LA-UR-98-1094

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE: A MAXIMUM ENTROPY FORMALISM FOR DISENTANGLING CHAINS OF CORRELATED SEQUENCE POSITIONS

AUTHOR(S): A.S. Lapedes, T-16
B.G. Giraud, C.E.N. Saclay
L.C. Liu, T-2
G.D. Stormo, Univ of Colo

SUBMITTED TO: ISMB98, Montreal Canada

Los Alamos

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

FORM NO. 836 R4
ST. NO. 2629 5/81

DISCLAIMER

DISCLAIMER

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

A Maximum Entropy Formalism for Disentangling Chains of Correlated Sequence Positions

submitted to ISMB98, Montreal Canada

A.S. Lapedes * † #, B. G. Giraud ‡, L.C. Liu *, G.D. Stormo ○

** Theoretical Division*
MS B213
Los Alamos National Laboratory
Los Alamos, NM 87545
asl@t13.lanl.gov, liu@t2.lanl.gov

† Santa Fe Institute, 1399 Hyde Park Road
Santa Fe, New Mexico 87501
asl@santafe.edu

‡ Service Physique Théorique, DSM, C.E. Saclay
91191 Gif/Yvette, France
giraud@spht.saclay.cea.fr

○ Dept. of Molecular, Cellular and Developmental Biology
University of Colorado
Boulder, Colorado
Stormo@Colorado.edu

 # (to whom correspondence should be addressed)
[Keywords: correlated mutations, mutual information, maximum entropy]

Abstract

Covariation analysis of sets of aligned sequences of protein molecules is successful in certain instances in elucidating certain structural and functional links [Korber(1993)], 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 [Gobel(1994)], [Clarke(1995)], [Shindyalov(1994)], [Thomas(1996)], [Taylor(1994)], [Neher(1994)]. In contrast, 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 [Gutell(1992)]. The goals of this paper are to (1) present the problem, (2) develop the mathematical formalism for solving the problem, and (3) validate the resulting algorithms on simulated data. Extensive application to biological sequences will be presented elsewhere.

1 Introduction

Analysis of sets of aligned sequences, such as RNA or protein sequences, is a common procedure in bioinformatic analysis. Often the assumption is made that sequence positions are independent of each other. However, recent work [Gutell(1992)], [Korber(1993)] has considered the possibility that mutations in various positions are not independent, but covary. "Mutual information", a measure of covariation for discrete symbols [Cover(1991)], is a convenient measure to quantify the amount of covariation of mutations between positions in biological sequences [Gutell(1992)], [Korber(1993)]. Mutual information can be expressed in numerous equivalent ways. In this paper we will use the following definition

$$M = \sum_{ab} P_{ab} log(P_{ab}/P_a P_b)$$

where P_{ab} denotes the pairwise probability distribution for symbols in a pair of sequence positions, and a and b represent the possible base or amino acid symbols of the sequence. P_a is the single site probability distribution for the first member of the pair, and P_b is the single site probability distribution for the second member of the pair.

To apply this formula one needs to estimate from the data the individual pairwise and single site probability distributions. Given a set of sequences which are assumed to be *i.i.d* (independent and identically distributed) samples from a probability distribution, then one can independently estimate each pairwise probability distribution for every pair of positions by frequency counting – this estimate results from a maximum likelihood analysis independently applied to each pair of positions (see [Lapedes(1997)] for an approach for dealing with *non i.i.d.* sequences which are related by a phylogenetic tree). Marginalizing the estimate of the pairwise distribution yields the estimate for single site probabilities.

It is natural to hypothesize [Gutell(1992)], [Korber(1993)] that direct physical interactions, for example those stemming from spatial proximity, is the cause of observed correlations between positions in biological sequences. This hypothesis holds relatively well for RNA sequences [Gutell(1992)]. However the situation with proteins is more complex, due to the diffuse nature of protein interactions. In contrast to RNA interactions, where once a Watson-Crick type base pair is formed, many additional interactions do not usually form, amino acid residues typically have diffuse, non-saturating interactions in which one amino acid can loosely interact with many others, leading to long chains of interaction. For example, sequence position 3 may be correlated with position 23 because these positions are spatially close in the folded structure. For proteins, position 23 may also be correlated with position 33 because these positions are close in the folded structure. Similarly, position 33 may be correlated with position 43 because these positions are close in the folded structure. Sequence position 3 would then typically be correlated with sequence position 43 due to the chaining of correlations between the two positions. However, sequence position 3 and sequence position 43 need not be spatially close in the folded structure, and an

inference that they were close based on significant covariation between the positions can be in error.

We address the question: how can one use single site and pairwise probability information (as embodied in e.g. correlation measures) to estimate the contact matrix of local physical interaction? Initial analysis of this problem was reported in [Lapedes(1997)]. Physicists will recognize the "chaining effect" as "correlation at a distance" in spin systems [Stanley(1971)] [Binney(1992)]. Statisticians will recognize this problem as being related to the inference of the parameters of a discrete multivariate probability distribution (for the sequence as a whole), given just estimates of the first and second order moments of the distribution. The problem of determining a probability distribution given a finite number of moments is ill-posed – there are many solutions. The additional constraint of maximal entropy makes the solution unique, and may be viewed as the plausible restriction that the distribution results in the observed correlations, but is otherwise as "flat", or as "simple", as possible. It is this simplicity constraint which allows one to deduce a small set of local interaction parameters which can account for nonlocal correlations induced by the "chaining effect", as we demonstrate in model simulations.

2 Modeling Extended Chains of Local Correlations

Although protein sequences can be hundreds of amino acids long, typically a much smaller number of amino acids display significant covariation. In this section we will consider, for reasons of pedagogical simplicity, models with only ten potentially interacting sites, and we will also restrict consideration to two-state "amino acids". Then, in the following section we extend this formalism to handle multi-states, i.e., four-state for RNA/DNA, or twenty-state for amino acid sequences. The algorithms developed here scale reasonably with the number of potentially interacting amino acids, and with the number of states per site, but in general the algorithms require non-trivial amount of computation time. Hence, for the following model simulations, which we use to explain and to validate the algorithms, we report results for smaller systems where results can be obtained easily.

To provide a framework for the development and validation of the algorithms we need a model of evolution that incorporates certain aspects of interaction between sites. Whether this specific evolution model is correct or not (it almost certainly is incorrect) is beside the point – it need only exhibit the general principle that correlation can occur between distant sites by extended chains of local correlations. The model we use is a Monte Carlo model where sites are chosen at random for mutation, and the probability to mutate to a new symbol at a site depends on two contributions: (1) contributions that are independent of the symbols at other positions (similar to Dayhoff [Dayhoff(1978)]), and (2) contributions that depend on the symbols in other positions (specifically those positions that interact with the given position).

In analogy to "threading potentials" of protein sequence analysis [Sippl(1990)] we

define a model "energy" function E:

$$E = \sum P_{ij} C_{ij} X_i X_j + \sum_i H_i X_i$$

where C_{ij} is a "contact matrix" whose elements, 1 or 0 , denote contact or not, respectively, between sequence positions i and j, and P_{ij} is an "interaction" potential that gives the contribution to E when amino acids at positions i, j are in contact. P is a fixed matrix (here, a two by two matrix, the analogue of the twenty by twenty matrices derived in threading analyses) which gives the contribution of the contacting amino acids at sites i, j. H_i are parameters that influence the final equilibrium probabilities of symbols at the various sites, and for simplicity we set H_i equal to zero (it does not effect interactions). X_i denotes the state of the two-state "amino acid" at each site, and assumes the values $(1, -1)$. Later we present the extension of the formalism to multi-states, including four states (DNA/RNA), twenty states (proteins), or any other number of states.

Our model of "evolution" is defined as the process of picking a site at random, making a trial mutation of the "amino acid" at that site, evaluating the energy of the mutated system compared to the unmutated system, and accepting the mutation (1) if the energy after mutation was lowered, or (2) if the energy after mutation was raised, then the mutation is accepted with (the exponentially low) probability proportional to $exp(-\Delta E)$ where ΔE is the difference in energy between the mutated and unmutated systems. This prescription for evolution of an interacting system will be recognized as the classic Monte Carlo "Metropolis" algorithm for evolution of a spin system [Binney(1992)].

Results of a typical model simulation are reported below. Five hundred two-state sequences of length ten were evolved to equilibrium according to the above Monte Carlo algorithm, using the following connnectivity matrix, C_{ij}, (which was chosen at random among matrices having average connectivity of three):

Connection Matrix

$$\begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
. & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
. & . & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
. & . & . & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
. & . & . & . & 0 & 0 & 1 & 0 & 1 & 1 \\
. & . & . & . & . & 0 & 0 & 1 & 0 & 1 \\
. & . & . & . & . & . & 0 & 1 & 0 & 0 \\
. & . & . & . & . & . & . & 0 & 0 & 0 \\
. & . & . & . & . & . & . & . & 0 & 0 \\
. & . & . & . & . & . & . & . & . & 0
\end{pmatrix}$$

The two by two interaction matrix, P, for a two-state system contains three independent components, which to use the language of spins, can be described as: up-up, up-down (and equivalently down-up), and down-down. A "ferromagnetic"

interaction matrix, which we use for this example, has the up-up and down-down values assigned positive one, and the up-down (equivalently down-up) value assigned negative 1.

The correlation matrix, $< (x_i - \overline{x}_i)(x_j - \overline{x}_j) >$, and the contact matrix are represented graphically in Fig. (1). For two-state systems the correlation matrix captures the same information as does mutual information. Solid lines between pairs of sites represent those pairs which are connected via the contact matrix (above). Dashed lines between pairs of sites represent pairs which had an absolute value of correlation over 0.3. This threshold value for representation was chosen because values over 0.3 occurred less than one time in one hundred, as computed in one hundred additional simulations of evolution which used independent evolution of sites (no contribution from the C_{ij} matrix). Hence, dashed lines between sites represent correlations that are very improbable to have occurred in the null model of independent evolution of sites, and yet are not caused by direct connections between sites. The appearance of non-zero correlation values in the null model of independence is solely a finite data effect. Correlations will tend towards zero in the null model in the limit of infinite data.

Note that significant covariation exists between many pairs of sites that are not physically connected. Sites (1,6), as well as sites (2,8), see Fig. (1) are examples. Note that sites (1,2) are physically connected, as are sites (2,6), and hence a chain of covariation, (1,2) (2,6), can form which leads to significant correlation between disconnected sites such as (1,6). Longer chains also exist, such as (8,1) (1,2) (2,9) leading to correlation between disconnected sites (8,9). Consider the observed correlation between disconnected sites (2,8). Although we prefer to call the general mechanism by which disconnected sites can co-vary, "chained correlation", there is also another interpretation. The disconnected sites (2,8) can be interpreted to display significant correlation because sites (1,2) are physically connected, as are sites (1,8). In this situation, correlation can exist between sites (2,8), even though they are not physically connected, because of a common "driving cause", which is the connection of both site (2) and site (8) to site (1). Site (1) "drives" both (2) and (8), resulting in correlation between these disconnected sites.

In situations where there can be extended interactions between sites, such as in proteins and in our model above, extended, complicated chains of correlation can occur which leads to correlations between sites that are not physically connected. To disentangle the chains of covariation one must solve the classic conundrum of "causation versus covariation". As we show below, an effective solution for the types of problems encounted in sequence analysis is to estimate the parameters of the "simplest" (i.e. maximal entropy) probability distribution which yields the observed correlations.

3 Maximum Entropy Formalism: Two-state

Classic Problem: Given estimates of the first and second moments of a probability

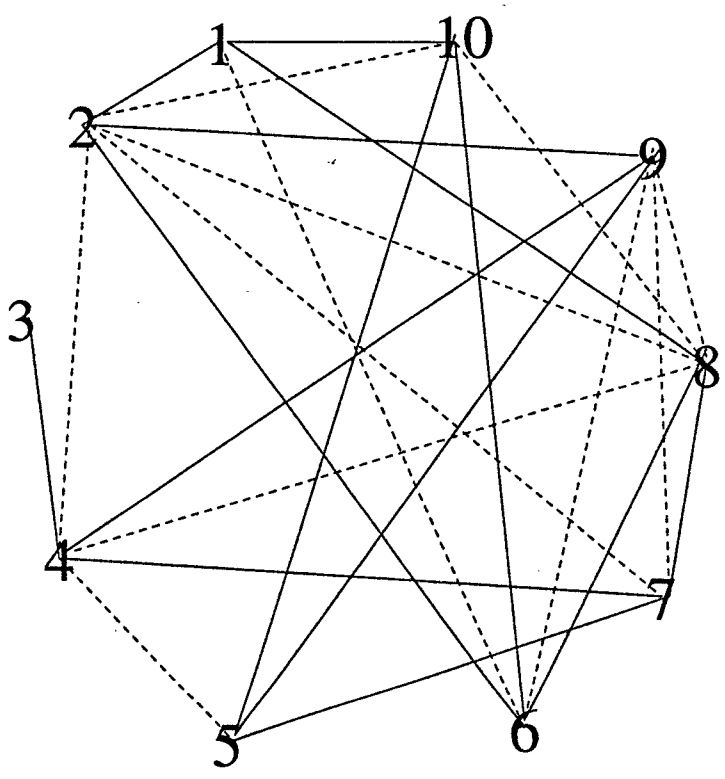

Figure 1: Statistically significant correlations between sites (i.e. correlation values greater than 0.3 in absolute value – see text) are represented by dashed lines. Numerous pairs of sites that are not physically connected ($C_{ij} = 0$) display statistically significant correlation. Sites which are physically connected ($C_{ij} = 1$) are represented by solid lines and also exhibit statistically significant correlations. The value of correlations between sites which have physical connections (solid lines) range from -0.5 to 0.6 (only two physically connected pairs of sites exhibit correlations as high in absolute value as 0.6). The value of correlations between sites which do not have physical connections (dashed lines) range from -0.5 to 0.5. There is significant overlap in the values of correlation between sites which are physically connected and those which are not, which prevents accurate identification of connections based on correlation value alone. The maximum entropy approach developed here solves this problem.

distribution (as used to estimate the correlation, above) determine the probability distribution which has maximal entropy, i.e. which is the "flattest" or "simplest" distribution satisfying the observed 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.

Classic Solution: Maximizing the entropy subject to the constraints of given first and second moments results in the classic [Tikochinsky(1984)] [Levine(1979)] form for P:

$$P(x) = \frac{exp - (\sum_i \lambda_i x_i + \sum_{ij} \lambda_{ij} x_i x_j)}{Z}$$

where the $\lambda's$ are Lagrange multiplier used 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 $\lambda's$:

$$F = logZ + \sum_i \lambda_i \overline{x_i} + \sum_{ij} \lambda_{ij} \overline{x_i x_j}$$

where $\overline{x_i}$ and $\overline{x_i x_j}$ represent the observed first and second order moments, respectively. Direct differentiation of F with respect to the $\lambda's$ verifies that the $\lambda's$ minimizing F will implement the desired constraints. For related investigations with a similar functional form see [Heumann(1995)].

F is a nonlinear function of the variables λ. It is possible to prove that F has a unique, global minimum by using standard inequalities of information theory [Cover(1991)]. Evaluating the minimum of F by e.g. gradient descent with respect to the variables λ, results in an expression involving the first and second order moments of the model distribution evaluated at intermediate (i.e. non-extremal) values of λ. Thus, the numerical procedure to solve for the parameters λ involves successive rounds of Monte Carlo evolution (with a sufficient number of steps in each round to reach equilibrium at each intermediate value of the $\lambda's$), followed by a small change in the $\lambda's$ in the gradient direction, followed by more Monte Carlo to evaluate the new expectations etc. This process converges (i.e. the gradient is zero) when the numerically computed expectations agree with the specified expectations $\overline{x_i}$ and $\overline{x_i x_j}$. Evaluation of the first and second order moments at each intermediate value of the $\lambda's$ would be prohibitively expensive if all states were enumerated exhaustively. However, standard techniques of importance sampling, long familiar to physicists performing Monte Carlo simulation of spin systems [Binney(1992)], and now popular in bioinformatic investigations [Lawrence(1993)], are an efficient alternative to exhaustive enumeration of all states. In addition, approximation techniques such as Mean Field Theory [Lapedes(1998a)], also provide an alternative to exhaustive enumeration.

Application to Model Simulation Comparing the resulting form of $P(x)$, above, to the contact potential used to generate the simulated data, we see that the reconstructed parameters λ_{ij} should be zero for non-connected sites, and equal to the appropriate element of the potential matrix, P, for connected sites. When applied to the above model simulation, the reconstructed parameters are:

Reconstructed λ_{ij} Parameters

$$\begin{pmatrix}
\ldots & \mathbf{1.0} & 0.3 & 0.1 & 0.2 & 0.3 & -0.1 & \mathbf{0.9} & 0.1 & \mathbf{1.3} \\
\ldots & \ldots & 0.1 & -0.2 & 0.1 & \mathbf{1.0} & 0.1 & -0.1 & \mathbf{0.9} & 0.0 \\
\ldots & \ldots & \ldots & \mathbf{0.9} & 0.0 & 0.0 & 0.0 & 0.3 & 0.1 & 0.2 \\
\ldots & \ldots & \ldots & \ldots & -0.2 & -0.2 & \mathbf{1.0} & 0.2 & \mathbf{1.1} & 0.1 \\
\ldots & \ldots & \ldots & \ldots & \ldots & 0.0 & \mathbf{0.9} & 0.2 & \mathbf{1.0} & \mathbf{1.0} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & -0.0 & \mathbf{1.0} & -0.3 & \mathbf{1.1} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \mathbf{1.1} & 0.3 & -0.3 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 0.2 & -0.1 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & -0.0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots
\end{pmatrix}$$

Even though the values of the correlations between non-connected sites, represented by the dashed lines in Fig. (1) can be as high as between most of the connected sites, the reconstructed parameter values, above, are low between non-connected sites and are high (bold face values) between connected sites. The maximum entropy procedure identifies the dashed lines of Fig. (1) as a chaining phenomenon and the solid lines of Fig. (1) as direct physical connections, i.e. as high values of λ_{ij}. Finite sample effects accounts for the remaining "noise" in elements of the matrix which should have value zero (because of zero connectivity), and in the elements which should have absolute value of one (due to nonzero connectivity and the values used in the ferromagnetic potential matrix, P).

The maximum entropy formalism performs an implicit search over C_{ij}, which even in the simple example considered here involves an implicit search over 2^{45} or approximately 10^{13} discrete contact matrices. This illustrates the power of the formalism. Other heuristic search techniques could also be used, such as genetic algorithms or use of Monto Carlo methods to search over the large space of discrete contact matrices, C_{ij}.

4 Maximum Entropy Formalism: Multi-state

Application of the above maximum entropy analysis to real biological sequences requires an extension beyond two-state to four-state (DNA/RNA), or to twenty-state (amino acid) sequences. In this section we extend the two-state "spin formalism", above, to multi-states.

Recall that the two state formulation assigns a *scalar variable*, X_i, at each site, assuming discrete values in the set {-1,1}, i.e. $X_i = 1$ or $X_i = (-1)$. A sequence of length L is denoted by X_i where i ranges from 1 to L, and X is either 1 or -1 at each site i. To generalize to four, twenty, or any other number of states, we let X_i be a *vector variable* at each site, taking values from a fixed set of possible vectors. Representation of the four bases of DNA utilizes a four-dimensional vector at each site, where e.g. A is represented by the vector (1000), C is represented by the vector (01000), G is represented by the vector (0010), and T is represented by the vector (0001). In this notation a DNA sequence, for example the arbitrarily

chosen six long sequence *ACCTGA*, would be represented as a sequence of vectors $X_1^{(1)}X_2^{(2)}X_3^{(2)}X_4^{(4)}X_5^{(3)}X_6^{(1)}$, where each X is a vector chosen from the set of four possible vectors at each site, as denoted by the superscript. We will denote a sequence of vectors at sites i as, $X_i^{(\alpha)}$, where i ranges from 1 to L and α labels which of the unary vectors is present at site i. To reference an individual component, b, of a vector at a sequence position we could write: $X_{b,i}^{(\alpha)}$. However, since the b^{th} component of the α^{th} unary vector is zero unless $b = \alpha$, and since polynomial expressions in X vanish for all but nonzero components, we will use the shorthand notation that X_i^α refers to the non-zero α component of the vector at position i.

It is important to note that due to the unary nature of the allowed vectors that $\sum_{\alpha=1}^{S} X_i^\alpha = 1$, where S denotes the number of states. Hence, any polynomial expression occurring in parameterized probability distributions, e.g.,

$$\sum_{\alpha\beta ij} \lambda_{ij}^{\alpha\beta} X_i^\alpha X_j^\beta + \sum_{\alpha i} \lambda_i^\alpha X_i^\alpha$$

involving the components of the vectors (λ are parameters), can always have one component of each vector (by convention we will always use the S^{th} component) replaced by a summation over the other $S-1$ components of the set. Therefore algebraic expressions involving vectors at sequence positions do not involve S independent components of the vectors. Only $S-1$ independent components of the possible S components appear. It is this fact that makes the representation of two-state objects appear to involve scalar variables, X_i, and not vector variables, X_i^α. The range of α is from 1 to 1, and therefore X_i^α is treated as a scalar (the difference between (0,1) notation and (-1,1) notation is a trivial linear shift). It is important to note that even though only $S-1$ components of the possible S components appear in any algebraic expression, that sums over the complete multi-state space (for example, as used to normalize a probability distribution), do involve sums over *all* S vectors.

The maximum entropy distribution of Section (3) can now be extended to represent multi-states by writing

$$P(X) = \frac{exp(\sum_{\alpha\beta ij} \lambda_{ij}^{\alpha\beta} X_i^\alpha X_j^\beta + \lambda_i^\alpha X_i^\alpha)}{Z}$$

We again emphasize that *sums over state components*, for example the sums over indices $\alpha\beta$ in the $\sum_{\alpha\beta ij}$ expression in the exponent above, range from 1 to $S-1$ (i.e. do not include the S component), while *sums over possible state vectors*, for example in the partition function Z which normalizes the probability (above):

$$Z = \sum_{X_i^\alpha} exp(\sum \lambda_{ij}^{\alpha\beta} X_i^\alpha X_j^\beta + \lambda_i^\alpha X_i^\alpha)$$

range over all S possible vectors, and does include the S^{th} vector.

The general probability distribution describing L-site, S-state systems requires $S^L - 1$ independent parameters (one parameter is fixed by the normalization condition). Simple extension of the above notation to terms of higher order than the

second, e.g. extension to third order terms like $\sum_{\alpha\beta\gamma ijk} \lambda_{ijk}^{\alpha\beta\gamma} X_i^\alpha X_j^\beta X_k^\gamma$ in the exponential, and so on for all higher orders up to the L^{th} order, involves the requisite number of parameters as may be seen by enumeration: the number of terms in a polynomial expression of M^{th} order, analogous to the above third order expression, involves $\binom{L}{M}(S-1)^M$ terms. Adding up the number of terms over all possible orders from 1 to L yields

$$\sum_{M=1}^{M=L} \binom{L}{M}(S-1)^M = S^L - 1$$

for the total number of free parameters, excluding the normalizing constant. Hence the formalism above, for representing multi-state objects, is general. Various degrees of approximation may be introduced by truncating at various orders the expansion of the polynomial appearing in the $log(probability)$. The order of truncation determines the order of correlation captured in the the parameterized probability distribution.

5 Conclusions:

Covariation between pairs of sequence positions that are not spatially proximate can result from possibly long chains of co-variation, as well as from "common cause" effects, and need not result from causation (i.e. from direct structural links). Chained covariation makes the prediction of structural links in proteins, using naive application of covariation analysis to sequence data, prone to error. A technique involving maximum entropy reconstruction of the parameters describing the probability distribution of the sequences was developed, and was validated in model simulations where accurate recovery of the structural links of the model was achieved. We remark that in application to real sequence data, further errors will probably remain even after addressing chaining effects. The origins of such errors can be diverse, such as possibly critical relationships between certain amino acids that are required to maintain the folding pathway. However, identifying and addressing the chaining problem should improve accuracy of prediction of spatial contacts by covariation analysis of sequences from variable protein families. (After this research was completed we received an English translation of work published in Russian [Afonnikov(1997)], addressing a related issue: chained covariation between real-valued numbers representing amino acid physical properties, but using a significantly different methodology to that adopted here.)

The conclusion that causation (direct structural links) can be distinguished from covariation, by fitting parameters to an assumed model, stands independent of the particular models and simulations used here to illustrate the point. Our goal in this paper is to lay a new conceptual and mathematical foundation for protein structure determination via analysis of covarying mutations, and to test the mathematical formalism in model simulations. Extensive application to biological sequence data will be presented else where.

Acknowledgements The authors would like to thank the Santa Fe Institute, where part of this work was performed. Alan Lapedes thanks the Department of

Energy for financial support of this research.

References

[Gutell(1992)] Gutell, R.R., Power, A., Hertz, G.Z., Putz, E. and Stormo, G.D. Nucl. Acids Res. 20:5785-5795

[Korber(1993)] Korber, B., Farber, R., Wolpert, D., Lapedes, A. *Covariation of Mutations in the V3 Loop of Human Immunodeficiency Virus Type 1 Envelope Protein: An Information Theoretic Analysis* Proc. Natl. Acad. Sci. USA 90:7176-7180

[Gobel(1994)] Gobel, U., Sander, C., Schneider, R., Valencia, A. *Correlated Mutations and Residue Contacts In Proteins* Proteins: Structure, Function, Genetics, vol 18:309-317

[Clarke(1995)] Clarke, N. *Covariation of Residues in the Homeodomain Sequence Family* Protein Science 4:2269-2278

[Shindyalov(1994)] Shindyalov, I., Kolchanov, N., Sander, C. *Can Three Dimensional Contacts in Protein Structures be Predicted by Analysis of Correlated Mutations?* Protein Engineering 7:349-358

[Thomas(1996)] Thomas, D., Casari, G., Sander C. *The Prediction of Protein Contacts From Multiple Sequence Alignments* Protein Engineering 9:941-948

[Taylor(1994)] Taylor, W., Hatrick, K. *Compensating Changes in Protein Multiple Sequence Alignments* Protein Engineering 7:341-348

[Neher(1994)] Neher, E. *How Frequent are Correlated Changes in Families of Protein Sequences?* Proc. Natl. Acad. Sci. USA 91:98-102

[Cover(1991)] Cover, T., Thomas, J. *Elements of Information Theory* Wiley Series in Telecommunications, John Wiley and Sons

[Stanley(1971)] Stanley, H. *Introduction to Phase Transitions and Critical Phenomena* The International Series of Monographs on Physics Oxford University Press Inc. Oxford and New York

[Dayhoff(1978)] Dayhoff, M., Schwartz R., Orcutt, B. *A Model of Evolutionary Change in Proteins* pps. 345-352 in Atlas of Protein Sequence and Structure, vol. 5, suppl. 3, Natl. Biomedical Res. Found., Silver Spring, Maryland.

[Sippl(1990)] Sippl, M., *Calculation of Conformational Ensembles from Potentials of Mean Force*, J. Mol. Biol. 213:859 - 883

[Tikochinsky(1984)] Tikochinsky, N., Tishby, N., Levine, R. *Alternate Approach to Maximum Entropy Inference* Physical Review A 30:2638-2644

[Levine(1979)] Levine, R., Tribus, M. *The Maximum Entropy Formalism* MIT Press, Cambridge MA Physical Review 106:620

[Heumann(1995)] Heumann, J., Lapedes, A., Stormo, G. *Alignment of Regulatory Sites Using Neural Networks To Maximize Specificity* Proceedings of the 1995 World Congress on Neural Networks II, 771-775

[Binney(1992)] Binney, J., Dowrick, N., Fisher, A., Newman, M. *The Theory of Critical Phenomena* Oxford University Press, Oxford

[Lawrence(1993)] Lawrence,C. Altschul, S., Boguski, M., Liu, J., Nuewald, A., Wooton, J. *Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment* Science 262:208-214

[Lapedes(1997)] Lapedes A., Giraud B.,Liu L., Stormo G., *Correlated Mutations In Protein Sequences: Phylogenetic and Structural Effects*, Santa Fe Institute working paper (1997), to be published in Proceedings of the AMS Conference on Statistics in Mathematical Biology, Seattle WA, (1997)

[Lapedes(1998a)] Lapedes A. *Mean Field Approximations to Probability Distributions of Interest in Biological Sequence Analysis* manuscript in preparation

[Afonnikov(1997)] Afonnikov D., Kondrakhin Y., Titov I., Kilchanov N. *Detecting Direct Correlations Between Positions in Multiple Alignments of Amino Acid Sequences* published in Russian in J. of Molecular Biology, accepted for publication German Conference on Bioinformatics Sept 1997, Special Issue of CABIOS (to be published).