Abstract: We report on the continuing research program at the Santa Fe Institute that applies complex systems methodology to computational molecular biology. Two aspects are stressed here: (1) the use of coevolving adaptive neural networks for determining predictable protein structure classifications, and (2) the use of information theory to elucidate protein structure and function. A “snapshot” of the current state of research in these two topics is presented, representing the present state of two major research thrusts in the program of Genetic Data and Sequence Analysis at the Santa Fe Institute.

INTRODUCTION

We address two topics: (1) a novel algorithm for using coevolving, adaptive networks to define and predict new classes of protein secondary structure, and (2) the use of concepts from information theory to elucidate protein structure and function.
The first topic describes the construction of a neural network algorithm that uses two coevolving neural networks to create new definitions of protein secondary structure that are highly predictable from primary sequences. Accurate prediction of the conventional secondary structure classes: alpha helix, beta chain, and coil, from primary sequence has long been an important, unsolved problem of computational molecular biology, with many ramifications, including multiple sequence alignment, prediction of functionally important regions of sequences, and prediction of tertiary structure from primary sequence. Our ability to use coevolving adaptive networks to evolve new and highly predictable definitions of secondary structure represent an example of the utility of new notions of complex systems theory, such as coevolution. This work was performed in collaboration with Robert Farber (External Faculty, Santa Fe Institute and staff member, Complex Systems Group, Los Alamos National Laboratory), and Evan Steeg (Department of Computer Science, University of Toronto, Toronto Canada).

The second topic concerns the use of information theory to detect correlations among positions in protein sequences. In earlier work we used the mutual information between codon positions in exons to define new features that allowed a neural network to distinguish between coding and noncoding regions of DNA with high accuracy. Subsequently, we applied the concept of mutual information to a set of aligned sequences of the V3 loop of HIV-1. Statistically significant correlations, as evidenced by high-mutual information values were observed in positions that were widely separated along the sequence, and experimental evidence shows that a subset of these positions are functionally linked. One hypothesis that would account for increased mutual information between functionally linked yet distant positions (as measured along the sequence), is that they are actually close in three-dimensional space. We present evidence from examples of protein secondary structure elements, such as alpha helices and beta chains, that structural constraints of protein secondary structure can be reflected in correlations between sequence positions. Evidence from beta sheets also shows that tertiary effects of protein structure can be reflected in weak correlations between nonlocal sequence positions. These statistically significant correlations indicate that ghosts of tertiary structure information are manifested in sequence data as weak, nonlocal correlations. Work is continuing to remove phylogenetic artifacts due to shared ancestry that can cause spurious correlations among sequence positions that are not proximate in space. This is work in collaboration with Bette Korber (Santa Fe Institute and staff member, Theoretical Biology Group, LANL), Robert Farber (external Faculty, Santa Fe Institute and staff member, Complex Systems Group, LANL), and David Wolpert (postdoctoral fellow, Santa Fe Institute).
COEVOLVING NETWORKS FOR DEFINING AND PREDICTING PROTEIN SECONDARY STRUCTURE

PREDICTION OF CONVENTIONAL SECONDARY STRUCTURE CLASSES

Prediction of secondary structure classes of proteins from amino acid sequence has evolved from attempts to construct a useful tool that can, e.g., aid prediction of protein tertiary structure, to a "numbers game," where researchers employ increasingly sophisticated algorithms to achieve incremental improvements in accuracy. With due respect to those researchers who have tried (and we are in that category), the bottom line is that presently no one can predict protein secondary structure with sufficient accuracy to be of much use, and it is immaterial whether the Q3 coefficient is, e.g., 62% or, say, 68%.27

The "secondary structure" of proteins are those classifications of structure that can be defined using only a local stretch (a short "window") of structural information about the protein. Structural information is available in databases like the Brookhaven database which contains structures of many proteins determined from x-ray diffraction. There have been numerous attempts to predict these locally defined secondary structure classes using only a local window of sequence information. It has become conventional to use the Kabsch and Sander definitions/software15 to define three classes of secondary structure: alpha helix, beta strand, and a default class called random coil. The prediction methodology ranges from a combination of statistical and rule-based methods4 to neural net methods.17,24,27

A major reason that prediction of secondary structure is of interest is that a successful prediction of secondary structure from amino acid sequence may be used in tertiary structure prediction algorithms to constrain their search space.26 For example, Skolnick26 has found that biasing amino acids towards assuming the measured secondary structure, when coupled to his global tertiary structure prediction codes, greatly increase the agreement of the global tertiary structure prediction with the experimentally determined structure. However, his test of the value of knowing the secondary structure classes used the actual, experimentally determined secondary structure classes, and not error prone predictions of secondary structure classes. His method, and others, are not successful if they attempt to use predictions of secondary structure classes at the current level of inaccuracy.

A widely accepted definition of protein secondary structure classes is that of Kabsch and Sander.15 Their definitions are implemented in a software that is widely available. In Figure 1 I illustrate the Kabsch and Sander software defining secondary structure classes, depicted as a "black box" on the right, and also a neural network that attempts to learn the secondary structure classes from the from amino acid sequence on the left. The Kabsch and Sander "black box" first defines hydrogen bonding patterns from the structural information, and then uses the hydrogen bonding patterns to define classes of secondary structure. This picture represents the standard approach to training a neural network to classify secondary structure from amino acid sequence.27 A local window of structure information obtained from,
e.g., x-ray diffraction data in the Brookhaven database, is input to the right-hand Kabsch and Sander black box. The box outputs the secondary structure class of the fragment, using the Kabsch and Sander definitions. For example, if one were dichotomizing all the windows of structure information into “alpha helix,” “not-alpha helix,” then the right-hand box will emit a “1” if the fragment is alpha-helix, and emit a “0” otherwise. The left-hand neural network “sees” the corresponding window of sequence information as input, and attempts to adjust its synaptic weights so that the output neuron of the neural network agrees with the output state of the Kabsch and Sander black box. Hence, if the input sequence adopts an alpha-helix state according to Kabsch and Sander, then the output neuron of the network should change state to “1.” Conversely, an input sequence fragment not in an alpha helix should cause the state of the output neuron to change “0.”

We consider in this exposition just two classes of structure—the extension to multiclases is trivial, but will not be made explicit for reasons of clarity. We won't discuss details concerning construction of a representative training set, or details of conventional neural network training algorithms, such as back-propagation. These are well-studied subjects that are addressed by e.g., Storloz, Yuan, and Lapedes,27 in

\[ E = \sum_{P} (O^{(P)}_{\text{net}} - O^{(P)}_{K&S})^2 \]

Kabsch and Sanders rules map Phi,Psi to secondary structure.
Note: the secondary structure classes, α helices, B sheets, and coils were derived visually. These rules are not necessarily optimal for prediction.

Network which maps AA sequence to secondary structure.

**FIGURE 1** Neural net learns Kabsch and Sander rules.
the context of protein secondary structure prediction. We note in passing that one can clearly employ more complicated network architectures, more output neurons (e.g., three neurons for predicting alpha helix, beta chain, random coil) etc. (c.f., Kneller, Cohen, and Langridge,\textsuperscript{17} Qian and Sejnowski,\textsuperscript{24} and Skolnick, Yuan, and Lapedes\textsuperscript{27}).

DEFINITION AND PREDICTION OF NEW SECONDARY STRUCTURE CLASSES

The key ideas of this section are contained in Figure 2. In this figure the right-hand black box implementing the Kabsch and Sander rules is replaced by a second neural network. This right-hand neural network therefore sees a window of structural information, while the left-hand neural network sees the corresponding window of sequence information. Note that the right-hand neural network can implement extremely general definitions of secondary structure. For example, if the weights in the right-hand network are set to arbitrary values, then the right-hand network will correspondingly produce an arbitrary classification of the structures that are input to it. On the other hand, one could train the weights of the right-hand network to perform structure classification according to, say, the Kabsch and Sander rules. To demonstrate the generality of the procedure we have done the latter, and have successfully captured the Kabsch and Sander structural definitions in the right-hand network with high accuracy. The representation of the structure data in the right-hand network uses phi-psi angles. Problems due to the angular periodicity of the phi-psi angles (i.e., 360 degrees and 0 degrees are different numbers representing the same angle) are eliminated by utilizing the sine and cosine of each angle.

POINT (1). One can replace the right-hand black box of Figure 1 with a neural network (see Figure 2). A neural network on the right-hand side is an equally valid implementation of a set of rules defining secondary structure as a piece of software. We have explicitly demonstrated this by training a neural network to reproduce the Kabsch and Sander rules with high accuracy.

POINT (2). The right-hand network need not be restricted to implementing the Kabsch and Sander rules for secondary structure. The right-hand neural network is capable of representing a very general set of rules, of which the Kabsch and Sander rules are but one choice.

To define new rules one merely changes the synaptic weights. Arbitrary synaptic weights would define arbitrary rules, and there would be little chance that these new classes would be either predictable or meaningful.

POINT (3). A requirement on the rules is needed. The necessary requirement is that the "secondary structure" classes defined by the right-hand net be predictable from the corresponding amino acid sequence of the left-hand network.
In other words, the only requirement is that the synaptic weights be chosen so that the output of the left-hand network and the output of the right-hand network agree for each sequence-structure pair that is input to the two networks.

To achieve this, both networks are trained simultaneously, starting from random initial weights in each net, under the sole constraint that the outputs of the two networks agree for each pattern in the training set. The mathematical implementation of this constraint is described in various versions below. This coevolution of the two networks is clearly a more difficult computational problem than the conventional approach (Figure 1) that employs fixed targets. Each net now chases a moving target during training, and additional numerical difficulties occur. Our preliminary results (below) show that these difficulties are surmountable (we achieved interactive runtimes using the CM2 Connection Machine). Therefore this procedure is a very general, effective method of evolving predictable secondary structure classifications of experimental data.
COEVOLVING ADAPTIVE NETWORKS

The requirement that the two networks coevolve states that they evolve from random initial conditions into a cooperative phase, in which each network is able to predict the output of the other network. Neither network has a fixed target to which it may be trained—the only requirement is that the outputs of both networks agree for each pattern. A naive method to require that the two networks evolve weights allowing cooperation is suggested by analogy to conventional back-propagation. We present the naive method first, and then refine the method to an effective procedure in the following section. In back-propagation one performs gradient descent in the synaptic weights of the error function, \( E \):

\[
E = \sum_p (\text{Left}O^{(p)} - t^{(p)})^2
\]

where \( t^{(p)} \) is the target output value for the \( p \)th pattern, and \( \text{Left}O^{(p)} \) is the output of the left network for the \( p \)th pattern. \( \text{Left}O^{(p)} \) is a function of the synaptic weights. Gradient descent in the synaptic weights will decrease the error, \( E \), evaluated on the training set by forcing the output of the network, \( \text{Left}O^{(p)} \), to agree with the target output, \( t^{(p)} \), for each pattern.

Note that in Figure 1 the target value for the left-hand network is given by the fixed rules implemented in the right-hand Kabsch and Sander black box. These targets of the conventional approach, \( t^{(p)} \), are therefore fixed constants, i.e., either "0" or "1" for each pattern. In Figure 2 one might consider using the same error function, Eq. (1), but replacing the previously fixed target values for each pattern by the variable output of the right-hand network. Hence the new error function, whose minimization will enforce agreement of the left-hand and right-hand networks is

\[
E = \sum_p (\text{Left}O^{(p)} - \text{Right}O^{(p)})^2.
\]

The difficulty with this naive idea is that there is a trivial way for the two nets to agree. They merely need to decrease their synaptic weights to the inputs to zero, so each will stay in the "0" state, regardless of the input pattern (it is also possible to have each stay in the "1" state). The output of the two nets would agree, as demanded by minimizing Eq. (2), but the result is trivial. The outputs remain either "on" or "off" regardless of the input data, and are completely uninformative. One might consider adding a variance term to Eq. (2) to require the networks to respond to their inputs, i.e., to impose variation in the network outputs as the input patterns change. However, a cleaner approach is to demand that the outputs of the networks co-vary by modifying Eq. (2) to maximize the mutual information or correlation between the network outputs.
COEVOLUTION: TRAINING WITH CORRELATION MEASURES. The standard correlation measure, $C$, between two objects, $LeftO^{(p)}$ and $RightO^{(p)}$ is

$$C = \sum_p (LeftO^{(p)} - \bar{LeftO})(RightO^{(p)} - \bar{RightO})$$

where $\bar{LeftO}$ denotes the mean of the left net's outputs over the training set, and respectively for the right net. $C$ is zero if there is no variation, and is maximized if there is simultaneously both individual variation, and joint agreement. In our situation it is equally agreeable to have the networks maximally anticorrelated, as it is for them to be correlated. (Whether the networks choose correlation, or anticorrelation, is evident from the behavior on the training set.) Hence the minimization of the following expression will ensure that the outputs are maximally correlated (or anticorrelated)

$$E = -C^2.$$  \hspace{1cm} (4)

Minimizing this expression forces the correlation of the two outputs, considered as a set of real values, to tend to either perfect correlation or perfect anticorrelation. Note that Eq. (4), as opposed to Eq. (2), does not allow the situation of unchanging outputs to be a local minimum. This would give a value of 0.0 to $E$. However, $E$ in Eq. (4) will be negative under even the slightest correlation given the random initial weights, and the dynamics of gradient descent will continue to decrease $E$. Thus the system is forced away from unchanging outputs, solving the problem associated with the naive approach of Eq. (2) above.

PREDICTING WITH CORRELATION MEASURES. The procedure for predicting the structure of a new sequence pattern is different when the correlation measure is used for training. Because one explicitly trained the network using Eq. (4), then the output of the networks are only significant in relation to their mean value over the training set. It is not just the state of the left-hand net that determines the prediction of the secondary structure of the pattern on the right, but rather it is whether the state of the left-hand net is above or below its mean value in the training set. It is therefore necessary to subtract from the output of the right-hand network its mean value as calculated over the training set. This simple change to the usual method of prediction is an easy-to-implement offset to the value of the output.

OTHER TRAINING MEASURES, ALGORITHMS, AND ARCHITECTURES. Other training measures forcing agreement of the left and right networks may be used. Particularly suitable for the situation of many outputs (i.e., more than two classes discrimination) is “mutual information.” This version of the idea is closely related to the IMAX algorithm of Becker and Hinton. The mutual information is defined as

$$M = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_i p_j}$$

\hspace{1cm} (5)
where \( p_{ij} \) is the joint probability of occurrence of states \( i \) and \( j \) of the left-hand and right-hand networks, and the \( p_i \) and \( p_j \) are the marginal probabilities. In previous work\(^{27}\) we showed in general how \( p_{ij} \) and the marginals may be defined in terms of neural networks. Minimizing \( E = -M \) maximizes \( M \). Preliminary simulations show that \( M \) is more prone to local maxima than \( C \). Initializing the network weights using \( C \), and then switching to maximizing \( M \), is a procedure well worth testing in view of the useful properties of \( M \) (we won't discuss details of information theory in relation to sequence analysis here, see, e.g., Farber and Lapedes;\(^9\) Korber et al.;\(^{18}\) Lapedes et al.;\(^{20}\) and Storloz, Yuan, and Lapedes\(^{27}\)). Finally, we point out that since a common quantity measuring predictive performance is the Mathews correlation coefficient (see, e.g., Storloz, Yuan, and Lapedes\(^{27}\)), then it is reasonable to train the two networks to maximize this measure. The maximum achievable Mathews correlation coefficient is 1.0. The Mathews coefficient is designed to be a single number that incorporates measure of both over-prediction, and under-prediction. Intensive investigation of the effect of network architecture on the derived structural classes, using all the error measures, is in progress.

As noted earlier, this problem is extremely computer intensive, requiring use of the CM2 Connection Machine (on which we've achieved 3 gigaflop throughput in our preliminary investigations). An alternative classification algorithm may not only run faster, but may also uncover different structural classes. We have been working with Melanie Mitchell, of the Santa Fe Institute and University of Michigan, to develop the ideas presented here using "genetic algorithms,"\(^{13,21}\) which are a alternative machine learning algorithm to neural networks. Genetic algorithms are powerful adaptive algorithms that may have some advantages for this problem. Investigations are in progress.

RESULTS

Best results so far have been obtained with the Mathews objective function. Random initial conditions are necessary for the development of interesting new classes— if one uses initial conditions appropriate for predicting the standard Kabsch and Sander classes then the local minima is so deep that nothing much else happens. Naturally, one can "gang" together objective functions as soon as one gets the network out of the initial local minima. Thus, one can start training with the correlation objective function, and then finish with the more precise mutual information function.

Our best results so far involve two class discrimination using the Mathews correlation function. If one assigns the name "Xclass" (for want of a more descriptive word at this stage of the investigation) to the newly defined structural class, then the network already classifies local windows of structure into a "Xclass/NotXclass" dichotomy with higher predictability than prediction of conventional secondary structure classes.
The Mathews coefficient on a disjoint prediction set of the new classes is 0.425. The Q3 of the new classes, which we emphasize is not a particularly informative quantity, (but is often quoted) is 73%. (Note that Q3 can be essentially 100% for a ridiculously simple algorithm that classifies most examples as "Xclass," and then uses a default for prediction. This is not happening here.) For comparison, the Mathews coefficient for dichotomization into the standard secondary structure classes, Alpha/NotAlpha, Beta/NotBeta, and Coil/NotCoil, for the same data is 0.33, 0.26, and 0.39, respectively. Given the minimal amount of optimization we have performed so far, the 0.53 Mathews correlation coefficient of the new class dichotomization is most encouraging.

Are the new classes simply related to the more conventional classes of alpha helix, beta, and coil? Although more precise analysis needs to await visual examination of examples of the newly defined classes one can conclude immediately that the relationship is not necessarily simple. We classified (using the Kabsch and Sander definition) the conventional secondary structure classes for the newly defined classes. Thus, all patterns labeled "Xclass" by the new code were classified into alpha, beta and coil according to Kabsch and Sander definitions. The new classes turn out to be a mixture of the conventional classes, and are not dominated by either alpha, beta or coil; although there is some relationship between Xclasses and helices. It will be most interesting to see if structural features of the new classes, which we emphasize are more predictable from amino acid sequence than the Kabsch and Sander defined classes, exhibit striking visual features.

FIGURE 3 A segment of 2ACT, sulphydryl proteinase. First line is labeled H = helix, B = beta chain, C = coil. Second line is target Xclass. Third line is predicted Xclass.
In Figure 3 I compare the assignment of structural features into Xclass/ NotXclass categories, with the conventional assignment of structural features into alpha helix, beta chain, and coil, for the protein "2act—Actinidin."

Comparison of the second and third lines of Figure 3 illustrates the accuracy of XClass prediction for this protein.

Comparison of the third and first lines illustrates the relation between the conventional secondary structure classes and the new Xclass categories.

Note that the Xclass category of secondary structure bears some relationship to Helix, but that significant differences also exist. Clarification of the new classes awaits detailed analysis, including visual inspection using molecular modeling of examples of the new Xclass categories.

A primary goal of this investigation is to evolve very predictable secondary structure classes that can then be used to constrain tertiary structure prediction. The above results, although preliminary, are most encouraging. Our goal is to significantly improve accuracy still further.

INFORMATION THEORY ANALYSIS OF PROTEIN STRUCTURE AND FUNCTION
BEYOND CONSENSUS SEQUENCES

A common, and intuitive, approach to characterizing important regions of sequence data, e.g., those regions containing regulatory signals, is to attempt to find a motif of mostly contiguous, conserved sites across many aligned sequence examples that contain the region of interest. The motif disappears, and the usual assumption is that information about the region also vanishes, for sequence positions that vary so much across examples that the dominant symbol in these positions become uninformative. However, the mere existence of variation in a position doesn’t necessarily mean that information about the position is no longer characterizable. It is quite possible that variations in different positions are linked, and that although a single position might appear to be varying randomly, it is in fact varying in a correlated fashion with changes in another position. Correlations between real, i.e., floating point variables are easy to measure by the usual linear correlation analysis. Correlations between discrete variables can be analyzed by using the concept of mutual information from information theory.

In previous work we’ve used mutual information to quantify the degree of correlation between positions in sequence data. We found that there exists nontrivial and statistically significant mutual information between the neighboring codons of exons in DNA, which allowed us to develop neural net algorithms of great sensitivity that distinguished between exons and introns in unannotated DNA sequences. In other work we discovered correlations in transcriptional promoters of E. coli that also aided computational identification of these regions. In more recent work
we’ve analyzed a set of aligned sequences of the V3 loop of HIV-1 and discovered nontrivial, statistically significant mutual information between nonlocal sequence positions.\textsuperscript{18}

**MUTUAL INFORMATION**

A formal measure of variability\textsuperscript{19} at position \( i \) is the Shannon entropy, \( H(i) \). \( H(i) \) is defined in terms of the probabilities, \( P(s_i) \), of the different symbols, \( s \), that can appear at sequence position \( i \) (e.g., \( s = A, S, L \ldots \) for the twenty amino acids: Ala, Ser, Leu ...). \( H(i) \) is defined as:

\[
H(i) = - \sum_{s=A,S,L\ldots} P(s_i) \log P(s_i). \tag{6}
\]

Mutual information is defined in terms of entropies involving the joint probability distribution, \( P(s_i, s'_j) \), of occurrence of symbol \( s \) at position \( i \), and \( s' \) at position \( j \). The probability, \( P(s_i) \), of a symbol appearing at position \( i \) regardless of what symbol appears at position \( j \), is defined by \( P(s_i) = \sum_{s'_j} P(s_i, s'_j) \) and similarly, \( P(s'_j) = \sum_{s_i} P(s_i, s'_j) \). Given the above probability distributions, one can form the associated entropies:

\[
H(i) = - \sum_{s_i} P(s_i) \log P(s_i),
\]

\[
H(j) = - \sum_{s'_j} P(s'_j) \log P(s'_j),
\]

\[
H(i,j) = - \sum_{s_i, s'_j} P(s_i, s'_j) \log P(s_i, s'_j).
\]

The mutual information \( M(i, j) \) is defined as:

\[
M(i, j) = H(i) + H(j) - H(i, j).
\]

Mutual information is always non-negative and achieves its maximum value if there is complete covariation. The minimum value of 0 is obtained either when \( i \) and \( j \) vary completely independently, or when there is no variation.\textsuperscript{3,19}

The above formulae assume true probability distributions are known. In practice, however, they are not known and must be estimated from a finite data set. Two effects require consideration. First, since mutual information is always non-negative, the mutual information between any single pair of truly independent positions is consistently overestimated, while the mutual information of a covarying position can be either overestimated or underestimated, depending on the nature
of the fluctuations in the finite data set. One must therefore assess statistical significance of single pairs in the light of small sample bias. Secondly, one must consider problems caused by selection effects. Typically, one selects a pair of positions that exhibits a large mutual information value, compared to other pairs in the sequence, as "interesting." One must therefore assess the probability that out of many such estimated mutual information values (one for each pair of positions in the sequence) a high-estimated value might be achieved by chance.

THE V3 LOOP OF HIV-1

The V3 loop of the HIV-1 envelope protein (env) has been the focus of intense research efforts because it is a potent epitope for neutralizing antibodies (NAb)\textsuperscript{10,14,33} and T cells,\textsuperscript{12,28} and plays a role in determining cell tropism and viral growth characteristics. While there is some propensity to conserve amino acid side chain chemistry in the different positions in the loop this conservation often breaks down upon inclusion of phylogenetically distant viruses (12). Such variation presents a difficult challenge for those attempting to design broadly reactive V3 loop based vaccines.\textsuperscript{14,38} Our goal was to quantify the degree of covariation of mutations at different sites by analyzing the available database\textsuperscript{22} of V3 amino acid sequences using mutual information, a concept from information theory (15–18). All pairs of positions in an alignment of 308 distinct V3 loop sequences were compared.

An algorithm which employed multiple randomizations of the initial data set was used to determine the statistical significance of the estimated mutual information values using a very conservative measure that addresses both small sample bias and selection effects (for general reviews of methods of this type, see Efron\textsuperscript{6,7,8}). Highly statistically significant mutual information scores were obtained for several pairs of sites, some on opposite sides of the V3 loop.

High-mutual information between certain sites suggests that functional studies of the V3 loop using site directed mutagenesis may depend upon simultaneously altering amino acids on both sides of the loop. Indeed, this has been shown to be the case for some of the positions linked through mutual information analysis—de Jong et al.\textsuperscript{5} showed that simultaneous mutations were required at sites 10 in conjunction with sites 21 through 24, located across the loop, to get a complete conversion in viral phenotype from nonsyncytium inducing, low replicating, to syncytium inducing, high replicating. Our analysis indicated sites 10, 23, and 24 were covariant. Virus viability as well such phenotypic “switches” may require simultaneous mutations in covarying sites. When sites related by high mutual specific information were compared with alignments of V3 regions of viruses with distinct tropism and cytopathicity, several of the positions that appear to be significant in terms of phenotype were also seen to covary. This correlation supports the hypothesis that mutual information can identify functionally interactive sites. While several of the sites we predicted to be mutually interactive were substantiated by experimental evidence, additional linkages were observed that also may be relevant
for the generation of viable V3 loops with specific phenotypes. These positions may have been missed in experiments to date, due to their relative conservation among cloned samples used for experiments in culture.

CORRELATIONS AND PROTEIN STRUCTURE

One possible explanation for why positions with covarying mutations seem to be associated with functional sites, is that these covarying positions, although possibly distant along the sequence, are proximate in space. The closeness in physical space, due to an underlying conserved structure associated with the functional region, might constrain the sequence mutations sufficiently to result in covariation. There is a precedent for associating covariation with structure in analysis of families of variable RNA sequences. Recent work of Stormo et al.\(^{11}\) show that covarying positions in RNA families are associated with both secondary and tertiary structural features. A structural basis is but one possibility for why covarying sites may be associated with functional sites. Other possibilities include interaction with the protein’s environment such as specific requirements due to protein/protein interactions, or the necessity to specify certain types of amino acids at particular sequence positions to help define the folding pathway. We emphasize that the tertiary structure of the V3 loop is not known in detail, and that we are not suggesting that the observed correlations among positions in the V3 loop are necessarily a reflection of structural constraints. Never-the-less, it is tempting to investigate in a setting where structure is known, if structural and physico-chemical constraints give rise to correlations among sequence positions.

CORRELATIONS INDUCED BY PROTEIN SECONDARY STRUCTURE

Protein secondary structure elements, such as alpha helices and beta chains, define regular elements of protein structure.\(^{15}\) It is possible to extract from the Brookhaven data base numerous, examples of secondary structure elements with different sequences. We took a database of alpha helices (previously used to train a neural net to distinguish helices, beta chains, and coils\(^{27}\)) and computed the mutual information between all pairs of positions in a window 13-residues long, which was centered on each successive residue that participated in an alpha helix. In Figure 4 I represent the mutual information between positions in these windows of alpha helix. Two features are clear: there is correlation between positions at a spacing of two residues, and also between those residues at a spacing of four residues. The latter correlation is gratifying. Alpha helices, by definition\(^{15}\) involve residues that have hydrogen bonding along the backbone across intervals of four residues. The structural and hydrogen bonding constraints implicit in the definition of alpha helices therefore seem reflected in the correlations between residues spaced four apart. The correlations between residues at a spacing of two is less clear. One possible
explanation is that the database contains a significant component of amphiphatic helices. Such helices have one side facing solvent, and the other facing the protein interior. Therefore one expects alternation of hydrophobic and hydrophilic residues between one side of the helix and the other. Since the pitch of an alpha helix is approximately 3.7 residues per turn, the spacing of two residues between correlated residues is in approximate agreement. While this research was being performed, we received a preprint\textsuperscript{16} that included related research on correlations between residues in secondary structure elements.

A similar calculation can be performed for beta strands. The mutual information is presented in Figure 5. Clearly, there is higher mutual information between residues at a spacing of two residues. Here again, we have taken windows of sequence, 13 residues long, centered on successive residue participating in a beta chain. The higher mutual information between residues at a spacing of two residues
is in accord with structural constraints on beta chains. Beta chains have an alternating hydrogen bonding pattern along the backbone which is presumably being reflected in the correlations seen between residues at a spacing of two.

The absolute magnitude of the mutual information presented in these figures is not particularly informative. Finite sample effects cause a shift in the absolute magnitude of the mutual information depending on sample size. Also, for visual clarity, we have nonlinearly scaled the values to improve contrast for reproduction. However, relative comparisons of mutual information between positions in an alpha helix or a beta chain, respectively, are informative. Randomization experiments to test statistical significance (in analogy to the V3 loop analysis, above) were performed, and verified the significance of the enhanced mutual information at various spacings between residues.

**CORRELATIONS INDUCED BY PROTEIN TERTIARY STRUCTURE**

Beta sheets, composed of beta chains that are hydrogen bonded together (and hence proximate in space), provide an example of tertiary structural information being reflected in mutual information between nonlocal sequence positions. Beta sheets are intrinsically nonlocal objects as far as sequence considerations are concerned, but are extremely localized in space. For example, a sequence can adopt a beta chain configuration early in the sequence, subsequently change to an extended region of coil, then later in the sequence adopt another beta chain configuration, which happens to be parallel and close to (in three-dimensional space) the first beta chain. Two, or in general more than two, spatially close regions of beta chain, can support interchain hydrogen bonding which contributes to the tertiary stability of the protein. There is also the usual intrachain hydrogen bonding with a spacing of two residues (see above) within each individual beta chain. An analogy to the spatial positioning of beta chains, and the interchain bonding necessary to make a beta sheet, is the sticking together of distant regions of a piece of multiply folded cello-tape. One might hope to see correlations induced among those residues in each beta chain that may be well separated along the sequence, but which are brought into close spatial proximity by the interchain hydrogen bonding defining the combination of chains into sheets.

We constructed a database of paired, antiparallel, beta chains participating in a beta sheet, by extracting a centered window of residues, 13 residues long, around each position in one chain, and also for the respective partner in the other chain of the beta sheet. For example, if residues numbered 5, 6, 7, 8, 9 were participating in a beta chain, and residues 21, 22, 23, 24, 25 were the corresponding partner chain, then we construct a database consisting of a window around each residue from one chain concatenated with the corresponding window from the other chain. Thus, windows in each chain that are, e.g., three-residues long, result in concatenated windows that are six residues long. These six residue long windows would be the concatenation of the centered windows (a) and (b) below:
(a) 4,5,6  (b) 20,21,22  
(a) 5,6,7  (b) 21,22,23  
(a) 6,7,8  (b) 22,23,24  
(a) 7,8,9  (b) 23,24,25  
(a) 8,9,10  (b) 24,25,26

In Figure 5 I show the mutual information between the positions in the paired windows of antiparallel beta chains of length 13 (concatenated window length it therefore 26). There is the usual increased mutual information between residues at a spacing of two within in each chain. In addition, there is increased mutual information between the residues in separate chains that are in close spatially, but distant along the sequence. This is evidenced by the increased mutual information along the antidiagonal in the lower right-hand corner of the plot. This provides
evidence that spatial proximity limits the amino acids that may occur in nearby positions, resulting in statistically significant mutual information between those positions.

The next step will be to extend this analysis to families of proteins with variable sequence, and known crystal structure, such as globins, immunoglobulins, MHC molecules, kinases, serine proteases, etc. Success in this endeavour, could result in a formalism that could be powerful in defining components of functional or structural domains which are distant in terms of linear sequence, yet may be working coordinately in intact proteins. It could then be applied to proteins which have defied attempts at crystallization, and serve as a guide for molecular biologist who are mapping functional domains through exchange of restrictions site fragments and deletion and mutational analysis. Our preliminary tests on globin sequences identified several important functional sites, but the identification of structural correlates was complicated by the nonmonomeric structure of hemoglobin. It will be necessary to construct linked sequences expressing the full tetrameric structure.
of hemoglobin in order to test the structure hypothesis. Indeed, most of the variable protein families are nonmonomeric, and construction of the linked sequence segments is a nontrivial, but resolvable, issue.

CONCLUSION

Techniques from adaptive network theory, information theory, and concepts such as coevolution, form the core of research in complex adaptive systems. In this article we have shown how this core of new ideas can be used to give fresh insight into some outstanding problems in computational molecular biology. It seems only fitting that theories and techniques developed to analyze and understand a wide variety of complex systems, have as a prime example of their value, the successful analysis of life itself.

REFERENCES


