Improvisation in evolution of genes and genomes: whose structure is it anyway?

Published in final edited form as:
Curr Opin Struct Biol. 2008 June ; 18(3): 375–381. doi:10.1016/j.sbi.2008.02.007.

Improvisation in Evolution of Genes and Genomes: Whose Structure is it Anyway?

Boris E. Shakhnovich[1] and **Eugene I. Shakhnovich**[2]

[1]Department of Molecular and Cellular Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138

[2]Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138

Abstract

Significant progress has been made in recent years in a variety of seemingly unrelated fields such as sequencing, protein structure prediction, and high-throughput transcriptomics and metabolomics. At the same time new microscopic models were developed that made it possible to analyze evolution of genes and genomes from first principles. The results from these efforts enable, for the first time, a comprehensive insight into the evolution of complex systems and organisms on all scales – from sequences to organisms and populations. Every newly sequenced genome uncovers new genes, families, and folds. Where do these new genes come from? How does gene duplication and subsequent divergence of sequence and structure affect the fitness of the organism? What role does regulation play in the evolution of proteins and folds? Emerging synergism between data and modeling provide first robust answers to these questions.

Introduction

Dramatic increase in number of known genome sequences and crystallized proteins lead to many insights into the global structure of the protein universe such as power-law distributions on various scales [1–4]. Several phenomenological models were proposed to explain these observations. These were either graphical models as in the protein domain universe graph [2,5] or diffusion-like models as in birth death innovation models [6]. While insightful in their own right, these models should be treated with some caution as they are often based on strong assumptions. Furthermore, while early models reproduced the overall shape of gene family distributions, they lacked the detail and specificity to predict the behavior of specific gene families, as well as their function and evolution in genomes. Another serious limitation of phenomenological models was in their somewhat abstract character whereby proteins were treated as nodes of some evolving graph without regard for their sequence-structure-function relationship.

Lack of appropriate tools and data has, until recently, limited our ability to investigate the role of protein structure in evolution of organisms in populations. Recently, there has been a concerted effort to sequence many, closely related species. Such high-density sequencing allows data-driven investigations into the role of structure in evolution of gene families and

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

organisms. On the other hand, ab-initio sequence-based modeling of protein structural evolution within organismal constraints started to emerge [7–11] providing a conceptual framework for interpretation of seemingly disparate observations. Such models, in a dramatic departure from phenomenological approaches [2–4,12] treat sequence-structure relationship explicitly (albeit often at the expense of significant simplification of the protein model) by evaluating directly impact of sequence mutations on protein structure and stability and in some cases also on folding kinetics [13].

The emergence of new data and progress in our understanding of protein biophysics fuelled a renewed focus into the role of structure in gene family evolution. Through integration of diverse experimental data and modeling, we can now evaluate the relative impact of function, structure and other characteristics in constraining protein sequence evolution. A related area of research has focused on the relationship between regulation and structure. Moreover, results from high-throughput genomic studies have inspired several models to investigate the relationship between the structural proteome and organismal fitness on a population level. Integration of experimental results and theoretical modeling has opened many new areas of research into the role of structure and folding in evolution of genes, organisms and populations. These models are becoming increasingly accurate and relevant to the underlying biology.

Protein evolution through duplication and divergence - data driven approaches

The decade since the publication of the *S. cerevisiae*[14] genome saw sequencing of an additional 17 yeast genomes. Fungi represent a great model system with which to study evolutionary dynamics due to the diversity in evolutionary distances and phenotypes. One interesting aspect of the yeast phylogeny is that it exhibits clear evidence of a whole genome duplication event[15]. Sequencing of *K. polysporus*, a species that represents the most distant lineage from S. cerevisiae since the duplication event allowed Wolfe et. al to use likelihood models to better understand the dynamics of gene loss[16]. They found that while initially, gene duplicates are lost randomly, that pattern becomes more systematic with time. Finally, a recent paper from the same authors has outlined intriguing evidence for speciation through precipitous gene loss [17]. Along with the yeast phylogeny, recent sequencing efforts have produced high-coverage genomes for 12 Drosophila species [18]. A series of related papers discuss new insights into evolution of gene families. For example, closely related Drosophila species seem to maintain lineage specific protein families whose evolutionary origin is still poorly understood. However, many of these families are involved in functions important for Drosophila development and survival. For example, Sackton et. al. showed positive selection acting on drosophila protein families involved in innate immune response[19]. Finally, Hahn et. al. focused on divergence of families that function in sexual reproduction of flies and showed that gene family size for these proteins is highly variable [18]. This, in turn, led the authors to hypothesize that speciation events correlate with expansions and contractions in many of the Drosophila gene families.

Furthermore, expansion of certain gene families in genomes correlates with multicellularity. Due to co-evolution between structure, and function[20,21], many of these gene families share a common fold. For example, Chothia and Vogel found that increased presence of IG-folds and certain zinc fingers whose functions are associated with signaling, and transcription factor activity correlate with increased organismal complexity [22]. Apart from duplication and divergence, diversity in sequence and function can also be achieved through other evolutionary mechanisms such as alternative splicing. In fact, Yanai et. al. recently provided evidence, on a whole genome level, for an inverse relationship between alternative splicing (AS) and gene duplication (GD) suggesting that divergence can be achieved by either of the two strategies[23]. In a particularly intriguing follow-up study that draws a relationship between evolutionary mechanism and protein structure, Talavera et. al. showed that evolution through AS and GD impacts protein folds in different ways. Specifically, AS results in more drastic changes in fewer parts of the protein as opposed to the evenly spaced point mutations that appear as a result of GD as measured by both sequence divergence and 3-dimensional analysis [24]. Finally, analysis of duplication, divergence and loss of paralogs in yeasts and flies have shown the importance of whole genome duplication events, chromosomal proximity, function, and regulation in preferential retention of duplicates [25,26].

With so many variables and biological mechanisms producing and maintaining sequence variation, determining the primary determinants of protein evolution is challenging. For example, structural properties of proteins such as designability (number of sequences that can stably fold into a structure [27,28]) have been tied to mutational plasticity[29]. Designability, in turn can be shown to correlate with gene family size[30] (reviewed in detail elsewhere). However, the role of designability in constraining evolution of duplicates and sequence divergence is under debate [31–33]. On the other hand, sequence neighborhood has been shown to affect evolutionary dynamics not only of proteins[30–34], but also of small collections of proteins e.g. viruses. Using a brilliantly simple experimental setup, Burch and Chao recently reported that mutants of φ6 viruses that had many advantageous mutations available to them consistently evolved towards a higher fitness maximum as compared to other variants that had mostly deleterious mutations available [35].

Even before the availability of genome sequences from closely related species, Lynch and coworkers, in several seminal papers, derived quantitative models of gene duplication and divergence including estimates of birth and death rates[36,37]. In these papers, the authors suggest that gene duplication occurs at rates on the order of single site mutations and that these duplications are sufficient to induce speciation. However, the majority of the duplicate genes become silenced. Davis and Petrov later showed that genes that had duplicates are more conserved. However, it was not entirely clear whether genes that were conserved for other reasons duplicated more often or if conserved genes had lower death rates and higher retention rates[38]. As a way of reconciling the two scenarios, Shakhnovich and Koonin showed that while duplication rates were largely independent of the strength of selection on the duplicates, retention of paralogs was indeed much higher for more conserved genes[34]. In the same paper, the authors show that longer retention of duplicates allows gene families under strong selection to explore more sequence space and diverge farther. This results in an interesting dynamic where gene families that evolve slower have farther diverged duplicates.

Co-evolution of Proteins and their Regulation

One of the interesting findings presented by Shakhnovich and Koonin [34] was that paralogs that were under evolutionary pressure were able to diverge farther not only in protein sequence but also in their regulatory regions as well. The same patterns were observed in drosophila as well with a positive correlation between sequence divergence and expression[39,40]. The question is what drives retention of duplicates: sequence, function or regulation divergence. Using a clever methodology of comparing post whole-genome duplication duplicates to their pre-duplication counterparts, Tirosh et. al. found 43 cases of asymmetric divergence in expression. [41] They interpreted this to mean that for these paralogs, neofunctionalization occurred via divergence in regulation rather than sequence. There have now been several experiments that compare expression profiles between orthologs in different species. One common thread that has emerged is the high evolutionary divergence of proteins responsible for stress response and promoters that are regulated by

TATA boxes[42]. The same pattern is observed for variants from replicate mutation accumulation lines[43]. This is consistent with the observation that TATA boxes increase the overall level of expression which has been shown to be an important determinant of protein sequence and structure evolution[44]. While these data together suggest a link between protein structure and regulatory divergence, there has been little progress in quantifying the extent of their inter-dependence and co-evolution.

Duplication and divergence of a gene clearly affects the concentration or dosage of that protein in the cell. Papp et. al. hypothesize that changes in dosage negatively affect protein complexes and confirm this by noting that few genes from large families participate in complexes[45]. This line of reasoning does not necessitate causality, but may be transitive through a shared characteristic of dosage sensitive proteins. A similar scenario is hypothesized to be the case for the correlation between the number of protein interactions and evolutionary rate[44,46]. Recently, Lukatsky et. al. developed a theory explaining increased propensity of similar structures to form complexes[47,48]. Because of this, the authors note that most protein complexes likely evolved through duplication and divergence. On the other hand, to avoid problems associated with nonspecific binding, proteins that form complexes have to differ significantly in their sequence [49]. This, in turn, suggests that to avoid aggregation, duplicated genes have to diverge in the timing of transcription very quickly. This view is supported by the observation that duplicated protein complexes diverge not only in their specificity, but also in regulation [50]. Thus, regulation plays a role not only in neofunctionalization of duplicates but also in maintenance of protein complexes.

Multiscale models of evolution – from proteins to organisms

Perhaps, the most ubiquitous constraint on gene evolution is the requirement of stability and reliable folding of encoded proteins. While certainly a minimalistic one, this constraint is universal (with exception of natively unfolded proteins which are present mostly in eukaryotes [51]). The importance of folding and stability constraints for sequence evolution of gene families was first demonstrated when in silico modeling of sequence-selection for stability was able to reproduce amino acid conservation patterns in many populated fold families [52–54]. A widely held view is that proteins must have some optimal stability for optimal functioning [55,56]. A usual argument in support of this view is that observed stability of real proteins is not very high. This argument stems from the notion that molecular properties can evolve to achieve highest fitness without confronting opposing factors. The reasoning goes on to claim that if there were selective advantage in more stable proteins then evolution would have resulted in emergence of super-stable proteins which is apparently not what is observed in reality, ergo higher stability confers selective disadvantage. However there is no experimental support to this view. Indeed Arnold and coworkers showed that extra stabilization of a protein, cytochrome P450 does not diminish its activity but makes stabilized proteins more conducive to evolution of new function [57]. In an earlier study, Akanuma and coauthors showed that increase of stability of an enzyme 3-isoprpopylmalate dehydrogenase from B.subtillus actually leads to an increase in its catalytic activity [58]. Given these empirical observations, the question remains which factors limit stability of natural proteins? Goldstein[59] suggested that the main factor opposing excessive stabilization of proteins is sequence entropy: there are much more sequences of less stable proteins than more stable ones [60–62]. Recent studies from our lab further highlighted the role of sequence entropy as an important factor determining evolution of protein stability. In a recent paper [11] we assumed that fitness landscape is locally flat with respect to protein stability – (de) stabilization does not confer fitness advantage or disadvantage as long as proteins remain folded. However when essential proteins lose stability as a result of accumulated mutations they cannot function - conferring lethal phenotype to the carrier genome.

There is also a strict limit on how far stabilization can go as there are fewer and fewer sequences corresponding to proteins of higher stabilities and finally there is a stability cutoff below which no sequences can be found [62]. Fitness landscape in this model represents a multidimensional hypercube (the number of dimensions being the number of essential genes) - see Fig.1. Evolution of protein stability in this model corresponds to diffusion in such hypercube with adsorbing boundaries at lower stabilities (corresponding to the notion of lethal phenotype as proteins lose stability). The parameters of the diffusion process can be derived from average impact of point mutations on protein stability known from numerous independent protein engineering experiments and collected in ProTherm database [63]. The analysis of this model predicts the distribution of stabilities of all proteins which is peaked at moderate value of 5kcal per mole and exponentially decaying at higher stabilities – in complete quantitative agreement with empirical data.

Further, the model predicts that populations go extinct at mutation rate which exceeds roughly six mutations per genome per generation. This imposes a “speed limit” on evolution which can be observed in species without DNA repair mechanisms like RNA viruses [64] or under influence of mutation inducing drugs [65]. In a related study Zeldovich et al simulated evolution of multigene organisms using a simple microscopic model of protein folding under constraints which related the death rate of evolving organisms to stability of their proteins [10]. The authors find that successful evolution runs which resulted in population growth were observed in a “Big-Bang” like scenario when structural diversity abruptly collapsed and few stable sequence-structure combinations were found. This simple model of protein evolution based on biophysical principles is very successful at explaining evolution at high mutation rates e.g. pre-biotic or viral, and predicts quick diffusion and divergence of the protein repertoire.

In line with this view, while viral genomes are a simple system to study evolution and selection, they are also extremely varied[66] [67]. The variability in viral genomes represents a significant technical challenge for evolutionary comparative genomics. For example, the first phylogenetic phage tree revealed that there isn't a single protein – identified by sequence homology - that is shared among all phages[68]. Moreover, a recent effort to identify and catalog the protein families in all phages revealed interesting properties of the phage protein universe. For example, viruses have twice the percentage of ORFans, and interconnected groups of proteins that divide along phage nucleotide type[69] which supports a theory of common ancestry for dsDNA phages[70].

One of the key observation from microscopic studies of protein evolution is that structures evolve much slower than sequences [10,54]. This observation suggests a possible approach to challenging problem of genomics of organisms with high mutation rates such as RNA viruses based on the structural repertoire of the proteome. Deeds et. al. [71] and Borne et. al. [72] developed methods for prokaryotic phylogeny reconstruction based on domain content in proteomes. In a related paper, Goldstein and coauthors analyzed accuracy and reliability of phylogenetic reconstruction methods by running a prototypical evolutionary simulation using off-lattice threading model of proteins under the assumption that fitness depends on the stability of a protein in a specified target structure [73]. This approach helps to analyze and troubleshoot standard methods of phylogenetic reconstruction within a set of fully controlled assumptions. The combination of phylogeny reconstruction based on domain content and modeling accuracy can be used to reconstruct the tree for fast-evolving organisms like viruses.

Conclusions

Recent efforts at high-density sequencing have provided a structured and contextualized view of the protein structure universe. This data can be used to validate comprehensive models relating genetics and evolutionary mechanisms to protein structure and function. The goal is to understand the role of structure in evolution of organisms and populations. Despite progress in this arena many important questions remain unanswered. For example, the relative role of structure and function as evolutionary constraints is not well understood, because the models that take function into account remain rudimentary. Furthermore, we have only begun to scratch the surface of how constraints on structure and function mediate the duplication and divergence of genes. Moreover, while some intriguing hypotheses exist, we do not fully understand how structure and fold co-evolve with regulation. Finally, while we recently began to model organismal fitness based on biophysical considerations of the proteome, these models are still in their infancy and are relatively simple. However, with increasing understanding of the biophysics of protein structure, and additional data from the sequencing projects, we can hope to shed light on the connections between molecular and population evolution.

References

1. Huynen MA, van Nimwegen E. The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol. 1998; 15:583–589. [PubMed: 9580988]
2. Dokholyan NV, Shakhnovich B, Shakhnovich EI. Expanding protein universe and its origin from the biological Big Bang. Proc Natl Acad Sci U S A. 2002; 99:14132–14136. [PubMed: 12384571]
3. Qian J, Luscombe NM, Gerstein M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J Mol Biol. 2001; 313:673–681. [PubMed: 11697896]
4. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature. 2002; 420:218–223. [PubMed: 12432406]
5. Roland CB, Shakhnovich EI. Divergent evolution of a structural proteome: phenomenological models. Biophys J. 2007; 92:701–716. [PubMed: 17071665]
6. Karev GP, Berezovskaya FS, Koonin EV. Modeling genome evolution with a diffusion approximation of a birth-and-death process. Bioinformatics. 2005; 21(Suppl 3):iii12–19. [PubMed: 16306387]
7. Taverna DM, Goldstein RM. The evolution of duplicated genes considering protein stability constraints. Pac Symp Biocomput. 2000:69–80. [PubMed: 10902157]
8. Bloom JD, Raval A, Wilke CO. Thermodynamics of neutral protein evolution. Genetics. 2007; 175:255–266. [PubMed: 17110496]
9. Taverna DM, Goldstein RA. The distribution of structures in evolving protein populations. Biopolymers. 2000; 53:1–8. [PubMed: 10644946]
10. Zeldovich KB, Chen P, Shakhnovich BE, Shakhnovich EI. A First-Principles Model of Early Evolution: Emergence of Gene Families, Species, and Preferred Protein Folds. PLoS Comput Biol. 2007; 3:e139. [PubMed: 17630830]
11. Zeldovich KB, Chen P, Shakhnovich EI. Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc Natl Acad Sci U S A. 2007; 104:16152–16157. [PubMed: 17913881]
12. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV. Birth and death of protein domains: A simple model of evolution explains power law behavior. BMC Evol Biol. 2002; 2:18. [PubMed: 12379152]
13. Tiana G, Shakhnovich BE, Dokholyan NV, Shakhnovich EI. Imprint of evolution on protein structures. Proc Natl Acad Sci U S A. 2004; 101:2846–2851. [PubMed: 14970345]
14. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. Life with 6000 genes. Science. 1996; 274:546, 563–547. [PubMed: 8849441]

15. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. Nature. 1997; 387:708–713. [PubMed: 9192896]

16. Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci U S A. 2007; 104:8397–8402. [PubMed: 17494770]

17. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature. 2006; 440:341–345. [PubMed: 16541074]

18. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007; 450:203–218. [PubMed: 17994087]

19. Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. Dynamic evolution of the innate immune system in Drosophila. Nat Genet. 2007; 39:1461–1468. [PubMed: 17987029]

20. Lerman G, Shakhnovich BE. Defining functional distance using manifold embeddings of gene ontology annotations. Proc Natl Acad Sci U S A. 2007; 104:11334–11339. [PubMed: 17595300]

21. Shakhnovich BE, Dokholyan NV, DeLisi C, Shakhnovich EI. Functional fingerprints of folds: evidence for correlated structure-function evolution. J Mol Biol. 2003; 326:1–9. [PubMed: 12547186]

22. Vogel C, Chothia C. Protein family expansions and biological complexity. PLoS Comput Biol. 2006; 2:e48. [PubMed: 16733546]

23. Kopelman NM, Lancet D, Yanai I. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nat Genet. 2005; 37:588–589. [PubMed: 15895079]

24. Talavera D, Vogel C, Orozco M, Teichmann SA, de la Cruz X. The (in)dependence of alternative splicing and gene duplication. PLoS Comput Biol. 2007; 3:e33. [PubMed: 17335345]

25. Wapinski I, Pfeffer A, Friedman N, Regev A. Natural history and evolutionary principles of gene duplication in fungi. Nature. 2007; 449:54–61. [PubMed: 17805289]

26. Heger A, Ponting CP. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. Genome Res. 2007; 17:1837–1849. [PubMed: 17989258]

27. Finkelstein AV, Gutin AM, Badretdinov A. Boltzmann-like statistics of protein architectures. Origins and consequences. Subcell Biochem. 1995; 24:1–26. [PubMed: 7900172]

28. Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. Science. 1996; 273:666–669. [PubMed: 8662562]

29. England JL, Shakhnovich EI. Structural determinant of protein designability. Phys Rev Lett. 2003; 90:218101. [PubMed: 12786593]

30. Shakhnovich BE, Deeds E, Delisi C, Shakhnovich E. Protein structure and evolutionary history determine sequence space topology. Genome Res. 2005; 15:385–392. [PubMed: 15741509]

31. Shakhnovich BE. Relative contributions of structural designability and functional diversity in molecular evolution of duplicates. Bioinformatics. 2006; 22:e440–445. [PubMed: 16873505]

32. Lin YS, Hsu WL, Hwang JK, Li WH. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. Mol Biol Evol. 2007; 24:1005–1011. [PubMed: 17264066]

33. Bloom JD, Drummond DA, Arnold FH, Wilke CO. Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol. 2006; 23:1751–1761. [PubMed: 16782762]

34. Shakhnovich BE, Koonin EV. Origins and impact of constraints in evolution of gene families. Genome Res. 2006; 16:1529–1536. [PubMed: 17053091]

35. Burch CL, Chao L. Evolvability of an RNA virus is determined by its mutational neighbourhood. Nature. 2000; 406:625–628. [PubMed: 10949302]

36. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000; 290:1151–1155. [PubMed: 11073452]

37. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000; 154:459–473. [PubMed: 10629003]

38. Davis JC, Petrov DA. Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol. 2004; 2:E55. [PubMed: 15024414]

39. Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM. Common pattern of evolution of gene expression level and protein sequence in Drosophila. Mol Biol Evol. 2004; 21:1308–1317. [PubMed: 15034135]

40. Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol. 2005; 22:1345–1354. [PubMed: 15746013]

41. Tirosh I, Barkai N. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. Genome Biol. 2007; 8:R50. [PubMed: 17411427]

42. Tirosh I, Weinberger A, Carmi M, Barkai N. A genetic signature of interspecies variations in gene expression. Nat Genet. 2006; 38:830–834. [PubMed: 16783381]

43. Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. Genetic properties influencing the evolvability of gene expression. Science. 2007; 317:118–121. [PubMed: 17525304]

44. Drummond DA, Raval A, Wilke CO. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol. 2006; 23:327–337. [PubMed: 16237209]

45. Papp B, Pal C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. Nature. 2003; 424:194–197. [PubMed: 12853957]

46. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. Science. 2002; 296:750–752. [PubMed: 11976460]

47. Lukatsky DB, Shakhnovich BE, Mintseris J, Shakhnovich EI. Structural similarity enhances interaction propensity of proteins. J Mol Biol. 2007; 365:1596–1606. [PubMed: 17141268]

48. Lukatsky DB, Zeldovich KB, Shakhnovich EI. Statistically enhanced self-attraction of random patterns. Phys Rev Lett. 2006; 97:178101. [PubMed: 17155509]

49. Wright CF, Teichmann SA, Clarke J, Dobson CM. The importance of sequence diversity in the aggregation and evolution of proteins. Nature. 2005; 438:878–881. [PubMed: 16341018]

50. Pereira-Leal JB, Teichmann SA. Novel specificities emerge by stepwise duplication of functional modules. Genome Res. 2005; 15:552–559. [PubMed: 15805495]

51. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J Proteome Res. 2007; 6:1882–1898. [PubMed: 17391014]

52. Shakhnovich E, Abkevich V, Ptitsyn O. Conserved residues and the mechanism of protein folding. Nature. 1996; 379:96–98. [PubMed: 8538750]

53. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol. 1999; 291:177–196. [PubMed: 10438614]

54. Dokholyan NV, Shakhnovich EI. Understanding hierarchical protein evolution from first principles. J Mol Biol. 2001; 312:289–307. [PubMed: 11545603]

55. DePristo MA, Weinreich DM, Hartl DL. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet. 2005; 6:678–687. [PubMed: 16074985]

56. Camps M, Herman A, Loh E, Loeb LA. Genetic constraints on protein evolution. Crit Rev Biochem Mol Biol. 2007; 42:313–326. [PubMed: 17917869]

57. Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. Proc Natl Acad Sci U S A. 2006; 103:5869–5874. [PubMed: 16581913]

58. Akanuma S, Yamagishi A, Tanaka N, Oshima T. Serial increase in the thermal stability of 3-isopropylmalate dehydrogenase from Bacillus subtilis by experimental evolution. Protein Sci. 1998; 7:698–705. [PubMed: 9541402]

59. Taverna DM, Goldstein RA. Why are proteins marginally stable? Proteins. 2002; 46:105–109. [PubMed: 11746707]

60. Shakhnovich EI, Gutin AM. A new approach to the design of stable proteins. Protein Eng. 1993; 6:793–800. [PubMed: 8309926]

61. Shakhnovich EI. Protein design: a perspective from simple tractable models. Fold Des. 1998; 3:R45–58.

62. Shakhnovich EI. Protein Folding Thermodynamics and Dynamics: Where Physics, Chemistry, and Biology Meet. Chem. Rev. 2006; 106:1559–1588. [PubMed: 16683745]

63. Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 2006; 34:D204–206. [PubMed: 16381846]

64. Drake JW, Holland JJ. Mutation rates among RNA viruses. Proc Natl Acad Sci U S A. 1999; 96:13910–13913. [PubMed: 10570172]

65. Crotty S, Cameron CE, Andino R. RNA virus error catastrophe: direct molecular test by using ribavirin. Proc Natl Acad Sci U S A. 2001; 98:6895–6900. [PubMed: 11371613]

66. Ackermann HW, Kropinski AM. Curated list of prokaryote viruses with fully sequenced genomes. Res Microbiol. 2007; 158:555–566. [PubMed: 17889511]

67. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Tatusova T. FLAN: a web server for influenza virus genome annotation. Nucleic Acids Res. 2007; 35:W280–284. [PubMed: 17545199]

68. Rohwer F, Edwards R. The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol. 2002; 184:4529–4535. [PubMed: 12142423]

69. Lima-Mendez G, Toussaint A, Leplae R. Analysis of the phage sequence space: the benefit of structured information. Virology. 2007; 365:241–249. [PubMed: 17482656]

70. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci U S A. 1999; 96:2192–2197. [PubMed: 10051617]

71. Deeds EJ, Hennessey H, Shakhnovich EI. Prokaryotic phylogenies inferred from protein structural domains. Genome Res. 2005; 15:393–402. [PubMed: 15741510]

72. Yang S, Doolittle RF, Bourne PE. Phylogeny determined by protein domain content. Proc Natl Acad Sci U S A. 2005; 102:373–378. [PubMed: 15630082]

73. Williams PD, Pollock DD, Blackburne BP, Goldstein RA. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol. 2006; 2:e69. [PubMed: 16789817]

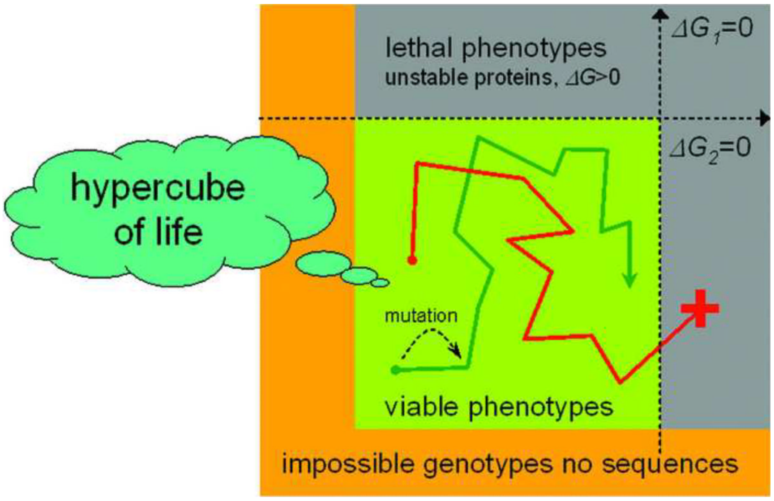

Figure 1.

Schematic representation of fitness landscape for evolution of protein stability as a constrained diffusion in a hypercube. Two gene organisms are shown as an example. Two evolutionary trajectories are shown corresponding to survival of a progenitor and death through mutational destabilization of an essential protein.