REVIEW

The interface of protein structure, protein biophysics, and molecular evolution

David A. Liberles,[1]* Sarah A. Teichmann,[2]* Ivet Bahar,[3] Ugo Bastolla,[4] Jesse Bloom,[5] Erich Bornberg-Bauer,[6] Lucy J. Colwell,[2] A. P. Jason de Koning,[7] Nikolay V. Dokholyan,[8] Julian Echave,[9] Arne Elofsson,[10] Dietlind L. Gerloff,[11] Richard A. Goldstein,[12] Johan A. Grahnen,[1] Mark T. Holder,[13] Clemens Lakner,[14] Nicholas Lartillot,[15] Simon C. Lovell,[16] Gavin Naylor,[17] Tina Perica,[2] David D. Pollock,[7] Tal Pupko,[18] Lynne Regan,[19] Andrew Roger,[20] Nimrod Rubinstein,[18] Eugene Shakhnovich,[21] Kimmen Sjölander,[22] Shamil Sunyaev,[23] Ashley I. Teufel,[1] Jeffrey L. Thorne,[14] Joseph W. Thornton,[24,25,26] Daniel M. Weinreich,[27] and Simon Whelan[16]

[1]Department of Molecular Biology, University of Wyoming, Laramie, Wyoming 82071

[2]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, United Kingdom

[3]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15213

[4]Bioinformatics Unit. Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Universidad Autonoma de Madrid, 28049 Cantoblanco Madrid, Spain

[5]Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109

[6]Evolutionary Bioinformatics Group, Institute for Evolution and Biodiversity, University of Muenster, Germany

[7]Department of Biochemistry and Molecular Genetics, School of Medicine, University of Colorado, Aurora, Colorado

[8]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, North Carolina 27599

[9]Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín, Martín de Irigoyen 3100, 1650 San Martín, Buenos Aires, Argentina

[10]Department of Biochemistry and Biophysics, Center for Biomembrane Research, Stockholm Bioinformatics Center, Science for Life Laboratory, Swedish E-science Research Center, Stockholm University, 106 91 Stockholm, Sweden

[11]Biomolecular Engineering Department, University of California, Santa Cruz, California 95064

[12]Division of Mathematical Biology, National Institute for Medical Research (MRC), Mill Hill, London NW7 1AA, United Kingdom

[13]Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, Kansas 66045

[14]Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27695

[15]Département de Biochimie, Faculté de Médecine, Université de Montréal, Montréal, QC H3T1J4, Canada

[16]Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, United Kingdom

[17]Department of Biology, College of Charleston, Charleston, South Carolina 29424

[18]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

[19]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven 06511

[20]Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS, Canada

[21]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138

[22]Department of Bioengineering, University of California, Berkeley, Berkeley, California 94720

Correspondence to: David A. Liberles, Department of Molecular Biology, University of Wyoming, Laramie, WY 82071. E-mail: liberles@uwyo.edu or Sarah A. Teichmann, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB20QH, UK. E-mail: sat@mrc-lmb.cam.ac.uk.

Grant sponsor: NSF EF; Grant number: 0905606.

[23]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115
[24]Howard Hughes Medical Institute and Institute for Ecology and Evolution, University of Oregon, Eugene, Oregon 97403
[25]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637
[26]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637
[27]Department of Ecology and Evolutionary Biology, and Center for Computational Molecular Biology, Brown University, Providence, Rhode Island 02912

Received 2 March 2012; Revised 22 March 2012; Accepted 23 March 2012
DOI: 10.1002/pro.2071
Published online 30 March 2012 proteinscience.org

Abstract: Abstract The interface of protein structural biology, protein biophysics, molecular evolution, and molecular population genetics forms the foundations for a mechanistic understanding of many aspects of protein biochemistry. Current efforts in interdisciplinary protein modeling are in their infancy and the state-of-the art of such models is described. Beyond the relationship between amino acid substitution and static protein structure, protein function, and corresponding organismal fitness, other considerations are also discussed. More complex mutational processes such as insertion and deletion and domain rearrangements and even circular permutations should be evaluated. The role of intrinsically disordered proteins is still controversial, but may be increasingly important to consider. Protein geometry and protein dynamics as a deviation from static considerations of protein structure are also important. Protein expression level is known to be a major determinant of evolutionary rate and several considerations including selection at the mRNA level and the role of interaction specificity are discussed. Lastly, the relationship between modeling and needed high-throughput experimental data as well as experimental examination of protein evolution using ancestral sequence resurrection and *in vitro* biochemistry are presented, towards an aim of ultimately generating better models for biological inference and prediction.

Keywords: evolutionary modeling; domain evolution; sequence-structure-function relationships; protein dynamics; protein thermodynamics; gene duplication; protein expression; ancestral sequence reconstruction

Introduction

At the interface of protein structure, protein biophysics, and molecular evolution there is a set of fundamental processes that generate protein sequences, structures, and functions. A better understanding of these processes requires both biologically realistic models that bring structural and functional considerations into evolutionary analyses, and similarly incorporation of evolutionary and population genetic approaches into the analysis of protein structure and underlying protein biophysics. A recent meeting at NESCent (National Evolutionary Synthesis Center in Durham, NC) brought together evolutionary biologists, structural biologists, and biophysicists to discuss the overlap of these areas. The potential benefits of the synergy between biophysical and evolutionary approaches can hardly be overestimated. Their integration allows us not only to incorporate structural constraints into improved evolutionary models, but also to investigate how natural selection interacts with biophysics and thus explain how both physical and evolutionary laws have shaped the properties of extant macromolecules.

Fitness is a biological concept that describes the degree to which an individual is likely to contribute to future generations, and to thereby pass on traits (such as gene sequences) that it carries. Genetic variants may confer greater fitness and therefore selective advantage to individuals that carry them, or they may confer lower fitness and thus carriers will be at a selective disadvantage. Hence those variants conferring greater fitness are likely to replace other variants (become fixed) through positive selection, whereas those that confer a decrease in fitness are likely to be eliminated. This occurs against a backdrop of neutral genetic drift. Although simple to describe, the idea that variants may confer greater or lesser fitness in this genetic paradigm involves many layers of complexity. There is a long chain of molecular and physiological interactions linking the genetic variation and resulting individual molecular phenotypes to changes in the probability that an individual organism survives and reproduces.

Molecular phenotype is characterized by properties that affect protein function such as protein structure, protein stability, protein binding specificity, and

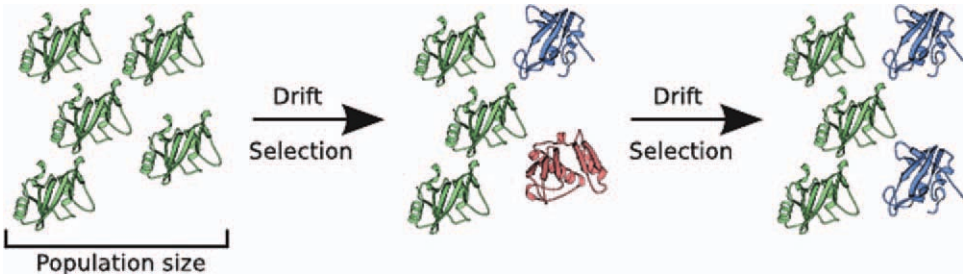

Figure 1. Evolution of proteins under selection for folding to maintain a function. The proteins exist in a population, the size of which determines the relative influences of drift and selection. The ancestral allele (green) is modified by mutation to deleterious (red) and nearly-neutral (blue) derived alleles, which are ultimately eliminated or fixed by selection or by drift randomly. Ancestral alleles are not always lost and derived alleles not always fixed. The process is stochastic rather than deterministic, described by the interplay of the strength of selection and population level dynamics. The figure is derived from PDB structures 1D4T (chain A), 1QG1 (chain E), and 1JD1 (chain A), which are used for illustrative purposes. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

protein dynamics. Ultimately, protein functions include specific processes, such as binding, catalysis, or transport. These functions generate questions that need to be answered to better understand protein evolution and enable downstream applications. What is the relationship between the above properties, protein function, and organismal fitness? As folding specificity is defined, what are the relevant thermodynamic properties necessary for folding? Misfolded, alternatively folded, and aggregate states are all possible but which are selected against? How large is the necessary energy gap between the native state and possible alternative conformations and what is the corresponding selective pressure? Is there a selective pressure against being too stable or is metastability a neutral emergent property of the evolutionary process? What then are the selective pressures on intrinsically unfolded proteins? Is it possible to derive general principles, or do the answers to these questions depend on the specific protein, organism, and environment?

Preliminary answers to some of these questions can be found in the literature. The long-standing observation that natural proteins are not excessively stable (typical stabilities of a protein domain range between 3 and 7 kcal/mol or from 5 to 10 kT units[1]) has been interpreted as evidence for selection against functionally detrimental over-stabilization of proteins.[2] Such a view reflects a selectionist paradigm, which posits that every observed trait has been optimized by selection. An alternative view is that the observed marginal stability of proteins is a result of mutation-selection balance[3–6] on a fitness landscape where stability is a neutral trait as long as it exceeds a certain threshold value. Simulations and analytical studies have shown that a realistic distribution of protein stabilities can be obtained on such a neutral landscape with the majority of proteins showing stability around 5 kcal/mol.[5] In this scenario the stability of protein domains is established as a result of a balance between mostly destabilizing mutations and selection against highly unstable proteins.

Comparative approaches have also been used to understand the targets of selection in proteins. Proteins of intracellular bacteria are estimated to be less stable with respect to misfolding (and possibly aggregation) than orthologous proteins of free living relatives. This can be interpreted as reduced selection due to the population size reductions (bottlenecks) that occur during transmission from host to host.[7] The predicted stability of misfolded structures is significantly larger for real protein sequences than for shuffled sequences due to destabilizing frequent contacts and correlated contact pairs.[8] Native contacts of short proteins are better optimized than those of large proteins, which are expected to undergo weaker selection since the number of intrachain contacts per residues is higher.[9]

As the field moves forward, it is clear that different models are needed to address different questions. For any model, rigorous assessment of its validity is required, either through simulations or comparison to empirical data. Models must generally conform to observed properties of proteins, such as the observations that surface residues of globular proteins undergo substitution more rapidly than those in the core, and that roughly 80% of nonsynonymous mutations are purged by selection in excess of the expectation of those eliminated by neutral drift.[10] Another potential benchmark for theoretical models is the observed coevolution of residues in structured proteins. In the next sections, we will survey the evolutionary models and the different ways of assessing these models based on evolution influenced by protein structure and biophysics (Fig. 1).

Common models for protein sequence evolution

Explicit probabilistic models of sequence change have a central role in the study of molecular evolution. Probabilistic models are attractive both because they allow qualitative exploration of protein evolution through simulation and because they permit parameter estimation and hypothesis evaluation via

likelihood-based statistical techniques. Evolution occurs within populations of organisms but widely employed inter-specific models of protein evolution often represent the proteins in a population with a single protein or codon sequence. In addition, these probabilistic models are usually site-independent and Markovian with respect to time. In other words, the models have the future of an evolutionary lineage depend on its current state (i.e., sequence) but not on earlier states visited in the history of a lineage. For example, the pioneering work of Halpern and Bruno[1] represented protein evolution as a Markovian process operating on one sequence in each instant in time as a simplification of the long term behavior of protein evolution in a population.

Evolutionary biologists commonly rely on models of sequence change that assume changes at one sequence position have no impact on whether other positions will change. The likelihood of the nucleotides or amino acids in an individual column of a multiple sequence alignment can be determined with the pruning algorithm of Felsenstein.[12] This assumption of independence between sites allows the probability of an observed set of aligned sequences at the tips of an evolutionary tree to be expressed as the product over alignment columns of the observed nucleotides or amino acids in those columns.[12] This independence assumption is simplistic, throwing away biological information, and can be shown statistically to be problematic, but permits computationally convenient likelihood-based inference.[13] Building upon this computational convenience, complex models that allow for lineage-specific rate shifts have been developed to phenomenologically (nonmechanistically) treat signal that may originate from site-interdependence.[14,15]

Relaxing assumptions of site-independence in models of sequence evolution

Understanding the coevolution of residues within protein structures is important for both the protein structure and evolutionary biology communities. There is an emerging strategy for achieving this understanding. To avoid assumptions of site-independence, the protein sequences are typically mapped to some phenotypic property, such as thermodynamic stability, folding ability, or some assay for functionality, and the substitution rate is expressed as a function of the resulting change in this property. These models have been developed for two specific goals. The first goal has been the investigation of the relationship between protein structure, function, foldability, and evolution, as well as realistic sequence simulation. Some of the early work in this area has relied upon extremely simple models of proteins, such as representing the structure as a self-avoiding walk on a cubical lattice, or reducing the amino acid alphabet to as few as two different residues. Several early models subsequently moved to protein structures with full amino acid alphabets.[16–21]

More recent efforts to model and simulate protein evolution have addressed thermodynamic properties of proteins, involving calculation of protein stability or binding affinity, requiring the use of some effective potential function that includes not only enthalphic terms (hydrogen bonding, van der Waals interactions) but also entropic terms (hydrophobicity, side-chain, and back-bone conformational entropy). In general, two broad classes of models have been developed, so called informational (knowledge-based) models that use pairwise statistical potentials[18,22,23] and so called physical models that apply a force field to a coarse-grained approximation of amino acid side chains[24,25] (see[26] for a comparison). These physical models are quite similar to models used in automatic "protein design"[27] and differ from each other in the degree of physical approximation used. A pioneering study by Dahiyat and Mayo[27] used a detailed description of the proteins and searched for the optimal position of all side-chains with an automatic design algorithm. In later studies flexibility of the backbone has also been included in the protein design programs,[28] but this may be computationally impractical to implement in an evolutionary context.

In the physical models, the terms have weights that are used to optimize the function. Variations in the force field used include weights derived from all PDB structures versus weights optimized from a single structure, side chain optimization with a fixed backbone versus no geometric side chain optimization, inclusion or exclusion of a binding (intermolecular association) interaction as part of the fitness function, and an energy gap that can include the unstructured state, explicit alternative folds, or a random contacts model.[29]

Both the informational models and the physical models have been developed with a balance of computational speed and accuracy in mind, but neither is yet accurate enough to be useful for questions that involve explicit sequence-structure-function evolution. In neither approach do biologically observed sequences score well. Aspects of negative design (selection against alternative folding/binding states) that are poorly understood might account for the poor explanation of native sequences, but fundamental problems with the assumptions of the thermodynamic model are a more likely explanation. In simulation work by Grahnen *et al.*,[25] an informational model that averages interaction propensities across all PDB structures and contexts shows changes in the frequencies of hydrophobic residues in the core and surface during simulation. Additionally, the substitution process lacks protein context specificity and support for a covarion model of substitution is never attained from sequences simulated using this particular model and simulation scheme. It is conceivable

that alternative informational models and implementations (for example, a model with more context-dependence in the interaction potentials) might have different evolutionary properties. The physical model used in the same work retains more detailed features of the protein including a hydrophobic core and appears to progress from an equal rates to a rates across sites to a covarion model during the simulation (a complex model of sequence evolution with shifting rates that retains the assumption of site-independence, see Ref. 14 for a review). However, lack of fit to the native state is still problematic, and examination of the structural model reveals a poorly packed protein with the approximate amino acid side chain representation inadequate. Improvements to the models are clearly needed to make them more useful for phylogenetic and sequence simulation purposes.

These structural/thermodynamic models define static interactions of amino acids within a structure without sufficient molecular flexibility or structural optimization upon mutation. A better model would more clearly connect the targets of selection to evaluated parameters in the model. How much selection acts directly on protein folding thermodynamics is unclear. Clearly, proteins are selected to function adequately. Their function may require them to bind specifically to interaction partners, to catalyze a reaction, or to transport what has been bound. How does the requirement to function interplay with folding thermodynamics in terms of selective pressures? Further, binding, catalysis, and transport are all governed by biophysical parameters, but how constant are these parameters across evolution given that members of pathways and networks are known to coevolve? What selective pressures do avoiding aggregation and requiring binding specificity place on sequences?

Given that our understanding of these issues is not yet complete, a sequence evolution model, that is, site-interdependent, but averages phenomenologically over some of these processes, may be a step forward (see for example[30]). Currently, existing mechanistic models cannot handle insertion and deletion (indel) events and models that deal mechanistically with the insertion and deletion processes are needed. Improved models for molecular evolution are needed to handle the functional and structural divergences that occur frequently following gene duplication events (i.e., if a phylogeny is to be estimated for a multigene family composed of paralogous groups). Such models will need to include the changing functional roles that occur at homologous sites, lineage- and site-specific rate variation, in addition to insertions and deletions relative to the common ancestor. If a phylogeny is to be estimated for an individual domain, but members of the family span different multi-domain architectures, the model will need to include domain architecture rearrangements. Simulation studies that can effectively model these complex evolutionary events would be very useful in elucidating the robustness of existing phylogenetic methods to handling these data.

Selection against alternative states (also termed folding selectivity and negative design) will also be an important aspect of models of protein evolution. The standard way to take into account misfolded structures of real proteins is through gapless threading[31,32] which involves explicit alternative states or use of a random contacts model[29] that typically averages over alternative states with the same amino acid composition and contact density.

Role of population genetic parameters

Protein evolution is not only dependent upon biophysical parameters. Underlying parameters associated with the mutation and fixation processes are also important. These include the mutation rate, the recombination rate, and the effective population size. There is a complex interplay between these parameters and the biophysical parameters associated with selection.[33] The effective population size is important in influencing the ability of selection to overcome stochastic neutral genetic drift. The link between the strength of selection and the actual number of individuals in the population is complex, especially when the actual population size is nonconstant. Several recent studies have begun to look specifically at the role of population genetic parameters in protein folding.[6,34]

Halpern and Bruno[11] were able to reconcile population genetics and protein evolution for the special situation where mutation rates are sufficiently low to have each new mutation be fixed (i.e., survive and eventually spread to all members of a population) or lost before the next one occurs. In this case, recombination can be ignored because linked sites are unlikely to be simultaneously polymorphic. For the low mutation rate situation, Kimura[35] derived a diffusion approximation for the probability that a new mutation is fixed. The approximation has the fixation probability be a function of the product of population size and the difference in relative fitness caused by the new mutation. These products have been referred to as "scaled selection coefficients".[36] Halpern and Bruno recognized that, if an evolutionary model has parameters that correspond to mutation and others that reflect natural selection, the Kimura fixation approximation could be used to convert parameter estimates to estimates of scaled selection coefficients.

Statistical inference with evolutionary models where sequence sites do not change independently

For statistical inference from sequences related by a phylogenetic tree, the pruning algorithm of Felsenstein[12] has been extensively employed for statistical

inference with models of protein evolution with the assumption that sequence positions (or individual codons within a sequence) change independently. But, conventional inference approaches become computationally impractical when sequences cannot be decomposed into short independently evolving units. For a data set of protein-coding DNA sequences, the goal might be to determine (or at least approximate) the probability of the observed sequence data at the tips of the tree conditional upon the evolutionary model, the tree, and values of parameters in the model. The challenge is that only the data at the tips of the tree are observed whereas the sequence at the root of the tree and the subsequent evolutionary events are not directly observed. Therefore, calculating the likelihood of the observed data entails an integration of probability densities over all possible root sequences and all possible subsequent histories of evolutionary events. Such an integration is most often computationally intractable for models of sequence change with dependence among sites.

Fortunately, evaluation of the probability density of individual substitution histories can be computationally feasible in many cases where integrating over all possible histories is prohibitive. Jensen and Pederson[37,38] exploited this fact to perform likelihood-based inference when models of sequence change have evolutionary dependence among sites. The basic idea is to augment the observed sequence data with a possible substitution history and to then use Markov chain Monte Carlo techniques to perform a random walk over histories that are consistent with the observed data.

Inspired by the approach of Parisi and Echave[18] for simulating protein evolution, Robinson *et al*.[39] adapted the ideas of Jensen and Pederson to statistical inference under a model of protein-coding DNA evolution that had codons change in a dependent fashion due to natural selection on protein tertiary structure (or any other aspect of phenotype for which the effect of a mutation can be predicted). More recent work[40–42] has greatly improved both the computational tractability of the inference procedure and the treatment of protein structure in these evolutionary models. An appealing feature of this line of research is that the predicted phenotypic effect of a mutation can be converted into a predicted substitution rate.

Models of protein evolution that incorporate protein structure have been shown to fit data better than the corresponding models that ignore protein structure.[41] However, an even better fit to the data could be achieved with state-of-the-art site-independent codon models.[42] Despite their having parameters with biologically meaningful explanations, the lackluster statistical fit of dependent site models is clearly disappointing. A silver lining for some phylogenetic applications could be that complicated biophysics-based models may not always be required. Lakner *et al*.[43] used simple measures of sequence-to-structure fit to study phylogenetic likelihood calculations under site-independent models. They calculated pseudo-energies for ancestral sequences from pairwise contact potentials, solvent accessibility terms and threading, and assessed specificity by considering a library of decoy structures. They found that likely substitution histories on phylogenetic trees mostly contain sequences that are consistent with the tertiary structure. The difference between these results and the less satisfactory results of Grahnen *et al*.[25] and Kleinman *et al*.[24] is likely due to the shorter evolutionary distances and the corresponding end point constraints that restricted paths through intermediates.

Phenomenological models

Problems with the structural models described may be due to problems in accurately describing protein thermodynamics, but they may also be due to a lack of understanding of the underlying biological fitness functions. One alternative is to use purely phenomenological models that attempt only to fit (and regenerate for sequence simulation) observed sequence data without considering underlying processes. Such models are typically judged by likelihood scores, Q-Q plots, and other measures of goodness of fit as the only benchmarks. A potential problem with these models is in their biological use and interpretation. For example, without models that adequately describe underlying biological processes, the phylogenetic tree estimate may not be reflective of the ancestral history of the sequences (a typical goal of phylogenetic tree reconstruction). In these cases, other signals, such as protein structure, protein function, and constraints at other levels of biological organization, override the ancestral history information in the sequences and result in inaccurate tree estimates.[44]

Sequence alignment

The context underpinning sequence alignment is a factor, that is, frequently overlooked when bringing together amino acid sequence and protein structure. The alignment represents a series of associations between the amino acids, which can then be interpreted either from a structural or evolutionary perspective (reviewed in Ref. 45). The structural perspective implies that corresponding amino acids are playing structurally corresponding roles, whereas the evolutionary perspective builds upon the assumption that amino acids have a shared common ancestor and that one can track nucleotide substitutions in their codon over time. In some cases these two perspectives coincide, and models that describe protein evolution in terms of function make structural sense. In other cases, there may be conflict

between evolutionary and structural homology (descent from common ancestry at the level of a position within a structure rather than a column in a multiple sequence alignment), that is, not accounted for in the model. It is unclear how these conflicts will affect downstream analyses and if the simpler evolutionary models or the structural models that do not model evolution make more accurate statements about common ancestry.

A second concern regarding sequence alignment is that recent research has shown that the outputs of different sequence alignment methods tend to produce different results that are not consistent,[46] and that sequence alignment accuracy degrades sharply with increasing evolutionary divergence.[47] If datasets are restricted to orthologs from closely related taxa (or to slow-evolving genes), sequence divergence may be less problematic, but if datasets include highly divergent taxa or span functionally divergent paralogous groups, alignment errors become increasingly likely and may cause significant errors in phylogenetic accuracy.[48]

There are several possible general solutions to this problem. One is to incorporate insertion and deletion events into models of protein evolution, which will make sophisticated models even slower and more complex computationally.[49–51] It is clear that affine gap penalties, like phenomenological models of sequence evolution, are not sufficiently reflective of underlying mechanistic processes.[52,53] Insertions and deletions are seldom modeled, since their effect on stability is more difficult to predict, particularly for insertions where new sequence is added in addition to changes in the orientation of existing structural elements. It is becoming increasingly evident that large novelties in protein evolution are produced by large insertion events in which an entire "domain" is added or deleted to a protein.[54] The proper modeling of insertion and deletion events will be a crucial step towards more realistic models of protein evolution.

Protein evolution at the level of the domain

Many biological systems, such as metabolic pathways, signaling pathways, and gene regulatory networks show a high degree of modularity with respect to the protein domains from which they are constructed. Domains are often autonomous and can be re-used in different contexts, with the potential to create high molecular functional diversity from a small number of operations. The modularity of domain recombination allows for swift changes to an organism's functional repertoire and the potential for rapid adaptation.[55] Domain rearrangement is a rare event, with rates much lower than the rates of amino acid substitution. Using structural domain assignments with hidden Markov models, Apic *et al*.[56] showed that a tiny fraction of the combinatorial potential of domain rearrangements is observed in the protein universe. The pairwise domain combinations have a scale-free network structure.[57] However, there are pairs and triplets of domains that act as evolutionary modules, and can be viewed as "supra-domains".[58] Which domain combinations have been discovered is probably a consequence of mutational opportunity, drift, and selection.

Ultimately, the wealth of available genomic data presents an unrivalled opportunity to study the functional importance of these molecular innovations, which can be retraced by comparative genomics with some accuracy. For instance, it was demonstrated that the evolution of domain architectures could primarily be explained by a simple scenario consisting of the addition or deletion of a single domain at the N or C-termini.[59,60] One notable exception to this rule is found in the case of repeating domains, that often are copied (or deleted) multiple domains at a time and at least equally frequent at the central region of a gene as at the termini.[61]

In Arthropods, the majority of new domain arrangements can be explained by simple, single-step modular rearrangement events dominantly at the N and C-termini of the proteins.[62] Modular rearrangements strongly impact all levels of the cellular signaling apparatus and thus have strong adaptive potential. Furthermore, emerging domains are predominantly found as single domains, thus most likely resulting from neighboring genomic regions.[63] A comparison with plant genome evolution reveals that the dynamics are qualitatively similar but with very different rates of emergence of novel domains.[64] Presumably, this is related to the complex interplay of domain rearrangements with the frequent whole genome duplication events observed in plant lineages.

Intrinsically disordered proteins

Of course, not all domains of proteins fit in the traditional model of a folded structure. Although there is some controversy over the fraction of proteins that show intrinsic disorder, their existence is consistent with the expectation that folding stability is not a target of natural selection for all proteins. Disordered proteins have shorter half-lives, reducing their potential for aggregation and misassembly,[65] which may affect selective pressures. Proteins that are partially or totally disordered in the native state should be accounted for in models of the evolution of protein stability (see[66] for an early review). Roughly 30% of human proteins are predicted to contain large unstructured regions.[67] These proteins are frequently involved in regulatory processes and contain short linear motifs, which exploit their ability to form very precise transient interactions, with high specificity but low affinity, and they often acquire structure only when they interact with other proteins or nucleic acids.[68] Disordered proteins are

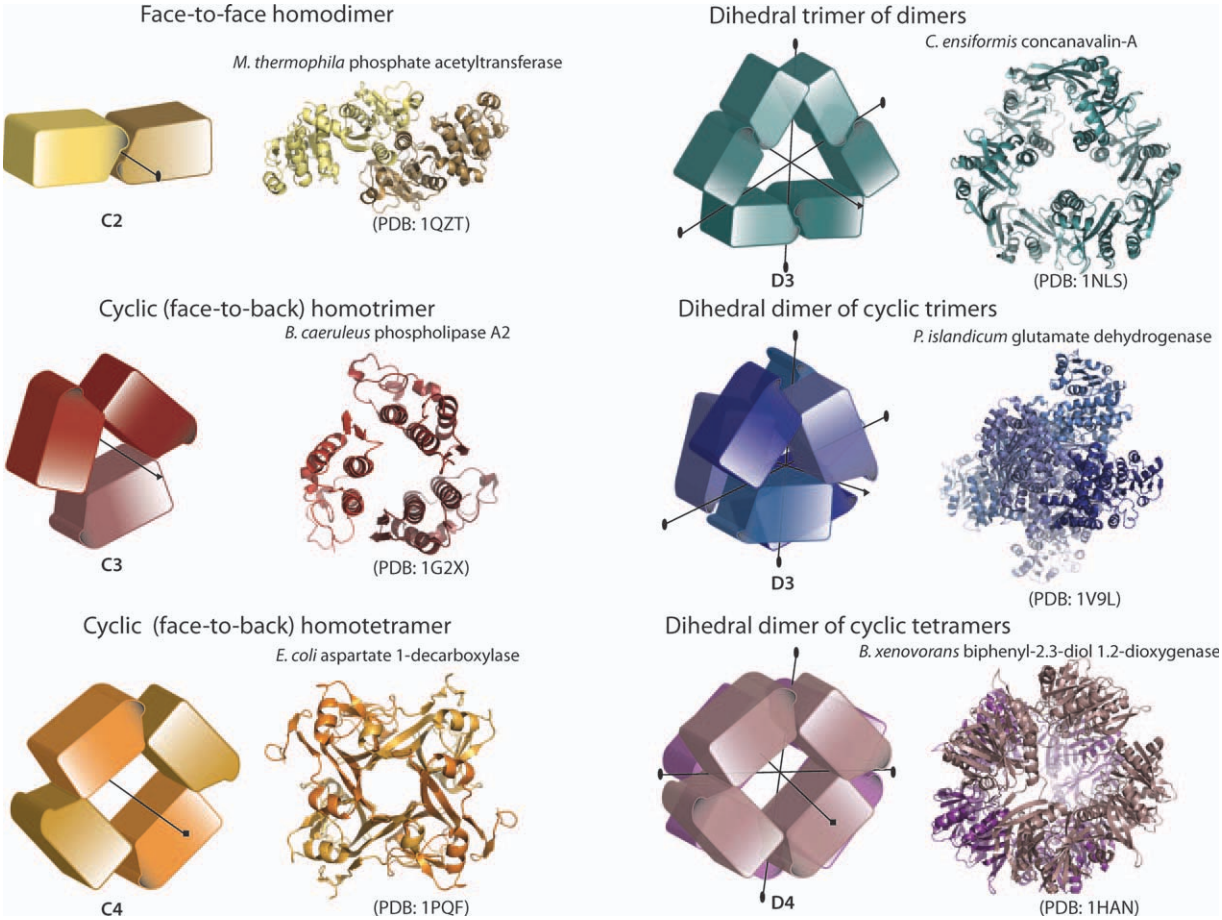

Figure 2. Symmetries of homomeric protein complexes. Complexes on the left hand side have cyclic symmetry (C_n), which means all subunits are related by rotation around a single n-fold rotation axis. Complexes on the right hand side have dihedral symmetry (D_n), which means they have an n-fold rotation axis that intersects a 2-fold rotation axis at right angles. Homomers have either symmetric face-to-face (e.g., a C2 homodimer, PDB:1QZT), or asymmetric face-to-back interfaces (e.g., a C3 homotrimer, PDB: 1G2X, or a C4 homotetramer, PDB: 1PQF). Symmetric interfaces result in complexes with dihedral symmetry, while asymmetric interfaces imply homomeric complexes with cyclic symmetry. Symmetric interfaces evolve more readily than asymmetric ones and thus there are more dihedral than cyclic complexes (see text). During the course of evolution, proteins can evolve multiple interfaces and form higher oligomers, such as trimers of dimers (D3, PDB: 1NLS); or dimers of trimers (D3, PDB: 1V9L) or tetramers (D4, PDB: 1HAN). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

often coupled to phosphorylation processes, which enhance their intrinsic flexibility even further and allow them to adapt to multiple interaction partners, thus enhancing their molecular complexity. Disordered proteins can be incorporated in biophysically aware models of evolution as an extreme case of flexibility, although perhaps one with greater evolutionary constraint than would normally be observed for extremely flexible regions.

An intriguing relationship between protein disorder and organism population size is provided by the study of proteins that form the centrosome, a large macromolecular complex that regulates animal cell differentiation and division. These proteins are predicted to be more phosphorylated than structured proteins from the same organism. Intrinsic disorder was found to increase in evolution along branches of the phylogenetic tree that lead to an increase of the number of cell types and a decrease in effective population size, mainly due to large insertions of new disordered regions, at a rate, that is, larger for centrosomal than for control proteins.[34,69] Thus, explicit consideration of population genetics is likely to be as important in understanding the evolution of disordered proteins as it is for ordered proteins.

Evolution of homomers

Analyses of all proteins of known three-dimensional structure,[70,71] functional genomics experiments,[72,73] and bioinformatic analyses of protein-protein interaction networks[74] show that the majority of proteins oligomerize (Fig. 2). Furthermore, they show that about half of cellular complexes are homomers, or complexes of self-interacting copies of the same gene product. There are numerous examples for how oligomerization benefits protein function and/or stability (reviewed in Refs. 75 and 76). However, for an oligomeric interaction to contribute to fitness, the protein

complex first needs to be significantly populated.[77] Therefore, the ubiquity of protein oligomers is not simply due to their adaptiveness but also to the evolutionary pathways by which they emerge. Rewiring of protein interactions is a common evolutionary event[78] and on average, only two mutations are sufficient to turn a protein surface into an interface.[79]

In the case of homomers, another major factor that enables evolvability of interactions is their symmetry. Andre *et al*.[77] modeled a random pool of protein complexes with low energy binding modes, and showed it is significantly enriched in symmetric interfaces. Structural symmetry enables a single mutation to have a two-fold impact[80] so a symmetrical, face-to-face interface is statistically more probable to emerge.[81] Symmetric interfaces result in complexes with dihedral symmetry, while asymmetric (face-to-back orientation within one plane) interfaces imply homomeric complexes with cyclic symmetry. A homomeric complex can have both types of interfaces, and many dihedral complexes can be described as stacks of cyclic complexes. Since symmetric interfaces evolve more easily than asymmetric ones in the first place, and are selected for a number of functional reasons, dihedral complexes are more abundant than cyclic complexes.[69]

One of the benefits of oligomerization is the increased stability due to the additional buried surface area of interface atomic groups. Destabilizing mutations can expose hydrophobic residues which can rapidly lead to aggregation into amorphous or amyloid aggregates.[82–85] There is significant selection pressure to avoid protein aggregation and deleterious gains of function.

Although burying additional protein surface in interfaces is not the main evolutionary strategy to increase the overall stability of the proteome,[86] oligomerization can compensate for a loss in stability, since protein interface formation and protein folding are governed by the same biophysical principles. This overlap is best illustrated by domain-swapped homomers.[87] Recent exhaustive analysis of available protein structures revealed that about 10% of protein folds, and 5% of protein families contain domain-swapped structures.[88] Moreover, proteins belonging to the same evolutionary family can have different domains swapped. Domain swapping can emerge as a compensatory response to a destabilizing mutation, which can cause a protein subdomain to unfold from the rest of the protein.[89] Unlike aggregation, domain swapping may preserve protein function, which may be the reason why these domain swapped proteins are observed in nature.[90]

Role of expression level and nonprotein selection on evolutionary rate

Not all attributes of protein sequence evolution can be explained by protein structure. Gene expression has been described as an important constraint on the evolutionary rate of proteins.[91] Several hypotheses explain this observation. Drummond and Wilke have explained this as selective pressure for translational robustness because levels of mistranslated proteins increase as gene expression increases.[92] Another explanation is that of selective pressures to prevent spurious associations.[93–96] As the concentration of a protein increases, the number of targets the protein can associate with as it diffuses through a cell also increases. This then places increasing selective pressure on binding interfaces to constrain sequences to those that will interact with favorable targets with high affinity, while eliminating the subset that can also interact with alternative targets that are deleterious. This might account for observations of increased constraint with increased protein concentration and gene expression and is a hypothesis that should be tested.

Constraints at the level of mRNA

Synonymous substitutions do not change the protein amino-acid sequence, and their rates have traditionally been regarded to be approximately constant along the sequence and to approximate the neutral substitution rate. There is strong evidence which shows that protein coding genes encode DNA and RNA level specific functions other than the amino acid sequence to be translated. Some such functions may be related to the encoded protein, whereas others may be independent of it. Examples of the former include codon bias[97–100] and mRNA secondary structures which may affect both the rate of the translation process and its accuracy.[101,102] Examples of the latter include overlapping genes,[103–105] nucleosome binding regions,[106,107] and *cis* regulatory elements such as exonic splicing enhancers,[108–111] and functional RNAs such as antisense RNAs[112,113] and micro-RNAs (annotated in miRBase[114]). Clearly, such functions would be perturbed by synonymous mutations and are hence expected to be under selection.

These situations require more realistic modeling of the evolutionary process in protein coding genes. Specifically, variable degrees of selective pressures at the DNA and RNA layers should be accounted for, and such models are being developed. For example, Pond and Muse[115] and Mayrose *et al.*[116] have modeled among-site-rate variation of both the synonymous and the non-synonymous substitution rates. However, these models limit the synonymous selective pressures to follow the reading frame, whereas DNA and RNA functions may be independent of the reading frame. For example, a functional RNA secondary structure may be maintained by the first and third positions of a certain codon that encodes an amino-acid site, that is, under weak purifying selection. In such a case the first and third codon

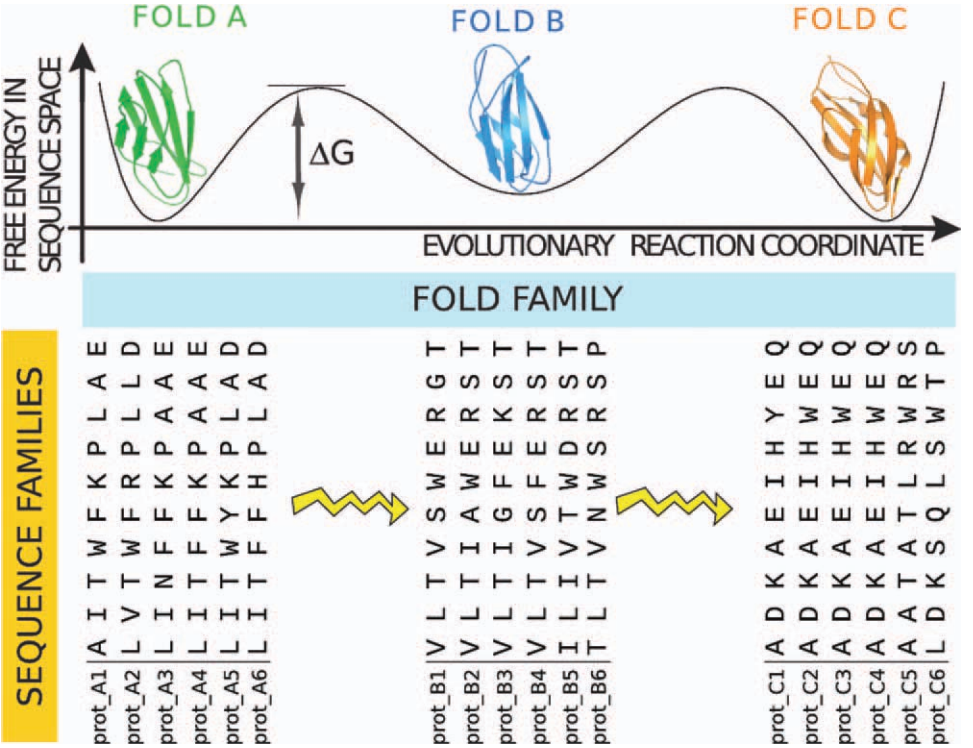

Figure 3. Hypothetical evolution of sequences and folds. On short time scales, mutations and selection due to protein fold (A) cause emergence of a closely related family of protein sequences (A1-A6). On longer time scales, sequences occasionally cross (yellow arrows) the larger free energy barriers that separate related folds (B, C) in sequence space and establish novel sequence families (B1-C6). This figure is modified from a figure published in Ref. 149. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

positions would be conserved while the second one would be variable. Limiting the modeling of the DNA and RNA level selective pressures to the reading frame compromises detection in such a scenario. Rubinstein *et al.*[117] have relaxed the reading-frame dependency by allowing a baseline DNA/RNA substitution rate to vary among individual codon positions and among codons. This relaxed model was shown to better explain substitution patterns of a large fraction of protein coding genes in that study relative to the simpler traditional models that do not account for DNA/RNA level selection, indicating its potential to detect DNA and RNA level encoded functions. In addition, it has revealed that accounting for the DNA/RNA level selective pressures has a dramatic effect on the inference of positive selection. In summary, modeling of substitution patterns in protein coding genes at the codon level is crucial for understanding protein function and structure, and should be articulated in biologically realistic terms.

Fold transitions and divergence over much longer evolutionary timescales

The sequence-structure relationship is perhaps one of the most intriguing problems driven by fundamental principles of molecular evolution. Proteins lacking sequence similarities (at the level of two random sequences that have saturated in substitutions) may still share a similar structure. This structure-function relationship has been addressed from a biophysical point of view, where sequences of many proteins correspond to folds that exist in a cellular environment and context. Hence, it was hypothesized and further demonstrated that amino acid conservation in a given fold family is driven by the contribution of each individual amino acid to the thermodynamic stability of folds.[118,119] In fact, in families of closely related proteins, one can observe conservation of individual amino acids, while in families of more distantly related proteins one can observe conservation at the level of amino acid positions. Dokholyan and Shakhnovich[119] have suggested that difference in time scales drive such mosaic conservation in divergent evolution. On shorter time scales families of homologs appear due to simpler mutagenesis, and on longer time scales sequences diverge enough that one cannot distinguish them from unrelated sequences while nevertheless maintaining fold integrity (Fig. 3). Eventually, the structures diverge enough that one can no longer identify relationships among them. This is the complex coevolutionary process in action, where site-interdependence becomes a stronger signal in evolutionary divergence data.[120] Using graph theoretical approaches, Dokholyan *et al.*[121] have constructed a protein domain universe graph (PDUG) that consisted of nodes, corresponding

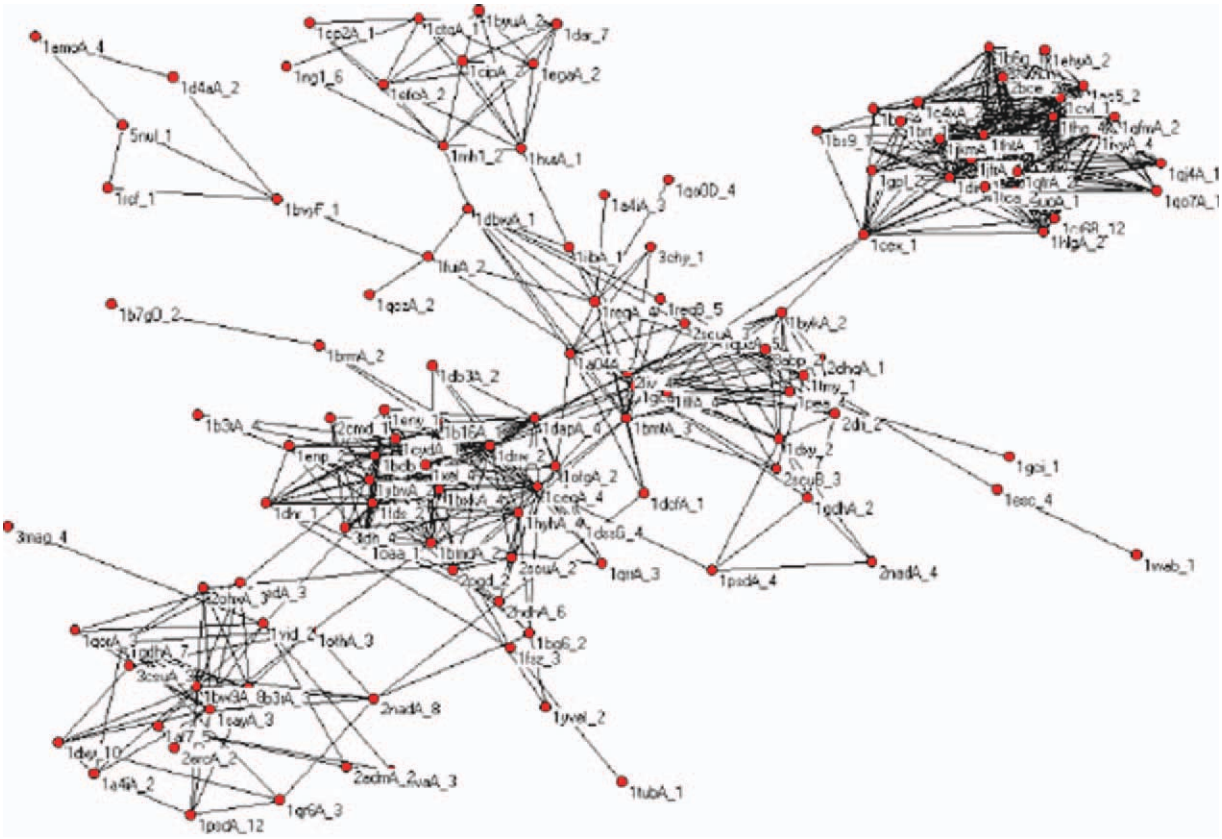

Figure 4. The largest component of the Protein Domain Universe Graph (PDUG) shows the structure of domain relationships and its interconnectedness based upon structural geometries. This figure has previously been published in[149] and is reproduced with copyright permission from Landes Bioscience and Springer Science+Business Media. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

to protein domains, and edges between structurally similar domains. Interestingly, just by looking at the PDUG, one can often connect two seemingly unrelated protein structures through an elaborate network of intermediate protein domains (Fig. 4). At such longer time scales, folds undergo significant change and the resulting pattern of conserved amino acids is lost.

Appearance of new folds suggests that protein thermodynamic stability of a specific fold is not the evolutionary driving force at the times scales of divergence of fold families.[122] Stability maintains the structural integrity of individual folds and fitness may drive fold appearance and divergence in quests for functional adaptation to emerging environments.

This leads to the question, how many sequences correspond to thermodynamically stable folds? This question was addressed with an analytical expression for the number of sequences which fold with given stability into a given structure.[123] Dokholyan[124] estimated the number of sequences that correspond to a stable 100-residue protein structure to be about 10^{47}. This observation suggests several important conclusions. It is clear that evolutionary processes have not resulted in equilibrium in sequence sampling under such thermodynamic constraint and current representation of the sequences corresponding to a given fold is severely biased and under-sampled.[125] However, the search for sequences of stable proteins should be feasible with reasonably good force fields and search algorithms. The estimate for the number of "designable" sequences was also provided in Ref. 124. Upon simulating molecular evolution using thermodynamic stability as a guide,[126] one should not expect full recovery of "native" sequences. However, if one fixes the protein backbone, sequence recovery can reach 60% in the core of the protein, which is also the most conserved because of the core's substantial contribution to stability. Ultimately, while there are a large number of sequences corresponding to a stable fold, the number of all possible sequences is much larger ($\sim 10^{130}$ for a 100 residue protein).

Evolution of protein dynamics

The relationship between protein dynamics and protein evolution has emerged as an important topic of study in this field, complementing analysis of solved structures. From the simple but necessary required flexibility of a ligand-binding site to the coherent conformational transitions of allosteric proteins, proteins must move to function. To gain insight into the dynamics-function relationship it is worthwhile to

study how protein motions evolve. Complementing the large body of evolutionary research dealing with other protein characteristics such as sequence, structure, or stability, comparative studies focused at understanding the evolution of protein dynamics are very recent and still scarce. Yet, there have been some advances on this front, which we will attempt to outline in this section.

There is one system in which the dynamics-function relationship has been informed by evolutionary studies: adaptation to extreme environments. Specifically, the study of a possible role of changes of flexibility in regulating enzymatic activity for organisms adapted to cold or hot environments.[127] Outside this system, only recently backbone flexibility (as conveyed by B-factor profiles) has been used to perform systematic studies, which have shown that flexibility diverges slowly so that it is significantly conserved at family and superfamily levels.[128,129]

Comparative flexibility studies are significant, but lack the necessary detail to deal with comparative studies of large-scale coherent motions. The standard way of analyzing protein motions uses normal modes, which describe independent intrinsic vibrations. Each mode has an associated energy and amplitude, which are related (the square amplitude is the inverse of the energy). There are several ways of obtaining the normal modes, from all-atom Molecular Dynamics (MD) simulations to coarse-grained Elastic Network Models (ENM). A detailed description of these methods is outside the focus of this section. For our purpose, it is enough to highlight that all methods give very similar results, especially for the low-energy large-amplitude motions, which are the most interesting.[130,131]

Specifically, in many cases the low-energy normal modes are presumed to be related to protein function.[132] For example, functional transitions between ligand-free and ligand-bound conformations, allosteric transitions, and so forth, can usually be described using one or a few low-energy normal modes. This functional importance prompted studies of normal-mode conservation. Low-energy normal modes have been found to be conserved in several case-studies.[133–135] A systematic study of a large dataset of proteins representative of all structural classes and folds shows that this is a general trend: the low-energy large-amplitude normal modes are the most evolutionarily conserved.[136]

The issue naturally arises of whether the collective normal modes are more conserved because of their functional relevance or for other reasons. Most case studies mentioned before assume, explicitly, or implicitly, the functional interpretation. Some interesting studies compare the divergence of sequence or structure with that of motions and connect this to functional aspects (see for example Refs. 137, 138). Casting doubt on the functional interpretation, structural similarity seems to grant dynamical similarity, as was found for nonhomologous proteins with the same architecture[139] or even for completely unrelated proteins.[136] An alternative explanation has been proposed that the low-energy normal modes are just more robust with respect to mutations.[136] This view is supported by preliminary studies using perturbed Elastic Network Models.[140,141] The null model should take into account that the low-energy normal modes would be conserved even under no selective constraints and a neutral evolutionary baseline for changes in normal modes needs to be established.

Relationship between protein dynamics and structural divergence

The ensembles of conformations that result from evolutionary divergence are very similar to those produced by thermal fluctuations. This similarity between the evolutionary and dynamical deformations was demonstrated in the pioneering work[142] and confirmed further in other studies.[143,144] An interpretation, put forward already in Ref. 142 and embraced by others, is that this has its origin in and is evidence of the functional relevance of the low-energy normal modes.

To better understand the observed connection between evolutionary deformations and dynamical deformations, a model was proposed in which perturbation of Elastic Network Models accounts for the effect of mutations on equilibrium conformation.[140] This model predicts that the equilibrium conformation will diverge along the low-energy normal modes even under random unselected mutations, which casts doubt on the functional interpretation. If the perturbed ENM is correct, dynamical deformations (normal modes) should govern not only evolutionary divergence, but also the structural change due to perturbation. Further support to the idea of functional signal in ENM perturbation comes from the observation that the same pattern variation along normal modes is found for unselected engineered mutants and for structures of the same protein determined in different experimental conditions.[141]

To say that even under random mutations a protein would diverge along the lowest normal modes is not to say that such modes are nonfunctional or that selection plays no role in molding structural divergence. It is possible that natural selection increases or decreases the contribution of a certain normal mode to structural variation. However, a careful assessment demands the use of a null model that takes into account the dominant effect of the lowest normal modes even in the absence of selection. There is some work that suggests that this could be the case for proteins that experience large functional conformational transitions.[145] Disentangling the effects of natural selection from those of drift on the

patterns of structural divergence is a subject on which further research is needed.

Missing datasets, an experimental wishlist, and experimental testing

A third factor that adds constraints to evolutionary processes, in addition to protein biophysics and population/evolutionary mechanisms, is the functional requirements of the molecule as it interacts with other molecules dynamically in networks and pathways (systems biology). To enable evaluation of these effects and to better define both structure and function, a number of datasets will be desirable. From multiple species and across multiple gene families including those that interact with each other, a better understanding of how, when, and where individual proteins are post-translationally modified is important. As protein function is defined quantitatively, physical and enzyme constants, like k_d, k_{cat}, K_m across multiple species will be tremendously important for studying the evolution of protein functions (for example, inter-molecular interactions) and the constraints they impose.

Techniques such as ancestral sequence reconstruction and detailed experimental studies of mutational epistasis can also shed light on the relationship between sequence coevolution, structure, and function. In studying the evolution of steroid hormone receptors and their affinities for various ligands, it was observed that a very small number of historical mutations are sufficient to cause most of the changes in function that have occurred, although further smaller effect mutations also optimized these new functions.[146] In some lineages, permissive and restrictive mutations (those that have little or no primary effect but are epistatically required for the ancestral or derived states to be tolerated) played a key role in the evolutionary process, opening up pathways to new functions and closing off others. X-ray crystallography and molecular dynamics analyses identified the biophysical mechanisms by which new functions evolved and epistatic mutations caused their effects. These mechanisms are not limited to the well-recognized paradigm of effects on global protein stability, but include dramatic conformational changes that alter the network of interactions between ligand and receptor, the introduction of new contacts that cause ligand-specific frustration, and changes in local protein stability that allow the protein to tolerate specific mutations in specific regions of the protein. Such mechanisms are not incorporated into current models of protein evolution.

In recent experimental studies, mutations in an essential gene folA coding for dihydrofolate reductase were introduced directly in *E. coli* under an endogenous promoter and their fitness effect as well as effect on biophysical properties of the protein (T_m, k_d, k_{cat}, K_m) were evaluated.[147] The analysis uncovered unexpected mechanisms whereby mutated proteins escape unfolding and loss of function by forming symmetric homodimers. Further, it becomes clear that the cell homeostasis machinery (chaperonins and proteases) plays a crucial role in determining the fitness effects of destabilizing mutations, by determining the effective concentration of active proteins in the cytoplasm through their effect on protein turnover. These experiments suggest that steady state description of dynamic processes in cytoplasm is much more relevant than just stability determining equilibrium distribution between the folded and unfolded states of a protein, according to Boltzmann's law. Further experiments along these lines will elucidate the relative importance of physical and physiological factors in sculpting fitness landscape of simple organisms

The TEM-1 family of beta-lactamases is another model system that has been used for several reasons, including the ease of reverse genetics and phenotypic assays, lack of participation in a metabolic pathway almost ensuring that mutational effects on phenotype are mediated by changes in the enzyme itself, and the relative ease of purification and characterization of biophysical and biochemical properties of this enzyme. Specifically, work with the 16 protein-coding alleles defined by all combinations of four missense mutations known to jointly increase drug resistance by over four orders of magnitude has shown that mutational interactions among these mutations (what the evolutionary biologist means by epistasis) sharply constrains the opportunities for adaptive evolution in this enzyme because many mutations are only beneficial in some combination.[148] More recently all 16 protein coding variants were purified and their kinetic and native-form folding stabilities characterized (Jennifer L. Knies and DMW, unpublished results). Interestingly, variation in k_{cat}/K_m among alleles accounts for ~80% of the variance for drug resistance, but native-form folding stability is almost entirely uncorrelated. Moreover, all alleles have ΔG in excess of -4 kcal/mol, challenging the notion that evolution is a balance between structure and function. Finally, there is almost no epistasis for either of these mechanistically more proximal traits. While this is a simple system to decompose mutational effects on fitness (using drug resistance as a proxy), we have been unable to do so, reflecting gaps in our understanding. In this case, mutations of profound evolutionary importance affect T_m by less than 5 degrees C, and 3D structure may be perturbed by less than 1-2 Å RMSD. And after accounting for kinetics, 20% of the variance in drug resistance remains a sort of mechanistic dark matter.

Concluding Thoughts

The evolution of biomacromolecules is complex and there is a constant tension between generating

simple models and embracing the complexity of molecular evolution. As models that describe mechanistic processes and fit data well/offer explanatory power are generated, our corresponding understanding of protein evolution and protein biophysics will increase. Bridging the gap between protein biophysics and molecular evolution is critical to the advancement of this understanding. It has been argued that evolution lies at the heart of biology, while reductionism draws biology into the realm of physics. This new synthesis aims to combine both lines of thinking.

References

1. Privalov PL (1979) Stability of proteins: small globular proteins. Adv Protein Chem 33:167–241.
2. DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet 6:678–687.
3. Taverna DM, Goldstein RA (2002) Why are proteins marginally stable? Proteins 46:105–109.
4. Bloom JD, Raval A, Wilke CO (2007) Thermodynamics of Neutral Protein Evolution. Genetics 175:255–266.
5. Zeldovich KB, Chen P, Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc Natl Acad Sci USA 104:16152–16157.
6. Wylie CS, Shakhnovich EI (2011) A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc Natl Acad Sci USA 108: 9916–9921.
7. Bastolla U, Moya A, Viguera E, van Ham RCHJ (2004) Genomic determinants of protein folding thermodynamics in prokaryotic organisms. J Mol Biol 343: 1451–1466.
8. Noivirt-Brik O, Horovitz A, Unger R (2009) Trade-off between positive and negative design of protein stability: from lattice models to real proteins. PLoS Comput Biol 5:e1000592.
9. Bastolla U, Demetrius L (2005) Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds. Protein Eng Des Sel 18:405–415.
10. Roth C, Liberles DA (2006) A systematic search for positive selection in higher plants (Embryophytes). BMC Plant Biol 6:12.
11. Halpern AL, Bruno WJ (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol 15:910–917.
12. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376.
13. Tuller T, Mossel E (2011) Co-evolution is incompatible with the markov assumption in phylogenetics. IEEE/ACM Trans Comput Biol Bioinform 8:1667–1670.
14. Gaucher EA, Gu X, Miyamoto MM, Benner SA (2002) Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci 27: 315–321.
15. Whelan S, Blackburne BP, Spencer M (2011) Phylogenetic substitution models for detecting heterotachy during plastid evolution. Mol Biol Evol 28:449–458.
16. Shakhnovich E, Abkevich V, Ptitsyn O (1996) Conserved residues and the mechanism of protein folding. Nature 379:96–98.
17. Michnick SW, Shakhnovich E (1998) A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. Fold Des 3:239–251.
18. Parisi G, Echave J (2001) Structural constraints and emergence of sequence patterns in protein evolution. Mol Biol Evol 18:750–756.
19. Taverna DM, Goldstein RA (2002) Why are proteins so robust to site mutations? J Mol Biol 315:479–484.
20. Bastolla U, Roman HE, Vendruscolo M (1999) Neutral evolution of model proteins: diffusion in sequence space and overdispersion. J Theor Biol 200:49–64.
21. Bornberg-Bauer E (1997) How are model protein structures distributed in sequence space? Biophys J 73:2393–2403.
22. Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534–552.
23. Bastolla U, Farwer J, Knapp EW, Vendruscolo M (2001) How to guarantee optimal stability for most representative structures in the protein data bank. Proteins 44:79–96.
24. Kleinman CL, Rodrigue N, Lartillot N, Philippe H (2010) Statistical potentials for improved structurally constrained evolutionary models. Mol Biol Evol 27: 1546–1560.
25. Grahnen JA, Nandakumar P, Kubelka J, Liberles DA (2011) Biophysical and structural considerations for protein sequence evolution. BMC Evol Biol 11:361.
26. Rastogi S, Reuter N, Liberles DA (2006) Evaluation of models for the evolution of protein sequences and functions under structural constraint. Biophys Chem 124:134–144.
27. Dahiyat BI, Mayo SL (1997) De novo protein design: fully automated sequence selection. Science 278:82–87.
28. Yin S, Ding F, Dokholyan NV (2007) Modeling backbone flexibility improves protein stability estimation. Structure 15:1567–1576.
29. Goldstein RA, Luthey-Schulten ZA, Wolynes PG (1992) Optimal protein-folding codes from spin-glass theory. Proc Natl Acad Sci USA 89:4918–4922.
30. Le SQ, Gascuel O (2010) Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. Syst Biol 59:277–287.
31. Bastolla U, Porto M, Eduardo Roman H, Vendruscolo M (2003) Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. J Mol Evol 56:243–254.
32. Goldstein RA (2008) The structure of protein evolution and the evolution of protein structure. Curr Opin Struct Biol 18:170–177.
33. Huzurbazar S, Kolesov G, Massey SE, Harris KC, Churbanov A, Liberles DA (2010) Lineage-specific differences in the amino acid substitution process. J Mol Biol 396:1410–1421.
34. Fernández A, Lynch M (2011) Non-adaptive origins of interactome complexity. Nature 474:502–505.
35. Kimura M (1962) On the probability of fixation of mutant genes in a population. Genetics 47:713–719.
36. Nielsen R, Yang Z (2003) Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol Biol Evol 20:1231–1239.
37. Jensen JL, Pedersen A-MK (2000) Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv Appl Probab 32:499–517.
38. Pedersen A-MK, Jensen JL (2001) A dependent-rates model and an MCMC-based methodology for the

maximum-likelihood analysis of sequences with overlapping reading frames. Mol Biol Evol 18:763–776.
39. Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL (2003) Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20:1692–1704.
40. Rodrigue N, Lartillot N, Bryant D, Philippe H (2005) Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347:207–217.
41. Rodrigue N, Philippe H, Lartillot N (2006) Assessing site-interdependent phylogenetic models of sequence evolution. Mol Biol Evol 23:1762–1775.
42. Rodrigue N, Kleinman CL, Philippe H, Lartillot N (2009) Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. Mol Biol Evol 26: 1663–1676.
43. Lakner C, Holder MT, Goldman N, Naylor GJP (2011) What's in a likelihood? Simple models of protein evolution and the contribution of structurally viable reconstructions to the likelihood. Syst Biol 60:161–174.
44. Castoe TA, de Koning APJ, Kim H-M, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD (2009) Evidence for an ancient adaptive episode of convergent molecular evolution. Proc Natl Acad Sci USA 106:8986–8991.
45. Anisimova M, Cannarozzi G, Liberles DA (2010) Finding the balance between the mathematical and biological optima in multiple sequence alignment. Trends Evol Biol 2:e7.
46. Wong KM, Suchard MA, Huelsenbeck JP (2008) Alignment uncertainty and genomic analysis. Science 319:473–476.
47. Blackburne BP, Whelan S (2012) Measuring the distance between multiple sequence alignments. Bioinformatics 28:495–502.
48. Sjölander K, Datta RS, Shen Y, Shoffner GM (2011) Ortholog identification in the presence of domain architecture rearrangement. Brief Bioinform 12: 413–422.
49. Löytynoja A, Goldman N (2009) Uniting alignments and trees. Science 324:1528–1529.
50. Thorne JL, Kishino H, Felsenstein J (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. J Mol Evol 33:114–124.
51. Suchard MA, Redelings BD (2006) BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics 22:2047–2048.
52. Qian B, Goldstein RA (2001) Distribution of indel lengths. Proteins 45:102–104.
53. Chang MSS, Benner SA (2004) Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. J Mol Biol 341:617–631.
54. Weiner J, Bornberg-Bauer E (2006) Evolution of circular permutations in multidomain proteins. Mol Biol Evol 23:734–743.
55. Moore AD, Björklund ÅK, Ekman D, Bornberg-Bauer E, Elofsson A (2008) Arrangements in the modular evolution of proteins. Trends Biochem Sci 33:444–451.
56. Apic G, Gough J, Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 310:311–325.
57. Dokholyan NV (2005) The architecture of the protein domain universe. Gene 347:199–206.
58. Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA (2004) Supra-domains: evolutionary units larger than single protein domains. J Mol Biol 336:809–823.
59. Weiner J, Moore AD, Bornberg-Bauer E (2008) Just how versatile are domains? BMC Evol Biol 8:285.
60. Weiner 3rd J, Beaussart F, Bornberg Bauer E (2006) Domain deletions and substitutions in the modular protein evolution. FEBS J 273:2037–2047.
61. Björklund ÅK, Ekman D, Elofsson A (2006) Expansion of protein domain repeats. PLoS Comput Biol 2:e114.
62. Moore AD, Bornberg-Bauer E (2012) The dynamics and evolutionary potential of domain loss and emergence. Mol Biol Evol 29:787–796.
63. Kersting AR, Bauer EB, Moore AD, Grath S (2012) Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. Genome Biol Evol 4:316–329.
64. Veron AS, Kaufmann K, Bornberg-Bauer E (2007) Evidence of interaction network evolution by whole-genome duplications: a case study in MADS-box proteins. Mol Biol Evol 24:670–678.
65. Gsponer J, Futschik ME, Teichmann SA, Babu MM (2008) Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science 322:1365–1368.
66. Siltberg-Liberles J, Grahnen JA, Liberles DA (2011) The evolution of protein structures and structural ensembles under functional constraint. Genes 2: 748–762.
67. Schad E, Tompa P, Hegyi H (2011) The relationship between proteome size, structural disorder and organism complexity. Genome Biol 12:R120.
68. Tompa P, Fuxreiter M (2008) Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. Trend Biochem Sci 33:2–8.
69. Nido GS, Méndez R, Pascual-García A, Abia D, Bastolla U (2011) Protein disorder in the centrosome correlates with complexity in cell types number. Mol BioSyst 8:353–367.
70. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA (2006) 3D complex: a structural classification of protein complexes. PLoS Comput Biol 2:e155.
71. Levy ED, Erba EB, Robinson CV, Teichmann SA (2008) Assembly reflects evolution of protein complexes. Nature 453:1262–1265.
72. Kühner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T, Bader S, Beltran-Alvarez P, *et al.* (2009) Proteome organization in a genome-reduced bacterium. Science 326:1235–1240.
73. Tarassov K, Messier V, Landry CR, Radinovic S, Molina MMS, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW (2008) An in vivo map of the yeast protein interactome. Science 320:1465–1470.
74. Ispolatov I, Yuryev A, Mazo I, Maslov S (2005) Binding properties and evolution of homodimers in protein–protein interaction networks. Nucl Acids Res 33: 3629–3635.
75. Devenish SRA, Gerrard JA (2009) The role of quaternary structure in (β/α)8-barrel proteins: evolutionary happenstance or a higher level of structure-function relationships? Org Biomol Chem 7:833–839.
76. Marianayagam NJ, Sunde M, Matthews JM (2004) The power of two: protein dimerization in biology. Trends Biochem Sci 29:618–625.
77. André I, Strauss CEM, Kaplan DB, Bradley P, Baker D (2008) Emergence of symmetry in homooligomeric biological assemblies. Proc Natl Acad Sci 105: 16148–16152.
78. Beltrao P, Serrano L (2007) Specificity and evolvability in eukaryotic protein interaction networks. PLoS Comput Biol 3:e25.

79. Levy ED (2010) A simple definition of structural regions in proteins and its use in analyzing interface evolution. J Mol Biol 403:660–670.
80. Monod J, Wyman J, Changeux JP (1965) On the nature of allosteric transitions: a plausible model. J Mol Biol 12:88–118.
81. Lukatsky DB, Shakhnovich BE, Mintseris J, Shakhnovich EI (2007) Structural similarity enhances interaction propensity of proteins. J Mol Biol 365:1596–1606.
82. Ding F, LaRocque JJ, Dokholyan NV (2005) Direct observation of protein folding, aggregation, and a prion-like conformational conversion. J Biol Chem 280: 40235–40240.
83. Chen Y, Dokholyan NV (2005) A single disulfide bond differentiates aggregation pathways of β2-microglobulin. J Mol Biol 354:473–482.
84. Khare SD, Ding F, Gwanmesia KN, Dokholyan NV (2005) Molecular origin of polyglutamine aggregation in neurodegenerative diseases. PLoS Comput Biol 1: e30.
85. Khare SD, Dokholyan NV (2007) Molecular mechanisms of polypeptide aggregation in human diseases. Curr Protein Pept Sci 8:573–579.
86. Robinson-Rechavi M, Alibés A, Godzik A (2006) Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of thermotoga maritima. J Mol Biol 356:547–557.
87. Bennett MJ, Schlunegger MP, Eisenberg D (1995) 3D domain swapping: a mechanism for oligomer assembly. Protein Sci 4:2455–2468.
88. Huang Y, Cao H, Liu Z(in press) Three-dimensional domain swapping in the protein structure space. Proteins.
89. Ding F, Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI (2002) Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. J Mol Biol 324:851–857.
90. Ding F, Prutzman KC, Campbell SL, Dokholyan NV (2006) Topological determinants of protein domain swapping. Structure 14:5–14.
91. Pál C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. Nat Rev Genet 7:337–348.
92. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352.
93. Deeds EJ, Ashenberg O, Gerardin J, Shakhnovich EI (2007) Robust protein–protein interactions in crowded cellular environments. Proc Natl Acad Sci USA 104: 14952–14957.
94. Zhang J, Maslov S, Shakhnovich EI (2008) Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. Mol Syst Biol 4:210.
95. Heo M, Maslov S, Shakhnovich E (2011) Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. Proc Natl Acad Sci USA 108:4258–4263.
96. Liberles DA, Tisdell MDM, Grahnen JA (2011) Binding constraints on the evolution of enzymes and signalling proteins: the important role of negative pleiotropy. Proc R Soc B 278:1930–1935.
97. Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34.
98. Sharp PM, Li W-H (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucl Acids Res 15:1281–1295.
99. Akashi H (1994) Synonymous codon usage in drosophila melanogaster: natural selection and translational accuracy. Genetics 136:927–935.
100. Zhou T, Gu W, Wilke CO (2010) Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol Biol Evol 27:1912–1922.
101. Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko L (2006) Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science 314:1930–1933.
102. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. Science 324:255–258.
103. Miyata T, Yasunaga T (1978) Evolution of overlapping genes. Nature 272:532–535.
104. Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV (2002) Purifying and directional selection in overlapping prokaryotic genes. Trends Genet 18:228–232.
105. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98–108.
106. Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang J-PZ, Widom J (2006) A genomic code for nucleosome positioning. Nature 442: 772–778.
107. Warnecke T, Batada NN, Hurst LD (2008) The impact of the nucleosome code on protein-coding sequence evolution in yeast. PLoS Genet 4:e1000250.
108. Baek D, Green P (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. Proc Natl Acad Sci USA 102:12813–12818.
109. Pagani F, Raponi M, Baralle FE (2005) Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc Natl Acad Sci USA 102: 6368–6372.
110. Xing Y, Lee C (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. Proc Natl Acad Sci USA 102:13526–13531.
111. Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G (2006) Comparative analysis identifies exonic splicing regulatory sequences—the complex definition of enhancers and silencers. Mol Cell 22: 769–781.
112. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, *et al.* (2005) Antisense transcription in the mammalian transcriptome. Science 309:1564–1566.
113. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW (2008) The antisense transcriptomes of human cells. Science 322:1855–1857.
114. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2007) miRBase: tools for microRNA genomics. Nucl Acids Res 36:D154–D158.
115. Pond SK, Muse SV (2005) Site-to-site variation of synonymous substitution rates. Mol Biol Evol 22:2375–2385.
116. Mayrose I, Doron-Faigenboim A, Bacharach E, Pupko T (2007) Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. Bioinformatics 23:i319–i327.
117. Rubinstein ND, Doron-Faigenboim A, Mayrose I, Pupko T (2011) Evolutionary models accounting for layers of selection in protein-coding genes and their impact on the inference of positive selection. Mol Biol Evol 28:3297–3308.

118. Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 291:177–196.
119. Dokholyan NV, Shakhnovich EI (2001) Understanding hierarchical protein evolution from first principles. J Mol Biol 312:289–307.
120. Pollock DD, Thiltgen G, Goldstein RA (in press) Amino acid coevolution induces an evolutionary Stokes shift. Proc Natl Acad Sci USA.
121. Dokholyan NV, Li L, Ding F, Shakhnovich EI (2002) Topological determinants of protein folding. Proc Natl Acad Sci USA 99:8637–8641.
122. Zeldovich KB, Chen P, Shakhnovich BE, Shakhnovich EI (2007) A first-principles model of early evolution: emergence of gene families, species, and preferred protein folds. PLoS Comput Biol 3:e139.
123. Shakhnovich EI (1998) Protein design: a perspective from simple tractable models. Fold Des 3:R45–R58.
124. Dokholyan NV. Protein designability and engineering. In: Gu J, Bourne PE, Ed. (2009) Structural bioinformatics. Hoboken, NJ: Wiley-Blackwell, pp 961–982.
125. Povolotskaya IS, Kondrashov FA (2010) Sequence space and the ongoing expansion of the protein universe. Nature 465:922–926.
126. Ding F, Dokholyan NV (2006) Emergence of protein fold families through rational design. PLoS Comput Biol 2:e85.
127. Papaleo E, Riccardi L, Villa C, Fantucci P, De Gioia L (2006) Flexibility and enzymatic cold-adaptation: a comparative molecular dynamics investigation of the elastase family. BBA Proteins Proteom 1764:1397–1406.
128. Maguid S, Fernández-Alberti S, Parisi G, Echave J (2006) Evolutionary conservation of protein backbone flexibility. J Mol Evol 63:448–457.
129. Pandini A, Mauri G, Bordogna A, Bonati L (2007) Detecting similarities among distant homologous proteins by comparison of domain flexibilities. Protein Eng Des Sel 20:285–299.
130. Ahmed A, Villinger S, Gohlke H (2010) Large-scale comparison of protein essential dynamics from molecular dynamics simulations and coarse-grained normal mode analyses. Proteins 78:3341–3352.
131. Rueda M, Chacón P, Orozco M (2007) Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. Structure 15:565–575.
132. Bahar I, Lezon TR, Yang L-W, Eyal E (2010) Global dynamics of proteins: bridging between structure and function. Annu Rev Biophys 39:23–42.
133. Carnevale V, Raugei S, Micheletti C, Carloni P (2006) Convergent dynamics in the protease enzymatic superfamily. J Am Chem Soc 128:9766–9772.
134. Marcos E, Crehuet R, Bahar I (2010) On the conservation of the slow conformational dynamics within the amino acid kinase family: NAGK the paradigm. PLoS Comput Biol 6:e1000738.
135. Pang A, Arinaminpathy Y, Sansom MSP, Biggin PC (2005) Comparative molecular dynamics—similar folds and similar motions? Proteins 61:809–822.
136. Maguid S, Fernandez-Alberti S, Echave J (2008) Evolutionary conservation of protein vibrational dynamics. Gene 422:7–13.
137. Münz M, Lyngsø R, Hein J, Biggin PC (2010) Dynamics based alignment of proteins: an alternative approach to quantify dynamic similarity. BMC Bioinformatics 11:188.
138. Raimondi F, Orozco M, Fanelli F (2010) Deciphering the deformation modes associated with function retention and specialization in members of the Ras superfamily. Structure 18:402–414.
139. Hollup SM, Fuglebakk E, Taylor WR, Reuter N (2011) Exploring the factors determining the dynamics of different protein folds. Protein Sci 20:197–209.
140. Echave J (2008) Evolutionary divergence of protein structure: the linearly forced elastic network model. Chem Phys Lett 457:413–416.
141. Echave J, Fernández FM (2010) A perturbative view of protein structural variation. Proteins 78:173–180.
142. Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz AR (2005) An analysis of core deformations in protein superfamilies. Biophys J 88:1291–1299.
143. Friedland GD, Lakomek N-A, Griesinger C, Meiler J, Kortemme T (2009) A Correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family. PLoS Comput Biol 5:e1000393.
144. Velázquez-Muriel JA, Rueda M, Cuesta I, Pascual-Montano A, Orozco M, Carazo J-M (2009) Comparison of molecular dynamics and superfamily spaces of protein domain deformation. BMC Struct Biol 9:6.
145. Mendez R, Bastolla U (2010) Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins. Phys Rev Lett 104:228103.
146. Bridgham JT, Ortlund EA, Thornton JW (2009) An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 461:515–519.
147. Bershtein S, Mu W, Shakhnovich EI (2012) Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. Proc Natl Acad Sci USA 109: 4857–4862.
148. Weinreich DM, Delaney NF, DePristo MA, Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312: 111–114.
149. Dokholyan NV, Shakhnovich EI. Scale-free evolution: from proteins to organisms. In: Koonin EV, Wolf YI, Karov GP, Ed. (2006) Power laws, scale-free networks and genome biology. Boston, MA: Springer, pp 86–105.