequence evolution. A histogram is compiled for the resulting mutual information values which are calculated between all pairs of sequence positions. Such mutual information values will be different from zero, even though the sites are evolving independently, due to (a) finite sample size effects (the mutual information is a positive semi-definite quantity and any fluctuation due to finite sample size can therefore only result in positive mutual information) and (b) effects of the phylogenetic tree (the bifurcations of a typical phylogenetic tree tend to amplify finite sample fluctuations).

The null model procedure described above determines a threshold mutual information value, such that if any mutual information value calculated for the real sequence data exceeds the threshold value, then it is very unlikely that such a value could have arisen from the null model of "given phylogenetic tree and independent evolution of sites". However, the conclusion that the mutual information between a pair of sites was unlikely to have arisen from the null model of independence does not necessarily mean those sites are directly physically interacting. A second procedure is needed which is able to disentangle long chains of correlation to determine which sites are correlated due to direct interaction, and which sites are correlated due to (possibly long) indirect chains of interaction.

Problem 2: Consider the following situation: Site A physically interacts and covaries with site B; site B physically interacts and covaries with site C; but site A and site C do not physically interact. Site A can covary with site C in spite of no physical interaction between A and C. This effect of chained covariation is known as "correlation, or order at a distance" in the analysis of interacting spin systems. How can one disentangle causation from chained covariation? We developed a maximum entropy approach to this problem and validated it in simulation.

Given estimates of first and second moments of a probability distribution (as used to estimate correlations) what is the probability distribution which has maximal entropy, i.e. which is the "flattest" or "simplest" distribution satisfying the observational constraints? This problem would be ill-posed without the additional constraint of "simplicity" i.e. maximum entropy - many probability distributions exist which agree with any given moments. Maximizing the entropy subject to the constraints of given first and second moments results in the classic form for P :

$$P(x) = \frac{\exp(-(\sum_i \lambda_i x_i + \sum_{i,j} \lambda_{ij} x_i x_j))}{Z}$$
 where the λ 's are to be determined to implement the constraints, and Z normalizes P to unity. The constraints are satisfied at the minimum of the following function F , considered to be a function of the λ 's:

$$F = \log Z + \sum_i \lambda_i \bar{x}_i + \sum_{i,j} \lambda_{ij} \bar{x}_i \bar{x}_j$$
 where $\bar{x}_i$ and $\bar{x}_i \bar{x}_j$ represent the observed first and second order moments, respectively. Ideally, the reconstructed parameters λ should be zero for non-connected sites, and equal to the appropriate element of the potential matrix for connected sites. This was validated in simulation, where correlations existed in the model simulation between numerous disconnected sites, and hence these sites would erroneously be predicted to be connected under a naive application of covariation analysis. Application of our maximum entropy formalism correctly distinguished sites that are covarying because of a physical link, from sites displaying covariation but which are not structurally linked.

This research was published in: "Correlated Mutations in Protein Sequences: Phylogenetic and Structural Effects", Bertrand Giraud, LonChang Liu, Gary Stormo, Alan Lapedes, to be published in the Proceedings of the AMS "Conference on Statistics in Molecular Biology", Seattle WA July 1997.

References

- [AHO93] K. Asai, S. Hayamizu, and K. Onizuka. HMM with protein structure grammar. In *Proceedings of the Hawaii International Conference on System Sciences*, pages 783–791, Los Alamitos, CA, 1993. IEEE Computer Society Press.
- [B.96] Hazes B. The (q_{xw})₃ domain: A flexible lectin scaffold. *Protein Sci.*, 5(8):1490–1501, August 1996.
- [Bat96] A. Bateman. Fibronectin type III domains in yeast detected by a hidden Markov model. *Current Biology*, 6(12):1544–1547, December 1996.
- [Bat97] A. Bateman. The structure of a domain common to archaeobacteria and the homocystinuria disease protein. *Trends in Biochemical Sciences*, 22(1):12–13, January 1997.
- [BCHM94] P. Baldi, Y. Chauvin, T. Hunkapillar, and M. McClure. Hidden Markov models of biological primary sequence information. *PNAS*, 91:1059–1063, 1994.
- [BEC96] A. Bateman, S. Eddy, and C. Chothia. Members of the immunoglobulin superfamily in bacteria. *Protein Sci.*, 5:1939–1941, 1996.
- [BL93] S. H. Bryant and C. E. Lawrence. An empirical energy function for threading protein sequence through the folding motif. *Proteins: Structure, Function, and Genetics*, 16(1):92–112, May 1993.
- [BLE91] J. U. Bowie, R. Lüthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170, 1991.
- [Chu89] G. A. Churchill. Stochastic models for heterogeneous DNA sequences. *Bull Math Biol*, 51:79–94, 1989.
- [CS92] L. R. Cardon and G. D. Stormo. Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *JMB*, 223:159–170, 1992.
- [CS97] L. Lo Conte and T. Smith. Visible volume: a robust measure for protein structure characterization. *JMB*, 1997. to appear.
- [DMHM97] J.Z. Dalggaard, M.M. Moser, R. Hughey, and I.S. Mian. Statistical modeling, phylogenetic analysis and structure prediction of a protein splicing domain common to inteins and hedgehog proteins. *Journal of Computational Biology*, 4:193–214, 1997.
- [Edd95] Sean Eddy. Multiple alignment using hidden Markov models. In Christopher Rallings et al., editors, *ISMB-95*, pages 114–120, Menlo Park, CA, July 1995. AAAI/MIT Press.
- [Edd96] S.R. Eddy. Hidden Markov models. *Curr. Opin. Struct. Biol.*, 6(3):361–365, 1996.
- [EMD95] S.R. Eddy, G. Mitchison, and R. Durbin. Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.*, 2:9–23, 1995.
- [GFL95] T. Grossman, R. Farber, and A. Lapedes. Neural net representations of empirical protein potentials. In *Proceedings of Intelligent Systems for Molecular Biology*, pages 154–61. AAAI Press, 1995.
- [GLE90] M. Gribskov, R. Lüthy, and D. Eisenberg. Profile analysis. *Methods in Enzymology*, 183:146–159, 1990.
- [GME87] Michael Gribskov, Andrew D. McLachlan, and David Eisenberg. Profile analysis: Detection of distantly related proteins. *PNAS*, 84:4355–4358, July 1987.
- [Gre97] L. Gregoret, 1997. Personal communications with Lydia Gregoret, Assistant Professor of Chemistry, UCSC.
- [HH91] Steven Henikoff and Jorja G. Henikoff. Automated assembly of protein blocks for database searching. *NAR*, 19(23):6565–6572, 1991.
- [HMBC97] T. Hubbard, A. Murzin, S. Brenner, and C. Chothia. Scop: a structural classification of proteins database. *NAR*, 25(1):236–9, January 1997.
- [HS93] Liisa Holm and Chris Sander. Protein structure comparison by alignment of distance matrices. *JMB*, 233(1):123–138, 5 Sept 1993.
- [Hub94] Tim J.P. Hubbard. Use of β -strand interaction pseudo-potentials in protein structure prediction and modeling. In *Proceedings of the 27th Annual Hawaii International Conference on System Sciences*, pages 336–344, 1994.

- [JC90] R. L. Jernigan and D. G. Covell. Conformations of folded proteins in restricted spaces. *Biochemistry*, 29(13):3287-94, April 1990.
- [JH96] K. Fasman J. Henderson, S. Salzberg. Finding genes in human dna with a hidden Markov model. In *ISMB-96*, St. Louis, 1996. AAAI.
- [KBM94] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *JMB*, 235:1501-1531, February 1994.
- [KMH94] A. Krogh, I. S. Mian, and D. Haussler. A Hidden Markov Model that finds genes in *E. coli* DNA. *NAR*, 22:4768-4778, 1994.
- [LAB93] C. E. Lawrence, S. Altschul, M. Boguski, J. Liu, A. Neuwald, and J. Wootton. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262:208-214, 1993.
- [Lap94] A. Lapedes. A complex systems approach to computational molecular biology. In *Proceedings of the Santa Fe Institute Workshop on Integrative Themes*. volume 21. Addison Wesley, 1994.
- [Lat97] R. Lathrop, 1997. Personal communications with Richard Lathrop, Professor of Information and Computer Science, UC Irvine.
- [LG87] E. S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 84:2363-2367, 1987.
- [LR90] C. E. Lawrence and A. A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7:41-51, 1990.
- [MC94] V. N. Maiorov and G. M. Crippen. Learning about protein folding via potential functions. *Proteins*, 20(2):167-73, October 1994.
- [Mur96] A. G. Murzin. Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.*, 6(3):386-94, June 1996.
- [NJD96] Goldman N., Thorne J.L., and Jones D.T. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *JMB*, 263(2):196-208, October 25 1996.
- [Pen96] Elizabeth Pennisi. Teams tackle protein prediction. *Science*, 273:426-428, 26 July 1996.
- [PMA96] S. Pascarella. F. Milpetz, and P. Argos. A databank (3d-ali) collecting related protein sequences and structures. *Protein Engr.*, 9(3):249-51, March 1996.
- [Rab89] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257-286, February 1989.
- [Sip90] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology*, 213(4):859-83, June 1990.
- [Sip95] M. Sippl. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, 5:229-235, 1995.
- [SWS93] C. M. Stultz, J. V. White, and T. F. Smith. Structural analysis based on state-space modeling. *Protein Sci.*, 2:305-315, 1993.
- [Tau96] Gary Taubes. Software matchmakers help make sense of sequences. *Science*, 273:588-590, 2 August 1996.
- [WMS94] J. V. White, I. Muchnik, and T. F. Smith. Modeling protein cores with Markov random fields. *Mathematical Biosciences*, 124(4):149-79, December 1994.
- [WR93] Shoshana J. Wodak and Marianne J. Rooman. Generating and testing protein folds. *Curr. Opin. Struct. Biol.*, 3:247-259, 1993.
- [WW93] D. H. Wolpert and D. R. Wolf. Estimating functions of probability distributions from a finite set of samples; part 2: Bayes estimators for mutual information, chi-squared, covariance, and other statistics. Technical Report LA-UR-93-833, TR-93-07-047, Los Alamos National Lab, Santa Fe Institute, 1993.
- [WW95] D. H. Wolpert and D. R. Wolf. Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841-54, December 1995.
- [BHK97] Christian Barrett, Richard Hughey, and Kevin Karplus. Scoring hidden Markov models. *CABIOS*, 13(2):191-199, 1997.

- [DMHM97] J. Z. Dalgaard, M. J. Moser, R. P. Hughey, and I. S. Mian. Statistical modeling, phylogenetic analysis and structure prediction of a protein splicing domain common to inteins and hedgehog proteins. *Jour. Comp. Biol.*, 4(2):193-214, 1997.
- [GHS97] J. A. Grice, R. Hughey, and D. Speck. Reduced space sequence alignment. *CABIOS*, 13(1):45-53, February 1997.
- [HK96] Richard Hughey and Anders Krogh. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *CABIOS*, 12(2):95-107, 1996.
- [Kar95] Kevin Karplus. Regularizers for estimating distributions of amino acids from small samples. In *ISMB-95*, Menlo Park, CA, July 1995. AAAI/MIT Press.
- [KKB97] Kevin Karplus, Kimmen Sjölander, Christian Barrett, Melissa Cline, David Haussler, Richard Hughey, Liisa Holm, and Chris Sander. Predicting protein structure using hidden Markov models. *Proteins: Structure, Function, and Genetics*, 1997. to appear.
- [Lap97] Bertrand Giraud, LonChang Liu, Gary Stormo, and Alan Lapedes. Correlated Mutations in Protein Sequences: Phylogenetic and Structural Effects to be published in the Proceedings of the AMS "Conference on Statistics in Molecular Biology", Seattle WA July 1997
- [SKB96] K. Sjölander, K. Karplus, M. P. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler. Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology. *CABIOS*, 12(4):327-345, August 1996.
- [TH97] C. Tarnas and R. Hughey. Reduced space hidden markov models, August 1997. submitted for publication.