

CHAPTER 1

Structural Perspectives on Protein Evolution

Eric Franzosa* and Yu Xia**

Contents		
	1. Introduction	3
	2. Determinants of Evolutionary Rate	4
	3. Theoretical Advances	5
	3.1. Key concepts	5
	3.2. Theory: designability	7
	3.3. Theory: evolvability	7
	3.4. Modeling structure and evolution	8
	4. Empirical Results: Single Proteins	10
	4.1. Approaches	10
	4.2. Physical properties	10
	4.3. Constitutional properties	11
	4.4. Protein domains	13
	4.5. Function	14
	5. Empirical Results: Higher Order Properties	14
	5.1. Interfaces	14
	5.2. Protein–protein interaction networks	15
	5.3. Protein complexes	16
	6. Summation	17
	Acknowledgments	18
	References	18

1. INTRODUCTION

Nothing in biology makes sense except in the light of evolution [1]. These words, written by famed evolutionary biologist Theodosius Dobzhansky, have a special place

* Bioinformatics Program, Boston University, 24 Cummington Street, Boston, MA 02215, USA

** Department of Chemistry, Boston University, 24 Cummington Street, Boston, MA 02215, USA. E-mail: yuxia@bu.edu

in the life sciences. They serve not only as a reminder of the central role evolution plays in the processes of life, but also as a paradigm under which research in biology should be conducted. When we think about evolutionary phenomena, it is important to remember that they—like all natural phenomena—can be reduced to events at the molecular scale. Evolutionary change is fueled by mutations: changes in the molecular structure of the genetic material. When these mutations are expressed, they result in changes to biomolecular structure and energetics which may in turn alter the abundance or interactions of proteins. Novel organismal traits stemming forth from these molecular scale changes are then judged by natural selection. Considering this perspective, it seems that nothing in evolution makes sense except in the light of biomolecular structure and energetics. It should therefore not come as a surprise that the relationship between evolution and physical phenomena like these has been an area of interest for some time. Here we review some of the major theories, models, and empirical evidence relevant to the relationship between protein structure and evolution at various scales.

2. DETERMINANTS OF EVOLUTIONARY RATE

Protein evolutionary rates are known to vary widely. In the genome of the model organism *Saccharomyces cerevisiae* (baker's yeast), evolutionary rates among the roughly 6,000 genes are spread out over three orders of magnitude [2]. Since the advent of the molecular biology age scientists have been interested in the way that homologous genes and proteins accumulate changes. It has been observed that the sequences and structures of some proteins are highly conserved, even when comparisons are made between distantly diverged species (for example, the histone proteins that package DNA, or the ribosomal proteins responsible for protein translation) [3]. Other proteins evolve rapidly, either due to relaxed constraint or positive selection for novel features (for example, proteins involved in immune systems) [4]. While theories explaining these differences originated alongside their observation, extracting general determinants of evolutionary rate variation has only become possible with the advent of the bioinformatics age. Statistical and machine learning techniques, when applied to massive genomic and phenotypic datasets (such as protein structures, interaction networks, and expression profiles), have been able to isolate some of the forces driving evolution at the molecular level (for general reviews of evolutionary determinants, see [2,5,6]).

Features that are directly connected to protein structure have been shown to explain roughly 10% of the variation in evolutionary rate [7]. This result seems initially surprising, given that structure mediates all aspects of a protein's existence. In contrast, expression—the frequency and scale at which a protein is manufactured—may explain up to 40% of evolutionary rate variation [6]. Expression and evolutionary rate vary inversely, with highly expressed proteins tending to evolve at very slow rates. A protein's dispensability (effect on cell growth when absent) and the number of interactions in which it participates explain additional

components of evolutionary rate variation; random noise may also be a large contributor [8]. It is worth noting that protein function, historically considered to be a major target of selection, does not seem to be a good general predictor of evolutionary rate [6]. While much progress has been made in the identification and ranking of evolutionary determinants, some disputes in this area still remain. We believe that considering structure is particularly important because of its role as both a determinant of evolution in its own right, and a medium through which other determinants act.

As an example, we will consider the role that structure plays in the apparent dominance of expression in the determination of evolutionary rate. Although several hypotheses have been proposed to explain the significance of expression [9,10], genomic evidence seems to best support the following conclusion:

Errors made during protein translation can result in misfolded proteins, which represent a burden to the cell. Mutations that make a protein more susceptible to error-induced misfolding will result in a loss of fitness. If the mutation occurs in a highly expressed protein, then translational errors (and misfolding events) will be more common, resulting in a larger fitness loss. Hence, protein expression will scale directly with selective constraint, and inversely with evolutionary rate [11].

Are errors in translation really so common that they can have a profound influence on the evolutionary trajectory of a protein? Although the machinery of translation operates with 99.95% accuracy (measured as correctly inserted amino acids), even a small potential for error becomes rapidly compounded given the enormous work load it must handle [12]. If we assume that an average protein is composed of 400 amino acids, roughly 20% of these proteins will contain at least one translational error. Robustness against error-induced misfolding (i.e., structural robustness) would presumably be beneficial for any protein, but more so for the highly expressed among them. Thus, structure, a characteristic of all proteins, plays an even more critical role in their evolution than is apparent at face value. A theoretical treatment of the sequence-structure relationship sheds light on the role of structure in this and other evolutionary phenomena.

3. THEORETICAL ADVANCES

3.1 Key concepts

The essence of most theoretical ideas governing protein structure and evolution begin with the following relationship:

$$\begin{array}{ccc} \text{Genotype} & \rightarrow & \text{Phenotype} \\ \text{(sequence)} & & \text{(structure)} \end{array}$$

Genotype yields phenotype. This is a general biological idea. In the case of proteins, it can refer to the DNA sequences (genotypes) that encode amino acid chains that fold to produce three-dimensional proteins (phenotypes). Alternatively we can bypass the genetic component of the picture and think of the translated sequence of amino acids as a genotype, with the notion of a phenotype remaining

the same. The relation above is simple, but profoundly important. In one sense it can be thought of as a statement of the central dogma of molecular biology (the elucidation of which is among the greatest scientific achievements of the 20th century) [13]. In another sense this relation is a statement of the protein folding problem, one of the largest challenges facing researchers in computational biology today [14].

The space of protein phenotypes observed in nature is surprisingly small. Current estimates place the number of stable folds in the neighborhood of 1000 to 10,000 [15,16]. We can imagine many other possible folds in the configuration space of an amino acid chain, but these have either (a) yet to occur in evolution or (b) been thermodynamically or selectively disfavored. The complete space of possible genotypes is assumed to be very large. Constraining ourselves to the size of an average protein (400 amino acids), there are 20^{400} ($\approx 2.6 \times 10^{520}$) possible protein sequences. Obviously evolution has sampled only a small fraction of these sequences, and an even smaller fraction persists on the planet today. Nevertheless, the mapping of genotypes to phenotypes remains many-to-one, with sets of genes and amino acid sequences producing largely identical protein structures [17]. We make the assumption that, under a given set of conditions, a single sequence maps to exactly one structure (governed by the minimization of free energy). This is a reasonable assumption for natural protein sequences, which tend to have a marked free energy minimum [18].

These observations are extremely important in light of the neutral theory of evolution. Simply put, this theory states that the majority of accepted changes that occur at the genotype level do not have a pronounced effect on the phenotype [19]. Silent substitutions in DNA are one example of this phenomenon. DNA codons **GGA** and **GGG** both encode the amino acid glycine, and hence a **GGA** \rightarrow **GGG** mutation would produce a genotype change, but not a phenotype change. Note how this idea fits naturally with the observation of the many-to-one mapping of protein sequences to structures. Since a given structure may be generated by multiple sequences, mutations that interconvert those sequences do not have phenotypic consequences, and are therefore selectively neutral. Some caution is warranted here, as no mutation is likely to be neutral across all environments [20]. We can imagine that sequences which produce identical structures under one temperature regime might produce two different structures under another. Even the canonical silent DNA polymorphisms can evolve non-neutrally in situations where one synonymous codon is preferred over another for purposes of transcriptional or translational efficiency [21]. In this review, we frequently employ the approximation that for most mutations, protein structure directly dictates protein function, i.e. mutations that preserve a protein's structure will also preserve its function. This is not always the case, as certain mutations which conserve structure may have significant functional consequences (for example, if they result in changes to key residues in the active site of an enzyme). In spite of these complications, the notion of *structural neutrality* in a *particular* environment or genetic background remains a useful concept in the study of protein evolution.

3.2 Theory: designability

There are two, somewhat conflicting perspectives from which we can consider the relationship between genotype (sequence) and phenotype (structure). The first of these is called *designability*. We noted that the relationship mapping sequences to structures is many-to-one. What must also be observed is that the sequences are not evenly partitioned across the structures. Some structures can be generated by folding any number of a very large set of sequences; other structures are more specialized, and can only be built up from a few sequences [22]. The structures with many generative sequences are said to be more *designable* than those with fewer generative sequences. Recall that the number of protein folds observed in nature is relatively small. These folds are likely to vary amongst themselves in terms of designability; more importantly, designability is expected to vary between the observed folds and “imaginary” folds. In fact, increased designability may contribute to the dominance of the observed folds [23,24], a fact that we illustrate by example.

Let us consider a hypothetical world with two folds: one which has a useful structure, but can only be generated by a single sequence (low designability), and another which is useless, but can be generated by many sequences (high designability). If selection strongly favors utility, then clearly the first fold will propagate by virtue of its functional advantage. Designability becomes important when we introduce mutations into our model. Although the first fold has a selective advantage in its native form, it is not robust against mutations. Any change in its underlying sequence will result in a loss of its useful structural characteristics. Because the second fold is designable, it is robust against mutations, but selectively disadvantaged because it is useless. For the first fold to remain dominant, its selective advantage must be strong enough to compensate for the losses due to mutation. Now imagine a third fold, one that is both useful and designable. Selection will favor this fold like the first, because it can fill a functional role. By virtue of its designability, this fold will maintain a useful structure even while its underlying sequence accumulates mutations. All else being equal, this fold will come to dominate the population.

The example above assumed that selection acts on structure as a trait in and of itself. Hence, many mutated variants of the third fold were assumed to be selectively neutral, simply because they result in the same structure. Selection will also act on a protein's function, which may be more sensitive to specific changes in the underlying sequence. As noted above, we have employed structure as a surrogate for function, but in reality both features are important. Protein structural properties are more easily generalized than functional properties, and so the former feature tends to be more amenable to theoretical treatments.

3.3 Theory: evolvability

We can also view the genotype-phenotype (sequence-structure) relationship from the perspective of *evolvability*. Simply put, evolvability is concerned with the generation of new phenotypes from existing phenotypes—a phenomenon that is central to the evolution of species [25]. This is in stark contrast to designability, which stressed the importance of maintaining a single phenotype—avoiding

change. Change, however, is at the heart of evolution. How does evolvability, and hence change, relate to the sequence-structure relationship?

Let us consider another hypothetical world for illustrative purposes. There are two dominant folds in this world, both of which result from many-to-one sequence-structure mappings. If we were to sample populations of either fold, we would find that its underlying genotypes were widely varied. We can say that these genotype sequences belong to the fold's *neutral network*—the set of all genotype sequences which produce the fold [26]. Mutating from one genotype to another within the neutral network does not change the phenotype, an example of neutral evolution. The two folds will differ in terms of their evolvabilities.

To understand what this means requires an understanding of the relationships between neutral networks. Some mutations in a given sequence result in a new sequence that remains within the neutral network of the original; other mutations result in a new sequence which belongs to a different neutral network. Mutations of the first type do not result in a change of phenotype (the fold remains constant), while mutations of the second type produce a new phenotype (the fold changes). We are interested not only in the frequency of mutations that leave a given neutral network, but also in the distribution of new networks in which they land. Do the new mutations always lead to another single network, or one of an ensemble? This information allows us to establish notions of closeness between neutral networks, and this is the essence of evolvability. A neutral network which is close to other neutral networks is evolvable—its genotypes have the potential to mutate, producing new genotypes with potentially innovative phenotypes. Conversely, an isolated neutral network is not evolvable. These ideas have been extensively tested in the context of the RNA sequence-structure relationship [27–29], in which the size of neutral networks and transition frequencies can be readily computed. The underlying theory, however, is applicable to the mapping between protein sequences and structures.

Returning to our example, assume that the neutral network of the first fold is close to those of several other folds, while the neutral network of the second fold is relatively isolated (surrounded by the neutral network of unfolded proteins, perhaps). Now assume that a change in the environment occurs causing both folds to be heavily penalized by selection. In our model, the new environment acts as an agent of selection, but does not affect the genotype-phenotype relationship (hence, the neutral networks do not change). The descendants of proteins in the second neutral network are doomed—their mutations cannot produce an innovative solution to the new environment. For the first neutral network there is hope—some of the genotypes here are likely to mutate into the neutral networks of other folds, one or more of which may fair better in the new environment.

3.4 Modeling structure and evolution

Direct observation of these theoretical forces in action is difficult due to the long timescales over which evolution operates. Ideally we would be able to model an accelerated version of this process using computers, but there are difficulties inherent to this as well. The mapping from sequence to structure (protein folding)

is a spontaneous, natural process in living cells, but a major challenge in simulation. The ability to generate an accurate 3D structure of a protein by computation alone given only its amino acid sequence is the essence of the *protein folding problem* [14,30]. Despite massive amounts of work in this area, we still lack a general efficient solution.

Instead, physical models of protein evolution are usually conducted using simplified representations of proteins (such as strings of balls woven through a regular lattice) [31,32]. While these models obviously represent a gross simplification, they capture some of the physical and geometric constraints governing the sequence-structure relationship in real proteins. Results obtained in these simulations tend to be more relevant than those derived on a purely conceptual basis; the cost in terms of computational complexity is also greater. For simple models, the entire space of genotypes and phenotypes can be sampled [33], something that will likely never be possible for real proteins. Adding sophistication to these models boosts the biological relevance to their findings, but often at considerable computational costs [34–36].

Lattice models have generated a variety of interesting results, relevant to both the protein folding problem and the relationship between protein structure and evolution. The distribution of sequences in a neutral network has been explored as a function of the mutational and selective pressures on the corresponding fold [37, 38]. Similar approaches have concluded that evolution selects for sequences which can rapidly adopt their final structures [39,40]. Simulations also tend to predict small sets of dominant protein folds [24]—a result which matches our observations about the real world. The dominant simulated structures are shown to be both highly designable and thermodynamically stable, implying that a causal relationship may exist between these two quantities [18]. Whether or not these observations are generally true for real proteins is not known. The most convincing research in this area is able to pair model-based predictions with observations in a sample of real proteins. For further review of work in this area see [33].

The relationship between designability and evolvability is another area of interest currently being studied with model-based simulations [41,42]. Designable structures are advantageous because they are robust against change. Evolvable structures are advantageous because they have the ability to innovate. Do these forces oppose one another in evolution, or is there a hidden synergy between them? Lattice-based models have been used to explore the relationship between the evolution of new functions and the maintenance of stability, which is conceptually related to the designability/evolvability paradox [43,44]. While these two objectives are antagonistic under some circumstances, a period of enhanced selection for stability can promote subsequent gain of function. Other modeling approaches have also shown that evolvability is itself a selectable trait, favorable in times of rapid environmental change [45].

Clearly theories and models exploring the role of structure in protein evolution produce a wealth of fascinating ideas, many of which are supported by intuition, real world examples, and consistency with physical laws. However, observations on the role of structure in evolution which begin with real proteins in the natural

world are—almost by definition—the most relevant. Our focus now turns to these observations.

4. EMPIRICAL RESULTS: SINGLE PROTEINS

4.1 Approaches

Some of the most intriguing observations about natural proteins involve the relationships between explicit physical parameters (e.g., solvent accessibility) and evolutionary rate. While it is common to think of conceptual “forces” which influence evolution, these relationships hint at true physical forces that govern the allowable changes in proteins. All of these analyses are based on real world sequence data and structures, and thus their results are highly relevant. The structure of real proteins can be determined using X-ray crystallography or NMR techniques [46,47]. These procedures are time consuming and have low throughput, but provide extremely precise (to within angstroms) 3D glimpses at the structures of real proteins. The RCSB Protein Databank, a major repository for this information, contains over 50,000 structures (as of 10/14/2008) [48]. This is a lot of data, but it pales in comparison to the millions of protein coding sequences contained in the BLAST database [49]. Therefore, another common strategy is to build models based on known structures and use these to predict physical features of the sequences whose experimental structures are not known. Both the real and inferred structural parameters can be explored at a variety of scales; from smallest to largest these include: individual residues, secondary structure motifs, protein domains, whole proteins, protein complexes, and protein networks. We reserve discussion of the last two topics for the next section.

4.2 Physical properties

One of the oldest observations linking protein structure to evolution involved the influence of solvent exposure on residue mutations. The homologous proteins hemoglobin and myoglobin, whose atomic structures had been solved by 1965, were observed to differ far more dramatically on their surfaces than in their cores [50]. This has since become a well known general feature in protein evolution. Several recent studies have reexamined the situation using large sequence and structure datasets [7,51–53]. These studies universally support the notion that buried residues in a protein's core are under tighter constraint, and therefore evolve more slowly. A protein's core—and hence, buried residues—play an important role in stabilizing its folded structure. It is therefore believed that mutations in the core may result in structure destabilization, potential misfolding, and a consequent loss of fitness. How do things change when we consider the solvent accessibility of full proteins (rather than residues in a “protein free” context)? The “functional density hypothesis,” proposed in 1976, states that the selective constraint that a protein experiences should be proportional to the fraction of its residues involved with its function (e.g., catalytic activity) [54]. Although

proper folding is not typically thought of as a “function” of a protein, if the protein does not fold or folds improperly, then its conventional functions will certainly be impaired. A modified “fitness density hypothesis” posits that a protein’s rate of evolution should be constrained by the fraction of its residues that, if mutated, would result in a significant loss of fitness [5,11]. Buried residues would certainly seem to be among this fraction.

Intuition therefore suggests that a protein’s evolutionary rate should scale with the fraction of its residues that are buried (or inversely with the fraction of solvent exposed residues). A study by Bloom and colleagues reported an opposing trend: proteins with a large fraction of buried residues seem to evolve *more* rapidly [7]. Their explanation is that proteins with large, stable cores have greater freedom to accumulate surface mutations, and that these mutations significantly elevate the overall rate at which the protein appears to be evolving. This study also considered the effect of atomic contact density on evolutionary constraint. Solvent exposure and contact density convey similar information about the three dimensional structure of a protein, and hence correlate well with one another. The Bloom et al. results regarding contact density are consistent with their solvent exposure findings: proteins with higher average contact densities appear to be evolving faster. This result is particularly interesting in light of a proposed relationship in which proteins with high contact density are also more designable [55]. Because structures with high designability have a large sequence space to explore, we might expect them to demonstrate accelerated evolution at the sequence level (as this study has found).

A subsequent study by Lin and colleagues considered both known and predicted exposure patterns in proteins with variable alignment lengths [51]. They conjecture that restricting an analysis to proteins with large alignment lengths—as was the case in the Bloom et al. study—biases results against disordered proteins, which tend to have smaller alignment lengths. Their results for proteins with smaller alignment lengths demonstrate a positive correlation between the percent of residues predicted to be solvent exposed and evolutionary rate. Results for proteins with larger alignments were consistent with the Bloom et al. findings, using either predicted or known percent exposure. The Lin et al. study makes the general observation that “proteins with a high [percentage of exposed residues] may evolve slowly or fast, whereas proteins with a low [percentage of exposed residues] almost always have a low evolutionary rate” [51]. Their conclusions stress the importance of fitness density in constraining evolutionary rate, a force which opposes designability-driven sequence divergence.

4.3 Constitutional properties

Although solvent exposure gets the most attention as a driving force in protein evolution, it is not the only physical parameter to be studied in this context. Multiple studies have considered the role of protein sequence length in evolution (which corresponds to the final size of the folded protein) [7,56]. Simple organisms

have evolved with a strong evolutionary pressure for reduced genome size, which may have evolutionary implications for protein sequences [56]. A significant positive correlation between length and evolutionary rate does appear to exist, and it is much more pronounced in short proteins (less than 250 amino acids). Studying protein length also provides a good example of the complexity inherent to isolating determinants of evolution. Although length appears to correlate with evolutionary rate, it is also known to be correlated inversely with expression [56] and directly with contact density [7,56] (both of which are determinants of evolution in their own right). Carefully controlling or isolating individual factors can be a challenge, even with advanced statistical techniques.

A protein's amino acid composition should also be considered as a potential determinant of evolutionary rate. It is well known that mutations between amino acids do not occur with equal frequencies. For example, mutations that swap one hydrophobic residue for another are commonly observed, suggesting that these mutations are neutral or only slightly deleterious. In contrast, mutations between hydrophobic and hydrophilic residues are far less common, suggesting that such transitions are generally disfavored by selection. These types of observations are the basis for amino acid substitution matrices, such as PAM [57] and BLOSUM [58], which are key components of sequence alignment and other bioinformatics algorithms. Recent work has shown that the space of acceptable mutations widens with protein divergence; this result applies to both general patterns of substitution as well as the specific requirements in buried versus exposed regions [52]. An early analysis using a small set of sequences suggested that the evolutionary trajectory of a protein could be inferred based on amino acid composition alone [59]. A more recent study with a much larger dataset rejected this hypothesis, concluding that amino acid composition contributes only weakly to predictions of evolutionary rate [60]. Thus, as was the case with solvent exposure, properties at the residue level do not necessarily translate directly to whole protein behavior. As far as evolution is concerned, proteins appear to be more than just a sum of their parts.

The next level up in protein organization involves secondary structures motifs—small structural elements that show up repeatedly within many different protein folds. Well-known examples include helices, strands, loops, and turns. Work in yeast has shown that the secondary structure composition of a protein does not appear to influence its evolutionary rate [7]. However, in a study of mammalian proteins, residues in helices and strands were shown to evolve more slowly than those in the less ordered loops and turns [53]. This last result highlights another apparent influence on evolution: molecular disorder. Disordered regions of proteins are generally known to evolve more rapidly than their ordered counterparts [61]. Our discussion to this point has assumed that a useful structure and a stable fold are synonymous—this is not necessarily the case. In fact, many proteins perform functional roles in the cell despite the fact that they, either in whole or in part, fail to achieve a fixed three-dimensional fold [62]. The fact that these proteins also appear to experience relaxed selection raises interesting questions about their evolutionary potential.

4.4 Protein domains

Protein domains, the next level up in the hierarchy of protein constitution, are important enough to warrant a separate discussion. Domains can be defined based on function, structure, or sequence characterization; in many cases the different approaches are compatible. We naturally adopt a structure-based definition: a protein domain is a spatially distinct structure (or structural component) that could conceivably fold and function in isolation [63]. Some proteins consist of a single domain, while others are composed of multiple domains each folded separately from a subsection of the underlying amino acid chain. To this point, our notion of the genotype to phenotype relationship has been protein sequence \rightarrow protein structure. Given the discrete spatial nature of domains, protein *subsequence* \rightarrow domain would be an equally valid definition. In fact, the notion of a protein fold (as in, “this fold is highly designable”) translates naturally to the protein domain concept.

To our knowledge, “domain constitution” of a protein has not been considered as a determinant of evolutionary rate. This results from the fact that domains are large, discrete units of proteins, unlike fine scale properties like buried residues or secondary structure elements. Instead, domains are typically considered as evolutionary targets in their own right [64]. Domains share a natural history that is similar in many ways to the phylogenies describing evolution at the level of whole organisms [63]. Many modern domains are thought to have evolved and radiated from lineages of ancestral domains, which were in turn derived from primordial protein folds. Domains without a shared evolutionary history may have also acquired similar structures due to convergent evolution [65]. Note that this is highly compatible with notions from designability theory. Classifying domains based on these principles is the primary mission of databases such as Pfam [17], CATH [66], and SCOP [15]. Some relationships between domains can be inferred at the sequence level, but owing to the many-to-one mapping of sequences to structures, structure comparison methods are often critical for describing connections between domains from distantly diverged proteins.

The discrete nature of domains has played an important role in protein evolution. After the evolution of a handful of primordial domains, many new functions could be efficiently evolved through their combination and permutation [67]. This process is facilitated by genomic evolution, in which pieces of genetic material (such as those encoding amino acid subsequences responsible for protein domains) are readily duplicated, fused, shuffled, and deleted [68]. In cases of domain duplication, while the original template continues to fill its role in the cell, the duplicate has the freedom to explore sequence and structure space, possibly acquiring new functions in the process [64]. The distribution of domains and proteins produced in this process follows power law behavior [69,70], which is emerging as a common trait among large scale biological systems. While interesting, genomic perspectives on domain evolution take us too far afield from our structural focus.

4.5 Function

As mentioned in the introduction, it has been suggested that a protein's functional classification is generally not a good predictor of its evolutionary rate [6]. However, some basic functional attributes of a protein certainly have important structural and evolutionary implications. For example, Kimura and Ohta demonstrated as far back as 1973 that residues involved with binding the heme group in α and β globin (the protein constituents of the hemoglobin molecule) evolve at one tenth the rate of the background structure [71]. Residues like these contribute to a protein's functional density and hence to the revised fitness density as well. Unlike structural properties that are common to many (or all) proteins, structure-function properties tend to be highly specialized, and are better reviewed on a case-by-case basis. A great deal of literature linking specific structure-function relationships to evolution is available for the interested reader.

5. EMPIRICAL RESULTS: HIGHER ORDER PROPERTIES

5.1 Interfaces

We begin our discussion of higher order structural properties with a final single protein property: interfaces. Although interfaces are properties of the unique protein structure to which they belong, they form a variety of interesting larger structures—each with evolutionary significance—when we consider them together. Interfaces are intimately linked with the notions of solvent accessibility and burial discussed previously, and several studies have investigated both simultaneously. We consider this to be a preferred approach, as interfacial residues and surface area (and their evolutionary contributions) will be wrongly counted as exposed residues and surface area when proteins are considered independently.

An early study of the cytochrome *c* protein structure revealed that some portions of the surface seemed to be experiencing unusually high functional constraint [72]. These surface residues were determined to be sites of interaction with other proteins (interfaces). Subsequent studies have generally supported the notion that interfacial surfaces are more conserved than the remainder of the protein's solvent-exposed surface, and slightly less conserved than the protein's core [73]. Substitutions that *do* occur in the interface are heavily skewed toward more conservative changes [53], as defined by the Grantham classification scheme [74]. Exploiting the difference in evolutionary rate between interfacial and non-interfacial sections of a protein's surface has been proposed as a means by which to identify interfaces in newly characterized proteins; this has proven to be difficult in practice [75].

The notion that evolutionary rate of an "average interface" is intermediate to those of buried and solvent-exposed portions of a protein seems very intuitive. Interfaces will likely spend at part of their lives in a buried state (when interacting) and another part in a solvent-exposed state (when not interacting). One might therefore expect the rate of evolution at an interface to scale inversely with the proportion of time that it is active; indeed, this is precisely what has been found [76].

Evolutionary rate among residues belonging to transient interfaces is significantly higher than for those found in constitutive (permanent) interfaces; rates for both are intermediate to those of buried and solvent-exposed residues. Decreased evolutionary rate at constitutive interaction sites may also reflect specific structural constraints imposed by the protein's interaction partner [76]. This represents a case of *coevolution* between protein structures, an instance of a higher order structure-evolution relationship.

5.2 Protein–protein interaction networks

Studying the topological structure (not to be confused with molecular structure) of protein–protein interaction networks is a hot topic in systems biology research. In such a network, proteins are represented as vertices, and interactions between protein pairs are represented as edges. For our purposes, interactions can be thought of as direct physical connections between the involved proteins (such as those mediated by interfaces); other common notions of protein–protein interactions exist that are equally important, but they lack structural significance. Interactions in these networks tend to follow a power law distribution, such that a small number of proteins have a very high degree (many interactions) and a large number have a very low degree (few interactions) [77]. Proteins with many interaction partners (hubs) tend to be essential and evolve slowly; whether or not there is a functional dependence between the number of a protein's interaction partners and its selective constraint has been a topic of contention [78]. Integrating network topology with expression data has also shown that hub proteins can be divided into two classes based on the timing of their interactions:

- (1) *party hubs*, which interact with several partners simultaneously, and
- (2) *date hubs*, which interact in a “one-partner-at-a-time” fashion [79].

The importance of integrating structural information into biological networks has been recognized [80], but relatively few studies have actually taken this leap. One such study related the number and extent of interfacial surfaces on a protein to its behavior in a network [81]. This approach allowed hubs to be partitioned into multi-interface and *singlish*-interface classes, which act as structural analogs of the temporal party and date hub classifications, respectively (note: *singlish* implies 1 or 2 interfaces). While *singlish*-interface hubs can evolve a new interaction through duplication and divergence of a partner, new interactions in multi-interface hubs necessitate the creation of a new binding interface. Finally, the study concludes that the extent of a protein's surface area involved in interactions is a better predictor of evolutionary rate than its number of interaction partners, in agreement with previous proposals [82]. Other efforts have directly employed structural information in network construction. We mentioned previously that it is difficult to predict interfacial components of a protein's surface based on conservation alone. Another structural approach to interaction prediction involves the consideration of protein domains [83]. Recall that domains are large subsections of proteins, typically with well conserved, discrete structures. Imagine that Domain *A* in Protein 1 and

Domain *B* in Protein 2 are found to physically interact. Protein 3, having uncharacterized interaction potential, is found to contain Domain *B*, either by structure or sequence comparison. It is reasonable to hypothesize that an interaction between Protein 1 and Protein 3 may occur in the cell. Networks based on domain interactions have been created following this logic with great success [84,85]. These networks further highlight the importance of structural modularity in the evolution of single proteins and protein networks. Returning to the example, while it is *possible* for Proteins 1 and 3 to interact on the basis of conserved domain relationships, this interaction is not a given. Conserved domain pairs that violate this assumption are common, and typically only differ by a few surface mutations; these subtle changes are enough to dramatically decrease the stability of the interaction [86].

Disordered (unfolded) regions of a protein are known to perform important biological functions, in spite of relaxed constraint on their three dimensional structures. It has been shown that hub proteins, which are believed to be constrained by coevolution with their interaction partners, are also more likely to feature intrinsic disorder [87]. This provides another example of a pair of deterministic forces in evolution with a paradoxical relationship. A recent work by Kim et al. addresses this issue by considering the precise physical context of disordered regions that occur in interacting proteins [88].

5.3 Protein complexes

Proteins seldom perform their functions in isolation. Either for the purpose of building multi-component architectural structures or streamlining functions, proteins are often grouped in space as complexes. This higher level structure is metaphorically similar to the way in which domains are grouped to build more sophisticated individual proteins. Note however that the combination of discrete proteins into complexes is a purely physical process, whereas domains are linked both physically and genetically by the underlying protein sequence. This first definition of a protein complex is generally given to groups of proteins which all interact constitutively. Complexes in this sense are analogs of the party hubs in expression-based networks and multi-interface hubs in structure-based networks. Evolutionary insights about these hubs are equally applicable to complexes, and vice versa.

One of the interesting connections between evolution and structure in complexes relates to the *balance hypothesis*, as described by Papp et al. [89]. This hypothesis states that changes which affect the proportions of complex-forming proteins in a cell will be deleterious, and hence purged by selection. A reduction in availability (either whole or partial) of a complex component limits the number of complete complexes that the cell can build, which may have obvious fitness consequences. Perhaps less intuitive is the fact that an over-available component may also represent a fitness loss, either by disrupting the kinetics of proper structure assembly, or by carrying out some “unsupervised” activity in its lone state. Thus, evolution acts to maintain fixed stoichiometry among the components of important complexes. It is possible, however, for the genomic segment encoding the

entire complex to be duplicated, as in this case the stoichiometry among complex components is maintained.

6. SUMMATION

The role that structure plays in protein evolution is evident on many scales. Single residues feel differences in selective constraint according to the extent of their solvent exposure. Whole proteins have a freedom to diverge that varies with the degree of disorder in their native structures. Cassettes of independent structures evolve together in order to maintain strict interaction proportions. These are examples of observations that have been made by considering real protein structures. Above this level there exists an armamentarium of theory and models to describe the structurally significant trends or events in protein evolution that we have not yet been able to observe directly. Much has been learned, and much remains to be discovered. Several ideas will motivate future work at the interface of structural and evolutionary biology.

(i) More atomic level structure data is needed. While few would scoff at the set of structures available in current databases, this represents only a minute fraction of the proteins present in nature today. We have a notion that a small set of structures is likely to be dominant among the protein universe. This notion lends itself well to summary and classification, but not to exhaustive description of the protein universe. As we saw in the case of domain interactions, the difference of a few amino acids in otherwise identical folds can be enough to significantly differentiate their behaviors. Advances in structure determination methods will be useful for increasing the pool of solved lone and complexed protein structures, which can then be pipelined into theoretical and empirical studies.

(ii) New ways of considering structure must be developed. Protein structures mediate biological function, and their evolution is shaped by that relationship. Approaches that integrate structural information into biological analysis, particularly analysis at the systems scale, will produce a more complete picture of the mechanisms that drive living organisms. Also implicit in this idea is a need for new methods to describe structures. Designability, evolvability, and fitness density are significant quantitative structural measures that influence a protein's evolution. How to precisely define and determine quantities such as these in real proteins remains an open question.

(iii) Theoretical and empirical results must keep pace with one another. We have a wealth of theoretical ideas concerning structure-evolution relationships. While many of these are very intuitive and have a "sense" of relevance, true relevance must be earned through explorations in real world systems. Advances in modeling of the sequence-structure relationship—e.g., progress made in the protein folding problem—will facilitate more realistic *in silico* models of protein evolution. Better integration of available data and directed laboratory evolution experiments will also aid in this goal. From the opposite perspective, theoretical treatments of structural (and other) determinants of protein evolution must be advanced to better handle the wealth of data available to them. Thus far, the inherent

noise and vast interconnections among these biological variables have proven to be worthy adversaries to state-of-the-art analysis methods.

Current work in protein design and directed evolution promises to produce exciting new discoveries at the interface of structure and evolution. These approaches seek to simulate the evolutionary processes we have discussed here in a laboratory context, providing researchers with a realistically short evolutionary timescale (the advantage of theoretical work) and the relevance of working with real proteins (the advantage of empirical observation). For further review of these approaches, see [90,91].

With continued progress toward these research objectives, we expect that our knowledge of biology and evolution will continue to strengthen in the light of molecular structure.

ACKNOWLEDGMENTS

Y.X. is supported by a Research Starter Grant in Informatics from the PhRMA Foundation. E.F. is supported by an IGERT Fellowship through NSF grant DGE-0654108 awarded to the BU Bioinformatics Program.

REFERENCES

1. Dobzhansky, T. Biology, molecular and organismic. *Am. Zool.* 1964, 4, 443–52.
2. McInerney, J.O. The causes of protein evolutionary rate variation. *Trends Ecol. Evol.* 2006, 21(5), 230–2.
3. Kellis, M., Birren, B.W., Lander, E.S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 2004, 428(6983), 617–24.
4. Hurst, L.D. The Ka/Ks ratio: Diagnosing the form of sequence evolution. *Trends Genet.* 2002, 18(9), 486.
5. Pal, C., Papp, B., Lercher, M.J. An integrated view of protein evolution. *Nat. Rev. Genet.* 2006, 7(5), 337–48.
6. Rocha, E.P. The quest for the universals of protein evolution. *Trends Genet.* 2006, 22(8), 412–6.
7. Bloom, J.D., et al. Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.* 2006, 23(9), 1751–61.
8. Plotkin, J.B., Fraser, H.B. Assessing the determinants of evolutionary rates in the presence of noise. *Mol. Biol. Evol.* 2007, 24(5), 1113–21.
9. Rocha, E.P., Danchin, A. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol. Biol. Evol.* 2004, 21(1), 108–16.
10. Akashi, H. Translational selection and yeast proteome evolution. *Genetics* 2003, 164(4), 1291–303.
11. Drummond, D.A., et al. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. USA* 2005, 102(40), 14338–43.
12. Goldberg, A.L. Protein degradation and protection against misfolded or damaged proteins. *Nature* 2003, 426(6968), 895–9.
13. Crick, F. Central dogma of molecular biology. *Nature* 1970, 227(5258), 561–3.
14. Dill, K.A., et al. The protein folding problem: When will it be solved?. *Curr. Opin. Struct. Biol.* 2007, 17(3), 342–6.
15. Hubbard, T.J., et al. SCOP: A structural classification of proteins database. *Nucleic Acids Res.* 1997, 25(1), 236–9.
16. Chothia, C. Proteins. One thousand families for the molecular biologist. *Nature* 1992, 357(6379), 543–4.

17. Bateman, A., et al. The Pfam protein families database. *Nucleic Acids Res.* 2002, 30(1), 276–80.
18. Helling, R., et al. The designability of protein structures. *J. Mol. Graph Model* 2001, 19(1), 157–67.
19. Kimura, M. *The Neutral Theory of Molecular Evolution* Cambridge: Cambridge Univ. Press; 1983.
20. Wagner, A. Robustness, evolvability, and neutrality. *FEBS Lett.* 2005, 579(8), 1772–8.
21. Sharp, P.M., et al. DNA sequence evolution: The sounds of silence. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 1995, 349(1329), 241–7.
22. Wong, P., Frishman, D. Fold designability, distribution, and disease. *PLoS Comput. Biol.* 2006, 2(5), e40.
23. Li, H., Tang, C., Wingreen, N.S. Are protein folds atypical?. *Proc. Natl. Acad. Sci. USA* 1998, 95(9), 4987–90.
24. Li, H., et al. Emergence of preferred structures in a simple model of protein folding. *Science* 1996, 273(5275), 666–9.
25. Kirschner, M., Gerhart, J. Evolvability. *Proc. Natl. Acad. Sci. USA* 1998, 95(15), 8420–7.
26. Huynen, M.A., Stadler, P.F., Fontana, W. Smoothness within ruggedness: The role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA* 1996, 93(1), 397–401.
27. Stadler, B.M., et al. The topology of the possible: Formal spaces underlying patterns of evolutionary change. *J. Theor. Biol.* 2001, 213(2), 241–74.
28. Fontana, W., Schuster, P. Shaping space: The possible and the attainable in RNA genotype-phenotype mapping. *J. Theor. Biol.* 1998, 194(4), 491–515.
29. Schuster, P., et al. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Biol. Sci.* 1994, 255(1344), 279–84.
30. Itzhaki, L., Wolynes, P. The quest to understand protein folding. *Curr. Opin. Struct. Biol.* 2008, 18(1), 1–3.
31. Dill, K.A., et al. Principles of protein folding—A perspective from simple exact models. *Protein Sci.* 1995, 4(4), 561–602.
32. Shakhnovich, E.I. Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* 1997, 7(1), 29–40.
33. Xia, Y., Levitt, M. Simulating protein evolution in sequence and structure space. *Curr. Opin. Struct. Biol.* 2004, 14(2), 202–7.
34. Hinds, D.A., Levitt, M. From structure to sequence and back again. *J. Mol. Biol.* 1996, 258(1), 201–9.
35. Park, B.H., Levitt, M. The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* 1995, 249(2), 493–507.
36. Mirny, L., Shakhnovich, E. Protein folding theory: From lattice to all-atom models. *Ann. Rev. Biophys. Biomol. Struct.* 2001, 30, 361–96.
37. Xia, Y., Levitt, M. Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc. Natl. Acad. Sci. USA* 2002, 99(16), 10382–7.
38. Xia, Y., Levitt, M. Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution. *Proteins* 2004, 55(1), 107–14.
39. Mirny, L.A., Abkevich, V.I., Shakhnovich, E.I. How evolution makes proteins fold quickly. *Proc. Natl. Acad. Sci. USA* 1998, 95(9), 4976–81.
40. Gutin, A.M., Abkevich, V.I., Shakhnovich, E.I. Evolution-like selection of fast-folding model proteins. *Proc. Natl. Acad. Sci. USA* 1995, 92(5), 1282–6.
41. Wagner, A. Robustness and evolvability: A paradox resolved. *Proc. Biol. Sci.* 2008, 275(1630), 91–100.
42. Lenski, R.E., Barrick, J.E., Ofria, C. Balancing robustness and evolvability. *PLoS Biol.* 2006, 4(12), e428.
43. Bloom, J.D., et al. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. USA* 2006, 103(15), 5869–74.
44. Bloom, J.D., et al. Stability and the evolvability of function in a model protein. *Biophys. J.* 2004, 86(5), 2758–64.
45. Earl, D.J., Deem, M.W. Evolvability is a selectable trait. *Proc. Natl. Acad. Sci. USA* 2004, 101(32), 11531–6.
46. Drenth, J. *Principles of X-Ray Crystallography* New York: Springer; 1999.
47. Clore, G.M., Gronenborn, A.M. Determining the structures of large proteins and protein complexes by NMR. *Trends Biotechnol.* 1998, 16(1), 22–34.

48. Bernstein, F.C., et al. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 1977, 112(3), 535–42.
49. Altschul, S.F., et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 1997, 25(17), 3389–402.
50. Perutz, M.F., Kendrew, J.C., Watson, H.C. Structure and function of haemoglobin II. Some relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.* 1965, 13, 669–78.
51. Lin, Y.S., et al. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Mol. Biol. Evol.* 2007, 24(4), 1005–11.
52. Sasidharan, R., Chothia, C. The selection of acceptable protein mutations. *Proc. Natl. Acad. Sci. USA* 2007, 104(24), 10080–5.
53. Choi, S.S., Vallender, E.J., Lahn, B.T. Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes. *Mol. Biol. Evol.* 2006, 23(11), 2131–3.
54. Zuckerkandl, E. Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. *J. Mol. Evol.* 1976, 7(3), 167–83.
55. England, J.L., Shakhnovich, E.I. Structural determinant of protein designability. *Phys. Rev. Lett.* 2003, 90(21), 218101.
56. Warringer, J., Blomberg, A. Evolutionary constraints on yeast protein size. *BMC Evol. Biol.* 2006, 6, 61.
57. Dayhoff, M., Schwartz, R.M., Orcutt, B. Atlas of Protein Sequence and Structure Silver Spring: National Biomedical Research Foundation; 1978.
58. Henikoff, S., Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 1992, 89(22), 10915–9.
59. Graur, D. Amino acid composition and the evolutionary rates of protein-coding genes. *J. Mol. Evol.* 1985, 22(1), 53–62.
60. Tourasse, N.J., Li, W.H. Selective constraints, amino acid composition, and the rate of protein evolution. *Mol. Biol. Evol.* 2000, 17(4), 656–64.
61. Brown, C.J., et al. Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* 2002, 55(1), 104–10.
62. Wright, P.E., Dyson, H.J. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 1999, 293(2), 321–31.
63. Ponting, C.P., Russell, R.R. The natural history of protein domains. *Ann. Rev. Biophys. Biomol. Struct.* 2002, 31, 45–71.
64. Orengo, C.A., Thornton, J.M. Protein families and their evolution—A structural perspective. *Ann. Rev. Biochem.* 2005, 74, 867–900.
65. Lupas, A.N., Ponting, C.P., Russell, R.B. On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* 2001, 134(2–3), 191–203.
66. Orengo, C.A., et al. CATH—A hierarchic classification of protein domain structures. *Structure* 1997, 5(8), 1093–108.
67. Chothia, C., et al. Evolution of the protein repertoire. *Science* 2003, 300(5626), 1701–3.
68. Bjorklund, A.K., et al. Domain rearrangements in protein evolution. *J. Mol. Biol.* 2005, 353(4), 911–23.
69. Zhang, C., DeLisi, C. Estimating the number of protein folds. *J. Mol. Biol.* 1998, 284(5), 1301–5.
70. Dokholyan, N.V., Shakhnovich, B., Shakhnovich, E.I. Expanding protein universe and its origin from the biological Big Bang. *Proc. Natl. Acad. Sci. USA* 2002, 99(22), 14132–6.
71. Kimura, M., Ota, T. Mutation and evolution at the molecular level. *Genetics* 1973, 73(Suppl 73), 19–35.
72. Dickerson, R.E. The structures of cytochrome c and the rates of molecular evolution. *J. Mol. Evol.* 1971, 1(1), 26–45.
73. Valdar, W.S., Thornton, J.M. Protein–protein interfaces: Analysis of amino acid conservation in homodimers. *Proteins* 2001, 42(1), 108–24.
74. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* 1974, 185(4154), 862–4.
75. Caffrey, D.R., et al. Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* 2004, 13(1), 190–202.

76. Mintseris, J., Weng, Z. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc. Natl. Acad. Sci. USA* 2005, 102(31), 10930–5.
77. Albert, R. Scale-free networks in cell biology. *J. Cell. Sci.* 2005, 118(Pt 21), 4947–57.
78. Jordan, I.K., Wolf, Y.I., Koonin, E.V. No simple dependence between protein evolution rate and the number of protein–protein interactions: Only the most prolific interactors tend to evolve slowly. *BMC Evol. Biol.* 2003, 3, 1.
79. Han, J.D., et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* 2004, 430(6995), 88–93.
80. Aloy, P., Russell, R.B. Structural systems biology: Modelling protein interactions. *Nat. Rev. Mol. Cell. Biol.* 2006, 7(3), 188–97.
81. Kim, P.M., et al. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* 2006, 314(5807), 1938–41.
82. Fraser, H.B., et al. Evolutionary rate in the protein interaction network. *Science* 2002, 296(5568), 750–2.
83. Kiel, C., Beltrao, P., Serrano, L. Analyzing protein interaction networks using structural information. *Ann. Rev. Biochem.* 2008, 0, 0.
84. Deng, M., et al. Inferring domain–domain interactions from protein–protein interactions. *Genome Res.* 2002, 12(10), 1540–8.
85. Schlicker, A., et al. Functional evaluation of domain–domain interactions and human protein interaction networks. *Bioinformatics* 2007, 23(7), 859–65.
86. Kiel, C., Serrano, L. Prediction of Ras-effector interactions using position energy matrices. *Bioinformatics* 2007, 23(17), 2226–30.
87. Haynes, C., et al. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput. Biol.* 2006, 2(8), e100.
88. Kim, P.M., et al. The role of disorder in interaction networks: A structural analysis. *Mol. Syst. Biol.* 2008, 4, 179.
89. Papp, B., Pal, C., Hurst, L.D. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 2003, 424(6945), 194–7.
90. Pokala, N., Handel, T.M. Review: protein design—Where we were, where we are, where we're going. *J. Struct. Biol.* 2001, 134(2–3), 269–81.
91. Farinas, E.T., Bulter, T., Arnold, F.H. Directed enzyme evolution. *Curr. Opin. Biotechnol.* 2001, 12(6), 545–51.