Statistics in Molecular Biology
IMS Lecture Notes – Monograph Series (1999) Volume 33

CORRELATED MUTATIONS IN MODELS OF PROTEIN SEQUENCES: PHYLOGENETIC AND STRUCTURAL EFFECTS

BY ALAN S. LAPEDES[1], BERTRAND G. GIRAUD, LONCHANG LIU AND GARY D. STORMO

Los Alamos National Laboratory, Santa Fe Institute, Service Physique Théorique, DSM, C.E.N. Saclay and University of Colorado

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 et al. (1992)]. Covariation analysis of sets of aligned sequences for protein molecules is successful in certain instances in elucidating certain structural and functional links [Korber et al. (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 et al. (1994), Clarke (1995), Shindyalov et al. (1994), Thomas et al. (1996), Taylor & Hatrick (1994), Neher (1994)]. In this paper we identify two reasons why naive use of covariation analysis for protein sequences fails to reliably indicate sequence positions that are spatially proximate. The first reason involves the bias introduced in calculation of covariation measures due to the fact that biological sequences are generally related by a non-trivial phylogenetic tree. We present a null-model approach to solve this problem. The second reason involves linked chains of covariation which can result in pairs of sites displaying significant covariation even though they are not spatially proximate. We present a maximum entropy solution to this classic problem of "causation versus correlation". The methodologies are validated in simulation.

1. Introduction. Analysis of sets of aligned sequences, such as RNA or protein sequences, is a common procedure in bioinformatic analysis. Various methods have been developed to describe aligned sequences: "consensus" sequences which are determined by the most conserved symbol in each sequence position; "profiles" [Gribskov et al. (1987)] which represent the probability distribution of symbols in each position, and can also include inserts and deletes with fixed position independent penalties; and "hidden Markov models" [Krogh et al. (1994)], which represent single site probability distributions as well as position dependent probability distributions for insertions and deletions. Correlation analysis extends such methods to consideration of the probability distribution for pairs of symbols in all possible pairs of positions in the sequence. "Mutual information", a measure of correlation for discrete symbols [Cover &

[1]Research supported by the Department of Energy under contract W-7405-ENG-36
AMS 1991 *subject classifications.* Primary 62F03 ; secondary 62P10, 92D20.
Key words and phrases. Correlated mutations, phylogeny, interaction, structure from sequence.

Thomas (1991)], quantifies the covariation of mutations for pairs of positions in biological sequences Gutell et al. (1992), Korber et al. (1993)].

Mutual information can be expressed in numerous equivalent ways, some of which derive from information theory, hence the name. In this paper we will not use information theoretic expressions involving entropy [see e.g. Korber et al. (1993)], but will instead use the following formula

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

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. This expression may be interpreted as the log-likelihood ratio for the data for a specific pair of positions to arise from the independent (factorized) distribution versus a pairwise distribution.

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. Marginalizing the estimate of the pairwise distribution yields the estimate for single site probabilities.

In Section 3 we will examine the effect on estimates of mutual information when the sequences used to estimate each individual pairwise probability distribution are not themselves independent samples, but are instead related via shared ancestry described by a phylogenetic tree. Other work addressing phylogenetic effects may be found in references [Altschul et al. (1989), Sibbald & Argos (1990), Gerstein et al. (1994), Hennikoff & Hennikoff (1994), Thompson et al. (1994)].

Our approach is to define a null model, which is based on evolution down an assumed known phylogenetic tree with independent mutations in different sequence positions. By incorporating the tree into the hypothesis of independent evolution of sites, we can determine a threshold value of mutual information from the null model, such that any values of mutual information seen in the real data which are over threshold have a very low probability of coming from the null model. In other words, a threshold is determined such that pairs of sites yielding mutual information values over threshold probably did not result

from independent evolution down the phylogenetic tree, i.e. they really are correlated. This approach simultaneously deals with two issues, (1) since mutual information is a positive semi-definite quantity, any estimate from finite data can only overestimate the mutual information. Put differently, positive values of mutual information will result even if sites are independent, purely due to fluctuations inherent in a finite sample size, and (2) non-trivial phylogenetic trees amplify the finite sample size effect, hence independent evolution of sites down a non-trivial phylogenetic tree will result in higher mutual information values than evolution down a simple star phylogeny. The null model technique addresses both these issues.

After addressing finite sample and phylogenetic effects, another important effect remains. In Section 4 we address this problem. The problem can be stated in various ways. One statement is that there can exist "chains" of covarying pairs of positions. For example, sequence position 3 may be correlated with position 23 (because these positions are spatially close in the folded structure) , position 23 may be correlated with position 33 (because these positions are close in the folded structure), and 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. This "chaining effect" is the cause of many of the errors that occur when attempting to deduce spatially close positions in protein sequences using a covariation analysis. This effect, and the associated errors, is not as pronounced for RNA sequences because the specific bonding and saturation of Watson-Crick pairs tends to prevent chains of correlated mutations. We present a solution to the chaining problem for protein sequences which is validated in model simulations.

Physicists will recognize the "chaining effect" as "correlation at a distance" in spin systems [Stanley (1971), Binney et al (1992)], of which the one dimensional Ising spin model in a heat bath is a favorite example. The Ising model is a one dimensional chain of two-state spins, with each spin having a local physical interaction with only the spins on either side of it. Nevertheless, significant non-local correlations occur between spins that are separated by large distances even though the physical interaction is strictly a local nearest-neighbor interaction. More generally, one might have a spin system with physical interactions between designated sites described by a "contact matrix", C_{ij}. If the spins have a direct physical contact, and hence a direct interaction, then $C_{ij} = 1$, and otherwise $C_{ij} = 0$. A potential matrix, P, describes the energetic contribution of two con-

tacting spins at i and one at j. Typically, this matrix is not position dependent, i.e. two "up" spins always have the same energetic contribution no matter where they are located (and similarly "down" spins, or mixed "up/down" spins). In proteins this matrix will be a twenty by twenty symmetric matrix describing the energetic contributions of two amino acids in contact. The probability of a configuration of spins (or amino acids), as represented by the Gibbs distribution, is proportional to $\exp -(Energy)$, where $Energy$ is the energy of the configuration (obtained from P and C_{ij}). In Section 4, 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?

Statisticians will recognize this question as being related to the inference of parameters, i.e. P and C_{ij}, occuring in the discrete multivariate probability distribution representing the probability of the sequence as a whole, given just estimates of the first and second order moments of the distribution. Clearly, under the special assumption that each site evolves independently of other sites then it is easy to estimate the probability distribution for the sequence as a whole using maximum likelihood techniques. However, this assumption utilizes only the first order moments, and ignores the second order moments. To also include the second order moments (as embodied in the observed correlations) we develop a maximum entropy analysis in Section 4. This analysis determines the unique probability distribution for the sequence as a whole, which has the given first and second order moments (i.e. correlations) and also has maximal entropy.

The problem of determining a probability distribution given just a finite number of moments is ill-posed – there are many solutions. The additional constraint of maximal entropy makes the solution unique. The maximal entropy constraint 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 limits the number of non-zero coefficients, and 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. General models of evolution.

2.1. *Evolution with independent sites.* Models describing independent evolution of bases, such as the Jukes-Cantor [Jukes & Cantor (1969)] model and its variants [Kimura (1980)], can be extended to describe the independent evolution of amino acids [Kishino et al. (1990), Hasegawa & Fujiwara (1993)] by

incorporating PAM matrices [Dayhoff et al. (1978)] in definition of the transition rates. Let $P(t)$ be the probability distribution for the amino acids at a site evolving according to a Kimura style model of independent evolution of amino acid sites [Kimura (1980)]. To fix ideas, consider the simplest situation where $P(t)$ is a 20-vector satisfying the following simple equation (compare Hillis et al. (1995)):

$$dP(t)/dt = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,20} \\ a_{2,1} & a_{2,2} & \dots & a_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ a_{20,1} & a_{20,2} & \dots & a_{20,20} \end{pmatrix} P(t)$$

where

$$a_{ii} = (-19\alpha) \quad \text{and} \quad a_{ij} = (\alpha), i \neq j$$

The probability distribution for the sequence as a whole is the product of the independent single site probabilities, above. Kishino et al. (1990) and Hasegawa & Fujiwara (1993) extend the above model to incorporate the propensities of amino acids to mutate to those of a similar physico-chemical nature [Dayhoff et al. (1978)] and to have different equilibrium probabilities. The analytical solution to equations such as the above results in the familiar exponential time dependence of the probabilities, which can take a quite complicated form even though the equation satisfied by the probabilities is simple. In practice, sequences can be evolved numerically via Monte Carlo, not analytically, where the probability to accept a mutation is related to the values a_{ij}. The Monte Carlo procedure allows one to numerically simulate solutions to the general evolution equation, including situations where complicated interactions are introduced between sites (see below), and no analytic solution is possible.

2.2. *Evolution with interacting sites.* The Chapman-Kolmorgorov equation for jump processes (the "Master Equation" to physicists) generalizes the above simple stochastic evolution equation to nonindependent i.e. interacting sites. It subsumes all possible discrete state evolution models. Let x_i be 20-state objects representing amino acids located at sequence position i. The Master Equation balances the probability of transitioning into a configuration, x, of the system from a configuration y, with the probability of transitioning out of configuration x of the system to all possible configurations y.

$$\frac{dP(x_1, x_2, \dots, x_n)}{dt} = \sum_{y_1 \dots y_n} \omega(x_1 \dots x_n; y_1 \dots y_n) P(y_1 \dots y_n) - \sum_{y_1 \dots y_n} \omega(y_1 \dots y_n; x_1 \dots x_n) P(x_1 \dots x_n)$$

Here, $\omega(x_1 \dots x_n; y_1 \dots y_n)$ is the instantaneous transition rate from configuration $y_1 \dots y_n$ to configuration $x_1 \dots x_n$. Only one mutation occurs in any infinitesimal time interval and hence the configuration x differs from the configuration y at only one site. The $\sum_{y_1 \dots y_n}$ are sums over all configurations y differing from configuration x in (all) single positions. This of course does not mean that the sites are evolving independently. The above single site, independent Kimura model, or any other discrete state standard evolution model which assumes independence, is recovered when the transition rates $\omega(x_1 \dots x_n; y_1 \dots y_n)$ depend only on single sites or are constants. The Master Equation is the most general expression possible for discrete state evolution models, and it encompasses non-independent evolution assumptions by choice of a suitable $\omega(x_1 \dots x_n; y_1 \dots y_n)$, as we illustrate below. If sites do not evolve independently (where analytic solutions are possible), then a numerical solution of the Master Equation via Monte Carlo is possible – a method of solution familiar to physicists in Monte Carlo analysis of interacting spin systems.

2.3. *Defining the transition rates for interacting sites.* Assume that an "interaction energy" function, $E()$, exists which defines the energy of a sequence based on pairwise interactions. To motivate the concept of such an interaction energy for protein sequences one may think of the classic pairwise "contact potentials" used in threading and inverse folding investigations [Sippl (1990), Sippl (1993)] however the arguments given below are independent of the exact form of the potential. Such pairwise "contact potentials" are of proven utility in relating sequence to structure. A pairwise potential based on an assumed energy of interaction provides the simplest possible model of evolution with interacting sites and thus provides the simplest possible generalization beyond the standard assumption of independent evolution of sites. Transition rates are related to the energy, $E()$, of configurations by a standard argument from statistical mechanics [Stanley (1971), Glauber (1963)] which we won't repeat in detail here. We remark that it is clear that such a relation should exist from the following two observations:

- In equilibrium, where $\dfrac{dP(x_1, x_2, \dots, x_n)}{dt} = 0$, each configuration will occur with the Boltzman probability $\propto exp(-E)$.
- In equilibrium, where $\dfrac{dP(x_1, x_2, \dots, x_n)}{dt} = 0$, the Master Equation yields a relation between the transition rate ω and the equilibrium probability, which involves E.

The energy defined by *contact potentials* used in inverse folding/threading problems (incorporating simple physico-chemical characteristics of pairwise amino acids interactions) motivates the form of E we will explore below, however

the formalism developed here is not limited to such potentials. In analogy to pairwise contact potentials, we define a model energy as: $E = \sum_{ij} P(A_i^\alpha, A_j^\beta) C_{ij}$ where $P(A_i^\alpha, A_j^\beta)$ is a fixed potential matrix defining the interaction energy of amino acid α at site i and amino acid β at site j. In inverse folding/threading investigations the $20 * 20$ symmetric matrix, P, is derived from the statistics of contacting amino acids observed in x-ray crystal structure data [Sippl (1990)]. In our model simulations, below, this $20 * 20$ matrix is chosen to have random elements between -1 and 1. C_{ij} is a "contact matrix" describing the structure of a "protein' , with element $C_{ij} = 1$ if the amino acid at site i is in contact with the amino acid at site j, and zero otherwise. In inverse folding/threading investigations "contact" is typically defined by a condition such as: the distance between the C^α atoms of residue i and residue j is less than 8 Angstroms. In our simulation C_{ij} was chosen to be a random, symmetric matrix of zeros and ones, with the average number of "contacts" per "amino acid" user specifiable. Contact potentials derived from inverse folding studies, and contact matrices derived from x-ray crystal structures of real proteins will be investigated in later work.

3. Phylogenetic effects. Sequences related by a phylogenetic tree do not constitute *i.i.d* samples. Hence estimation of pairwise probabilities by a frequency counting approximation, resulting from a maximum likelihood analysis which (falsely) assumes independence of the sequence samples, can be biased. Note that there are two uses of the concept of "independence" in this paper: (1) the assumption that the individual biological sequences are *i.i.d*, and (2) the assumption that individual positions in the sequences evolve independently, i.e. with no interaction between the positions. Of course, these two uses are quite different, and should not be confused.

In this section we present a null-model approach to handle phylogenetic bias in estimation of covariation and validate 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 of Section 2. A histogram is compiled for the resulting mutual information values which are calculated between all pairs of sequence positions. These 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). At a bifurcation point of the tree the state of the sequence is duplicated, and the two copies are

subsequently independently evolved.

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

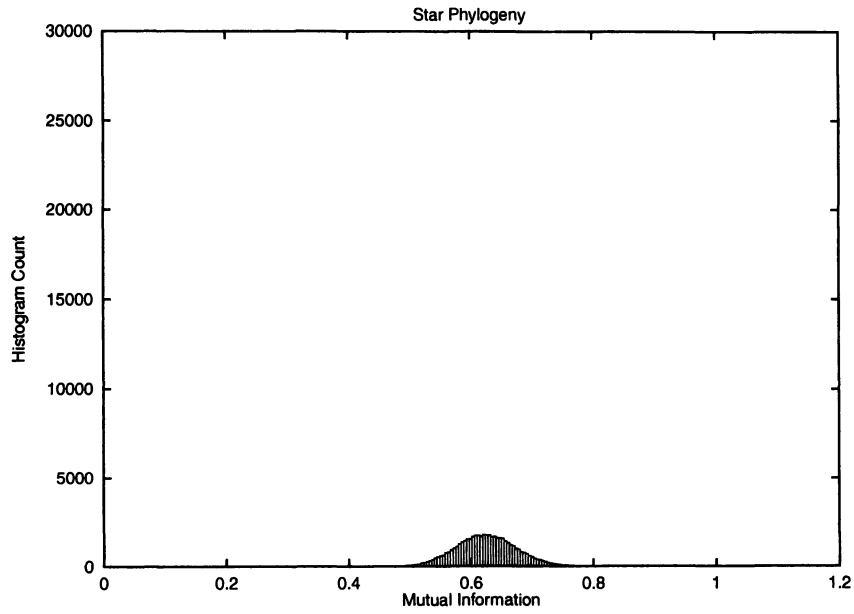

FIG. 1. *Histogram of mutual information between all pairs of sites in 256 "amino acid" sequences at the leaves of a star phylogeny. The sequences are 100 "amino acids" long. The total branch length from the root to the leaves is 64 time units. Ten separate runs with different random root sequences have been averaged to create the histogram.*

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. This procedure is introduced in Section 4.

3.1. *The null model: Independent evolution of sites down a given phylogenetic tree.* Consider a model simulation in which 100 amino acid long sequences are evolved using a Kimura style independent site evolution model [Kimura (1980), Kishino et al. (1990)] (see Section 2), with mutations occurring independently in different positions. Any amino acid can mutate with equal probability to

any other amino acid. We show that non-trivial phylogenetic trees "create" mutual information between sites, even when explicit interactions between sites is absent. This is because the topology of the tree magnifies the effects of finite sample size on estimation of mutual information. In the simulation we will consider a binary "tree" of 8 levels with each branch length 8 time units long,

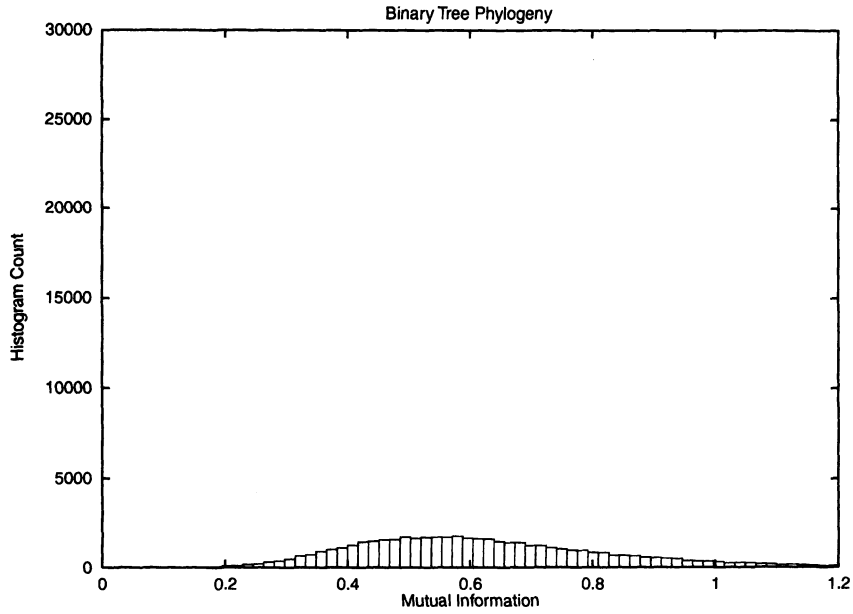

FIG. 2. *Histogram of mutual information between all pairs of sites in 256 "amino acid" sequences at the leaves of a tree phylogeny which is a balanced binary tree of eight levels. The sequences are 100 "amino acids" long. The individual branch lengths are of length eight, resulting in a total branch length from root to leaves which is identical to the star phylogeny of Figure 1 i.e. 64 time units. Ten separate runs with different random root sequences have been averaged to create the histogram.*

resulting in 256 different "amino acid" sequences of length 100 at the leaves. The total time from root to leaves is 64 time units. We use a binary tree merely as a familiar example of a tree with non-trivial topology which can be compared to a star phylogeny having a trivial topology. Phylogenetic trees of actual biological sequences will generally not be balanced binary trees and details will differ depending on the detailed tree topology. We will calculate the mutual information between all pairs of positions of the sequences at the nodes of the tree, and compare this calculation to that of a star phylogeny with 256 children and total branch length equal to that of the binary tree above, i.e. 64 time units. To evolve a sequence down a given phylogenetic tree the state of the sequence is duplicated at each bifurcation point of the tree, and the two copies are stochastically and independently evolved from the common ancestor.

The histogram of mutual information for the star phylogeny and the tree phylogeny are shown in Figures 1 and 2 respectively. It may be observed that there is a non-zero probability of achieving higher mutual information values (even though all sites are evolving independently) in the tree phylogeny, as opposed to the star phylogeny. A null model threshold based on the star phylogeny, i.e. based on an incorrect threshold which results from ignoring the real phylogenetic tree, would be too low and result in false conclusions of non-independence. On the other hand, if a null model threshold is chosen based on the correct phylogenetic tree, then sites which were truly independent, but evolved down the given tree and hence associated with an amplification of finite sample size effects, will be detected as being independent. This is quantified in the following section, where we introduce a specific interaction between sites, and show by explicit simulation that the specificity of predicting non-independent sites by evaluation of mutual information is increased when knowledge of the correct phylogenetic tree is used to create the null model.

3.2. *Validating the null model.* Various attempts to "weight" sequences in a manner related to the tree to correct for bias exist in the literature [Altschul et al. (1989), Sibbald & Argos (1990), Gerstein et al. (1994), Hennikoff & Hennikoff (1994), Thompson et al. (1994)]. However, to our knowledge such approaches have not been validated in model simulations where the interaction between designated sites is under the investigator's control. Here, we validate the null-model approach described above, in a model world where we can test the ability to predict interacting sites based on observed correlations. To create the model world:

(a) select a phylogenetic tree, here a balanced binary tree of eight levels with 256 leaves, such as used in Figure 2. Various branch lengths of the tree will be considered in separate runs, ranging from extremely short lengths, to lengths that are sufficiently long to have sequences evolve to equilibrium.

(b) evolve sequences via Monte Carlo using a *non-independent* model (see Section 2) with a selected C_{ij} and potential matrix P to generate sequences which play the role of "sequences observed in Nature", and which have sites that are truly interacting. The connection matrix for results reported here has every "amino acid" contacting three other amino acids chosen at random. The potential matrix P for results reported here was chosen to be a symmetric, twenty by twenty matrix, with elements chosen at random from a flat distribution between negative one and one.

(c) Calculate the mutual information between pairs of sites in these "real" sequences.

Next, we create the null model of "a given tree and independent evolution of

sites" by evolving sequences via Monte Carlo down the selected tree (always assumed known to the investigator), but using the independent model of evolution where each "amino acid" has equal probability to mutate to any other "amino acid". We determine the threshold of the null model to be such that mutual

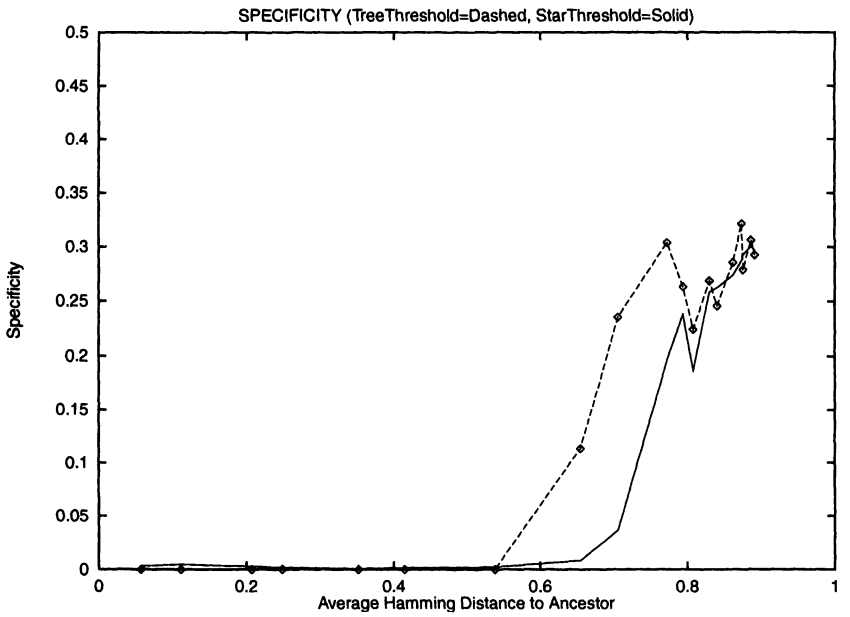

FIG. 3. *Specificity of the prediction of contacts based on over-threshold values of the mutual information is plotted as a function of the average Hamming distance of the 256 children to the root. Length 100 "amino acid" sequences were evolved down a phylogenetic tree with a binary tree topology of eight levels (256 child sequences). Dashed curve: threshold determination of the null model was correctly based on the binary tree topology. Solid curve: threshold determination was incorrectly based on a star phylogeny. If the topology of the balanced binary phylogenetic tree is ignored and a star phylogeny is incorrectly assumed, then the specificity of predictions is seen to be significantly lower than that achieved using a null model which correctly incorporates the phylogenetic tree.*

information values which exceed the threshold are very unlikely to have been generated by the null model (i.e., the threshold is in the far tail of the histogram of mutual information values). Hence, mutual information values for pairs of positions as calculated in the "real" data which exceed threshold are very unlikely to have been generated by the null model of independence of mutations.

Mutual information values between pairs of positions i and j in the "sequences observed in Nature" which are over the null model threshold are predicted to have contact matrix element, $C_{ij} = 1$, i.e. are predicted to be spatially close. To verify the predictions, an "experiment" may be performed in the model world to determine the "real" values of C_{ij}. Of course, this experiment is as simple as viewing the file containing the original values chosen for C_{ij} in

step (b) above, which was used in the Monte Carlo evolution that generated the "sequences from Nature".

This procedure results in "specificity" and "sensitivity" plots, Figures 3 and 4 for the prediction of $C_{ij} = 1$. Specificity is defined to be the percentage of predicted contacts that were actual contacts, i.e. that were defined in step (b) above to have $C_{ij} = 1$. Sensitivity is defined to be the percentage of actual contacts that were predicted to be contacts. Figures 3 and 4 show the result of numerous runs with varying branch lengths ranging from short (one time unit) to long (five hundred time units). The two separate extremes of very short branch lengths where essentially no evolutionary mutation takes place, and very long branch lengths where equilibrium can be reached within one branch of the tree, show little difference as should be expected. For intermediate branch lengths the effects of the phylogenetic tree become evident.

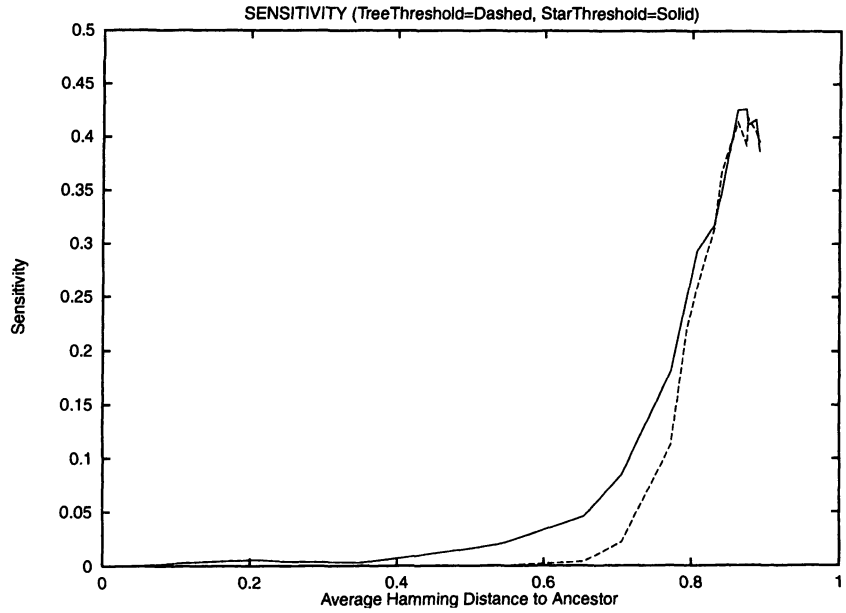

FIG. 4. *Sensitivity of the prediction of contacts based on over-threshold values of the mutual information is plotted as a function of the average Hamming distance of the 256 children to the root. Length 100 "amino acid" sequences were evolved down a phylogenetic tree with a star topology of 256 child sequences and total branch length equal to that of the binary tree phylogeny. Dashed curve: threshold determination of the null model was correctly based on the binary tree topology. Solid curve: threshold determination was incorrectly based on a star phylogeny. Choosing a threshold based on a null model correctly incorporating the binary phylogenetic tree, as opposed to incorrectly assuming a star phylogeny, enhances specificity at the expense of sensitivity.*

Note that choosing the threshold of the null model by using the topology of the given phylogenetic tree does significantly improve specificity, compared to a

null model threshold determined by ignoring tree topology and naively using a star phylogeny. However a significant number of errors clearly still remain. These are addressed in the following section. The sensitivity for predicting correct contacts is decreased, see Figure 4, if the correct phylogenetic tree is incorporated into the construction of the null model. A decrease in sensitivity with enhanced specificity, is preferable to enhanced sensitivity at the expense of specificity. In Section 4 we introduce a new methodology to use the observed first and second order moments to predict physical contacts which is not based on simple thresholding of mutual information or correlation measures.

4. Structural effects. The origin of the specificity errors in the simulation investigated in Section 3 are due to a covariation versus causation phenomenon. 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 at a distance" or "order at a distance" in the analysis of interacting spin systems [Stanley (1971), Binney et al (1992)]. How can one disentangle causation (direct physical interaction) from chained covariation (order at a distance)? We present in the following a maximum entropy approach to this problem.

4.1. *Maximum entropy analysis.* Although protein sequences can be hundreds of amino acids long, typically a much smaller number of amino acids display significant covariation, perhaps on the order of ten to twenty amino acids depending on the length of sequence examined. In this section we will consider, for reasons of simplicity only ten potentially interacting sites, and we will also restrict consideration to two-state "amino acids". The algorithms developed here scale reasonably with the number of potentially interacting amino acids, and with the number of states per "residue" (see below), but in general the algorithms require a 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. However, the algorithms do have a practical scaling behavior and are applicable to larger systems.

Consider the simplest situation of evolution down a star phylogeny. Phylogenetic tree effects due to more complicated tree topology can also easily be accommodated. Consider a star phylogeny with five hundred leaves containing two-state sequences of length ten, obtained by evolution to near equilibrium (10000 time steps) using the following connectivity matrix where on average each site is connected to three other sites:

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

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. Allowing more general potential matrices for two-state systems merely has the effect of adding new terms to the energy that are linear in the spins ("external magnetic fields" in spin language), which serve only to bias the final equilibrium single site probabilities and do not illuminate chaining phenomena.

The correlation matrix, $< (x_i - \overline{x}_i)(x_j - \overline{x}_j) >$, and the contact matrix are represented graphically in Figure 5. 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 achieved a correlation (absolute value) over 0.3. This threshold value for representation was chosen because such values occurred less than one time in one hundred, as computed in one hundred simulations of evolution down the same star phylgeny, but using independent evolution of sites. 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.

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 Figure 5 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). In larger systems one can have extended chains. Correlation between sites (2,8) is also significant even though they are not physically connected. 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 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

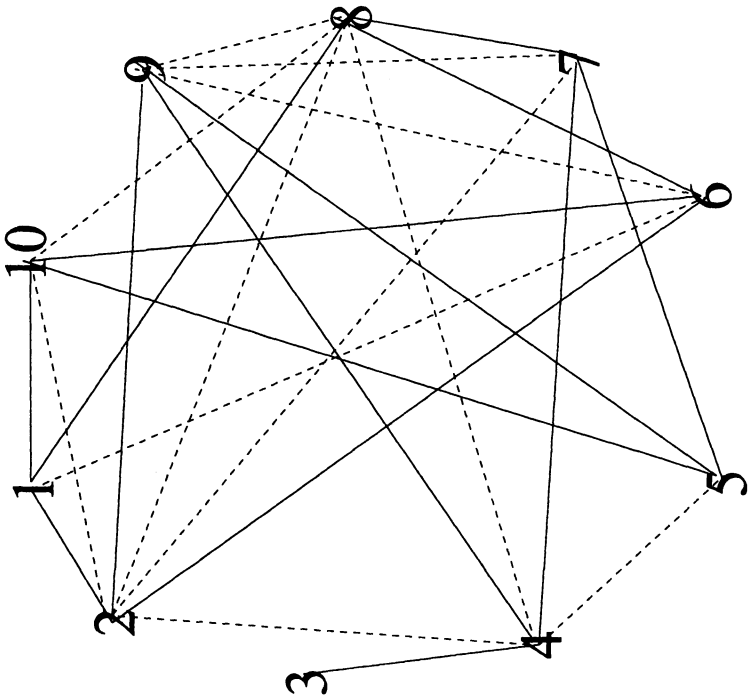

FIG. 5. *Correlations (dashed lines) between sites such as (1,6) and (2,8) occur because of the chaining effect. Sites (1,8), (1,2) and (2,6) have physical connections (solid lines). Sites (1,6) and (2,8) are not connected, yet still exhibit statistically significant correlation. Statistically significant values of correlation were computed by performing one hundred simulations of evolution down the same star phylogeny, but with independent evolution of sites. A correlation greater than* 0.3 *(absolute value) has less than a one in one hundred chance of occuring if sites are truly independent.*

not physically connected, because of a common "driving cause", which is the connection of both site (2) and site (8) to site (1). Such chained correlation effects are examples of the classic conundrum of "covariation versus causation".

It is clear that in situations where there can be extended interactions between sites, such as in proteins (but not so much in RNA, where once a base pair forms, it is unusual to form many other base pairs), then extended, complicated chains can occur which leads to correlations between sites that are not physically connected. Attempting to predict the connectivity matrix based on the correlation matrix would in general result in poor specificity. One must disentangle the chains of covariation and solve, in our specific case at least, the classic conundrum of causation versus covariation. As we show below, an effec-

tive solution is to estimate the parameters of the simplest model probability distribution which yields the observed correlations.

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

Classic Solution: Maximizing the entropy subject to the constraints of given first and second moments results in the classic [Tikochinsky et al. (1984), Levine & Tribus (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. For related investigations with a similar functional form see [Heumann et al. (1995)].

4.2. *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:

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

Note that even though the correlation between non-connected sites, such as (1,6) and also (2,8) can be high, that the reconstructed parameter values above are generally low between non-connected sites, and are high (bold face values) between connected sites. The maximum entropy procedure identifies the dashed lines of Figure 5 as a chaining phenomenon. 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).

4.3. *Practical considerations.* 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 & Thomas (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, followed by a small change in the $\lambda's$ in the gradient direction, followed by Monte Carlo to evaluate the new expectations etc. It may be seen by direct differentiation of F with respect to the $\lambda's$ that this process converges 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 et al (1992)], and now popular in bioinformatic investigations [Lawrence et al. (1993)], are an accurate and efficient alternative to exhaustive enumeration of all states.

4.4. *Conceptual considerations.* An assumed pairwise interaction energy may not accurately model the *fitness function* that is optimized in Nature. We emphasize that although we have motivated our discussion of correlated mutations using analogies to pairwise contact potentials (because we believe that some aspects of fitness are related to the match of sequence to structure as represented in pairwise contact potentials), the formalism is not limited to the protein contact potentials in use today. Indeed, the second order interactions we allow are perfectly general. In the maximum entropy formalism the second order interactions are determined by the observed correlations and as such provide the first logical step beyond independence of sites.

For practical reasons, it will be of interest (but is not necessary) to pursue the

utility of pairwise contact potentials in evolutionary analysis by using the form of pairwise potential matrices, P, to restrict the variability of the λ_{ij} parameters above. This can be done by assuming that the λ_{ij} parameters are the product of a fixed twenty by twenty amino acid interaction matrix (related to the potential of pairwise contact potential investigations), P, multiplied by a variable contact matrix, C_{ij}. 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}. We note that there is an assumption that the mutations do not change the C_{ij}, i.e. that the protein backbone remains relatively unchanged in spite of the amino acid mutations. Examples of this abound in Nature, including e.g. the hundreds of variable globin sequences which share a well conserved backbone structure.

Other issues deserving investigation include:

- the effects of mis-specification of the assumed model for the probability of a sequence: suppose that the "true" fitness function according to which real sequences are evolved in Nature includes, e.g., third order terms in addition to second order terms, and hence these terms will influence the observed values of second order correlations. If one observes just second order correlations then how accurately will the second order terms of the fitness function be recovered in a model which ignores third order terms? In other words, how "structurally stable" is the formalism?
- robustness of the reconstructed parameters to noise or sampling error in the original estimation of the moments from data: limited data will produce errors in the estimated moments. How stable are the reconstructed parameters to the presence of such errors?
- the assumption of evolution to equilibrium: if sequences in Nature are observed at times before equilibrium is reached, how will this affect the reconstructed parameters (obtained under an assumption of equilibrium)?

5. Conclusions. We have addressed two issues in covariation analysis of biosequences, (1) the effect of nontrivial phylogenetic trees on the estimation of mutual information, and (2) the effect of protein structure on propagation of correlations to sites that are not structurally linked. Regarding issue (1): Naive application of covariation analysis to biological sequences related by a phylogenetic tree can give misleading results. A non-trivial phylogenetic tree can amplify finite sample size fluctuations, making it appear that significant covariation exists between pairs of sites, when in fact all sites are evolving in-

dependently of each other. A null model procedure was introduced to address this problem. Regarding issue (2): Covariation between disconnected pairs of sites in sequences can result from possibly long chains of co-variation and from "common cause" effects, and not from causation (i.e. not from structural links). Chained covariation makes the prediction of structural links using naive application of covariation analysis prone to error. A technique involving maximum entropy reconstruction of the parameters for the probability distribution of the sequences was developed, and was validated in model simulations where accurate recovery of the structural links was achieved.

We remark that additional errors will probably remain even after addressing phylogenetic and 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, addressing the phylogenetic and chaining effects should go a long way towards improving accuracy of prediction of spatial contacts in families of varying protein sequences.

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 the conceptual foundation for protein structure determination via analysis of covarying mutations, by using models describing the probability distribution of the sequences as a whole, and by constraining the probability distributions with simplicity criteria such as maximum entropy.

Acknowledgments. The authors would like to thank the Santa Fe Institute, where part of this work was performed.

REFERENCES

ALTSCHUL, S., CARROLL, R. and LIPMAN, D. (1989). Weights for data related by a tree. *Journal of Molecular Biology* **207** 647–653.

BINNEY, J., DOWRICK, N., FISHER, A. and NEWMAN, M. (1992). *The Theory of Critical Phenomena.* Oxford University Press, Oxford.

COVER, T. and THOMAS, J. (1991). *Elements of Information Theory.* Wiley Series in Telecommunications, New York.

CLARKE, N. (1995). Covariation of residues in the homeodomain sequence family. *Protein Science* **4** 2269–2278.

DAYHOFF, M., SCHWARTZ R. and ORCUTT, B. (1978). 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.

GERSTEIN, M., SONNHAMMER, E. and CHOTHIA, C. (1994). Volume changes in protein evolution. *Journal of Molecular Biology* **235** 1067–1078.

GLAUBER, R. (1963). Time-dependent statistics of the Ising Model. *Journal of Mathematical Physics* **4** 294.

GOBEL, U., SANDER, C., SCHNEIDER, R. and VALENCIA, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, Genetics* **18** 309–317.

GRIBSKOV, M., MCLACHLAN, A. and EISENBERG, D. (1987). Profile analysis: Detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the USA* **84** 4355–4358.

GUTELL, R.R., POWER, A., HERTZ, G.Z., PUTZ, E. and STORMO, G.D. (1992). Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Research* **20** 5785–5795.

HASEGAWA, M. and FUJIWARA, M. (1993). Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor joining methods for estimating protein phylogeny. *Molecular Phylogenetics and Evolution* **2** 1–5 .

HENNIKOFF, S. and HENNIKOFF, J. (1994). Position-based sequence weights. *Journal of Molecular Biology* **243** 574–578.

HEUMANN, J., LAPEDES, A. and STORMO, G. (1995). Alignment of regulatory sites using neural networks to maximize specificity. In: *Proceedings of the 1995 World Congress on Neural Networks II*, 771–775.

HILLIS, D., MORITZ, C. and MABLE, B. (1995). *Molecular Systematics* (second edition). Sinauer Associates Inc., Sunderland, MA.

JUKES, T. and CANTOR,C. (1969). *Evolution of protein molecules.* In Munro, H. (ed.) *Mammalian Protein Metabolism*, Academic Press, New York.

KIMURA, M. (1980). A simple method of estimating evolutionary rate of base substitutions through comparative analysis of nucleotide sequences. *Journal of Molecular Evolution* **16** 111–120.

KIMURA, M. (1983). *The Neutral Theory of Evolution.* Cambridge University Press, Cambridge, England.

KISHINO, H., MIYATA, T. and HASEGAWA, M. (1990). Maximum likelihood inference of protein phylogenies and the origin of chloroplasts. *Journal of Molecular Evolution* **31** 151–160.

KORBER, B., FARBER, R., WOLPERT, D. and LAPEDES, A. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proceedings of the National Academy of Sciences of the USA* **90** 7176–7180.

KROGH, A., BROWN, M., MIAN, I., SJOLANDER, K. and HAUSSLER, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology* **235** 1501–1531.

LAWRENCE,C., ALTSCHUL, S., BOGUSKI, M., LIU, J., NUEWALD, A. and WOOTON, J. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* **262** 208–214.

LEVINE, R. and TRIBUS, M. (1979). The maximum entropy formalism. *Physical Review* **106** 620.

NEHER, E. (1994). How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences of the USA* **91** 98–102.

SHINDYALOV, I., KOLCHANOV, N. and SANDER, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering* **7** 349–358.

SIBBALD, P. and ARGOS, P. (1990). Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *Journal of Molecular Biology* **216** 813–818.

M. J. SIPPL (1990). 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** 859–883.

M. J. SIPPL (1993). Boltzmann's principle, knowledge-based mean fields and protein folding. *Journal of Computer-Aided Molecular Design* **7** 473–501.

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

TAYLOR, W. and HATRICK, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Engineering* **7** 341–348.

THOMAS, D., CASARI, G. and SANDER C. (1996). The prediction of protein contacts from multiple sequence alignments. *Protein Engineering* **9** 941–948.

THOMPSON, J. HIGGINS, D. and GIBSON, T. (1994). Improved sensitivity of profile searches through the use of sequence weights and gap excision. *CABIOS* **10** 19–29.

TIKOCHINSKY, N., TISHBY, N. and LEVINE, R. (1984). Alternate approach to maximum entropy inference. *Physical Review A* **30** 2638–2644.

THEORETICAL DIVISION
LOS ALAMOS NATIONAL LABORATORY
LOS ALAMOS, NM, 87545
ASL@LANL.GOV
LIU@LANL.GOV

SANTA FE INSTITUTE
1399 HYDE PARK ROAD
SANTA FE, NM 87501
ASL@SANTAFE.EDU

SERVICE PHYSIQUE THÉORIQUE
DSM, C.E.N. SACLAY
GIF/YVETTE, FRANCE 91191
GIRAUD@SPHT.SACLAY.CEA.FR

MOLECULAR, CELLULAR AND DEVELOPMENTAL BIOLOGY
UNIVERSITY OF COLORADO
BOULDER, COLORADO
STORMO@BEAGLE.COLORADO.EDU