Protein Evolution as a Parallel-distributed Process: A Novel Approach to Evolutionary Modeling and Protein Design

Sylvia B. Nagl*
Department of Biochemistry and Molecular Biology,
University College London,
Gower Street, WC1E 6BT
London, UK

Proteins are complex adaptive systems. Their functional and structural units, termed domains, are conserved and recombined during evolution. Domains are thermodynamically stable and fold independently within the context of the whole protein, and can arguably be seen as stable units of evolution. New domain functions evolve within the constraints of maintaining thermodynamic stability and autonomous folding capability. This gives rise to a complex interplay of molecular organization and evolutionary dynamics, which is still a largely unexplored area of research. The major aim of this paper is to approach this problem from a perspective informed by recent developments in complexity theory. This work employs distributed representation by neural networks in modeling protein domain evolution.

1. Introduction

1.1 The changing vision of modern biology

The history of science has provided wide-ranging evidence for the power of conceptual shifts, simultaneously transforming accepted modes of inquiry and standards of explanation in a given discipline. From our contemporary perspective on biology, two such shifts stand out above all others. The first was the formulation of the theory of evolution, marked by the publication of Darwin's *The Origin of Species* in 1859 and the second occurred a century later with the birth of molecular biology, precipitated by Watson and Crick's determination of the structure of DNA in 1953. During the second shift, as biology changed its focus from organisms to biomolecules, biological systems were conceptualized as chemical and mechanical devices whose components were thought to interact in a relatively simple fashion. Consequently, the functioning of the organism was to be explained on the basis of the chemistry

*Electronic mail address: nagl@biochem.ucl.ac.uk.

and physics of its constituent parts, and molecular interactions were charted in mostly linear pathways. Concomittantly, the evolution of these constituent parts, in other words, molecular evolution, became an overriding concern in evolutionary thought. This "molecular vision of life" [1] has proved tremendously successful for the identification of a vast number of components and mechanisms present in biological systems, at higher and higher levels of resolution. The sequencing of the human genome, the exponentially growing number of three-dimensional protein structures determined by crystallography and NMR, and the biochemical elucidation of metabolic and signaling pathways in cells are among the successes of this research focus.

As these endeavors reach their culmination, we are already in the middle of a third conceptual shift that is rapidly gaining momentum. A new "global biology" is emerging that aims at understanding the systems properties of cells and organisms at various levels. At present, attention has become focused on the level of the genome (genomics), the level of "transcribed messages," that is, expressed genes (transcriptomics), and the level of the complete set of proteins (proteomics) [2–5]. Some of the questions raised are, for example: What can global gene expression patterns in model organisms teach us about the order and logic of the genetic "program?" How do genomes and other biological systems evolve? How do large-scale networks of molecular interactions integrate biological signals within cells? It is expected that, in the future, the systems properties of biomolecules will also become a focus for new research.

Of paramount importance to this current conceptual shift to "systems thinking" is the focus on the information processing properties of biological systems. Simplistically put, biological entities are now seen as information processing machines. The central concepts underpinning this change are biological information, signaling, and complex interaction networks. The most ambiguous of these is biological information; it can be equally applied to a stretch of DNA (a gene), a hormone, or the information content of a molecular structure, however one may want to measure this. These issues aside, most generally, explanations are sought as to how the spatial and temporal organization of mixed populations of molecules gives rise to the "smart" properties of biological systems. Such smartness is in evidence at all levels of organismal complexity. It is evidenced, at the level of single molecules, by signal-mediated activation of proteins by conformational change, for example, and at the level of complex networks, by the complexities of signal processing during mammalian development, and by consciousness itself.

We are faced with profound challenges, both of an empirical and theoretical nature, as we attempt to study the information processing properties of biological systems. It is becoming widely accepted that biological systems are not merely complicated, but that they are com-

plex. Whilst all biological processes are consistent with the physical and chemical laws of our universe, and in this sense can ultimately be "reduced" to chemistry and physics, there is a growing awareness that biological phenomena require an appproach that equally addresses the problem of emergence. How do living systems emerge from the laws of physics and chemistry? In complex systems, emergent phenomena result from the rule-governed, nonlinear interactions of a large number of components occurring in a highly context-dependent manner. To come back to the example of consciousness, it arises out of the unimaginably densely connected interactions of billions of neurons (and their constituent molecules), and is not a property of any one brain region, let alone of the neurons themselves. Consciousness is an emergent property of the brain as a whole. Beyond this special case, information processing can be seen as an emergent property of complex biological systems in general. The enormous task before us then is nothing short of identifying the rules of the underlying interactions and the logic embedded in the organization of living matter.

However, a theoretical framework and methodology for the investigation of complexity and emergence in biology are still largely undeveloped. Whilst the study of complex systems has undergone vigorous expansion over the past decade, with contributions from a wide range of disciplines, biology has so far remained almost untouched by these developments. If a new "systems vision of life" is to be placed on a robust foundation, a concerted effort to bridge this gap is both timely and urgent.

One possible strategy, aimed at achieving such an integration, is to re-examine biological phenomena we think we already understand reasonably well from a perspective of complexity theory. This approach is grounded in the expectation that it will give us valuable knowledge about the kinds of questions that can be asked, and the kind of answers that can be arrived at, when working from within a "complex systems" framework. Importantly, it would allow analysis of whether these questions and answers are truly new, in the sense that they would be inaccessible from within other conceptual frameworks. If this turns out to be the case, the knowledge gained would support the development of new methods of investigation.

This strategy is here applied to proteins, and, more specifically, to the problem of protein evolution. Proteins are by far the most abundant and diverse class of biomolecules and mediate the vast majority of biochemical processes. The molecular evolution of proteins has been extensively studied and is most often conceptualized as a series of independent state changes at single sites. Consequently, investigations of functional evolution commonly focus on mutational changes in a closely circumscribed part of the protein structure, such as the catalytic site in enzymes or the hormone-binding site in a receptor, while other parts of the structure

are thought to make little or no contribution. There is, however, experimental evidence indicating that function can be diminished or altered by mutations distant from the classic "functional" site [6–8]. Another example of great clinical relevance, the emergence of drug resistance in human immunodeficiency virus 1 (HIV-1) protease, involves amino acid mutations distant from the active site where the inhibitor binds [9,10]. Importantly, the success of engineering new function can crucially depend on modifications of regions spatially distant from, but functionally linked to, the catalytic site [11–14]. The question then arises whether a deeper understanding of this phenomenon can be gained by studying protein functional evolution at the systems level. To address this, the aim of the work presented here was to develop a framework and methodology that allow the study of proteins as complex adaptive systems. General characteristics of complex systems as they apply to proteins will be discussed first. Protein structural evolution will then be conceptualized and modeled as a form of parallel-distributed information processing using a classic feedforward artificial neural network (ANN).

1.2 Proteins are complex adaptive systems

With the recent explosion of protein sequence data from all three kingdoms of life, the archea, prokarya, and eukarya, we have come to even more fully appreciate the modular nature of proteins, and the complex ways in which their functional and structural units, termed *domains*, are conserved and recombined during evolution. Domains are thermodynamically stable and fold independently within the context of the whole protein. Novelty in protein function often arises as a result of the gain or loss of domains, or by reshuffling existing domains along the linear amino acid sequence. Thus, protein domains can arguably be seen as stable units of evolution.

New domain functions evolve within the constraints of maintaining thermodynamic stability and autonomous folding capability. The work presented here employs a complex systems approach for the study of the interrelationships between functional diversification of homologous protein domains and conservation of thermodynamic stability. In order to approach domain evolution from this perspective, the first question to be addressed is whether protein domains can legitimately be classed as complex adaptive systems. Although a formal consensus on the characteristics of complex systems has yet to emerge, the following characteristics have found general agreement ([15], p. 3) and are present in protein domains [16].

- *Complex systems consist of a large number of elements.* At the atom level, protein domains typically consist of thousands of elements. At a higher level of description, the amino acid level, they are comprised of up to several hundred elements. Whilst description and modeling at the

atom level is computationally intractable at present, domain systems can be modeled at the amino acid level. In this work, the positions along the protein sequence, rather than the amino acids themselves, are defined as the elements (agents) of the systems. These elements can be in one of 20 different states (be filled by one of the 20 amino acids). The state of an element can change, that is, positions can mutate to a different amino acid.

- *The elements of a complex system interact in a dynamic fashion and these interactions change over time.* Dynamic interactions between amino acids (primary sequence positions in certain states) mediate the folding process and a stable pattern of interactions subsequently determines the three-dimensional fold of the domain. Dynamic interactions are also fundamental to domain functions that are mediated by conformational changes. During evolution, the pattern of interactions between fold positions changes as a consequence of amino acid substitutions (gain or loss of hydrogen bonds, salt bridges, or van der Waals interactions).

- *The interactions between elements are richly connected—any one element influences, and is influenced by, a large number of others.* In a domain fold, amino acid positions along the linear protein sequence are engaged in multiple local (involving positions that are close in the linear sequence) and nonlocal (involving positions that are distant in the linear sequence) physical interactions. With the exception of neutral positions, each fold position makes an individual fitness contribution and simultaneously affects the fitness of many other positions within the domain. Fitness is here defined as the capacity of the domain to maintain its structural integrity and to carry out a specific function(s).

- *The interactions between elements are nonlinear.* Small causes can have large results, and *vice versa*. Complexity results from the patterns of richly connected interactions between the elements. Complex systems exhibit so-called "emergent properties," properties that are only seen in systems of an equivalent degree of complexity. One of the key processes responsible for emergence is self-organization ([15], p. 89; [17], p. 115 and p. 225). This behavior results from the nonlinear interactions of system components which lead to collective effects. Self-organization also leads to spontaneous transitions into new collective states, at times as adaptive responses to changes in the environment. The nonlinearity of interactions between amino acid positions is a major reason why certain amino acid substitutions at only one or a few positions may unravel a domain fold. And, conversely, is a reason why amino acid sequences can at times diverge from homologous sequences beyond any statistically significant similarity, while the shared domain fold is still conserved intact. We are unable to explain or predict these phenomena (at least for now), and so they also illustrate how nonlinearity severely limits predictability. Another related issue is the persistent elusiveness of a solution to the "folding problem," despite three decades of intensive efforts.

- *The interactions between elements are relatively short-range.* Physical constraints and information are mostly transmitted between immediate

neighbors. However, this does not mean that there cannot be long-range influences. In a richly connected network, the path between two elements can usually be covered in a small number of steps. Influences can be enhanced, suppressed, or modulated in some way along the path. Amino acids in domain cores are packed in an engergetically favorable arrangement, and strong local constraints on amino acid variation are present. The network of amino acids that are in contact with each other collectively constrains mutational change. Although this mechanism is mediated by local interactions, it can propagate throughout the domain to distant sites *via* "chains of local interactions" [18]. Nonlinear constraint modulation along such interaction chains occurs due to the rich connectivity between elements (multiple physical interactions and mutual constraints).

- *There are recurrent interaction pathways.* The effects of a state change at one element can feed back on itself, either directly or *via* a number of intervening states. The feedback can be either enhancing or inhibiting. Depending on its nature, a mutation (state change) at one domain position may enhance or inhibit the probability of accepted mutations at coevolving positions. These subsequent mutations may in turn enhance or inhibit the likelihood of further accepted substitutions occurring at the first position.

- *Complex systems have a history.* They evolve through time, and their present state is constrained by their past. Present-day protein domains have evolved from ancestral domains. Domain evolution can only occur within the constraints of maintaining thermodynamic stability and autonomous folding capability.

2. Protein domain evolution as a parallel-distributed process

2.1 An information-theoretic approach to protein domain evolution

Partial gene duplication and recombination are thought to be the primary mechanisms for the generation of protein domain diversity. In this process, the portion of a gene encoding a given domain is duplicated, and, subsequently, one copy maintains the original function while the other is free to evolve new functions. The concept of a functional space, as an abstract representation of all possible functions that can evolve within the structural constraints of a domain fold, is useful for the investigation of domain evolution. Within this conceptual framework, the emergence of new functions can be understood as the result of adaptive walks in sequence space. During this adaptive evolution, duplicated gene portions accumulate successive mutations that progressively enhance the new function.

In this work, the positions along the linear amino acid sequence of the domain are conceptualized as the elements, or "agents," of the system that can each assume one of 20 different states (i.e., the 20 amino

Figure 1. The positions along the linear amino acid sequence of the domain constitute the agents of the system, and can each assume one of 20 different states. An arbitrary state sequence AYQ... is here shown as an example (amino acids in single letter code).

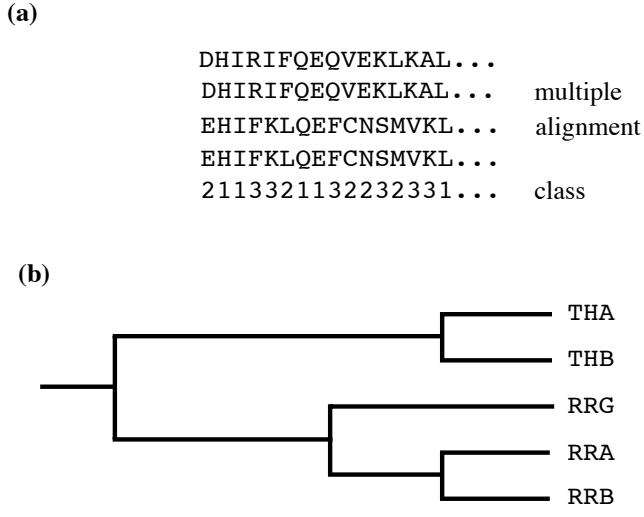


Figure 2. The evolutionary history of a protein domain family is contained in a multiple sequence alignment. (a) Short section of a multiple sequence alignment. Classes: 1, completely conserved position; 2, conserved physicochemical property (e.g., charge, hydrophobicity); 3, highly variable. (b) Example of a phylogenetic tree. Phylogenetic relationships can be ascertained from a multiple sequence alignment. This tree depicts the evolutionary relationships between ligand-binding domains of thyroid hormone and retinoic acid receptors.

acids) (Figure 1). Four classes of fold positions can be distinguished in domains that are descended from a common ancestral domain (Figure 2(a)): (i) positions with conserved amino acid identities; (ii) positions with conserved physicochemical properties; (iii) positions with highly variable physicochemical properties; and (iv) unconstrained positions accumulating neutral mutations. Positions in the coevolutionary distributed network to be modeled by neural networks belong to class (ii) or (iii) (section 3).

The evolutionary history of a domain, contained in a sequence alignment, is a record of successful mutagenesis experiments carried out by nature (Figure 2(a)). A multiple sequence alignment indicates the extent to which specific residues may be changed without destroying domain structure. At the same time the alignment can identify those residues that need to be changed in order to create a new function within a similar structural framework.

Coevolving positions can be identified from a sequence alignment of a domain family using mutual information, a measure of correlation for discrete symbols. A formal measure of variability 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 a sequence position (i.e., for amino acid sequences $s = 20$, for the 20 possible states of amino acid occurrence) [19]. $H(i)$ is defined as

$$H(i) = -\sum_{s} P(s_i) \log P(s_i). \tag{1}$$

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 associated entropies for each position i and j are

$$H(i) = -\sum_{s_i} P(s_i) \log P(s_i) \tag{2}$$

$$H(j) = -\sum_{s'_j} P(s'_j) \log P(s'_j). \tag{3}$$

And the joint entropy is defined as

$$H(i,j) = -\sum_{s_i, s'_j} P(s_i, s'_j) \log P(s_i, s'_j). \tag{4}$$

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

$$M(i,j) = H(i) + H(j) - H(i,j). \tag{5}$$

If the positions are independent, their mutual information is zero. If, on the other hand, the positions are correlated, their mutual information is positive and achieves its maximum value if there is complete covariation.

Given a set of sequences that are assumed to be independent and identically distributed samples from a probability distribution, one can independently estimate each pairwise probability distribution for every pair of positions by frequency counting. However, sequences belonging to a domain family are not independent samples, but are related through shared ancestry described by a phylogenetic tree (Figure 2(b)). If two mutations occur independently in an ancestral sequence and these are subsequently inherited by many of the descendants further down the

tree, the two positions involved will receive a high mutual information score. To estimate the mutual information content between position pairs that is created by tree inheritance alone, and not by covariation, a simulation experiment can be performed. [18,20]. This procedure simulates the evolution of sequences by random mutations along a phylogenetic tree obtained from the domain sequence alignment. Using the outgroup as a seed, random sequences are evolved following the phylogenetic tree obtained from the real data set. During simulated random mutation of sequences, the state of the sequence is duplicated at a bifurcation point in the tree, and the two copies are then independently evolved. Every amino acid can mutate with equal probability to any other amino acid. The procedure is repeated numerous times, and significance threshold values are determined from the frequency distributions of the mutual information scores in the control and real data sets. These threshold values indicate the probability of any given mutual information score not being due to inheritance through the tree. Depending on the problem to be addressed, values may be set at different levels.

2.2 An example study: Evolution of hormone-binding domains in nuclear receptors

Using the ligand-binding domains of steroid receptors as an example, this section will illustrate how a complex systems approach can further our understanding of the evolution of new functions in protein domain families. Ligand-binding sites in protein domains sharing a common ancestor can diverge greatly during evolution. This poses a particularly interesting problem in those cases where the ligand-binding site is situated in, or close to, the domain core, or where ligand-docking induces dramatic conformational changes. These features are present in many receptors and enzymes; the hormone-binding domain present in the receptors for steroids and retinoids exhibits both characteristics. How do binding sites for diverse ligands evolve in core regions of structurally dynamic domains? Are evolutionary changes locally restricted to the ligand-binding site, or are they distributed throughout the domain?

Steroid, thyroid, and retinoid hormones comprise the broadest class of gene-regulatory ligands known. Their receptors belong to the diverse superfamily of nuclear receptors (NRs) and are present in all metazoans from cniderians onward. NRs have had a central part in the evolution of biological complexity since the Cambrian explosion [21,22]. As ligand-inducible transcription factors, they play essential roles in the regulatory pathways that transmit signals, originating from the extra- and intra-cellular environment, to large genetic networks through a complex sequence of molecular interactions.

The ligand-binding domain of NRs possesses a unique fold that is partly disordered in the absence of ligand, termed the *antiparallel* α

helical sandwich. The helices are grouped into three layers around an internal ligand-binding core. Crystallographic studies of ligand-bound receptors suggest a structural role for ligand that is fundamental to the allosteric control mechanisms found in the ligand-binding domain. The ligand is completely buried within the domain interior and contributes to the hydrophobic core of the active conformation of the receptor (for references, see [20]). Therefore, ligand binding directs the alignment of the secondary structural elements critical for receptor function, and strongly constrains the conformational freedom of the ligand-binding domain.

During the evolution of the nuclear receptor superfamily, the ligand-binding pocket has evolved to allow binding of ligands possessing strikingly diverse chemical structures. In [21] it is proposed that the ancestor of the superfamily was an orphan receptor without ligand-binding capability. Their study of NR evolution suggests that liganded receptors have arisen relatively recently and have gained the ability to bind ligands independently. Since the ligand-contacting residues line the binding pocket in the domain core, they perform a dual role; a functional role in ligand recognition and a structural role as core residues. With respect to ligand recognition, they can be seen to constitute an "interior interaction surface." In principle, this would allow extraordinary scope for the evolution of the ligand-binding pocket. However, since the hydrophobic ligand is an integral part of the domain core in the active conformation, the ligand and the ligand-binding residues combined need to be able to maintain structural stability and domain dynamics (conformational changes). How is this potential conflict between structural constraints and functional diversity resolved within the domain fold? Is evolutionary change locally confined to the ligand-binding pocket or does it also involve distant coevolving positions?

Mutual information analysis can be used to reveal coevolutionary relationships between the amino acid positions in domain families. Figure 3(a) depicts a network of coevolving positions that are distributed throughout the NR ligand-binding domain. This network is characterized by a low noise-to-signal ratio due to a high confidence threshold set at 80 (80% confidence that the mutual information score was not due to tree inheritance). Interestingly, 72% of coevolving pairs involve positions in the ligand-binding pocket: 36% of the pairs involve positions that make direct ligand contacts, and a further 36% contain positions that are adjacent to ligand-contacting positions (Figure 3(b)). This suggests that the coevolutionary network in this domain family is closely associated with the evolution of ligand-binding. It was also observed that five out of a total of 36 covarying pairs show an $i, i+2$ or $i, i+4$ periodicity. Covariation between these sites may be due to local constraints at the level of secondary structure, as these types of correlations reflect the hydrophobic periodicity of amino acids seen in amphipathic α helices [23].

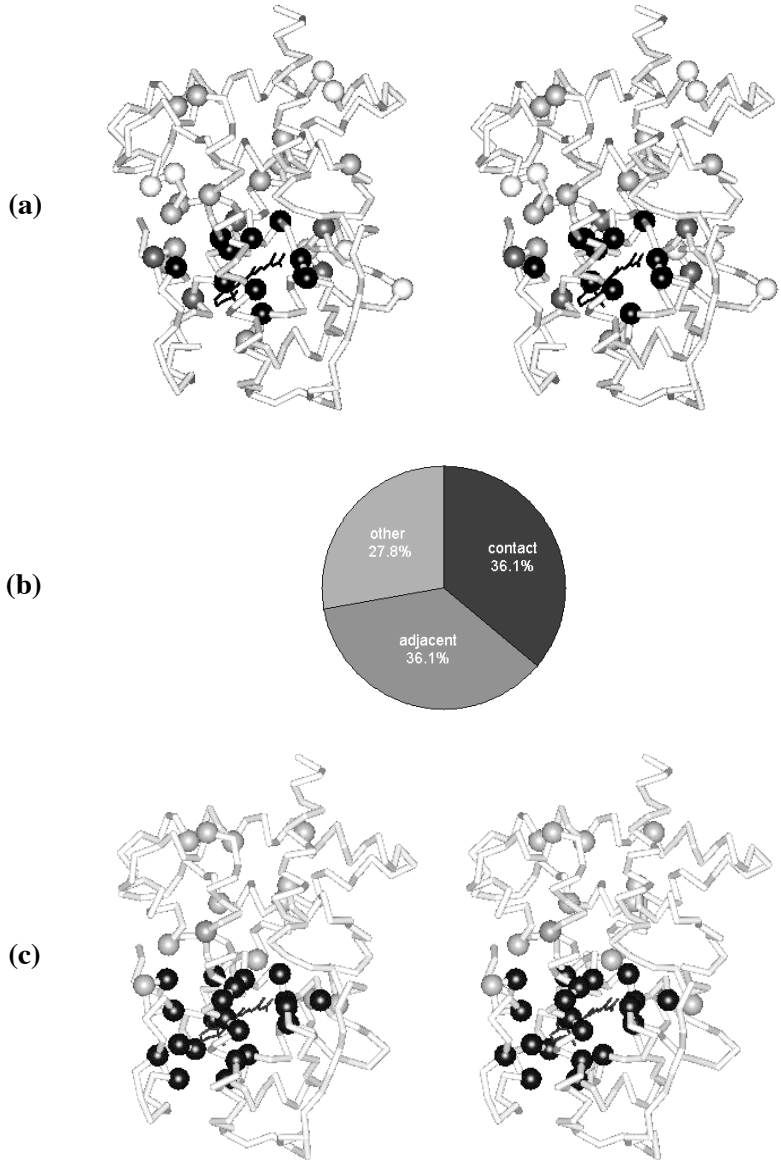


Figure 3. Two perspectives on coevolution in the nuclear receptor ligand-binding domain. (a) A distributed network of coevolving positions can be identified by mutual information analysis of the multiple domain alignment. For illustration, the network is shown mapped onto the retinoic acid receptor X-ray structure

Figure 3. (*continued*)
(stereo view, Protein Data Bank code 2lbd.pdb). The significance threshold value was set at 80. Only alignment positions having less than 40% amino acid identity, and displaying at least four "states" (four different amino acids), were included. (b) 72% of coevolving pairs involve positions in the ligand-binding pocket. 36% of the pairs involve positions that make direct ligand contacts, and a further 36% contain positions that are adjacent to ligand-contacting positions. Black, ligand-contacting positions (within 4.5 C) (α-carbons, spacefill mode); dark gray, positions adjacent to ligand contacts; light gray, covarying positions; white, other covarying positions (not linked to ligand pocket). The ligand is shown in black (stick mode). (c) The network with the significance threshold set at 60 to allow detection of covarying pairs that are restricted to NR functional subfamilies, mapped onto the retinoic acid receptor. Retinoic acid-contacting positions and first-order covarying positions in the ligand-binding domain of the retinoic acid receptor are shown. Ligand contacts are shown in black, covarying positions are shown in gray. See [20] for details on threshold calculation and higher-order coevolutionary relationships between domain positions.

It is important to remember that the high confidence threshold chosen to characterize this network may have excluded valid coevolving pairs. Whilst desirable from a statistical point of view, a high cut-off value may cause relationships that are linked to evolution in domain subfamilies to be overlooked. This is because mutual information scores for positions that coevolve only in a subset of family members will be lower than those for pairs coevolving in the entire family and may be classed as noise (false-negatives). To detect these relationships, the confidence threshold needs to be set at a lower level with a concomitant trade-off between sensitivity (minimizing false-negatives) and inclusion of false-positives that are due to tree inheritance, not coevolution. When the method is used to define the data set for modeling a coevolutionary network by ANNs, one may be prepared to tolerate a small proportion of false-positives in order to increase sensitivity (see section 3). The cost of this would be that the ANN size is increased by inclusion of these false-positives.

To increase sensitivity, the confidence threshold was next set at 60. Figure 3(c) depicts the resultant network, involving more than 90% of all known ligand-contacting positions. In an earlier report it was shown that the ligand-contacting residues in the hormone-binding pocket are evolutionarily linked to an extensive, hierarchically organized network of coevolving positions [20]. Coevolving positions are likely to compensate for the destabilization resulting from the binding of diverse ligands and to preserve the structural integrity and the conformational dynamics of the ligand-binding domain. In conclusion, a distributed evolutionary mechanism, involving the domain fold as a whole, is present in the

ligand-binding domains of nuclear hormone receptors. It is suggested that this mechanism maintains a thermodynamically favorable interplay between molecular organization and evolutionary dynamics.

3. Neural network models of protein domain evolution

Constraint satisfaction within coevolutionary networks can be understood as a form of biological information processing. Coevolutionary relationships can have very high interconnectivity, where each position in the coevolutionary network constrains, and is constrained by, many other positions. This was the case in the nuclear receptor ligand-binding domain [20]. Each amino acid is uniquely defined by its physicochemical properties, such as shape, volume, polarity, hydrophobicity, and charge among many additional, less well understood properties. Depending on the location of coevolving positions within the network, different physicochemical properties may be crucial in determining the pattern of coevolution. As has been learnt from homologous domain alignments, in many cases volume conservation is of paramount importance [24], while other properties are less constrained. In other cases, the hydrophobicity value or charge may be crucial; or a combination of several properties. Presumably, the greater the number of properties involved, or the more restricted the allowed range of a single property, the stronger will be the mutual constraints on allowed states for each position. All these factors combined result in constraints of high dimensionality. ANNs can represent the complexity, and the parallel-distributed nature, of this evolutionary process.

ANNs are computer algorithms that attempt to model information processing and adaptive learning executed by ensembles of biological neurons. One particularly valuable and intriguing characteristic of information processing in biological brains seems to also be present in ANNs—the ability to make decisions based on very complex, noisy, irrelevant, and/or partial information. An ANN is composed of a large number of highly interconnected processing elements that are analogous to neurons and are tied together with weighted connections that are analogous to synapses. Interconnected neurons, whether biological or artificial, have certain neuro-"logical" properties and can be seen as logic gates: They receive input signals from a large number of other neurons, process these signals according to specified transformation functions, and produce an output signal as a result of this processing. In *auto-association* or *associative memory* tasks the input sample is presumed to be corrupted, noisy, or a partial version of the desired output pattern. In *hetero-association*, the output pattern may be any pattern that is to be associated with a set of input patterns. The neural network architecture can either be tailored to an auto-associative mode or a hetero-associative mode.

Brains and ANNs represent information in a distributed fashion; information is encoded by the patterns of synaptic connection strengths (weights) between neurons. The distributed networks of neurons perform many transformation steps in parallel, a style of computation known as parallel distributed processing (PDP). When fully connected neural networks are used, a combination of a large set of connection weights and nonlinear transfer functions allows models of any complexity to be fitted between the response and the input parameters. Neural networks are therefore highly efficient nonlinear data modeling devices, and can be seen as universal models for information processing in complex systems.

Arguably, the evolution of functional sites within the coevolutionary network of a domain family can be conceptualized as a type of PDP. It should be well noted that this statement is not meant to imply a direct correspondence in architecture between the coevolutionary network and an ANN, but refers to an analogous information-processing mode. Furthermore, as all parallel-distributed computational steps are executed simultaneously, ANN models of domain evolution do not represent the historical sequence of stepwise mutation at coevolving sites over evolutionary time. This temporal aspect of coevolutionary networks can be analyzed and modeled by reconstruction of ancestral states by parsimony or maximum likelihood methods. Recently, analogies between protein evolution and neural networks (Hopfield nets) have also been noted in [25,26].

For the purpose of building an ANN model of a coevolutionary network, a protein is represented as a chain of agents in a linear sequence, each of which can take on one of 20 states (Figure 1). The agents are understood as mechanisms for mediating interactions ([17], p. 6), and state transitions in agents (mutations) lead to a modification in the patterns of interactions, sometimes resulting in a change in structure/function. The state transitions are constrained by rules ([17], p. 116), and all possible state sequences are the outcomes of a succession of transitions specified by these rules. In this way, the rules generate evolutionary novelty. Structure/function can now be reconceptualized as an emergent property, the result of context-dependent interactions, that changes over time. The state transition rules can be encoded in the values of the connection weights of the ANN model.

The evolution of new functional sites within the context of a coevolutionary network can be modeled by a classical fully-connected feedforward neural network (Figure 4) (for a detailed mathematical treatment of feedforward network properties and behavior, see, for example, [27–29]). The inbuilt directionality of this type of neural net corresponds to selection pressure on the domain for evolving new functions. During training, the network is presented with instances of functional sites (input) and associated amino acid identities at coevolving positions (out-

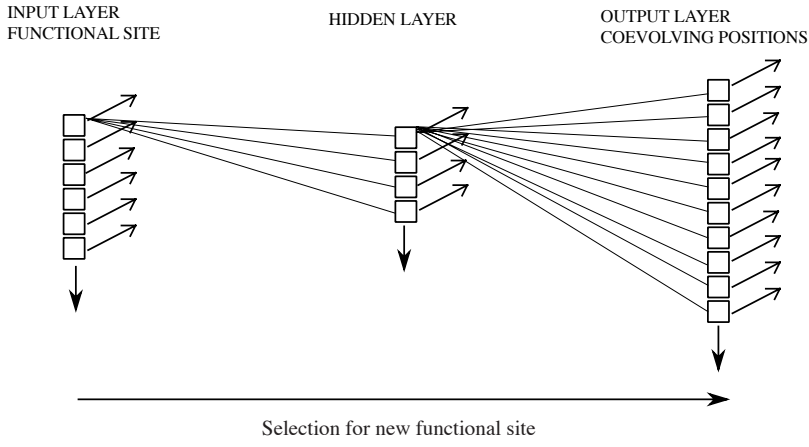


Figure 4. ANN model of the evolution of new functional sites in a domain family. The network architecture is that of a classic feedforward network whose size can vary depending on the coevolutionary network to be modeled (closed arrows). Sequence positions (agents) function as fully connected processing elements (squares). Each agent is represented as a binary vector (open arrows). Amino acids are encoded as bitstrings. A hetero-associative mapping is performed that maps the input vector matrix (agent states in the functional site) to the output vector matrix that ranges over a different vector space (states of coevolving agents). After training, the ANN encodes the state transition rules of the coevolutionary network.

put) taken from domain family sequence alignments, and trained to associate outputs with input patterns. When the network is subsequently used for modeling, it identifies the input pattern and tries to produce the associated output pattern.

The power of neural networks comes to life when a pattern that has no output associated with it, is given as input. In this case, the network predicts the output based on the rules learnt in the training phase. This property is responsible for the potential of ANNs in evolutionary modeling. When, for example, given an artificially designed functional site as input, it will predict compensatory states of the coevolving agents based on the learnt state transition rules.

3.1 Artificial neural network model of the first-order coevolutionary network in the nuclear receptor ligand-binding domain

To date, direct interactions with ligand have been characterized for 42 positions in the nuclear receptor ligand-binding domain ([20]; and references therein). When mapped onto crystallographic domain structures, these positions are shown to form the internal ligand-binding pocket.

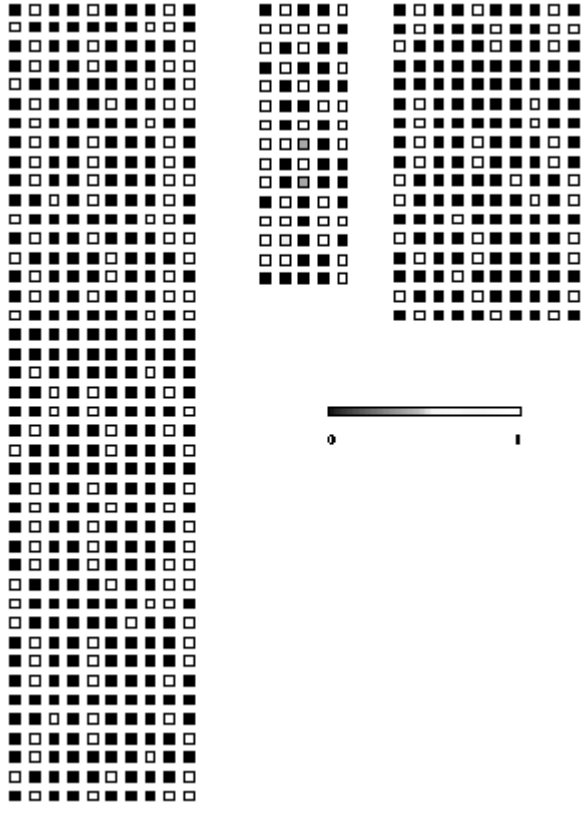
In different nuclear receptor types, subsets of pocket-forming positions comprise the ligand-specific contact residues.

The coevolutionary network made up of the 42 pocket-forming positions (input) and 17 first-order correlated positions (output) was modeled by a fully-connected 3-layer backpropagation network (Figure 5).[1] Input and output layers were encoded by two-dimensional binary amino acid identity matrices, based on the amino acid class hierarchy shown in Figure 6. A training set derived from 80 nuclear receptors, representative of the entire superfamily, and a validation set obtained from 20 additional receptors were employed for network training. To avoid local minima problems and poor convergence behavior, conjugate gradient descent was used as the network training paradigm. The mean square error for the training cycle (after 200 epochs) and the validation cycle was 0.0003.

4. Conclusions

Artificial neural network (ANN) modeling of coevolutionry networks in protein domains represents a novel approach with diverse applications in evolutionary studies and protein design. Firstly, the complex systems framework employed allows the reconceptualization of protein domain evolution as a form of biological parallel-distributed information processing. This is a novel perspective whose full implications require further exploration. Extensive analysis of the frequency of occurrence of coevolving networks in domain families, and of their statistical and spatial structures, is needed, and the neural network modeling procedure requires further refinement. Secondly, the ANN modeling procedure outlined in this paper is expected to be valuable for the design of novel functions, such as new ligand-binding capabilities, for a given domain fold. ANN modeling, based on the coevolutionary relationships between ligand-binding sites and coevolving positions, may enable one to overcome otherwise prohibitive limits to binding-site modifications. Predictions of required mutations located at coevolving positions throughout the domain may be used to maintain the stability of the modified fold. ANN domain modeling will realize its fullest potential in the "post-genome era," once more members of domain (super)families are sequenced and the full range of sequence variation within a superfamily is available for the training of the neural network. The larger the training set, the more sensitive will be the ability of the ANN to simulate domain evolution *in silico*.

[1] Stuttgart Neural Network Simulator 4.1 was used. This software simulator for neural networks was designed at the University of Stuttgart. The program is available at http://www.informatik.unistuttgart.de/ipvr/bv/projekte/snns/snns.html.



D	1000100010	V	0100100011	E	1000100001
F	0100010010	K	1000010010	W	0100010001
R	1000010001	Y	0100010011	H	1000010011
C	0100001000	N	1000001010	M	0100000100
Q	1000001001	A	0010100010	S	1000000110
G	0010100001	T	1000000101	P	0001000000
I	0100100010	X	0000000000	L	0100100001

Figure 5. The ANN model of the first-order coevolutionary network in the nuclear receptor ligand-binding domain, shown in a two-dimensional projection. Here, the activation states of the network elements for one validation pattern are depicted as an example (input layer, left; hidden layer, middle; output layer, right). Links between elements are not shown. See text for details. Each amino acid was represented by a 10-bit vector.

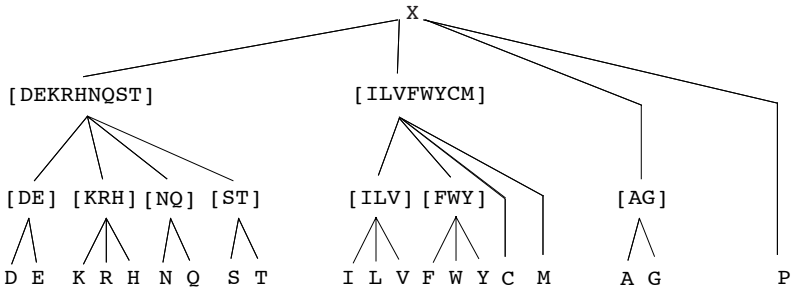


Figure 6. The amino acid class hierarchy from [30] was used for binary encoding of amino acid identities.

Acknowledgments

This research was supported in part by a fellowship grant from the Foundation of Research, Science, and Technology, New Zealand.

References

[1] L. E. Kay, *The Molecular Vision of Life: Caltech, The Rockefeller Foundation, and the Rise of the New Biology* (Oxford University Press, Oxford, 1993).

[2] C. A. Harrington, C. Rosenow, and J. Retief, "Monitoring Gene Expression using DNA Microarrays," *Current Opinions in Microbiology*, **3** (2000) 285–291.

[3] D. J. Lockhart and E. A. Winzeler, "Genomics, Gene Expression and DNA Arrays," *Nature*, **405** (2000) 827–836.

[4] A. Pandey and M. Mann, "Proteomics to Study Genes and Genomes," *Nature*, **405** (2000) 837–846.

[5] K. L. Williams, "Genomes and Proteomes: Towards a Multidimensional View of Biology," *Electrophoresis*, **20** (1999) 678–688.

[6] Y. Kitaoka, "Involvement of Amino Acids Outside the Active Site Cleft in the Catalysis of a Ricin Chain," *European Journal of Biochemistry*, **257** (1998) 255–262.

[7] L. Pritchard and M. J. Dufton, "Evolutionary Trace Analysis of the Kunitz/BPTI Family: Functional Divergence May Have Been Based on Conformational Adjustment," *Journal of Molecular Biology*, **285** (1999) 1589–1607.

[8] S. Oue, A. Okamoto, T. Yano, and H. Kagamiyama, "Redesigning the Substrate Specificity of an Enzyme by Cumulative Effects of the Mutations of the Non-active Site Residues," *Journal of Biological Chemistry*, **274** (1999) 2344–2349.

[9] D. Pillay, S. Taylor, and D. D. Richman, "Incidence and Impact of Resistance Against Approved Antiretroviral Drugs," *Review of Medical Virology*, **10** (2000) 231–253.

[10] J. W. Erickson, S. V. Gulnik, and M. Markowitz, "Protease Inhibitors: Resistance, Cross-resistance, Fitness, and the Choice of Initial and Salvage Therapies," *AIDS*, **13 Suppl A** (1999) S189–204.

[11] J. P. Kitchell and D. F. Dyckes, "A Synthetic 13-residue Peptide Designed to Resemble the Primary Binding Site of the Basic Pancreatic Trypsin Inhibitor," *Biochemical and Biophysical Acta*, **701** (1982) 149–152.

[12] E. Kohfeldt, W. Gohring, U. Mayer, M. Zweckstetter, T. A. Holack, M.-L. Chu, and R. Timpl, "Conversion of the Kunitz-type Module of Collagen VI into a Highly Active Trypsin Inhibitor by Site-directed Mutagenesis," *European Journal of Biochemistry*, **238** (1996) 333–340.

[13] J. A. Kraunsoe, D. T. W. Claridge, and G. Lowe, "Inhibition of Human Leukocyte and Porcine Pancreatic Elastase by Homologues of Bovine Pancreatic Trypsin Inhibitor," *Biochemistry*, **35** (1996) 9090–9096.

[14] B. L. Roberts, W. Markland, A. C. Ley, R. B. Kent, D. W. White, S. K. Guterman, and R. C. Lander, "Directed Evaluation of a Protein Selection of Potent Neutrophil Elastase Inhibitors Displayed on M13 Fusion Phage," *Proceedings of the National Academy of Sciences USA*, **89** (1992) 2429–2433.

[15] P. Cilliers, *Complexity & Postmodernism* (Routledge, London, 1998).

[16] S. B. Nagl, "Neural Network Models of Protein Domain Evolution," *Hyle: An International Journal for the Philosophy of Chemistry*, special issue on models in chemistry (in press).

[17] J. H. Holland, *Emergence* (Addison-Wesley, Reading, MA, 1998).

[18] A. S. Lapedes, B. G. Giraud, L. C. Liu, and G. D. Stormo, "Correlated Mutations in Protein Sequences: Phylogenetic and Structural Effects," *Proceedings of IMS/AMS 1997 International Conference on Statistics in Molecular Biology* (Seattle, 1997).

[19] B. T. M. Korber, R. M. Farber, D. H. Wolpert, and A. S. Lapedes, "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 USA*, **90** (1993) 7176–7180.

[20] S. B. Nagl, J. Freeman, and T. F. Smith, "Evolutionary Constraint Networks in Ligand-binding Domains: An Information-theoretic Approach," *Proceedings of the Pacific Symposium on Biocomputing*, (1999) 90–101. A full-text on-line version is available at http://www.smi.stanford.edu/projects/helix/psb99/.

[21] H. Escriva, R. Safi, C. Hanni, M.-C. Langlois, P. Saumitou-Laprade, D. Stehelin, A. Capron, and R. Pierce, "Ligand Binding was Acquired During Evolution of Nuclear Receptors," *Proceedings of the National Academy of Sciences USA*, **94** (1997) 6803–6808.

[22] V. Laudet, C. Hanni, J. Coll, F. Catzeflis, and D. Stehelin, "Evolution of the Nuclear Receptor Gene Superfamily," *EMBO Journal*, **11** (1992) 1003–1013.

[23] T. M. Klingler and D. L. Brutlag, "Discovering Structural Correlations in Alpha-helices," *Protein Science*, **3** (1994) 1847–1857.

[24] M. Gerstein, E. L. Sonnhammer, and C. Chothia, "Volume Changes in Protein Evolution," *Journal of Molecular Biology*, **236** (1994) 1067–1078.

[25] L. Pritchard and M. J. Dufton, "Do Proteins Learn to Evolve? The Hopfield Network as a Basis for the Understanding of Protein Evolution," *Journal of Theoretical Biology*, **202** (2000) 77–86.

[26] J. J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," *Proceedings of the National Academy of Sciences USA*, **79** (1982) 2554–2558.

[27] D. M. Skapura, *Building Neural Networks* (Addison Wesley, Reading, MA, 1995).

[28] K. Mehrotra, C. K. Mohan, and S. Ranka, *Elements of Artificial Neural Networks* (MIT Press, Cambridge, MA, 1996).

[29] D. J. Livingstone, D. T. Manallack, and I. V. Tetko, "Data Modelling with Neural Networks: Advantages and Limitations," *Journal of Computer-Aided Molecular Design*, **11** (1997) 135–142.

[30] R. F. Smith and T. F. Smith, "Automatic Generation of Primary Sequence Patterns from Sets of Related Protein Sequences," *Proceedings of the National Academy of Sciences USA*, **87** (1990) 118–122.