Chapter 6
Evolutionary Genomics: Linking Macromolecular Structure, Genomes and Biological Networks

Gustavo Caetano-Anollés

6.1 Introduction

What makes individuals, populations, species and organismal lineages unique? Are genetic complements enough to define phenotypic repertoires? Only 1.5% differences in nucleic acid sequence separate humans from chimpanzee, two species believed to have diverged from each other over six million years ago (Cheng et al. 2005). Yet humans differ notably from chimpanzees and other primates. Are nucleic acid sequence differences at the gene level important? A recent whole-genome analysis of concatenated gene sequences shows that higher organisms have been given more taxonomic resolution than microbes; organisms assigned to separate phyla in Eukarya would clearly belong to a same phylum in the prokaryotic classification (Ciccarelli et al. 2006). Yet they appear to be phenotypically more plastic expressing greater morphological diversity. We may be tempted to state that differences in phenotypes between species are due to limited sets of coding genes that make critical proteins, or to differential regulation of a larger number of protein coding genes. The discovery of a diverse modern RNA world with regulatory function could support the differential regulatory explanation (Bartel 2004). We could also argue that it is not the gene repertoire what counts but the encoded proteins. Protein sequence is extraordinarily diverse and so is the three-dimensional (3D) structure of proteins and their associated functions (Chothia et al. 2003). However, protein sequences encoded in the genomes of the millions of species that currently inhabit earth cover necessarily only a minute fraction (at most one in 10^{-300}) of the enormous permutational space defined by amino acid sequence. Yet the tools of structural genomics and protein structure determination reveal that this limited

G. Caetano-Anollés
Department of Crop Sciences, 332 NSRC, 1101 West Peabody Drive,
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
e-mail: gca@uiuc.edu

C.S. Nautiyal, P. Dion (eds.) *Molecular Mechanisms of Plant and Microbe Coexistence.* Soil Biology 15, DOI: 10.1007/978-3-540-75575-3

exploration of sequence space has uncovered considerable diversity in structure and biological function (e.g. enzymatic catalysis; Gutteridge and Thornton 2005). We could also argue that it is the unique modular structure of proteins that makes the difference. A substantial portion of proteins is made of multiple domains, units of compact structure that can combine in different ways to provide structural diversity (Vogel et al. 2004). Are differences at this level crucial?

In order to answer these and many other fundamental questions we need to draw from the vast information that has been accumulating since the first secrets of the genome were unveiled by the genomic revolution of this past decade. A first and fundamental recognition of modern biology is the need to survey component parts and their interaction. In fact, we have been very effective in this task (Fig. 6.1). Hundreds of genomes have been completely sequenced yielding tens of billions of base pairs, millions of protein sequences, and thousands of putative non-coding RNA molecules that serve a regulatory function and are likely to play important roles in species diversity. This effort outpaces structural genomics with its over 35,000 Protein Data Bank (PDB) entries of 3D molecular structure. A second important recognition following half a century of research into molecular evolution is that we can only understand the present if we can reconstruct our past effectively. Fundamental developments related to natural history reconstruction include the generation of a comprehensive tree of life, global phylogenetic analyses that help track evolutionary history at genome levels, and better understanding of evolution-

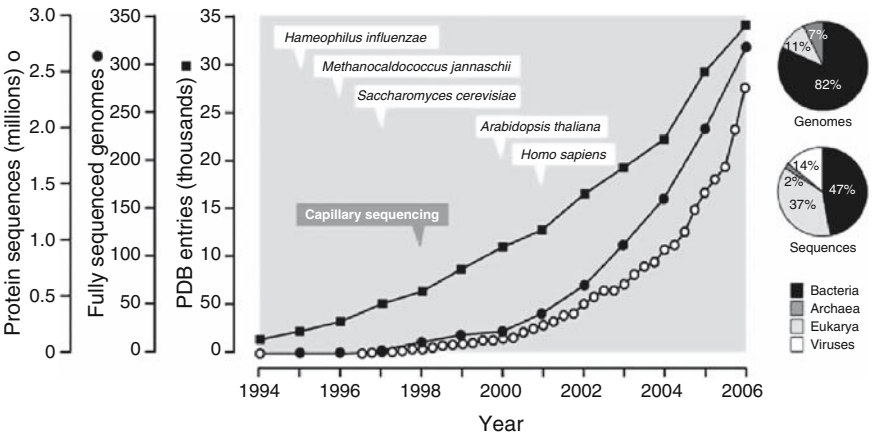

Fig. 6.1 The genomic revolution of this past decade provides hundreds of genomes, millions of protein sequences, and thousands of 3D models of molecular structure embedded in Protein Data Bank (PDB) entries. Fundamental milestones include the sequencing of the first bacterial, archaeal and eukaryotic genome, the genome of the first plant, and the human genome. All this was made possible by the technological development of capillary sequencing. Presently, the living world represented in genomes and sequences is highly biased towards microbial life. Data was retrieved from the PDB (http://www.rcsb.org/pdb), UniProt (http://www.ebi.ac.uk/uniprot), and GOLD (Kyrpides 1999) databases (April 25, 2006)

ary processes (Doolittle 2005; Kurland et al. 2006). A third fundamental recognition comes with the development of systems biology, with the tenet that cells and organisms are integrated systems and not collections of isolated parts. Currently, we use molecular survey components to define descriptive, graphical and mathematical models, confirm these models by perturbation (mutation, environment, etc.), and integrate information and models effectively (Kitano 2002; O'Malley and Dupré 2005). All this is made possible thanks to enhancements in computational power and development of efficient computational algorithms. Molecular survey, history reconstruction, and systems analysis are the fuel of evolutionary genomics and the three pillars of modern biology. Erected at the start of this new millennium, they promise deep understanding of life.

In this chapter I will discuss how evolutionary genomics is helping define new paradigms. Phylogenomic approaches will be described that take advantage of the opportunity to characterize unique sets of genes capable of defining lineages at different taxonomical levels, including species, populations, and organisms. I will also lay the principles of a general evolutionary framework capable of reconstructing evolutionary history directly from the structure of macromolecules.

6.2 Evolutionary Genomics, Networks and Systems

6.2.1 The Genomic Revolution

In the past few years, nucleic acid and protein sequences have been acquired in a massively parallel effort from a wide variety of organisms. Moreover, initiatives that seek to create a complete inventory of the structure of orthologous gene groupings across whole genomes, protein fold architectures from crystallographic data, and the tree of life itself offer unprecedented opportunities to understand genomic complexity (Zhang and Kim 2003; Doolittle 2005). A recent survey (April 25, 2006) showed there were 373 published genomes, with many being deposited in GenBank on a weekly basis (Fig. 6.1). Most genomes that were sequenced were prokaryotic (88%) and had small genome size. However, eukaryotic genomes represented a substantial portion (~38%) of the sequencing effort. There were also 46 finished and ongoing metagenomic projects that study genomic sequences present in a wide range of environments, including soil from Alaska and Minnesota, rice and poplar endophytic communities, root colonizing archaeal communities, and other complex environments. The number of ongoing genome sequencing projects (1605) was also an indicator of exponential increase in years to come. For the first time we have an opportunity to explore the evolution of entire sets of genomes representing a diverse range of organisms and environments, including plants and associated microbes, using the tools of computational biology, comparative genomics and molecular evolution Whole genome comparisons are now possible on a scale from which general principles of evolution can be derived. This has given rise to the new field of phylogenomics (Doolittle 2005).

6.2.2 Phylogenomics

Genomics has opened new avenues in evolutionary research. Evolutionary history has been reconstructed using combined or concatenated genomic sequences, and genomic features describing the survey (genomic demography) and arrangement (genomic topography) of genomic component parts (reviewed in Wolf et al. 2002; Delsuc et al. 2005; Doolittle 2005). In particular, phylogenomic (whole-genome) trees were built effectively from features describing the occurrence and distribution of protein folds in proteomes (Gerstein 1998, Gerstein and Hegyi 1998; Wolf et al. 1999, 2002; Lin and Gerstein 2000; House and Fitz-Gibbon 2002; Caetano-Anollés and Caetano-Anollés 2003; Yang et al. 2005). In one implementation of this strategy, we measured the popularity (number of occurrences) of each protein fold in sequenced genomes and used multi-state phylogenetic characters to reconstruct intrinsically rooted proteome trees invoking the concept that being popular at the molecular level is a favored evolutionary outcome (Caetano-Anollés and Caetano-Anollés 2003, 2005). We have recently taken these approach further and reconstructed phylogenies from features describing the content and arrangement of domains in proteins at a genomic level (Wang and Caetano-Anollés 2006). Phylogenetic characters are here drawn from a molecular topography that describes how evolutionary units of structure arrange in protein molecules and how popular these arrangements are within each proteome. The reconstructed universal tree suggests dramatic diversification events in the history of life (Fig. 6.2). It also shows that genomes in Eukarya were basal, suggesting a eukaryotic rooting of the tree of life. However, phylogenetic trees also revealed early reductive tendencies in the architectural repertoire of Archaea that suggest the very early split of this lineage (Wang et al. 2007). Almost all pan-domain phylogenies generated from genomic information support the tripartite (three-domain) nature of life already evident in trees reconstructed from ribosomal RNA molecules, confirm accepted lineage relationships within major organismal groups, support disputed or preliminary classifications, and reveal novel evolutionary patterns (Doolittle 2005).

6.2.3 Network Biology: Understanding the Wiring Diagram of Life

Network biology characterizes and describes quantitatively the networks of molecular interactions that operate in biological systems (Barabási and Oltvai 2004). These networks can be represented naturally as graphs and hypergraphs and their study is supported by graph and percolation theory. There is considerable interest in the processes underlying evolution of networks. Networks of different kinds appear at different levels of molecular evolution (Schuster and Stadler 2003). We can find networks embedded in biopolymer molecules through conformational spaces that are highly complex and multidimensional and describe molecular and

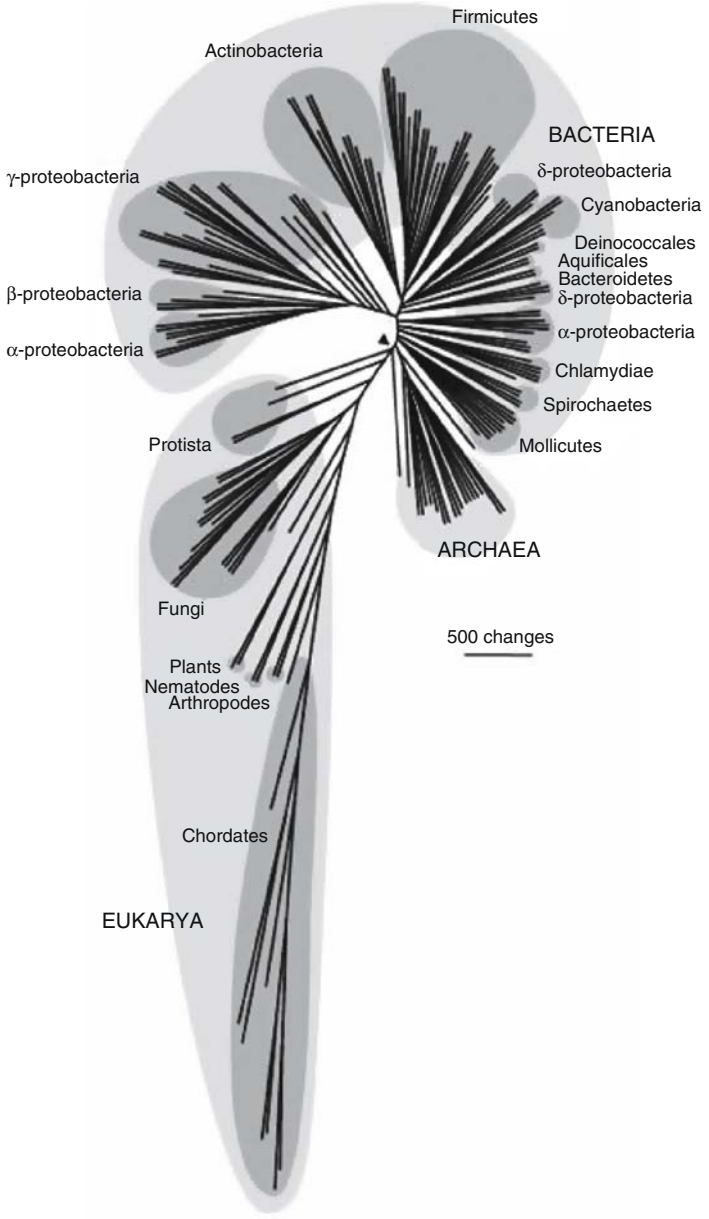

Fig. 6.2 Phylogenomic tree of life. The tree was reconstructed from an analysis of 35,559 domain combinations at fold superfamily level in proteins belonging to 185 organisms that have been completely sequenced. Only one optimal tree of 948,547 steps was obtained using maximum parsimony as the optimality criterion (CI = 0.2714; RI = 0.5375; RC = 0.1459; $g_1 = -1.0334$). Terminal leaves are not labeled as they would not be legible. The arrowhead shows the placement of the root. Note that character change is maximum in Eukarya, that Mollicutes, Spirochaetes and Chlamydiae are basal within the bacterial clade, and that plants and animals represent sister taxa

combinatorial diversity. Examples include RNA and multi-domain proteins. Thermodynamics and forces that stabilize molecular structure drive evolution of these networks through replication and mutation. We can also find inter-molecular networks expressed for example in metabolism, gene expression, protein-protein interaction, and signaling networks. These ubiquitous networks are generally scale-free (i.e. their degree distribution approximates a power law) and evolve by two fundamental processes, growth and preferential attachment (Barabási and Oltvai 2004). Growth arises when new network components (nodes) are added to the system, and preferential attachment results when nodes establish interactive connection (links or edges) preferentially with already well-connected nodes. Growth and preferential attachment are jointly responsible for the emergence of the scale-free ("rich get richer") property of complex networks, and probably, have an origin in duplication and mutational divergence of network components. Gene duplication has been postulated to drive evolution of networks in protein domain combinations (Rzhetsky and Gomez 2001), protein fold occurrence in genomes (Qian et al. 2001), gene expression (Bhan et al. 2002), and protein interactions (Pastor-Satorras et al. 2003). On the other hand, gene duplication may not be the only driver of evolution of networks, or the generator of power law behavior (Wagner 2003).

6.2.4 Molecular Mechanics and Evolution

Molecular machines made of protein and RNA can be considered the major operating components of the living world. The function of these molecules is largely determined by their structure. Consequently, structural conformations can be regarded as molecular phenotypes to which genotypes can be mapped. Because of their unique chemistries, the mapping of genotype (sequence) to phenotype (structure) in proteins and RNA biopolymers offers different challenges but share three properties: (i) the sequence-to-structure map is degenerate; i.e. there are orders of magnitude more sequences than structures; (ii) few common but many rare structures materialize in structure space; and (iii) extensive neutral networks that percolate sequence space define common structures and structural neighborhoods (Fontana 2002; Schuster and Stadler 2003). Because the distribution of sequences that fold into the same structure within neutral networks in RNA is approximately random, the mapping has "space covering" properties. This means that all structures can materialize within relatively few mutational changes in sequence space. This property has been confirmed experimentally using RNA functional switches (Schultes and Bartel 2000). Computational studies also predict the existence of neutral networks and space covering for polypeptides (Babajilde et al. 2001) and experiments support the model (Keefe and Szostak 2001). However, the sequence-to-structure mapping of proteins is much more complex and its landscape "holey", with protein conformations missing in vast segments of sequence space due to the effects of steric hindrance, hydrophobic and H-bonding interactions, and short-range dispersion forces.

6.3 Defining an Evolutionary Genomic Framework

Evolutionary genomics can be powerful when it interfaces with network biology thermodynamics, and molecular mechanics. The function of molecules is curved by evolution, generally resulting from natural selection operating at high levels of structural organization. We have therefore chosen to design a general evolutionary genomic framework that reconstructs evolutionary history directly from the structure of protein and nucleic acid molecules. In initial studies, structure, function, and genomic demography are embedded directly into phylogenetic analyses and molecules and genomes are compared at a wide range of evolutionary levels, from the subspecies analysis of laboratory strains of unicellular green algae to the universal tree of life (Caetano-Anollés 2001, 2002a,b, 2005; Caetano-Anollés and Caetano-Anollés 2003, 2005). This approach can be used to unravel evolutionary processes and uncover functional relationships in macromolecules, and the basis of molecular diversity and genome coexistence. The framework enables global bottom-up or top-bottom approaches of genomic analysis and is supported by three fundamental premises:

1. *Molecular structure is far more conserved than sequence and carries considerable phylogenetic signal*. Structure is directly linked to function and is therefore the subject of natural selection and strong evolutionary constraint (Bajaj and Blundel 1984; Vukmirovik and Tilghman 2000). Consequently, 3D structure is less prone to be affected by mutation than sequence and the information in structure is expected to persist longer than in primary sequence. Similarly, rare genomic processes such as intron indels, retrotransposon integrations and genome rearrangements can preserve deep phylogenetic information (Rokas and Holland 2000). Theoretical considerations suggest that sequence data may be inherently limited in its ability to uncover deep phylogenetic signatures and ancient relationships when the repeated accumulation of substitutions in nucleotide sites (site saturation) erases evolutionary history (Sober and Steel 2002; Penny et al. 2003; Mossell 2003). Convergent evolution of nucleotide sites, differing substitution rates among sites and lineages, and non-independent substitutions among sites, are just few of many other contributing factors (Philippe and Laurent 1998; Delsuc et al. 2005).
2. *Successfully implemented biological designs tend to be reused over and over again in nature*. Structural designs that had been successfully deployed will have more chances to be reused in other biological contexts, and consequently, are expected to become popular (Hartwell et al. 1999). Moreover, robust and well-evolved molecular designs have more chances of withstanding the effects of time. Evidence of this can be found in the redundant and modular nature of protein structure, where certain supersecondary structures and protein domains are highly ubiquitous (Söding and Lupas 2003).
3. *There is a universal tendency towards molecular order*. This very simple hypothesis of polarization depicts generalized trends applied to the structure of molecules, which have been supported by a considerable body of evidence. In

the case of RNA molecules, a tendency towards order was supported by: (i) the study of extant and randomized RNA sequences, showing that evolution enhances conformational order and diminishes frustration over that intrinsically acquired by self-organization (Stegger et al. 1984; Higgs 1993, 1995; Schultes et al. 1999; Seffens and Digby 1999; Gultyaev et al. 2002; Caetano-Anollés 2005); (ii) experimental verification of a molecular tendency towards order and stability using thermodynamic principles generalized to account for non-equilibrium conditions (Gladyshev and Ershov 1982); (iii) a large body of theoretical evidence that maps the structural repertoire of evolving RNA sequences from energetic and kinetic perspectives (Ancel and Fontana 2000; Higgs 2000; Fontana 2002); (iv) phylogenetic congruence in the reconstruction of trees generated from sequence, structure, and genomic rearrangements at different taxonomical levels (Billoud et al. 2000; Collins et al. 2000; Caetano-Anollés 2001, 2002a,b, 2005; Swain and Taylor 2003).

Bottom-up strategies unify phylogenetic analysis with structural biology using a cladistic approach based on shared and derived features descriptive of common descent that use features of molecular structure to generate phylogenetic trees. Cladistic methods offer explicit and general definitions of biological relationships proven to be powerful tools in phylogenetic systematics and molecular evolution. We applied this approach to the study of RNA molecules, generating histories of architectural and organismal diversification directly from their structure.

Top-bottom strategies study global diversification patterns in molecules using information embedded in entire genomic and proteomic complements. Since parsimony analysis has been one of the most widely used methods of phylogenetic inference and has mathematical attributes compatible with the complexity of these genomic datasets, we use this method to chart the protein world. We also explore the unique genomic regions that differentiate genomes from each other and shape, for example, the diversity of closely related species. The phylogenomic framework we have developed can be used to characterize the protein repertoire at the gene family, protein family, superfamily, and protein fold levels (Fig. 6.3), 'structuring' the evolutionary relationships between sequences and architectures, and revealing evolutionary patterns unique to individuals, species and organismal lineages.

6.4 Exploring the Evolution of Modern RNA

6.4.1 Diversity of Non-protein Coding RNA

RNA molecules are ubiquitous and highly sociable and exhibit defined structural, enzymatic and regulatory activities. They have been considered predecessors of DNA and protein in an ancestral RNA world (Gilbert 1986). In recent years, however, we came to realize that our 'modern' RNA world is not a 'relic' but a truly functional entity that is quite diverse (Eddy 2001; Storz 2002; Bartel 2004). Besides

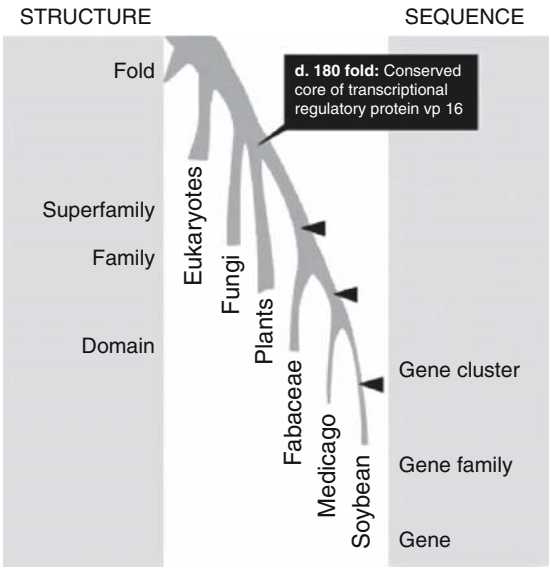

Fig. 6.3 Phylogenomic analysis of protein sequence and structure. Genes can be grouped into gene families and gene clusters using the tools of phylogenetic analysis. At the structural level, these families and clusters can be further defined by domains, and these can be unified into families, fold superfamilies and folds. This hierarchical scheme of molecular organization contains entities (from folds to genes) that can uniquely define organismal lineages. For example, there is currently only one protein fold that is unique to plants (d.180). We expect to find many superfamilies and families unique to individual plant lineages (arrowheads). In fact, the number of unique entities will increase at lower taxonomical levels

the classical three groups of molecules, tRNA, rRNA and mRNA, a repertoire of other RNA have been described. Collectively, these molecules have been termed non-protein coding RNA (ncRNA). ncRNAs are generally small. However, they range in size from ~21–25 nt (for regulatory RNAs) to ~10^3–10^4 nt (for ncRNAs involved in the maintenance of chromatin structure). ncRNAs play important roles in a number of cellular processes, such as those related to transcription, replication, RNA processing and modification, mRNA translation, and protein stability and translocation. Gene expression is modulated by micro RNA (miRNA) and small interfering RNA (siRNA). These molecules, discovered by their role in the control of developmental timing in *Caenorhabditis elegans* (Lee et al. 1993), are tiny and ubiquitous in animals and plants, and are present in all organismal domains (Bartel 2004). ncRNAs play roles in other cellular processes such as the translational tagging of proteins by tmRNA and the targeted mRNA degradation in RNA interference (RNAi) (e.g. Hutvágner and Zamore 2002). Other small ncRNA molecules are important for RNA processing, modification, and stability, such as the catalytic core of the universally conserved RNase P enzyme (~300–500 nt) that cleaves leader sequences from tRNA precursors (Frank and Pace 1998) or the small

nucleolar RNA (snoRNA) (~70–250 nt) that are required for cleavage and processing of rRNA precursors (Eliceiri 1999). ncRNA molecules are also involved in protein translocation across membranes. One example is the signal recognition particle (SRP) that targets nascent secretory and membrane proteins (Keenan et al. 2001). Finally, ncRNA molecules have also been implicated in post-transcriptional gene silencing (siRNA) (Baulcombe 2004). Many other ncRNA molecules have been discovered that play structural roles, mimic the structure of other nucleic acids, or have very specific catalytic activities (Storz 2002).

Holistic views of the universe of RNA structure are missing. This is in part due to difficulties related to the study of RNA (Eddy 2001). For example, genes are identified by the proteins they encode. So genes that encode other molecules remain 'computationally' intractable when using standard tools that scan genome sequences. Novel systematic gene-discovery approaches are therefore needed to uncover effectively the RNA-encoding component of genomes (Washietl et al. 2005). There are no RNA taxonomies and the study of the evolution of RNA structure is still incipient.

6.4.2 Phylogenetic Analysis of RNA Structure

In our laboratory we search for evolutionary patterns embedded in the structure of functional RNA (Caetano-Anollés 2002a,b, 2005). Structures are first characterized using attributes that describe the overall geometry ('shape') of molecules and 'statistical' parameters that describe stability and statistical mechanic features quantitatively. Shape attributes measure for example the nucleotide length of each and every spatial component of secondary structure, such as double helical stems and unpaired sequences, and the number of loops in coaxial stem tracts. Note that unpaired nucleotides can form unusual base-pairings or establish non-covalent interactions (Hermann and Patel 1999). These base pairs and interactions are involved in high-order three-dimensional motifs that are not considered in the structural models of our analysis. Statistical parameters include the Shannon entropy of the base-pairing probability matrix (Q), base-pairing propensity (P), and mean length of helical stems (S) (Fontana et al. 1993; Schultes et al. 1999; Ancel and Fontana 2000). Q, P and S define a complete molecular morphospace, in which Q measures the number of conflicting inter- and intra-molecular interactions (frustration) during RNA folding, and P and S describe how extensively folded and ramified (multifurcated) are molecules (Schultes et al. 1999). In phylogenetic analysis, attributes are considered 'characters', and the numerical values they display 'character states' (Page and Holmes 1998). Characters that are homologous (i.e. share common ancestry) and have been appropriately coded (i.e. provide maximum phylogenetic signal) are compared. Structural characters used in this study transform from one state to another in linearly ordered and reversible pathways 'polarized' by superimposing an evolutionary tendency towards structural order (described above). This tendency should be interpreted as an evolutionary lock-in triggered by the branching of lineages in the trees (cladogenesis), resulting in

molecules that are less plastic but more modular. Finally, hypotheses about character states and models of character evolution were transformed into hypothesis about evolutionary relationship of molecules using maximum parsimony (Page and Holmes 1998). Figure 6.4 describes the overall rationale.

We reconstructed structural phylogenies from several kinds of RNA, including tRNA, rRNA, spacer rRNA, SRP RNA, small mRNA molecules, and retroelements. We also generated a universal tree of life from the structure of rRNA that was rooted in the Eukarya (Caetano-Anollés 2002a,b). However, we decided to focus on tRNA, a molecule that bridges fundamental components of the translation machinery (Sun and Caetano-Anollés 2008). We analyzed the entire set of 571 tRNA molecules deposited as RNA sequences in the Bayreuth database. tRNA structural phylogenies placed tRNA molecules that coded for a group of four amino acids and harbored a variable loop ($tRNA^{Sec}$, $tRNA^{Ser}$, $tRNA^{Leu}$, and $tRNA^{Tyr}$) at the base of the tree of tRNA structure. These four amino acids were probably the first charged or coupled by tRNA in processes related to translation and/or RNA-world based replication that occurred before organismal diversification. Because our phylogenies did not reveal clearly the tripartite nature of life, or clear anticodon or amino acid-linked patterns, we used phylogenetic constraint to falsify alternative hypotheses about the origin of organismal diversification, amino acid specificities, and structural diversification in tRNA molecules. The results of these analyses suggest a sister-clade relationship between Bacteria and Archaea that is consistent with trees of life reconstructed from rRNA structure (Caetano-Anollés 2002a,b), protein fold architecture (Caetano-Anollés and Caetano-Anollés 2003), and domain combinations (Fig. 6.2). Results also show patterns of diversification of tRNA that developed once the cloverleaf structure was fully formed. Apparently, structural diversification preceded the establishment of amino acid and anticodon specificities, and these probably preceded organismal diversification.

We also designed a novel phylogenetic approach that reconstructs the evolution of substructural components of a molecule and generates *"contour maps"* capable of superimposing ancestral-derived relationships directly onto 3D RNA representations (Fig. 6.4). This involves defining new kinds of taxa (substructures) and characters (molecules), and a criterion of primary homology pertaining substructural repertoires based on molecular lineages. Phylogenetic trees of substructures describe here the evolutionary relationships of molecular substructural components that make up RNA molecules. These trees reveal evolutionary patterns of structural diversification, showing how RNA structure changes in the course of evolution. Patterns suggest by definition a structural origin and a relative timeline (a series of steps) describing how individual substructures are incorporated into the evolving RNA molecules. Analysis of tRNA molecules using this novel approach provided strong support to the 'two halves' hypothesis put forth by Maizels and Weiner (1994) that proposes that the anticodon/dihydrouridine domain constitutes a refinement that was incorporated later in evolution (Fig. 6.5). However, our structural trees also support a more detailed structural transformation sequence. In this model, the tRNA molecule evolves by gradual addition of nucleotide pairs to a primordial hairpin stem loop

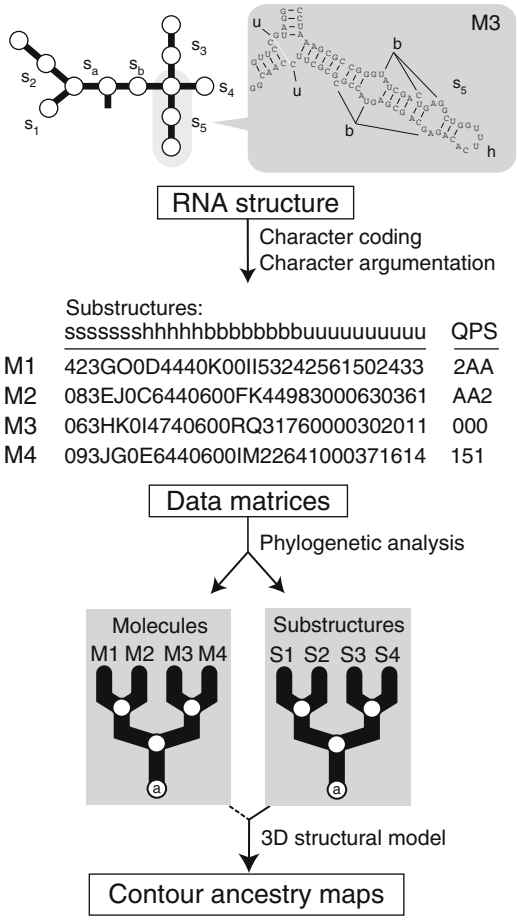

Fig. 6.4 Molecular structures (M1, M2, …) and substructural repertoires (S1, S2, …) of RNA molecules can be organized hierarchically in nested sets (*phylogenetic trees*) using cladistic principles. Trees describe structural diversification and allow identification of ancestors (e.g. nodes labeled *a*). The structure of an RNA molecule, such as signal recognition particle (SRP) RNA from rice (M3), can be decomposed for example into segments (S_1–S_5, S_a and S_b) and substructures (e.g., coaxial stem tracts and unpaired loop regions), and these substructural components studied using molecular features (*characters*) that describe their geometry [e.g., length of stems (s) and unpaired regions (h, b, and u)] or their stability and uniqueness (e.g., using morphospace parameters *Q*, *P* and *S*). These *shape* and *statistical* characters are coded and assigned 'character states' (in alphanumeric format) according to an evolutionary model that polarizes character transformation towards an increase in molecular order (character argumentation). Coded characters are arranged in data matrices and subjected to cladistic analysis, generating phylogenies of molecules and substructures. Rooted trees can be used to color 2D or 3D structural models of RNA (*contour ancestry maps*) that help infer models of structural evolution

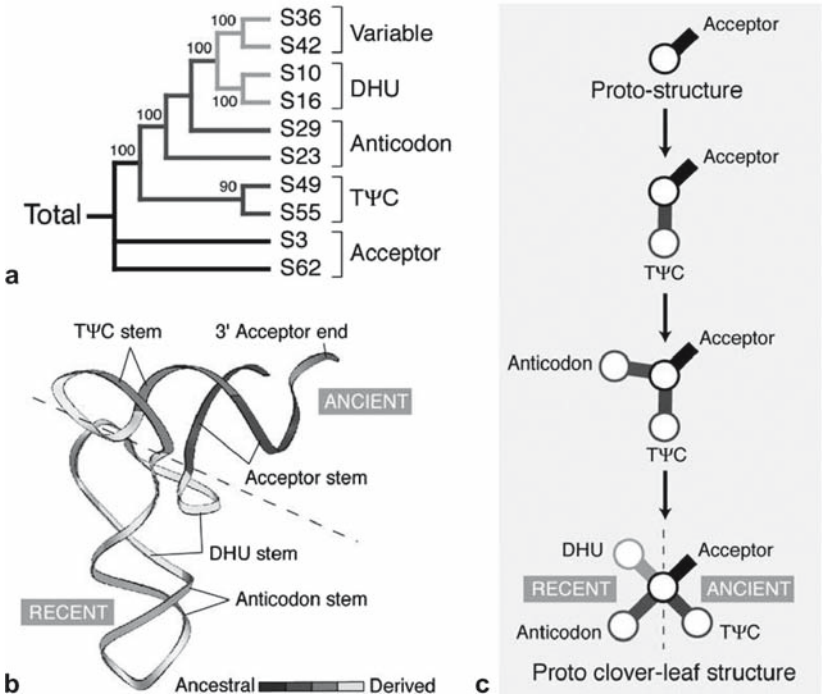

Fig. 6.5 Evolution of tRNA structure. **a** Trees of tRNA substructures show patterns of structural evolution inferred from the total tRNA dataset using maximum parsimony and branch-and-bound searches. Analysis of stabilizing stem characters produced two optimal trees of 4468 steps each (CI = 0.961; RC = 0.937; g_1=−1.25). The tree that is shown represents a strict consensus of these two trees and is labeled with bootstrap support values >50%. **b** Contour ancestry map showing the geometrical evolution of stem components that stabilize tRNA molecules. Trees were painted directly on the structural model using a color scale bar describing relative ancestry values. **c** A model of the early evolution of proto-tRNA molecules. The model is derived directly from trees of substructures and shows formation of substructures homologous to present-day acceptor, TΨC, anticodon and dihydrouridine arms. Substructures may have had different functions than those of extant tRNA molecules. Unpublished data from Sun and Caetano-Anollés (2008)

and then to its growing stems, ultimately resulting in a molecular arrangement that favors multiloop conformations and molecular multifurcation, an expected outcome when seeking to maximize molecular order.

We have extended our phylogenetic approach to the analysis of other interesting questions. For example, we used the structure of tRNA-derived transposable elements (SINEs) to study their evolution in plants (Sun et al. 2007). The exercise established a model of structural evolution of these transposable elements that explains the popularity of sequence families in the plant genome. We also found interesting patterns in the small (SSU) and large (LSU) subunits of rRNA (Harish and Caetano-Anollés, unpublished), including the ancestral placement of stem S49, the dominant SSU rRNA component of the subunit interface and the proposed ribosomal functional relay (Yusupov et al. 2001).

It is particularly noteworthy that ancient substructures were located in the middle of the rRNA ensemble and at the subunit interface. The origins of these ribosomal ancient substructures appear not associated with translation.

6.5 Exploring the Evolution of the Protein World

6.5.1 The Hierarchical Nature of Protein Structure

The protein world is extraordinarily diverse in sequence, structure and function (Ponting and Russell 2002). Most proteins (60%) fold compactly into more than one domain, and these domains can be repeated or combined in defined order. The number of available domains is considerable but appears finite (Chothia et al. 2003) and so does the repertoire of domain combinations in proteins (Vogel et al. 2004). When creating new functions, redundancy appears to be a favored outcome, with domains reused more often than discovered.

Domains are not only units of protein structure and function but also units of evolution (Riley and Labedan 1997). Taxonomies that attempt to provide a comprehensive description of structural and evolutionary relationship of proteins of known structure, such as the Structural Classification of Proteins (SCOP) (Murzin et al. 1995) and the CATH protein structure classification (Orengo et al. 1997) use these building blocks as units of classification. In SCOP, proteins that are evolutionarily closely related at the sequence level are clustered together into protein families. Proteins belonging to different families that exhibit low sequence identities but share structural and functional features suggesting a common evolutionary origin are further unified into fold superfamilies. Finally, fold superfamilies sharing secondary structures that are similarly arranged and topologically connected are unified into protein folds (Murzin et al. 1995). These folds sometimes have peripheral regions of secondary structure that differ in size and conformation and 'decorate' distinctly the central fold architecture.

While our knowledge of sequence space is far from complete (Kunin et al. 2003), it is apparent that protein diversity originated from a limited set of architectural designs (Koonin et al. 2000). Most proteins have been formed by gene duplication, recombination, and divergence and proteome evolution can be tracked by matching proteins of known folding structure to genome sequences (Chothia et al. 2003). While protein folds can be mapped onto about half of amino acid residues encoded in genome sequences, using hidden Markov models (HMMs) of structural recognition, it has become increasingly more difficult to find new folds in nature (Grant et al. 2004). Consequently, the world of protein molecules appears finite and its study feasible at global levels. However, fold categories should be regarded as "neighborhoods" defined by how much structural overlap exists between them (Harrison et al. 2002). In fact, some regions of the protein fold space represent a continuum for some architectural arrangements (sometimes linked by super-secondary motifs) while in other regions clearly distinct non-overlapping topologies are observed.

6.5.2 An Evolutionarily Structured Universe of Protein Architecture

A number of approaches have been used to characterize protein space, including fold family trees (Efimov 1997; Zhang and Kim 2000), a periodic table of structures (Taylor 2002), or taxonomies based on secondary structure (Przytycka et al. 1999). Recently, the metric comparison of structure similarity of proteins representing different protein fold categories provided measurements of distance between the different structures and a global representation of protein space (Hou et al. 2003). Four clear groups representing the α/β, α+β all-α, all-β protein classes were evident in this representation. These studies show that it is possible to generate global views of the protein universe. However, comparative genomic efforts have been largely confined to describing wide-encompassing features as similarities and differences. To be useful, however, strategies require methods capable of organizing the comparative data within an evolutionary perspective.

We recently reconstructed universal phylogenies of protein architecture (Caetano-Anollés and Caetano-Anollés 2003, 2005; Wang et al. 2006). These phylogenies depict the evolution of the protein world – they also bring a unique power to the identification of structurally orthologous gene families defining unique gene complements. The general strategy is depicted in Fig. 6.6. We counted the number of genes corresponding to particular protein architectures in genomes and used these measures of 'genomic demography' to map the world of proteins and track architectural and organismal history directly at the proteome level. Intrinsically rooted phylogenomic trees of proteomes and fold architectures were generated that described phylogenomic relationships, patterns of evolution, and information on the underlying evolutionary processes. Studies involved small and large subsets of protein folds, and complete datasets matching three releases of SCOP (1.39, 1.59 and 1.67). Figure 6.6 shows a tree of fold architectures generated using information embedded in 185 genomes.

6.5.3 Evolutionary Patterns and Transformation Pathways

The universal tree of protein architecture revealed interesting patterns. Folds that were widely distributed in nature were found at the base of the tree and were only missing in parasitic organisms with highly reduced genomes (Fig. 6.6). These organisms (e.g., *Mycoplasma*, *Nanoarchaeum, Encephalitozoon*) have discarded enzymatic and cellular machinery in exchange from resources provided by their hosts. In fact, the first nine folds to emerge in evolution are common to every genome analyzed and include folds widespread in metabolism. It is noteworthy that only 16 folds are universally shared and all of them originated deep in the tree. Similarly, all classes of protein architecture appeared very early in the tree of architectures. Folds in the α/β protein class arose first and were followed by those in the α+β, all-α, all-β, small, and multi-domain classes, in that order. These folds accumulated at different levels. The α/β folds occurred at relatively constant rates and were prevalent in the bottom half of the tree. In contrast, the α+β folds started to

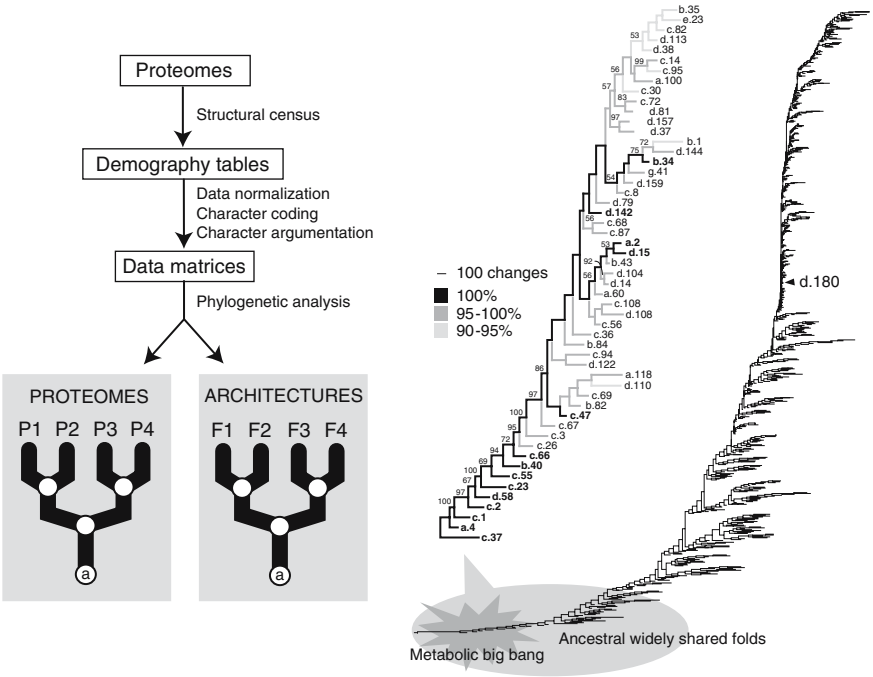

Fig. 6.6 Exploring the diversity of the protein world. The diagram (*left*) shows how fold architectures of extant proteins (F1, F2, …) and entire protein complements (P1, P2, …) can be used to generate hypotheses (phylogenies) about groups of folds and proteomes. This involves a structural census defined by advanced HMMs that assign domain structure to genomic sequences, normalization of data, and phylogenetic analysis. A phylogenomic tree of protein architecture generated from a protein domain census in 185 completely sequenced genomes (*right*) was recovered from an heuristic maximum parsimony search with branch swapping and 100 replicates of random addition sequence. The tree had 85,644 steps (CI = 0.043, RI = 0.770) and is well supported by measures of skewness in tree distribution ($g_1 = -0.138$; $P < 0.01$). Terminal leaves are not labeled except for SCOP fold d.180 (see Fig. 8.3), which is unique to plants. The subtree drawn above describes the evolution of the 53 most ancestral folds and has branches labeled with different shades indicating percentage of genomes sharing folds. The subtree shows folds labeled according to SCOP nomenclature and bootstrap support values >50% above nodes. It is noteworthy that folds shared by >90% genomes are missing almost exclusively in parasitic organisms with reduced genomes. Data from Wang et al. (2006)

accumulate significantly later but with increasing rates until these folds became the most prevalent class. Folds in all other classes followed this same pattern of accumulation but with lower rates. Maximum rates diminished in the order of fold appearance, i.e. all-α, all-β, small and multi-domain proteins. These patterns suggest that the most primitive proteins contained interspersed α-helical and β-sheet elements (as in the α/β class). In the course of evolution, these elements were first segregated within their structure (α+β class) and then confined to separate molecules (all-α and all-β classes). This hypothesis is consistent with the suggestion that

diversity in protein architecture originated by stochastic processes expressed both in protein sequence and structure (the random origin hypothesis; White 1994).

Remarkably, the most ancestral folds harbored interleaved β-sheets and α-helices and barrel structures (Fig. 6.7). Many important structural designs were derived in the tree, including polyhedral folds in the all-α class and β-sandwiches, β-propellers and β-prisms in the all-β class. Protein transformation pathways that describe

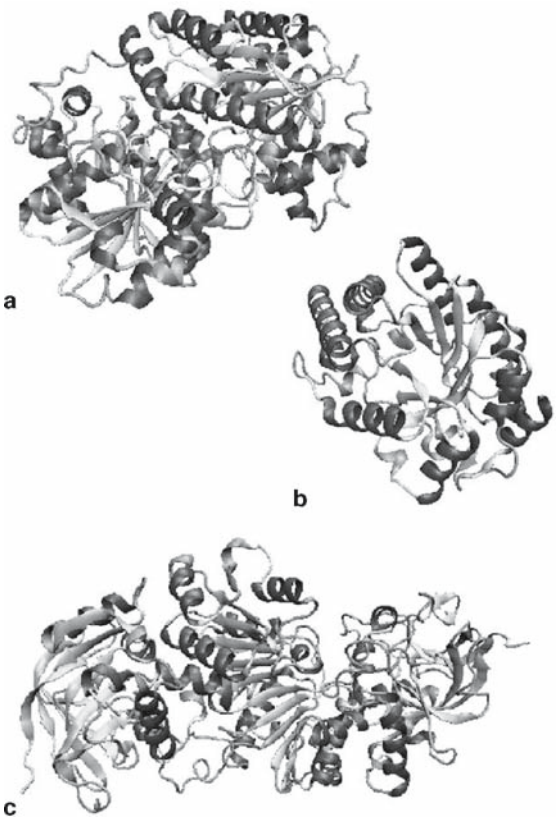

Fig. 6.7 Ancient protein folds share a common architecture of sheets and helices that form either barrels or are interleaved and are highly symmetrical. The structural models of selected structures show the arrangement of α-helices (described by dark helical ribbons) and β-strands (described by arrows that point towards the C-terminus of the protein). Structures were visualized in 3D using the new cartoon format of the VMD (Visual Molecular Dynamics; http://www.uiuc.edu/research/vmd) visualization package. **a** The nitrogenase iron protein from *Azotobacter vinelandii* (PDB entry 1fp6), an enzyme important for nitrogen fixation, harboring the P-loop hydrolase fold (c.37), the most ancient protein architecture with three layers in which a parallel or mixed β-sheet is sandwiched by α-helices. **b** The xylanase from *Penicillium simplicissimum* (1bg4), a protein exhibiting the TIM β/α-barrel fold (c.1), a α/β protein architecture with a parallel β-sheet closed barrel. **c** The gluthatione-dependent formaldehyde dehydrogenase enzyme from humans (1m6h) with the NAD(P)-binding Rossmann fold (c.2) that harbors two layers of α-helices sandwiching a parallel β-sheet of 6 β-strands. These three ancient architectures are very common in modern metabolism

likely scenarios of structural evolution (Murzin 1998; Grishin 2001) could be traced in our tree of architectures. For example, the conversion of an α-helix into a three-stranded β-meander causes Rossmann fold proteins to change to the FAD/NAD(P)-binding domain architecture. Both folds are ancient and are closely related, so this putative transformation must have already occurred very early during evolution. In contrast, circular permutations in protein phosphatases resulted in changes that were quite derived in the history of protein diversification. Figure 6.8 shows how an all-α protein containing a three-helical bundle transforms by indels and substitutions into a β-sheet structure that is part of an all-β barrel-like architecture, probably through an α+β protein intermediary. Transformations from all-α to all-β proteins may be quite common and follow general tendencies of architectural transformation (Caetano-Anollés and Caetano-Anollés 2003). Clear transformation pathways were also evident in structural families of fold architectures. For example, the popular β-barrels increased the tilt of the β-strands, the frequency of open barrel structures, and the complexity of strand topology. These tendencies suggest barrel architectures with increased curl and stagger of β-sheets (sensu Taylor 2002) are favored evolutionary outcomes.

6.5.4 Sharing Patterns of Fold Architecture in Life

We found that tracing features depicting organismal diversity along the branches of the evolutionary tree of protein architecture provided interesting information (Caetano-Anollés and Caetano-Anollés 2005). We were able to infer a relative timing for the emergence of prokaryotes, congruent episodes of architectural loss and diversification in Archaea and Bacteria, and a late and quite massive rise of architectural novelties in Eukarya probably linked to the rise of multicellularity. Folds associated with processes related to multicellularity (e.g. apoptosis, cell death, adhesion and recognition, and extracellular matrix remodelling) contained multiple domains and appeared both immediately after prokaryotic diversification (mostly folds common to all domains of life) and during eukaryotic diversification (mostly eukaryotic-specific).

Our observations indicate that protein novelties unique to organismal lineages appeared late and in defined order during evolution. The proteomes of these diversified organisms originated apparently from ancestors that shared already an arrangement of quite complicated molecular architectures and biological functions. This view is consistent with a proto-eukaryote (Poole et al. 1998; Kurland et al. 2006) responsible for ‘crystallizing’ diversified life (Woese 2000).

6.6 Exploring the Evolution of Networks

Our phylogenomic analysis is quite novel (Doolittle 2005) and offers the opportunity to identify and trace architectures unique to organisms or organismal groups, unique to functions and ontologies, and unique to biological networks. Since proteins

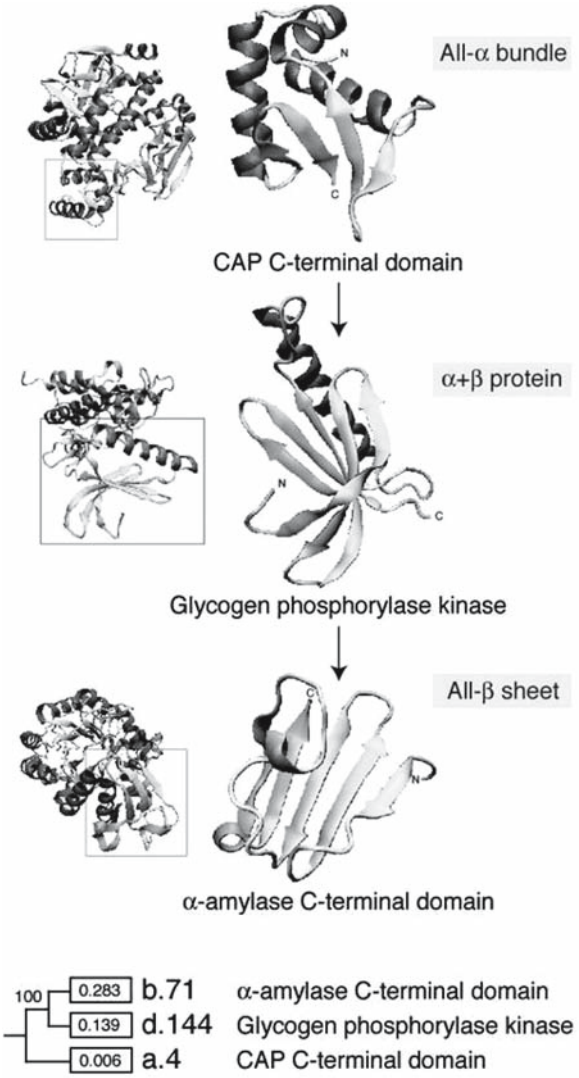

Fig. 6.8 Evolutionary transformation pathway from an all-α to an all-β protein architecture induced by indels and substitutions. The figure shows how the winged helix-turn-helix (HTH) domain characteristic of nucleic-acid-binding domains such as the C-terminal domain of the catabolite gene activator protein (CAP) (1cgp) transforms into the γ-subunit of the glycogen phosphorylase kinase (1phk), and this structure then transforms into the C-terminal domain of the G4-α-amylase (1,4-α-D-glucan maltotetrahydrolase) (2amg). The CAP C-terminal domain has a DNA/RNA-binding three-helical bundle fold (a.4), in which three α-helices form a partly opened right-handed bundle. The glycogen phosphorylase kinase has a protein kinase-like fold (d.144) with two α+β domains, one of which (the C-terminal) is almost α-helical. The α-amylase C-terminal domain has a glycosyl hydrolase domain (b.71) with a β-sheet that follows the catalytic β/α barrel domain. The entire multidomain proteins are shown in the left with the relevant domain enclosed by rectangles. The transformation from a three-helical bundle to a β-sheet seen in the structural models of the domains is confirmed by the phylogenomic tree shown below with terminal nodes indicating ancestry values of individual folds derived from the tree of fold architecture in Fig. 6.6

are generally components of biological networks, protein structure can be used to study network evolution.

Cellular metabolism is the best-studied biological network. It represents one of the greatest achievements of science, resulting from almost two centuries of biochemical research. However, we do not know its origin or how it has evolved. In an initial study, we explored the relationship between protein architecture and function by tracing the total number of enzymatic functions associated with folds in the tree of architectures (Caetano-Anollés and Caetano-Anollés 2003). As expected, the most ancestral folds had the most enzymatic functions associated with them. This supports the proposal that during metabolic evolution enzymatic multifunctionality was replaced by specialized function (Kacser and Beeby 1984). We also explored the origins and evolution of modern metabolism using phylogenomic information embedded in protein structure. We first painted the ancestries of enzymes derived from rooted phylogenomic trees directly onto over one hundred metabolic subnetworks in mesonetworks defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kim et al. 2006). This evolutionary tracing exercise involved linking metabolic enzymes to fold architectures and an analysis of 860,000 genomic sequences with HMMs (Fig. 6.9). To our knowledge, this represents the first global attempt to map evolutionary relationships directly onto biological networks. Careful analysis of evolutionarily painted subnetworks revealed patchy distribution patterns indicative of widespread enzymatic recruitment, consistent with previous evidence (Schmidt et al. 2003). It is noteworthy that the distribution of abundance of folds with various ancestries showed that mesonetworks differ in mean ancestry, with amino acids oldest and lipids and glycans youngest. We also revealed patterns of origin of modern metabolism (Caetano-Anollés et al. 2007). Apparently, a "big bang" of enzymatic diversification occurred at the base of the tree of protein architectures (Fig. 6.6). In fact, most enzymatic reactions at all levels of Enzyme Commission (EC) classification were associated with the nine most ancestral and widespread folds. Furthermore, phylogenetic trees reconstructed from enzymatic sharing of fold architectures and other information indicated that metabolism originated in the purine and pyrimidine subnetworks. Consequently, the first enzymatic take-over of a prebiotic chemistry involved the synthesis of nucleotides for the RNA world.

6.7 Evolutionary Genomics and Organismal Coexistence

An important focus of genomic research has been the identification of differences between genomes (Koonin et al. 2000) and the systematic grouping of hundreds of thousands of protein sequences into protein clusters based on sequence and structural similarities (Grant et al. 2004). Initial studies uncovered a diverse genetic repertoire and a large proportion of genes that were uniquely species-characteristic

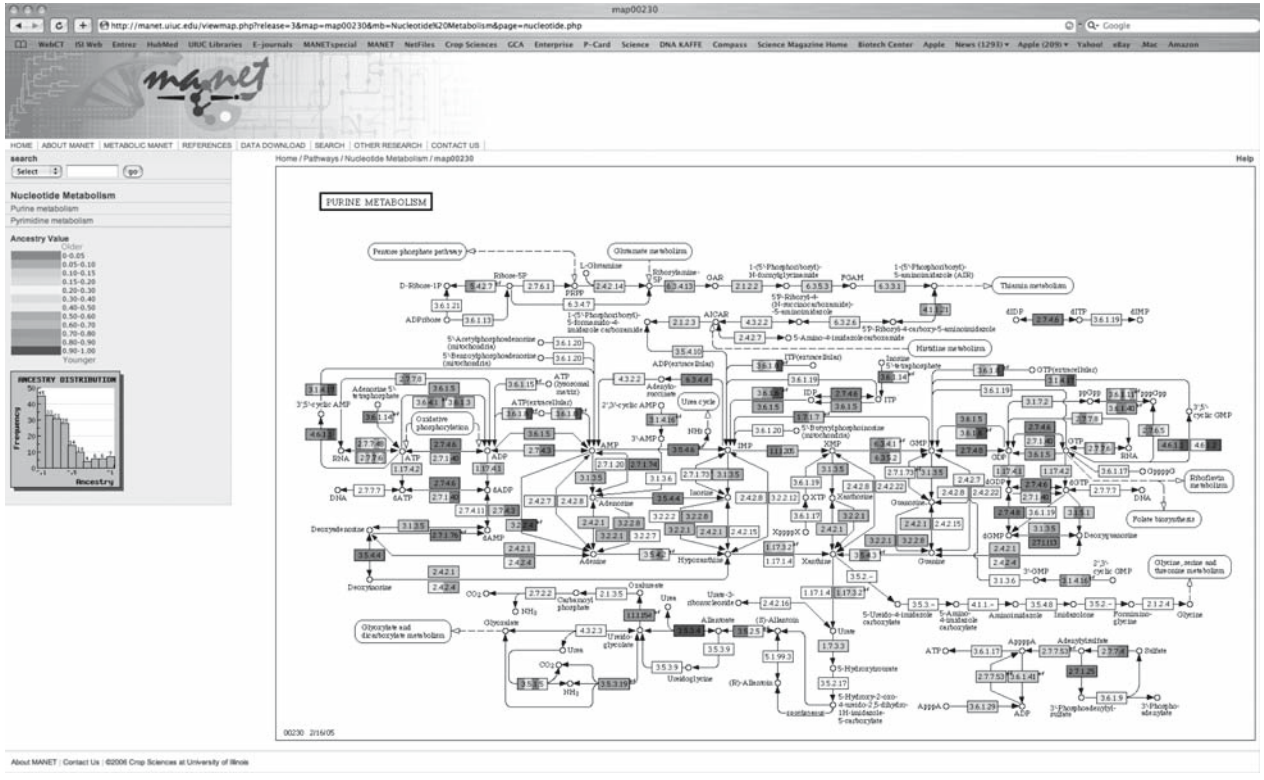

Fig. 6.9 Web page showing a representative subnetwork diagram in the metabolic MANET (Molecular Ancestry Networks) database (http://manet.uiuc.edu). A colored scale is used to assign binned ancestry values of protein folds to enzyme nodes named with Enzyme Commission (EC) numbers. Note that some enzymes have more than one structural assignment. They are multidomain proteins or exhibit different structures in different organisms. Colored enzymes with ancestry assignments from HMM-based predictions are labeled with an 'sf' marking at the top right of KEGG-hyperlinked rectangles depicting enzymatic nodes. Each subnetwork diagram also shows a frequency distribution plot of ancestries (Kim et al. 2006)

(Doolittle 2005). With the advent of evolutionary genomics, the focus of research now shifts heavily towards molecular evolution and the mechanisms that fuel genomic sequence and structural divergence.

Evolutionary genomics places the comparative relationship of organisms within an evolutionary perspective, and does so at the genomic level. The interaction between organisms and the interaction of organisms with the environment are curbed by ecology and evolution and are therefore expected to affect complements defined by the survey of genomic component parts. A substantial body of evidence suggests complex interactions of gene products are responsible for the establishment of pathogenic or symbiotic interactions. For example, plants and pathogenic microbes interact in an endless race to cause disease, and this interplay dominates many important issues in plant pathology (Schumann and D'Arcy 2006). However, our knowledge of how plant and microbial coexistence shapes genomic composition is limited (Ochman and Moran 2001). We know changes in microbes can be large and involve instances of lateral transfer events that exchange considerable genetic material and occur pervasively but not indiscriminately. The existence of fully sequenced genomes from pathogenic and non-pathogenic organisms as well as organisms that have different lifestyles now offer the opportunity to explore the specific effects of organismal coexistence on genomic repertoires. For example, in a recent study the proteomes of several parasites and symbionts exhibiting highly reduced genomes were compared (Chandonia and Kim 2006). The study showed that proteins performing essential functions closely related to transcription and translation exhibited a higher degree of fold usage than proteins in other functional categories. In a systematic and global study of 185 fully sequenced genomes exhibiting free-living, parasitic and obligate parasitic lifestyles, we revealed very specific effects of lifestyle on proteome composition at protein fold level (Wang et al. 2007). For example, Fig. 6.10 shows how protein folds are used and how fold abundance distributes along the tree of fold architectures in genomes from free-living (FL), parasitic (P) and obligate parasitic (OP) bacteria that establish interactions with plants.

The representative organisms analyzed illustrate the general tendency observed in genomes from organisms with P and OP lifestyles to diminish the number of folds used as well as their abundance, regardless of whether the lifestyle causes genomes to be reduced in size. These tendencies are general and are also observed in Archaea and Eukarya. Interestingly, even folds that are ancient and common to all fully sequenced genomes (ABE_o) were considerably under-represented in P and OP bacterial genomes. This and other evidence suggests strongly that establishing parasitic (or symbiotic) interactions results in either protein architectural specialization or the forfeit of protein architectures in exchange of resources from their hosts.

These and many other studies suggest biotic and abiotic interactions impact the makeup and evolution of genomes. I anticipate that patterns and processes uncovered by evolutionary genomics will explain these and other phenomena, benefiting the study of molecular diversity embodied in genome coexistence.

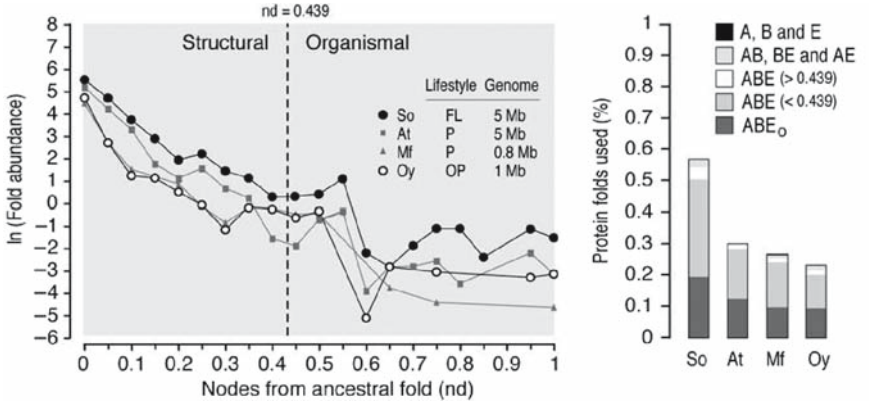

Fig. 6.10 Impact of organismal lifestyle on protein fold architectures in bacterial proteomes. The genomes of four representative bacterial species known to interact with plants and harbor either free-living (FL), parasitic (P) or obligate parasitic (OP) lifestyles were studied and both fold genomic abundance and fold use determined (Wang et al., in preparation). *Shewanella oneidensis* (So) is a FL bacterial species that is present in freshwater sediments and is known to inhabit a wide range of environments and utilize a wide variety of electron acceptors during anaerobic respiration. *Agrobacterium tumefaciens* (At) is a pathogenic bacteria that produces tumors (crown galls) on dicotyledoneous plants. *Mesoplasma florum* (Mf) is a mollicute that establishes P interactions with plants, insects and mammals and has a highly reduced genome. Finally, onion yellows phytoplasma (Oy) is an OP organism that inhabits phloem sieve elements causing a variety of plant diseases. The bacterium has a highly reduced genome and interestingly, lacks the phosphotransferase system, the pentose phosphate pathway and ATP synthases. Fold abundance was studied as a function of the ancestry of individual folds measured by the number of nodes from the most ancestral fold (nd) in the tree of fold architectures. Folds specific to Archaea (A), Bacteria (B) and Eukarya (E) start appearing at an nd value of 0.439 that signals the transition between architectural and organismal diversification in the evolution of the protein world (Caetano-Anollés and Caetano-Anollés 2005). Fold usage in individual genomes was depicted in bar diagrams as the percentage of protein folds used that are either common to all fully sequenced genomes (ABE_o), common to all organismal domains (ABE) or specific to individual or sets of domains

Acknowledgements I would like to thank Jay E. Mittenthal for invigorating discussions and team members Fengjie Sun, Minglei Wang, Hee Shin Kim, Ajith Harish, Liudmila Yafremava, and Vegeir Knudsen for their effort and support. Research was funded by the National Science Foundation (NSF), the Office of Naval Research, Illinois Council for Food and Agricultural Research (C-FAR), the International Atomic Energy Agency (IAEA) in Vienna, and the Critical Research Initiative of the University of Illinois. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the funding agencies.

References

Ancel LW, Fontana W (2000) Plasticity, evolvability, and modularity in RNA. J Exp Zool (Mol Dev Evol) 288:242–283

Babajilde A, Farber R, Hofacker IL, Inman J, Lapedes AS, Stadler PF (2001) Exploring protein sequence space using knowledge based potentials. J Theor Biol 212:35–46

Bajaj M, Blundell T (1984) Evolution and the tertiary structure of proteins. Annu Rev Biophys Bioeng 13:453–492
Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nature Rev 5:101–113
Bartel DP (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281–297
Baulcombe D (2004) RNA silencing in plants. Nature 431:356–363
Bhan A, Galas DJ, Dewey TG (2002) A duplication growth model of gene expression networks. Bioinformatics 18:1486–1493
Billoud B, Guerrucci MA, Masselot M, Deutsch JS (2000) Cirripede phylogeny using a novel approach: molecular morphometrics. Mol Biol Evol 17:1435–1445
Caetano-Anollés G (2001) Novel strategies to study the role of mutation and nucleic acid structure in evolution. Plant Cell Tissue Org Culture 67:115–132
Caetano-Anollés G (2002a) Evolved RNA secondary structure and the rooting of the universal tree of life. J Mol Evol 54:333–345
Caetano-Anollés G (2002b) Tracing the evolution of RNA structure in ribosomes. Nucleic Acids Res 30:2527–2587
Caetano-Anollés G (2005) Grass evolution inferred from chromosomal rearrangements and geometrical and statistical features in RNA structure. J Mol Evol 60:635–652
Caetano-Anollés G, Caetano-Anollés D (2003) An evolutionarily structured universe of protein architecture. Genome Res 13:1563–1571
Caetano-Anollés G, Caetano-Anollés D (2005) Universal sharing patterns in proteomes and evolution of protein fold architecture and life. J Mol Evol 60:484–498
Caetano-Anollés G, Kim H-S, Mittenthal JE (2007) The origins of modern metabolism inferred from phylogenomic analysis of protein architecture. Proc Natl Acad Sci USA 104:9358–9363
Chandonia J-M, Kim S-H (2006) Structural proteomics of minimal organisms: conservation of protein fold usage and evolutionary implications. BMC Struct Biol 6:7
Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Paabo S, Rocchi M, Eichler EE (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437:88–93
Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. Science 300:1701–1703
Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Towards automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287
Collins LJ, Moulton V, Penny D (2000) Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. J Mol Evol 51:194–204
Delsuc F. Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nature Rev Genet 6:361–375
Doolittle RF (2005) Evolutionary aspects of whole-genome biology. Curr Opin Struct Biol 15:248–253
Eddy SR (2001) Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2:919–929
Efimov AV (1997) Structural trees for protein superfamilies. Proteins 28:241–260
Eliceiri GL (1999) Small nucleolar RNAs. Cell Mol Life Sci 56:22–31
Fontana W (2002) Modelling 'evo-devo' with RNA. BioEssays 24:1164–1177
Fontana W, Konings DA, Stadler PF, Schuster P (1993) Statistics of RNA secondary structures. Biopolymers 33:1389–1404
Frank DN, Pace NR (1998) Ribonuclease P: unity and diversity in a tRNA processing ribozyme. Annu Rev Biochem 67:153–180
Gerstein M (1998) Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. Proteins Struct Funct Genet 33:518–534
Gerstein M, Hegyi H (1998) Comparing genomes in terms of protein structure: Surveys of a finite parts list. FEMS Microbiol Rev 22:277–304
Gilbert W (1986) The RNA world. Nature 319:618

Gladyshev GP, Ershov YA (1982) Principles of the thermodynamics of biological systems. J Theor Biol 94:301–343
Grant A, Lee D, Orengo C (2004) Progress towards mapping the universe of protein folds. Genome Biol 5:107
Grishin NV (2001) Fold change in evolution of protein structures. J Struct Biol 134:167–185
Gultyaev PA, van Batenburg FHD, Pleij CWA (2002) Selective pressures on RNA hairpins in vivo and in vitro. J Mol Evol 54:1–8
Gutteridge A, Thornton JM (2005) Understanding nature's catalytic toolkit. Trends Biochem Sci 30:622–629
Harrison A, Pearl F, Mott R, Thornton J, Orengo C (2002) Quantifying the similarities within fold space. J Mol Biol 323:909–926
Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 401:C47–C52
Hermann T, Patel DJ (1999) Stitching together RNA tertiary architectures. J Mol Biol 294:829–849
Higgs PG (1993) RNA secondary structure: a comparison of real and random sequences. J Phys I France 3:43–59
Higgs PG (1995) Thermodynamic properties of transfer RNA: a computational study. J Chem Soc Faraday Trans 91:2531–2540
Higgs PG (2000) RNA secondary structure: physical and computational aspects. Q Rev Biophys 33:199–253
Hou J, Sims GE, Zhang C, Kim S-H (2003) A global representation of the protein fold space. Proc Natl Acad Sci USA 100:2386–2390
House CH, Fitz-Gibbon ST (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: an update. J Mol Evol 54:539–547
Hutvágner G, Zamore PD (2002) RNAi: nature abhors a double-strand. Curr Opin Genet Develop 12:225–232
Kacser H, Beeby R (1984) On the origin of enzyme species by means of natural selection. J Mol Evol 20:38–51
Keefe AD, Szostak JW (2001) Functional proteins from a random-sequence library. Nature 410:715–718
Keenan RJ, Freymann DM, Stroud RM, Walter P (2001) The signal recognition particle. Annu Rev Biochem 70:755–775
Kim H-S, Mittenthal J, Caetano-Anollés G (2006) MANET: tracing evolution of protein architecture in metabolic networks. BMC Bioinformatics 7:351
Kitano H (2002) Computational systems biology. Nature 420:206–210
Koonin EV, Aravind L, Kondrashov AS (2000) The impact of comparative genomics on our understanding of evolution. Cell 101:573–576
Kunin V, Cases I, Enright AJ, de Lorenzo V, Ouzounis CA (2003) Myriads of protein families, and still counting. Genome Biol 4:401
Kurland CG, Collins LJ, Penny D (2006) Genomics and the irreducible nature of eukaryote cells. Science 312:1011–1014
Kyrpides N (1999) Genomes Online Database (GOLD): a monitor of complete and ongoing genome projects worldwide. Bioinformatics 15:773–774
Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–854
Lin J, Gerstein M (2000). Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res 10:808–818
Maizels N, Weiner AM (1994) Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. Proc Natl Acad Sci 91:6729–6734
Mossell E (2003) On the impossibility of reconstructing ancestral data and phylogenies. J Comp Biol 10:669–678
Murzin A (1998) How far divergent evolution goes in proteins. Curr Op Struct Biol 8:380–387

Murzin A, Brenner SE, Hubbard T, Clothia C (1995) SCOP: a structural classification of proteins for the investigation of sequences and structures. J Mol Biol 247:536–540
O'Malley MA, Dupré J (2005) Fundamental issues in systems biology. BioEssays 27:1270–1276
Ochman H, Moran NA (2001) Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. Science 292:1096–1098
Orengo CA, Michie AD, Jones S, Jones DJ, Swindells MB, Thornton JM (1997) CATH: a hierarchic classification of protein domain structures. Structure 5:1093–1108
Page RDM, Holmes EC (1998) Molecular evolution: a phylogenetic approach. Blackwell Science, Oxford
Pastor-Satorras R, Smith E, Sole R (2003) Evolving protein interaction networks through gene duplication. J Theor Biol 222:199–210
Penny D, Hendy MD, Poole AM (2003) Testing fundamental evolutionary hypotheses. J Theor Biol 223:377–385
Philippe H, Laurent J (1998) How good are deep phylogenetic trees? Curr Opin Genet Dev 8:616–623
Ponting CP, Russell RR (2002) The natural history of protein domains. Annu Rev Biophys Biomol Struct 31:45–71
Poole A, Jeffares DC, Penny D (1998) The path from the RNA world. J Mol Evol 46:1–17
Przytycka T, Aurora R, Rose GD (1999) A protein taxonomy based on secondary structure. Nat Struct Biol 6:672–682
Qian J, Luscombe NM, Gerstein M (2001) Protein family and fold occurrence in genomes: power-law behavior and evolutionary model. J Mol Biol 313:673–681
Riley M, Labedan B (1997) Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module. J Mol Biol 268:857–868
Rokas A, Holland PWK (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15:454–459
Rzetsky A, Gomez SM (2001) Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. Bioinformatics 17:988–996
Schmidt S, Sunyaev S, Bork P, Dandekar T (2003) Metabolites: a helping hand for pathway evolution? Trends Biochem Sci 28:336–341
Schultes EA, Bartel DP (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. Science 289:448–452
Schultes EA, Hraber PT, LaBean TH (1999) Estimating the contributions of selection and self-organization in RNA secondary structure. J Mol Evol 49:76–83
Schumann GL, D'Arcy CJ (2006) Essential plant pathology. APS Press, St Paul, Minnesota
Schuster P, Stadler PF (2003) Networks in molecular evolution. Complexity 8:34–42
Seffens W, Digby D (1999) mRNA have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res 27:1578–1584
Sober E, Steel M (2002) Testing the hypothesis of common ancestry. J Theor Biol 218:395–408
Söding J, Lupas AN (2003) More than the sum of their parts: on the evolution of proteins from peptides. BioEssays 25:837–846
Stegger G, Hofman H, Fortsch J, Gross HJ, Randles JW, Sanger HL, Riesner D (1984) Conformational transitions in viroids and virusoids: comparison of results from energy minimization algorithm and from experimental data. J Biomol Struct Dynam 2:543–571
Storz G (2002) An expanding universe of noncoding RNAs. Science 296:1260–1263
Sun F-J, Caetano-Anollés G (2008) The origin and evolution of tRNA inferred from phylogenetic analysis of structure. J Mol Evol 66:21–35
Sun F-J, Fleudépine S, Bousquet-Antonelli C, Caetano-Anollés G, Deragon J-M (2007) Common evolutionary trends for tRNA-derived SINE RNA structures. Trends Genet 23:26–33
Swain TD, Taylor DJ (2003) Structural rRNA characters support monophyly of raptorial limbs and paraphyly of limb specialization in water fleas. Proc R Soc London B 270:887–896
Taylor WR (2002) A 'periodic table' for protein structures. Nature 416:657–660

Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA (2004) Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol 14:208–216
Vukmirovic OG, Tilghman SM (2000) Exploring genome space. Nature 405:820–822
Wagner A (2003) How the global structure of protein interaction networks evolves. Proc R Soc Lond B 270:457–466
Wang M, Caetano-Anollés G (2006) Evolution inferred from domain combination in proteins. Mol Biol Evol 23:2444–2454
Wang M, Boca SM, Kalelkar R, Mittenthal JE, Caetano-Anollés G (2006) A phylogenomic reconstruction of the protein world based on a genomic census of protein fold architecture. Complexity 12:27–40
Wang M, Yafremava LS, Caetano-Anollés D, Mittenthal JE, Caetano-Anollés G (2007) Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res 17:1572–1585
Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol 23:1383–1389
White SH (1994) Global statistics of protein sequences: implications for the origin, evolution, and prediction of structure. Annu Rev Biophys Biomol Struct 23:407–439
Woese CR (2000) The universal ancestor. Proc Natl Acad Sci USA 95:6854–6859
Wolf YI, Brenner SE, Bash PA, Koonin EV (1999) Distribution of protein folds in the three superkingdoms of life. Genome Res 9:17–26
Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Genome trees and the tree of life. Trends Genet 18:472–479
Yang S, Doolittle RF, Bourne PE (2005) Phylogeny determined by protein domain content. Proc Natl Acad Sci USA 102:373–378
Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JHD, Noller HF (2001) Crystal structure of the ribosome at 5.5 Å resolution. Science 292:883–896
Zhang C, Kim SH (2000) A comprehensive analysis of the Greek key motifs in protein β-barrels and β-sandwiches. Proteins 40:409–419
Zhang C, Kim SH (2003) Overview of structural genomics: from structure to function. Curr Op Chem Biol 7:28–32