

No Molecule Is an Island: Molecular Evolution and the Study of Sequence Space

Erik A. Schultes, Peter T. Hraber,
and Thomas H. LaBean

Abstract Our knowledge of nucleic acid and protein structure comes almost exclusively from biological sequences isolated from nature. The ability to synthesize arbitrary sequences of DNA, RNA, and protein *in vitro* gives us experimental access to the much larger space of sequence possibilities that have not been instantiated in the course of evolution. In principle, this technology promises to both broaden and deepen our understanding of macromolecules, their evolution, and our ability to engineer new and complex functionality. Yet, it has long been assumed that the large number of sequence possibilities and the complexity of the sequence-to-structure relationship preempts any systematic analysis of sequence space. Here, we review recent efforts demonstrating that, with judicious employment of both formal and empirical constraints, it is possible to exploit intrinsic symmetries and correlations in sequence space, enabling coordination, projection, and navigation of the sea of sequence possibilities. These constraints not only make it possible to map the distributions of evolved sequences in the context of sequence space, but they also permit properties intrinsic to sequence space to be mapped by sampling tractable numbers of randomly generated sequences. Such maps suggest entirely new ways of looking at evolution, define new classes of experiments using randomly generated sequences and hold deep implications for the origin and evolution of macromolecular systems. We call this promising new direction sequenomics—the systematic study of sequence space.

1 Introduction

Almost a century ago, as it came to be understood that protein and RNA molecules were linear chain polymers, it also became clear that the evolution of these molecules occurred as natural selection chose from among a vast sea of sequence possibilities. Although extraordinary progress has been achieved in understanding the sequence, structure, function, and evolution of biological proteins and RNAs; this corpus of theory has had little to say about the properties of the much larger space of potential, yet unrealized sequences. Since our knowledge of molecular structure remains idiosyncratic to the vanishingly small, profoundly biased biological sampling of the enormous space of possible sequences, it has been impossible to draw

E.A. Schultes (✉)
Department of Computer Science, Duke University, Durham, NC 27708, USA
e-mail: schultes@hedgehogresearch.info

truly general conclusions about molecular structure and evolutionary history. For example, without unevolved sequences in the analysis, it is impossible to resolve properties of proteins and RNAs that must be true in principle (due to physiochemical properties), from those that are the result of historically contingent accidents propagated by evolutionary descent. Hence, it is impossible to understand the ultimate origins of complex protein and RNA structures and to account for evolutionary transformations from one structure to another.

The explanation for why unevolved sequences have remained so neglected can be found in the historical development of molecular evolutionary theory. This began when concepts of molecular evolution were being forged under the twin influences of Darwin's theory of evolution and the new science of structural biology at the molecular scale. Experimental discoveries in protein and RNA biochemistry revealed molecular structures of unprecedented size and complexity, and discovery of highly ordered yet aperiodic structure in these macromolecules needed an explanation that chemistry and physics at the time could not offer. As it was then understood, the conformational entropy of molecules of such size should overwhelm the folding process, and thus the 3-dimensional (3D) structures of proteins and RNAs should resemble amorphous materials, such as liquids or glasses, rather than the observed, highly-ordered structures, more reminiscent of organic crystals. In the absence of a physical explanation, Darwin's theory of natural selection was used to escape from this dilemma: vast periods of time (billions of years) and shear numbers of trials (global populations of organisms) allowed nature to find those exceedingly rare sequences that fortuitously had the ability to overcome conformational entropy and fold to stable 3D structures. Supplanting God as the creator of life, natural selection also became the accepted mechanism accounting for the "miracle" of folded proteins and RNAs. This early fusion of structural biology and Darwinian theory implied that unevolved sequences were disordered, and so they became scientifically irrelevant. Despite ample evidence to the contrary, this assumption took root early and continues to influence thinking today.

In what follows, we first outline the historical development of molecular evolutionary theory in order to trace the origins of the assumption that sequences coding for well-ordered molecular structures can only be located through the strenuous labors of natural selection and that unevolved sequences are predestined to be unstructured. From this point of view, much of what we have to say about proteins is true for RNA and vice versa, and we will interchangeably draw on examples from each. We then summarize a few recent approaches that have challenged this assumption and have not only legitimized the analysis of the structures of unevolved sequences, but have gone on to embrace the much larger space of sequence possibilities as a missing, though essential component of a complete molecular evolutionary theory. We then conclude with what we view as a promising new direction for molecular evolution in a post-genomic era—a research program we call *sequenomics*.

2 The Delicate Clockwork Hypothesis of Molecular Evolution

It is hard to believe that within living memory, there was a time when molecular sequence and structural information for protein and single-stranded RNA molecules was non-existent (for brief reviews, see [51, 68]). At the risk of oversimplification, it can be said that our current understanding of RNA and protein structure has emerged from three fundamental experimental discoveries (recognized by five Noble Prizes) all before circa 1960.

First, through a series of genetic experiments on the bread mold *Neurospora*, Beadle and Tatum [3] had established a correspondence between genes and protein enzymes, formulating the famous *one-gene-one-enzyme hypothesis*.

Second, using analytical centrifugation, Theodore Svedberg and his colleagues had established definitively that proteins were macromolecules of unprecedented size (not colloidal aggregates as had been conjectured) [64]. In conjunction with the first protein sequences that were chemically determined (Sanger's work on insulin, [55]) and the first protein structures that were solved by x-ray crystallography (Perutz's and Kendrew's work on haemoglobin [20] and myoglobin [28]), it became clear that *function at the molecular level (specific binding and catalysis) could be rationalized by atomic-scale molecular structure*.

Third, expanding on the ideas of Mirsky and Pauling [45], and working with the protein ribonuclease, Anfinsen demonstrated that the sequence information of the polypeptide chain can be sufficient to account for the acquisition of the functional, folded structure known as the *native* fold (1957, reviewed in [1]). As it was assumed that the native conformation was also the minimum free energy conformation (an assumption consistent with contemporary structural studies on crystals and small molecules), Anfinsen's proposal became known as the *thermodynamic hypothesis*.

Taken together, these seminal experiments link evolution, sequence information, folded structure, and biochemical function into a single, coherent framework that can be expressed via the well-known mantra: *sequence dictates structure, structure dictates function*. The relationship between sequence and structure implies that if the rules governing folding can be deduced, then it might be possible to predict structure (and function) from sequence information alone, a proposition referred to as the *folding problem*. Concurrent discoveries in nucleic acids also suggested an analogous sequence-structure folding problem for single-stranded RNAs [15]. It is notable, however, that these experiments say nothing regarding the structures of unevolved molecules.

At the time when the first RNA and protein x-ray structures became available, conformational studies of molecules had focused on low molecular weight organic structures having few conformational states. But as single macromolecules, it was unclear how conformational entropy could be overcome. Instead, proteins and RNA were expected to access a large number of conformations either dynamically (i.e., as "random-coils") or as a disordered collapsed (i.e., as a "glass") [9]. Yet, as Anfinsen had demonstrated, despite the vast number of possible conformations, nature could solve the folding problem in seconds.

As Kauffman reviews in his *Origins of Order* [31], natural selection had come to be seen, for a variety of different reasons, as the sole source of order in biology. In the absence of a physical theory of complex polymer folding, researchers adopted natural selection as an explanation for how proteins and RNAs acquired unique folded conformations. For example, Levinthal had noted early on that the number of possible conformations in protein polymeric chains is far too large to allow exhaustive conformational searches for minimum free energy structures—a conundrum that came to be known as *Levinthal's Paradox* [39, 40]. Levinthal noted that protein backbones have two, independent dihedral angles for each amino acid and additional rotations permitted for each side chain. Hence, for proteins of even modest length, say 150 amino acids (to use Levinthal's original formulation), there are 450 degrees of freedom with 10^{300} possible conformations. For RNA, the situation is worse, with six independent backbone angles, phosphate rotamers, distinct sugar conformations, and dozens of hydrogen-bond-mediated base interactions [4].

Levinthal's own answer to the Paradox was that the native, functional fold need not necessarily be a global minimum free energy structure (in contradiction to the thermodynamic hypothesis) and the exploration of conformation space by a protein need not be exhaustive. An exhaustive search implies that the energy associated with each conformation (the so-called energy landscape) is high (unstable) and approximately the same, thus permitting a random-walk. Levinthal postulated that in reality, the energy landscape must be structured into a kind of high-dimensional funnel with correlations between neighboring conformations that could constrain and expedite the folding process. Levinthal commented that this informed energy landscape was presumably an evolutionary adaptation, implying that unevolved proteins would have uncorrelated energy landscapes.

More recently, Frauenfelder and Wolynes [18], also starting with a “random-energy” model of protein folding, have described this evolutionary-induced modification of the energy landscape in their Principle of Minimum Frustration. In this case, the energy landscape is molded by evolution as a “random heteropolymer” is altered by mutation and selection in such a way as to minimize conflicting intramolecular interactions in the native fold. Folding funnels and minimally frustrated structures were used to explain folding in evolved proteins, however, uncorrelated, random energy landscapes were assumed to be reasonable approximations for unevolved molecules. This central role of natural selection in accounting for macromolecular conformations remains pervasive and continues to impact contemporary research. For example, the website for IBMs recent \$100M Blue Gene Initiative in protein folding states unequivocally that unevolved “Heteropolymers typically form a random coil in solution and do not ‘fold’ to any reproducible structure in experimentally accessible times” and that “Arbitrary strings of amino acids do not, in general, fold into a well-defined three-dimensional structure.”

The ability of a protein or RNA polymers to position thousands of atoms into energetically stable, kinetically accessible and functional configurations suggested that selection had to “work hard” to “engineer” the mysteriously complex solutions to the folding problem. Hence, evolved molecular structures came to be seen as delicate clockworks, carefully “designed” and “adjusted” as natural selection chose

from among sequence possibilities. Just as randomly swapping gears and sprockets in a clockwork mechanism would almost certainly be disastrous, the random sequence mutations that are the basis of molecular evolution were expected to be almost always deleterious, if not catastrophic to the Ångstrom-scale architectures. For our purposes here, we refer to this perspective as the *Delicate Clockwork Hypothesis* (DCH) of molecular evolution.

The DCH was a reasonable response to the unprecedented complexity of native protein and RNA structures, especially at this time when natural selection dominated thinking in biology [21]. However, by the late 1960s, molecular sequence data began to contradict the DCH. It had become possible to observe that homologous proteins and RNAs (functionally equivalent molecules from different organisms, e.g., hemoglobin from cow, pig, shark, fish, and worms) had recognizable similarities in their sequences, and that there tend to be more similarities between sequences from closely related species [74]. It turned out that the rare mutations that had been accepted by natural selection, and accumulated over time, could thus be used to define the evolutionary relationships of different species. Genealogical lineages could be mapped, mutation-by-mutation, and for molecules like ribosomal RNA that were common to all species, it was possible to construct universal phylogenetic trees [71].

This diversity among observed molecular sequences was difficult to explain if the DCH was an accurate description of molecular structure. This prompted Salisbury [54] to point out that “If life really depends on each gene being as unique as it appears to be, then it is too unique to come into being by chance mutations.” In an analysis of molecular evolution in sequence space that is curiously reminiscent of Levinthal’s approach to folding in conformation space, Salisbury calculated the probabilities (as estimates of time) of finding a particular protein sequence by random mutation in the enormous space of sequence possibilities. In what we might refer to as *Salisbury’s Paradox*, it was clear that under the prevailing idea of the uniqueness of the gene sequence space would be simply too large, and the number of sequences with folded structures too few for meaningful sampling by random mutation. The DCH assumes that there would be insufficient raw material (favorable mutations) for selection to work on.

The accumulating sequence data forced Salisbury to challenge what he called the “dogma of high gene specificity”. Referencing a previous mathematical analyses by Quastler [50] who speculated that with respect to the amino acid sequence, “a certain neighborhood of structurally related amino acid polymers. . . can perform the same function” and “identical functions can be associated with multiple neighborhoods that are structurally unrelated.” Salisbury concluded that biological functions could not be as rare or randomly distributed in sequence space as the DCH implied. Protein structures must somehow be insensitive to some, or even the majority, of amino acid substitutions, as if clockwork could be randomly rearranged and still keep perfect time. Apparently, in direct contradiction to the DCH, a significant, if unexpected, redundancy in protein sequences’ encoding of structure was the key to molecular evolution. As Levinthal’s Paradox had been solved with the idea of correlated energy landscapes, so was Salisbury’s Paradox solved with correlations in sequence space. Functional proteins were not islands in sequence space, but archipelagos of

interconnected sequences that folded to the same structures and having the same function. In either case, however, these isle refuges of coherent folding dotted a sea of disordered sequences.

Shortly thereafter, John Maynard Smith [62] offered a more rigorous solution to Salisbury's Paradox. Not only did Smith reconsider the fraction of molecular sequences that must be functional, but he also modeled the inherent organization of sequence space: "Suppose now that we imagine all possible amino-acid sequences to be arranged in a 'protein space', so that two sequences are neighbors if one can be converted into another by a single amino-acid substitution." Where A is the number of monomers ($A = 20$ amino acids for proteins, $A = 4$ nucleotides for RNA) and N is the length of the sequence, for any polymer sequence X , there are $(A - 1)N$ sequences that are single-step mutations. For functional proteins, Smith then defined the fraction, f , of these local neighbors that are at least as active as X . So long as the product of f and $19N$ is greater than 1, "meaningful proteins will form a network, and evolution by natural selection is possible." To demonstrate what he had in mind, Smith invoked the image of a popular word game whereby one word can be converted into another word by a series of one letter substitutions, each substitution creating a viable word. So, WORD can be converted into GENE as:

WORD
WORE
GORE
GONE
GENE

The analogy was that viable words were like functional proteins, and the substitution of letters like the substitution of amino acids. Depending on the size and extent of these networks of neutral sequence variants in sequence space (i.e., neutral networks), random mutations would always be able to access a sufficient number of viable sequences to ensure molecular diversification. As evidence that $f19N > 1$, Smith had cited the then recent paper by King and Jukes [30] that presented empirical evidence that a large fraction of amino acid substitutions are selectively neutral. Along with the work of Kimura [29], this observed preponderance of selectively neutral mutations resulted in the formulation of the then heretical *Neutral Theory* of molecular evolution. The Neutral Theory was in direct contradiction to the DCH and ignited surprisingly acrimonious debates that have been well documented (for example, see the Dibner Institute's excellent online resources for "Early Reception of the Neutral Theory" and "Ideology in the Neutralist-Selectionist Debates"). The vigor of these debates is testimony to the degree to which the DCH was held and defended. More recently, Meier and Özbek [44] acknowledge that even today "protein structures and their concurrent functions have remained largely mysterious, as the destruction of these structures by mutation seems far easier than their construction."

Although Smith's concept of neutral networks was consistent with Salisbury's and Quastler's analyses, it takes the extra step of considering sequence space, not as a grab-bag of probabilities, but as a network or graph, with definite relations, subject to mathematical analysis, metrics, and the possibility of establishing coordinate sys-

tems. Smith was envisioning molecular evolution as a diffusion process on neutral networks and molecular phylogenies as “samplings” of such neutral networks.

The diversity of molecular sequence data suggests that neutral networks must be extensive in sequence space. It has now been demonstrated that only 5–20% of a given protein’s amino acid sequence remains invariant during evolution [47] and that sequences with little if any measurable sequence identity can nonetheless fold into identical conformations [63, 66]. The same appears true for RNA: only seven nucleotides are strictly conserved among group I self-splicing introns, yet the secondary (and presumably the tertiary) structure of the ribozyme is preserved [42]. Because these disparate group I isolates have the same fold and function, it is thought that they descended from a common ancestor by taking distinct paths on the same neutral network.

In the last 20 years, well-established methods of nucleic acid and protein synthesis have permitted direct, quantitative evaluation of the DCH. Automated DNA synthesis can be used to construct arbitrary sequences or combinatorial pools of sequences that serve as templates for the genetically-encoded expression of RNA and protein molecules. Although this technology has been primarily used in the synthesis and analysis of evolved, biological sequences (or their mutated counterparts), it can also be adapted to the synthesis of random-sequence, unevolved RNA and proteins. For example, LaBean et al. [35–37], designed and synthesized DNA sequences that, when cloned, expressed unevolved proteins having sequence lengths and amino acid compositions matching small globular proteins found in nature. Surprisingly, they discovered ample evidence for solubility, specific secondary structure, and cooperative unfolding transitions, among these unevolved protein sequences. Davidson and Sauer [11, 12], Prijambada et al. [48], Chiarabelli et al. [8] and Doi et al. [14] have also expressed, purified, and analyzed the structures of unevolved, random-sequence proteins. Although these systems produce oligopeptides that are smaller than biological proteins and sometimes used a restrictive set of the 20 amino acids, they also obtained evidence for secondary structure. Schultes et al. [56] expressed and analyzed a set of 20 RNA molecules whose sequences had been randomly-generated in a computer. Using a battery of physical and chemical techniques for probing the folded conformations of these RNAs, they demonstrated sequence specific, magnesium-dependent folding to structures that were often as compact as evolved sequences having analogous size and nucleotide composition. Hence, for both proteins and RNAs, it appears that at least some elements of folded structure are common in sequence space and are independent of natural selection.

In addition to these structural studies, Schultes and Bartel [57] synthesized a series of RNA to verify the existence of RNA neutral networks for two catalytically active RNAs. This study was also able to demonstrate the close proximity of neutral networks to one another in sequence space (these results will be discussed in more detail in the next section).

Furthermore, by screening combinatorial pools of random-sequence proteins or RNAs for predefined function (i.e. *in vitro* selection), it has now been amply demonstrated that specific, biological-like binding and catalysis are at least as common in sequence space as one in 10^{10} to 10^{12} [70]. Moreover, since combinatorial pools

have been shown to produce functional sequences for particular, arbitrary physiochemical tasks, these same pools, by implication, must contain functional sequences for *any* physiochemical task [10, 26, 27]. Hence, from a space of 10^{60} possible sequences (RNA oligonucleotides having 100 randomized positions; $4^{100} \approx 10^{60}$ possibilities), a typical combinatorial pool containing as few as 10^{14} sequences (only 1 part in 10^{46} of the possible), nevertheless appears to contain every function found in the whole of sequence space.

Taken together, these experimental results demonstrate that sequence space is, contrary to the DCH, densely populated with sequences able to fold into stable, well ordered and even native-like conformations. Furthermore, the diversity of structures and functions available in relatively small samples of sequence space points toward correlations and symmetries within the complex multi-dimensional encoding of structure that have yet to be fully appreciated for their roles in the evolution of new biochemical functions. These results indicate the need for a new theory of molecular structure that is independent of natural selection, evolution, or even biology. This theory would certainly accommodate the dynamics of selection and mutation in accounting for the diversity and disparity among functional sequences and their structures, but this theory would also make specific predictions about how the much larger space of possible sequences facilitates evolution.

Starting from Smith's conception of a protein space, and its analogue for RNA, we have been developing some of the conceptual approaches and the computational and experimental tools of such a theory. We review some of this work in the next section.

3 Theory and Experiments with Unevolved Sequences

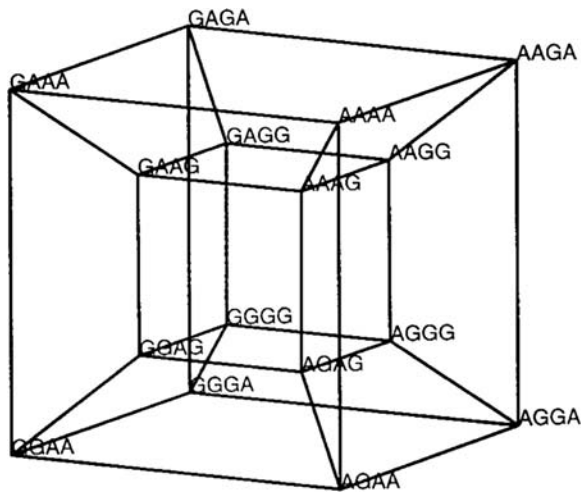
Smith's example made clear the formal organization of related sequences into neighborhoods where nearest neighbors are related by single-step amino acid substitutions. Smith restricted his discussion to proteins of a single length, N , ignoring for the sake of clarity, deletion, and insertion mutations. He also assumed a fixed alphabet size ($A = 20$ amino acids). As each amino acid position of a given sequence can mutate to any one of the other 19 amino acids, it is obvious that each sequence is connected to $19N$ neighboring sequences by single-step substitutions. As this is true for all A^N sequences, protein sequence space is a $(19N)$ -regular graph. In this representation, each node on the graph is a sequence and each single-step substitution is an edge. Generalizing to RNA which has only four monomeric building blocks ($A = 4$ nucleotides), RNA sequence space is a $(3N)$ -regular graph. Table 1 summarizes some of the fundamental mathematical properties of sequence spaces using well-known combinatorial formulas. Figure 1 represents a trivial sequence space for binary strings of length $N = 4$, drawn from the monomers G and A.

To make the concept of protein sequence space more concrete, imagine you are gazing with your $(19N)$ -dimensional eye at the regular graph of a protein sequence space. Since this is the totality of sequence possibilities, all of protein evolution must take place inside the finite boundaries of this graph. From this hyper-bird's-eye

Table 1 Fundamental properties of sequence space

Alphabet size	A , the cardinality of the set of monomers $\{a_1, a_2, a_3, \dots, a_A\}$
Sequence length	N
Number of sequence possibilities	A^N
Number of nearest neighbors (dimensionality of the space)	$(A - 1)N$
Number of neighbors k -steps away	$(A - 1)_N^k C_k$
Number of composition classes	$(N+(A-1))C_N$
Number of sequences per composition class	$N!/(a_1! \cdot a_2! \cdot a_3! \cdot \dots \cdot a_A!)$

Fig. 1 A Boolean hyper-cube is the sequence space for binary sequences, $N = 4$, where each sequence is connected to its four single-step mutational neighbors



view of evolution, imagine color-coding sequences according to various structural or functional properties. First, color all the nodes that are capable of folding into one specific, compact, globular conformation such as a native myoglobin fold (this is, in the context of an aqueous, buffered saline solution at room temperature, conditions typically used in in vitro experiments). Although Salisbury would be interested to know the fraction of sequences that have lit up, we might go on to ask how these colored nodes are distributed across the graph. Are they isotropically dispersed or clustered? Are there multiple clusters? If so, are the different clusters interconnected, or are they isolated by regions of sequence space devoid of myoglobin folds? Using a different color, now mark all those myoglobin sequences that have been actualized in the course of evolution. Are they interconnected by a single $(19N)$ -dimensional phylogenetic tree or do they belong to multiple, independent evolutionary lineages? Using a third color, light up the constellations of sequences capable of folding and functioning in the context of a fibrous protein such as collagen. Do the myoglobin and collagen distributions occupy distinct regions of sequence space, or are they interwoven? What happens to sequences “in between” the colors? Are they half-

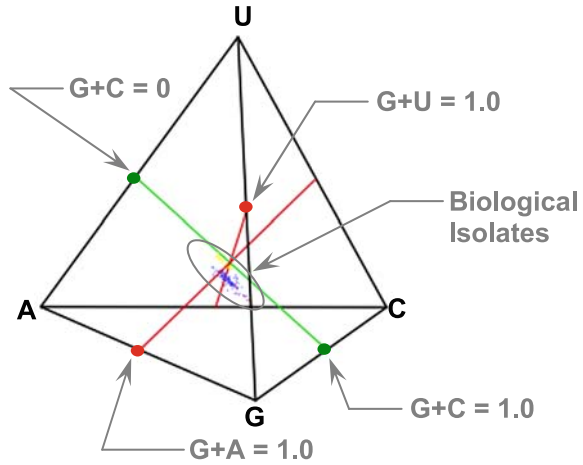
globular, half-fibrous, or something altogether different? Now imagine repeating these experiments while parametrically scanning the experimental conditions (e.g., changing temperature or salt concentration or pH). How sensitive are the two distributions of colors to these changes? Under what conditions do the distributions grow or contract or change shape? Are there certain critical values of conditions for which the distributions undergo abrupt changes in size and shape?

The purpose of this *gedankenexperiment* is to demonstrate the range and scope of fascinating and important questions that were not permissible under the DCH. Of course, in reality, we can never hope to directly view a high-dimensional space, but we can conceive of useful projections. Also, for even short polymers, it will never be possible to exhaustively analyze all possible sequences. However, although any particular sampling of sequence space will always be necessarily sparse, there are nonetheless, practical strategies for sampling sequence spaces that can yield meaningful data. As we will see, sequence spaces contain a number of profound symmetries and complex correlations that when understood will not only help to rationalize the evolution of RNA and protein molecules in nature, but will also serve as a powerful guide in laboratory-based searches for novel molecular structures and function.

Practical approaches to visualizing sequence space must involve a drastic reduction in dimensionality, from $19N$ or $3N$ down to just 2 or 3. Although there are a number of mathematical techniques one could use to perform this dimensional reduction, without proper constraints these projections would tend to confound visualization and obscure the patterns we are seeking. However, based on what we know from the theory of regular graphs and from the physiochemical properties of proteins and RNAs, a useful projection scheme immediately presents itself.

As can be seen in Table 1, all A^N sequence possibilities can be partitioned into a relatively small number of *composition* classes: sequences that have the same proportions of monomers. The number of sequences that belong to a composition class varies dramatically, and can be calculated using the multi-nominal function. The notion of composition class has particular relevance to real molecular sequences because each of the monomers are chemically and structurally distinct, and so sequences biased in one composition are expected to have structural properties that are different from sequences having other compositional biases. Extreme cases are the homopolymers, sequences composed of only a single type of monomer. These sequences are unique in that they contain no sequence information, and their physical properties reflect only the intrinsic propensities of the monomer itself. Depending on the monomer, the homopolymer may or may not be structured. For example, in the case of RNA, poly-guanine is expected to have a collapsed structure (primarily due to base stacking interactions) whereas poly-uridine is not. Indeed, poly-uridine is probably a good example of a true random-coil [56]. For proteins, the same could be said for poly-lysine at alkaline pH (ordered) and at neutral and acidic pHs (disordered). Homopolymers composed of the other amino acids would be more or less the same, but in any case, for both proteins and RNA, the homopolymer sequences act as internal references with information theoretic, physiochemical, and biological significance. Hence, homopolymers act as universal reference points for regular

Fig. 2 The RNA simplex. Chargaff's Axis (green line) defines the familiar gradient in GC content. The simplex also makes explicit additional composition parameters: $G + A$ and $G + U$ (red axes). Isoheteropolymers are located at the intersection of the red and green axes. Plotted are 147 16S rRNA biological isolates (oval): Archaea, red; Bacteria, blue; Eucarya, yellow (from [60])



graph of sequence space. Using the homopolymers, we could even “triangulate” distances (in single-step substitutions) to any arbitrary sequence.

To be concrete, the sequence space for RNAs having exactly 85 nucleotides has 255 dimensions and over 10^{51} sequence possibilities. This enormous number of sequences, however, is partitioned among only 109,736 distinct composition classes. Although sequence spaces have no absolute “center” or “inside/outside” sequences, the space of composition classes does. In the case of RNA, each of the composition classes can be geometrically arranged as points in the volume of a tetrahedron, where the four vertices are the composition classes of the four homopolymers, and the composition class at the center-of-gravity of the tetrahedron represents sequences having a uniform distribution of the four nucleotides (i.e., 25% each A, C, G, and U), Fig. 2. We refer to these maximum entropy sequences as the isoheteropolymers. Because all possible sequences are partitioned among the composition classes, the tetrahedron can be considered a 3-dimensional projection, or simplex, of the $(3N)$ -dimensional regular graph of RNA sequence space.

Because the four nucleotide bases are sterically and chemically distinct, different ratios of the four bases impose an anisotropic distribution of chemical properties among RNA with differing compositions. The most important anisotropy, due to Watson–Crick base-pairing (A pairs with U, C pairs with G), is the symmetry defined by the set of compositions extending from the mid-point of the CG-edge through the center of the simplex, to the mid-point of the AU-edge (green line in Fig. 2). Referred to as Chargaff's Axis (after Chargaff's Rule for base composition in genomic DNA, where molar fractions of $A = T$ & $C = G$ [7]), it is the locus of composition classes formally permitting the maximum possible Watson–Crick base-pairing in RNA. In this way, the RNA simplex projection combines the formal properties of sequence space with well established biophysical properties of RNA.

Taking a lead from the thought experiment described above, we can imagine that evolving RNA sequences map trajectories in sequence space as they undergo modification by mutation and selection. Trajectories involving changes in monomer

composition can be plotted and observed in the RNA simplex. This is done by simply extracting the frequency of nucleotide bases composing individual sequences, and then calculating their simplex coordinates as $G + A$, $G + U$ and $G + C$ contents. In our first analysis, we compiled 2800 distinct sequences representing 15 distinct functional classes of biological RNA [60]. Remarkably, we found that these diverse biological RNAs universally occupied a restricted volume of the simplex, forming narrow clouds that parallel Chargaff's Axis, yet are displaced (by $\sim 5\%$) toward the AG-edge (i.e., toward the purines), Fig. 2. The diversity in sequence and structure among the 15 functional classes of RNAs examined, imply that these molecules share little, if any, evolutionary history and that their coincidence in this region of the simplex was a heretofore unknown example of adaptive convergence. This was observational evidence that a universal principle of macromolecular structure was being independently exploited by different genealogical lineages of RNA.

At the time, we were completing these first analyses, the first complete genomes of free-living organisms were being published, and the idea of calculating and plotting base composition was seen by some to be a step backward to days before efficient sequencing methods had been developed. To the contrary however, by casting RNA evolution first in the context of sequence space, and then in the context of composition space, our analysis immediately revealed profound patterns that even the most comprehensive cladistic methods had failed to detect. This is because cladistic methods seek patterns within evolutionary lineages and are thus unable to resolve trends that are due to universal constraints from trends that merely reflect genealogical descent. Of course, phylogenetic analyses and the approach taken with the RNA simplex are entirely complementary, and either method is diminished in its explanatory power absent the other.

The displacement of the biological distributions from Chargaff's Axis reflects the structural constraints inherent to the folding of the linear phosphodiester backbone of RNA polymers under the influence of basepairing. The acquisition of arbitrarily complex folds in single-stranded RNAs emerges from the alternating composition of double-helical stem structures separated by unpaired single-stranded joining structures. As the bases in the canonical stem structures must, by definition, fulfill Chargaff's Rule (and, therefore, must lie on Chargaff's Axis), it is primarily the composition of unpaired bases in the joining structures that dictate the magnitude and direction of the displacement away from Chargaff's Axis.

For the biological isolates, the *magnitude* of the displacement from Chargaff's Axis is dictated by the compromise between the thermodynamic stability of the folds (favoring stems) and need for complex macromolecular structures (favoring single-stranded joining regions). The *direction* of the displacement of biological isolates from Chargaff's Axis could be, a priori, in any direction, yet the unpaired residues are universally purine-biased. This observed "purine-loading" in nature may reflect the unique chemical properties of purines contributing to stable and specific folding in RNA [34]. For example, X-ray crystallographic analyses of native-fold RNAs have demonstrated the ubiquity of purine-associated structural primitives including the A-minor motif, a tertiary interaction whereby an unpaired adenosine residue docks in the minor groove of a helical stem elsewhere in the macromolecular fold [46]. It has been established that the A-minor motif plays an essential role

in stabilizing overall three-dimensional folded structures and maintaining biological activity [13].

The patterns revealed by the RNA simplex, and the purine-loading observed in structural studies comprise circumstantial evidence for the existence of general principles governing the complex folds required for biological activity. To explain this coincidence of biological sequences in restricted volumes of the RNA simplex, it is necessary to analyze and compare the macromolecular folding of sequences that are distinct from biological isolates. For example, if purine-loading is a universal principle, then purine-depleted sequences would be less likely to acquire biologically relevant folds. Perhaps purine-depleted sequences tend to have lower thermodynamic stability or kinetic barriers that somehow prevent sufficiently complex or adaptable structures. Answering such questions require the synthesis and biophysical characterization of unevolved RNA sequences. Because in vitro methods are laborious even for well-behaved RNA structures [67], we began our probing of unevolved RNA sequences in silico. This required a fast RNA folding algorithm, and a strategy for meaningful, if sparse sampling.

Michael Zuker's program, *mfold*, employs a dynamic programming algorithm to compute the minimum free energy secondary structure of a specified RNA sequence [72, 73]. It also incorporates empirically derived thermodynamic parameters of base-pair interactions [43]. *mfold* is limited to secondary structure prediction (forgoing any attempt at three-dimensional structure prediction) which is to say that it is better at finding reverse-complementary sub-sequences than correct backbone topologies. Nonetheless, it is a robust and efficient algorithm for modeling sequence specific base-pair interactions, which contribute the bulk of the free energy of folding.

At first, the idea of sampling RNA sequence space seems futile. For even short RNAs of only 100 nucleotides have over 10^{60} sequence possibilities. Using *mfold*, it is a heroic task to fold 10^8 sequences, yet this is only one-part in 10^{52} . How could we possibly produce a meaningful analysis of sequence space with such an exceedingly sparse sample? This conundrum can be resolved by employing methods of "perfect sampling." Rather than sampling sequence space by generating random sequences from a uniform distribution of the four nucleotide bases (repeating the sampling from within or near the isoheteropolymers), we instead generate random sequences that have specific base frequencies spanning uniformly and systematically, the entire volume of the RNA simplex. In a method, we call *Constrained Independent Random Sampling* (CIRS), a given composition class is repeatedly and independently sampled producing a cohort of randomly generated sequences that are unrelated, but have identical base compositions. CIRS essentially "shuffles" a sequence within a given composition class, creating a random permutation that is one of many possible molecular isomers.

In our simulations, 100 arbitrary sequences (each having 100 nucleotide bases) were sampled from 1771 compositions classes differing by 5% composition intervals throughout the simplex [58]. The resulting thermodynamic free energies computed from the folded RNAs were then averaged for each composition class and plotted in the simplex (Fig. 3).

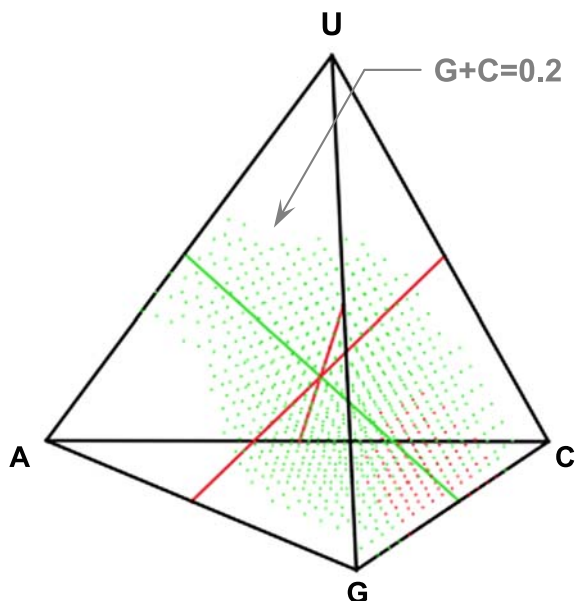


Fig. 3 Constrained Independent Random Sampling of the RNA simplex. Each composition class is colored according to the average computed thermodynamic free energy of folding for 100 arbitrary sequences. Due to the asymmetries of the CG and AU base-pair interactions, the most stable folds (*red*) tend to occur near the CG-edge of the simplex. Folds having intermediate stability (*green*) fan out along Chargaff's Axis toward the AU-edge. Only in CG-depleted sequences, are there ample opportunities for AU base-pair interactions. Blank space indicates composition classes sparse in Watson-Crick partners and, therefore, sparse in RNA polymers having stable, unique folds (from [60])

According to these computations, the most stable folds occur for sequences near the mid-point of the CG-edge (red points). Intermediate stabilities (green points) occur along Chargaff's Axis, where the potential for Watson-Crick base pairing is maximal. The "bottle neck" in this distribution near $G + C = 0.2$ reflects the frustration of stable G:C pair formation by the abundance of A and U residues (and A:U pairs). Blank space is occupied by RNAs lacking Watson-Crick partners and, therefore, stable secondary structure. This complex distribution of RNA folding with respect to composition, maps the spontaneous base-pairing propensity of RNA polymers throughout sequence space. As the RNA simplex makes clear, spontaneous macromolecular properties (spontaneous in the sense that they are independent of selection or rational design) can sometimes provide as much or even more ordered structure than those found in biology. This spontaneous ordering of macromolecular folds, is a purely physiochemical processes driven by the release of free energy of folding and has nothing to do with natural selection or biological evolution. Indeed, from this point of view, it is the preponderance of well-ordered macromolecular folds in sequence space that permits evolutionary adaptation [31]. In the context of sequence space, natural selection is seen merely as a culling mechanism, rather than as a creative force.

Rob Knight's group at the University of Colorado in Boulder used mfold and the RNA simplex to demonstrate that for unevolved sequences having the same base composition as biological isolates, the predicted structures had the same compositional preferences among their structural elements as observed among the biological isolates [61]. That is, like the biological RNAs, these unevolved sequences folded such that purines predominated the unpaired residues. Hence, purine-loading is not a consequence of natural selection, but is a manifestation of the self-ordering properties intrinsic to RNA polymers.

Curious as to how the complex distribution of folding in the RNA simplex might constrain or facilitate RNA evolution, we proposed a stochastic model for mapping evolutionary trajectories within the simplex. Making simple assumptions about selection (tending to maximize thermodynamic stability) and mutation (tending to randomize the sequence) we were able to evaluate the evolutionary "potential" for each composition class, creating the analogue of a "attractor-basin portrait" for RNA evolutionary dynamics in the simplex. Surprisingly, this simple model predicted the mean G + A and G + U values of 928 tRNAs and 382 5SrRNAs to within 3% [58, 59].

In a more sophisticated analysis of the distribution of specific RNA structural motifs in sequence space, Knight et al. [33] discovered (after folding several hundred million arbitrary RNAs on a computational grid) that the sequences capable of folding to the isoleucine aptamer and hammerhead ribozyme structures are most likely to be found in distinct regions of the simplex. This curious result also implied that the neutral networks for these two RNA structures probably has local variation in the degree of connectivity. Furthermore, this result suggests that random-sequence pools of RNAs used in vitro selection experiments could be optimized by compositionally biasing the pool synthesis, thereby focusing the sparse sampling of sequence space into the most promising composition classes.

mfold is explicitly a secondary structure prediction algorithm and ignores tertiary-level interactions completely, so although computational analyses have an important role to play in a survey of sequence space, they are inherently limited and must ultimately be supplemented with analogous empirical data. As mentioned previously, Schultes et al. [56] have implemented CIRS in vitro, synthesizing 10 arbitrary sequences (having 85 nucleotides) at two distinct composition classes: the isoheteropolymers (a useful reference point when considering base composition) and a composition corresponding to the genomic form of the Hepatitis Delta Virus (HDV) self-cleaving ribozyme. The HDV ribozyme and other model sequences were used as structural controls against which to compare the conformations of unevolved RNAs. The structures of these 20 unevolved RNA molecules were then probed using three independent methods: native gel electrophoresis; analytical ultracentrifugation; and lead(II) chemical probing. These experiments demonstrated that these unevolved RNAs had sequence-specific secondary structure configurations and compact magnesium-dependent conformational states comparable to those of evolved RNAs. But unlike evolved sequences, unevolved sequences were prone to having multiple competing conformations. So, by comparing structures of only two dozen RNAs, it was possible to begin teasing apart properties of RNA structure

that are dependent on natural selection from those that are independent of natural selection: Evolution appears necessary to achieve uniquely folding sequences, but not to account for the well-ordered secondary structures and overall compactness commonly observed in nature.

In CIRS, each sequence in constitutes an independent random sample and is therefore unable to address the local network architecture of sequence space. To gain insight into the local neighborhood structure of sequence space we use a different method of sampling called *Local Network Sampling* (LNS). In this case, a particular sequence (called the anchor) is used as the basis for generating a sample set of mutant sequences. In this way, the sample gathers data about how structural correlations between sequences are distributed in sequence space. For instance, LNS may generate all the single- and double-step substitution mutations of the anchor sequence (either in silico, or in vitro using combinatorial pools or microarrays). Such a LNS would provide a direct measurement of the fraction of neighbors that are neutral (f as defined by Smith [62]). In principle, the anchor may be an evolved or unevolved sequence. Using LNS in this way, we could compare the degree of neutrality among evolved and unevolved sequences and/or sequences having different monomer compositions.

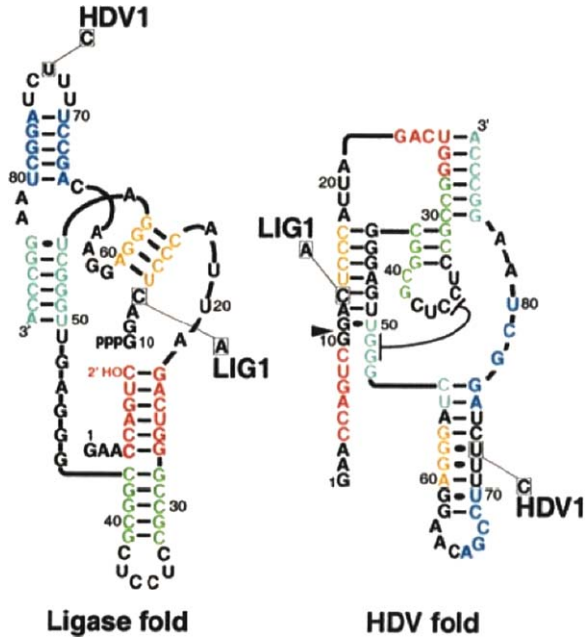
In a different application of LNS, a pair of sequences (an anchor and a target) is connected by a series of sequences that form a connected path through sequence space, linking the anchor and target by single- or double-step substitutions (double-step substitutions are particularly relevant in nucleic acids as they are basis of compensatory mutations in helical stems). The intervening sequences can be generated by random mutations or by mutations that preserve some aspect of structure (i.e., as neutral mutations, Grüner et al. [22, 23]), [17, 52]. For example, Schultes and Bartel [57] used rational design and methods of site-directed mutagenesis to experimentally implement a LNS that demonstrated the proximity of RNA neutral networks in sequence space.

In these LNS experiments, the anchor and target sequences were two different ribozymes: the class III ligase and the antigenomic form of the HDV self-cleaving ribozyme. The class III ligase is a synthetic ribozyme isolated from a pool of random RNA sequences. It joins an oligonucleotide substrate (5'-GAACCAGUC) to its 5' terminus using a 2'-5' linkage that is distinct from the typical 3'-5' linkage used in biological RNAs. The HDV self-cleaving ribozyme carries out the site-specific cleavage reactions needed during the life cycle of the virus (at the position indicated by the arrow in Fig. 5, near G10). The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance.

Using what was known about the structures of these two ribozymes, it was possible to rationally design a single sequence that simultaneously satisfied the base-pairing requirements of both the HDV and ligase ribozymes. Although this sequence design was initially done by hand, the procedure was later automated as a simple computer program (Graham Ruby, personal communication). Indeed, "on paper" a large number of such "intersecting" sequences can be conceived.

One such sequence (Fig. 4) was 42 mutational steps away from the ligase anchor (39 base substitutions, one point deletion, and two single-nucleotide insertions) and

Fig. 4 The intersection sequence: a single RNA sequence accommodating the base-pairing configuration of two different ribozyme folds. Sequence position is numbered and color coded (with respect to the ligase secondary structure), demonstrating that there are no two base-pairs in common between the two folds. Two single-step substitutions are indicated (LIG1 & HDV1) that stabilize one fold over the other (from [57])



44 mutational steps from the HDV ribozyme target (40 substitutions, one deletion, and three insertions). When this sequence was synthesized in the two formats depicted in the figure, catalytic activity significantly above the uncatalyzed rates were detected for both self-ligation and site-specific self-cleavage (Fig. 5). It was shown that ligation occurred with the regiospecificity of the class III ligase (forming a 2' linkage rather than the biological 3' linkage), indicating that the class III ligase fold was achieved by some of the molecules. Cleavage also occurred as expected, with formation of cyclic phosphate. Although ligation and cleavage rates were lower than for the anchor and target sequences, this single sequence could assume two completely different, catalytically active folds.

In designing the intersection sequence, an unavoidable substitution from A to C (at position 13) was required, a position known to be critical to the optimal functioning of the ligase. Substituting C13 with A (creating the LIG1 construct, see Figs. 4 and 5) simultaneously restored ligase activity and introduced a G:A mismatch in the context of the HDV fold. The C13A point substitution dramatically increased the ligation rate (90 times) and lowered the cleavage rate below detection. On the other hand, the U73C substitution (creating the HDV1 construct), which is expected to stabilize the HDV fold, substantially increased site-specific cleavage (120 times), while lowering the ligation rate twofold. The substantial improvement seen with single-nucleotide substitutions suggested that the intersection sequence might be very close to the neutral networks of both ribozymes. With only two additional mutations (again one stabilizing the ligase fold, the other stabilizing the HDV fold), it was demonstrated that two ribozyme sequences (LIG2 and HDV 2), having totally

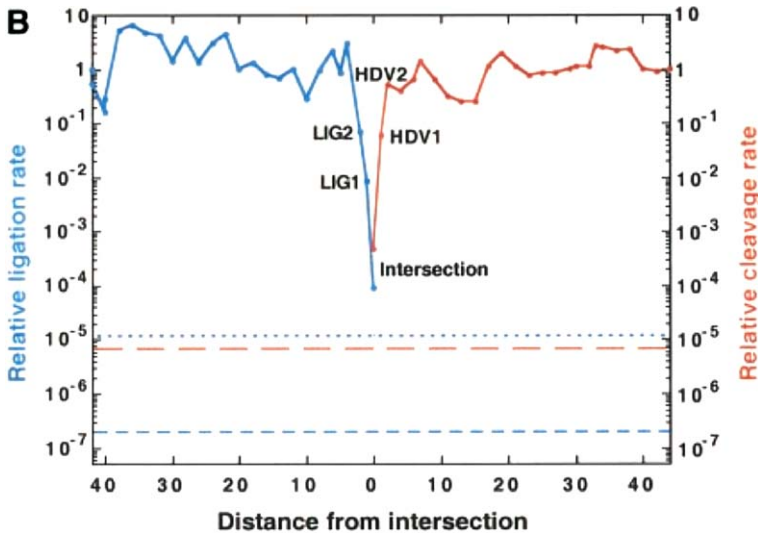


Fig. 5 The close apposition of two neutral networks in RNA sequence space. The *abscissa* indicates distance in single-step mutations from the intersection sequence (center) to the prototype ligase and HDV ribozyme sequences. The *curves* represent measured activities for ligation (*blue*) and cleavage (*red*) along the respective neutral networks. The intersection sequence demonstrated both ligation and cleavage activities (*horizontal dotted lines*). Note that LIG2 and HDV2, though separated by only 4 substitutions, are considered to be neutral for their respective functions (from [57])

different folds and near-prototype-level activities, were separated in sequence space by only four substitutions.

At this point, in order to confirm that both LIG2 and HDV2 are on the respective neutral networks of the anchor and target sequences, paths were designed in sequence space that link these minor variants of the intersection sequence to their prototype sequences. Each step along these paths changed no more than two residues, often as compensatory mutations. Very smooth paths could be designed (Fig. 5), gradually changing nearly half the ribozyme residues yet never falling from the prototype activities by more than sevenfold. The ease by which these paths were designed is consistent with the theoretical results suggesting that neutral networks are a common feature of RNA sequence space. Because the anchor and target ribozymes share no evolutionary history or structural features, it suggests that neutral networks for other pairs of ribozymes closely approach each other. Indeed, by virtue of the high-dimensionality of sequence space, it appears plausible that each ribozyme neutral network closely approaches all other ribozyme networks.

In a very different approach to LNS, Curtis and Bartel [10] employed combinatorial pools to gain insight into the proximity of different ribozyme folds in sequence space. Nucleic acid sequences having a known functionality (i.e., the anchor) can often be optimized by creating high-diversity, partially randomized, combinatorial libraries of sequence variants (often differing from the anchor sequence by 2–10%). These so-called doped pools, containing up to 10^{14} molecules, are then screened for

sequences having higher levels of activity than the anchor. In this case, however, Curtis and Bartel [10] used doped pool selections not to optimize the existing functionality, but as the basis for selecting an entirely new activity unrelated to that of the anchor.

First, a previously described, 90 nucleotide, self-aminoacylating ribozyme (anchor) was partially randomized (each of 65 bases were permitted to vary to one of the three other bases with a probability of 11%). Using methods to separate and amplify any sequence that could covalently link a phosphate group to itself, 23 distinct classes of self-kinasing ribozymes were eventually isolated. The kinase ribozyme activities were radically different than that of the parent (including the use of a trigonal bipyramidal transition state in contrast to the tetrahedral transition state of the parental self-aminoacylating ribozyme). Furthermore, the folds of each of the kinase ribozymes had little, if any correspondence with the fold of the parent. Hence, it was clear that many new folds and new activities can be found in close proximity of sequence space.

However, it was also found that the new kinase ribozymes were on average significantly farther from the self-aminoacylating ribozyme in sequence space (14 mutational steps) than was expected considering the statistical distribution of sequence variants in the initial doped pool (averaging only 7.5 mutational steps). Evidently, the more closely related sequences in the local mutational neighborhood of the anchor frustrate the formation of the alternative folds necessary for the alternative kinase function. Similar results have been obtained in analogous experiments using doped pools of nucleic acid aptamers [10, 26, 27], suggesting this is a general feature of the distribution of structure and function in sequence space. Although a close apposition of neutral networks (only 4 mutational steps) was demonstrated in the LNS experiment using rational design of individual mutants discussed above, the doped pool LNS experiments suggest that this may be relatively rare (though easily designed when the competing structures are understood). Although no conventional imagery can adequately capture the complexity of these high-dimensional neutral networks, it would appear that in some sense, RNA neutral networks are less like footpaths and more like expressways, where many different lanes of traffic are juxtaposed, yet remain distinct.

Prospects for a Protein Simplex and Exploring Protein Neutral Networks

Using the RNA simplex and very limited sampling (in vitro and in vivo), it is surprising how much we could learn about RNA sequence space. The same could be done for protein sequence space, although in the case of proteins there are 20 amino acids rather than 4 nucleotides, complicating the analysis (the composition of space for proteins is a 19-dimensional analogue of a 3-dimensional tetrahedron). Furthermore, the structure of the protein polymer backbone and the kinds of interactions that drive protein folding are different from RNA and necessitate the use of different algorithms, projections, and exploitation of different symmetries in order to

create maps that emphasize the most relevant features of protein folding and function. Hence, protein sequence spaces will require the development of alternative methods of representation. As with RNA, such approaches promise insights into deep problems of protein folding, structure, function, and evolution not otherwise resolvable.

As just one example, from analyses of the native folds of protein sequences found in nature, it has been estimated that the total number of protein folds may be as few as 3000 [41]. This remarkably small number of folds found in nature immediately raises a fundamental question that cannot be answered when referring to genomic data sets alone. Are these 3000 folds the limit of what sequence space has to offer or is this limited number a consequence of historically constrained genealogical lineages? As is the case with RNA, current theories of protein structural biology and evolution make no predictions about what lies beyond biological sequences or the 3000 well-defined folds identified among those sequences. Only by systematically sampling protein sequence space in silico and ultimately in vitro will it be possible to reach a definitive explanation for the observed redundancy of protein folds found in nature.

Not only does this silence about unevolved sequences and structures create an explanatory gap in understanding the distributions of biological proteins, but it also fails to account for the spectrum of complex protein behavior found in nature and the laboratory. For example, it has recently been recognized that entire proteins, or large segments of proteins lack well-structured, three dimensional folds and that the amino acid sequence of these disordered regions can be highly conserved. Some of these disordered segments have been shown to have specific function, such as folding upon binding to specific targets and providing linkers in macromolecular arrays [16]. Without a theory of protein structure extending beyond evolved sequences, it has not been possible to understand the evolutionary significance or the structural biology of these so-called intrinsically unstructured proteins. On the one hand, their disorder may be taken as models of the structures of unevolved sequences. On the other hand, their evolutionarily conserved sequences (and characteristic amino acid composition, [65]) suggests that this “disordered” state is actually evolutionarily derived and, therefore, just as “evolved” as the structures of highly ordered globular or fibrous proteins. Only by learning more about the folds of unevolved protein sequences will it be possible to formulate a coherent understanding of the entire spectrum of protein structure, from disordered homopolymers [53] and intrinsically unstructured proteins to meso-ordered molten globules [49] to insoluble “over-structured” protein aggregates such prions [69].

No Molecule Is an Island

This survey of recent theoretical and experimental findings leads to a very different conception of folding and evolution than had been assumed from the time of the earliest biochemical investigations of protein and RNA. Under the delicate clockwork hypothesis, functional molecules were thought to be rare, isolated islands of

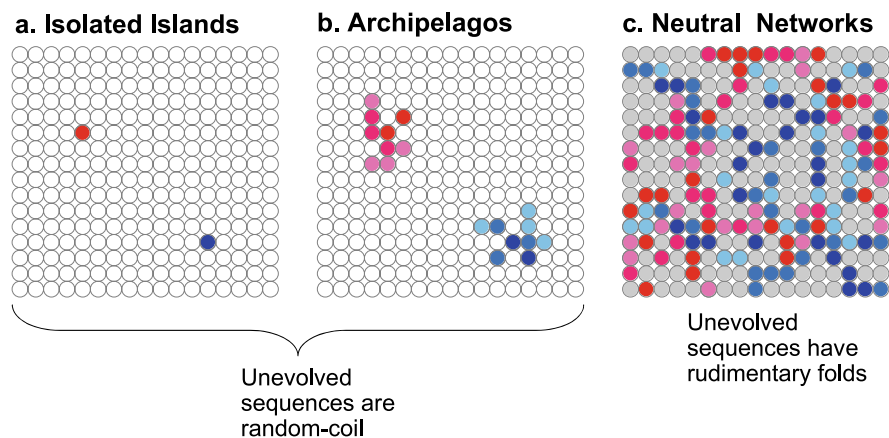


Fig. 6 Three theories of structure and function in sequence space. *Circles* represent different sequences and neighboring circles represent neighboring sequences. *Red colors* and *blue colors* represent two distinct molecular structures/functions. Different *shades of red and blue* represent slight variations in structure or function that are considered neutral. *White circles* indicate sequences without defined structures (e.g., random-coils). *Grey circles* indicate sequences with multiple competing conformations (rudimentary folds)

ordered structure in a sea of disordered random-coils (Fig. 6a). Although accumulating theory and data began to conflict with the DCH, suggesting that the distribution of function in sequence space was more like archipelagos (where isthmuses connected a small chain of islands, Fig. 6b), it remained axiomatic that the surrounding seas of unevolved sequences were disordered and irrelevant to understanding molecular structure and evolution. Since the 1990s, theoretical and experimental results suggest that RNA and protein sequence space is permeated by vast networks of sequences having neutral folds and functions. These networks span the entire sequence space, just as mountain ridges may span entire continents (Fig. 6c). As each fold is thought to correspond to a unique neutral network, these networks must be interwoven, densely “packed” in sequence space like a ball of string made of many different threads. Although it remains an open question as to whether arbitrary, unevolved sequences belong to extensive neutral networks, our experimental analyses of the structures of unevolved RNA and protein sequences suggest that a polymorphic *rudimentary* folding is prevalent among unevolved sequences throughout sequence space. We envision this rudimentary folding, for both RNA and protein, not as the absence of order, but rather as the superposition of multiple “ordered” structures, not unlike induced molten-globule states of native proteins. Indeed, the ubiquity of the molten-globule state (usually seen as a “broken” native fold), is from this point of view evidence of the intrinsic capacity of sequence space to generate ordered structure. In Fig. 6c, the rudimentary state depicted as grey circles, in contrast to the random-coil states (white circles, Figs. 6a and 6b) postulated in the DCH.

The hypothesis that unevolved sequences typically have rudimentary folds provides the coherent framework for structural biology and molecular evolution that

has been missing since the 1950s. Not only does rudimentary folding provide a context in which to understand the many non-native states that have been identified in RNA and proteins, but it also begins to rationalize the origins of functional folds and how one fold could evolutionarily transform into another. Rudimentary folding implies that evolution proceeds not by building up a unique folding pathway from sequences that fold poorly, but by the elimination of competing meta-stable structures from sequences that are already significantly folded. Molecular evolution is therefore not a laborious search for rare sequences, but a problem in negative design, whereby numerous competing folds are selected against. Hence, rather than seeing the vast space of unevolved sequences as a near-perfect vacuum of biological function, the ubiquitous rudimentary fold suggests that sequence space is a nursery of nascent functionality, incubating evolutionary diversification and transformation. Due to the high-dimensionality of sequence space, any unevolved sequence having a rudimentary fold is probably close to a native fold, if not many native folds (as in Fig. 6c, where a grey circle may be near both red and blue circles). This would explain the ease by which molten-globule states can be induced from native folds by mutation or alteration in solution conditions.

Extensive neutral networks of native folding proteins or RNAs explain the diversification of sequences seen in nature, but neutral networks and rudimentary folding also explain the origin of entirely new folds. Different neutral networks that have close apposition in sequence space allow one structure to transform into another with only a small number of mutations. Rudimentary folds may facilitate this transition, acting as intersection sequences, simultaneously maintaining functional capacity for both folds until suitable mutations can complete the transition to the new neutral network. Indeed, rudimentary folding may be indistinguishable from sequences that lie at the intersections of (or between the close apposition of) neutral networks. From this point of view, the molten-globule states of perturbed biological native folds may be seen as intersection sequences bridging two (or more) neutral networks. It would be interesting to see if known molten-globules can be stabilized into alternative yet native-like structures. If so, then it is most accurate to conceive of arbitrary unevolved sequences not as evolutionary dead ends, but as a rich matrix of superimposed structure that bridge the multidimensional space between numerous neutral networks of native folds. In a sequence space supporting ubiquitous rudimentary folding, no molecule need ever be an island unto itself.

Rather than being irrelevant to our understanding of macromolecular structure and evolution, unevolved sequences are the crucial missing link between the large amount of sequence and structural data and a predictive theory of molecular evolution and its intelligent biotechnological application. Learning what we have about navigating and sampling sequence space, and seeing where this could lead, we would like to propose a framework for a more unified and concerted research effort to map protein and RNA sequence space. In analogy to genomics (a comprehensive sampling of biological sequences), we refer to this research program as sequenomics, a comprehensive sampling of sequence space.

4 Sequenomics: Mapping the Universe of Sequence Possibilities

Sequenomics is the study of sequence space. Although this includes all nucleic acid and protein sequences found in nature, the primary concern of sequenomics is the much larger space of unevolved sequences. The goal of sequenomics is to systematically and comprehensively sample nucleic acid and protein sequence spaces, *in silico* and *in vitro*, producing “maps” and “atlases” that define biological distributions, intrinsic symmetries, and physiochemical “territories” in the context of the totality of sequence possibilities. Such maps will reveal previously hidden, universal rules governing molecular folding and evolution, stimulate new approaches in the development of structure prediction algorithms, and suggest novel experiments using individual sequences, pools and microarray technology to inform the search for functional RNAs and proteins.

Before describing our vision of sequenomics, it is helpful to describe what sequenomics is *not*. Sequenomics is not cluster analysis or cladistic analysis of functional sequences. Although many theoretical [2, 19, 32], bioinformatic [38, 44], and experimental [5, 6] research efforts are otherwise complementary to the goals of sequenomics, it is the referral to informative, random samples of sequence space that most distinguishes this new approach.

The *core questions* of sequenomics, for both RNA and protein are: How do the folded conformations of unevolved sequences compare to known biological structures? What is the spectrum of ordered and disordered folding states? How common is native folding? How many different folds are there? How does the distribution of physiochemical properties and the connectivity of neutral networks vary across sequence space? How do these distributions limit or enable evolution? How can we use this knowledge to advance *de novo* design and *in vitro* selection? How can we best search sequence space? The *core technical challenge* of sequenomics is the development of sampling methods that are efficient and informative despite being necessarily sparse (experimentally, we need to sample on the order of a googol sequences using only thousands of sequences).

Answering these core issues will require a concerted approach of theoretical analyses and laboratory experiments, involving the development of new computational tools and experimental technologies with high-throughput capabilities. Based on our own research experience investigating unevolved RNAs and proteins, we have identified four distinct but complementary activities within sequenomics: (1) establishing a theoretical framework; (2) creating interactive visualization tools as low-dimensional windows into high-dimensional sequence spaces; (3) the assembly of a unique database of sequence and structural information combining evolved and unevolved RNAs and proteins; (4) laboratory experiments dedicated to the synthesis and structural characterization of unevolved, random-sequence RNAs and proteins.

Theoretical Sequenomics

The theoretical foundation of sequenomics is the regular graph structure of sequence space. As such, combinatorial and information theoretic tools can be immediately brought to bear in order to reveal numerous symmetries intrinsic to sequence space and correlations that emerge from the physiochemical constraints of macromolecular folding.

These theoretical descriptions will guide the formulation of sampling strategies using computer simulations, and ultimately laboratory experiments, to characterize the folding and structures of unevolved sequences. As discussed above, there are two broad classes of sampling methods in sequenomics: *Constrained Independent Random Samples* (generating sets of sequences randomly and independently under constraints, such as a specified monomer composition) and *Local Networked Samples* (generating sets of sequences that are interconnected as single-step mutational variants of an a priori specified sequence). Although LNS samples can sometimes probe local mutational neighborhoods exhaustively (i.e., all single- or double-step mutations), this method can also be used to design networks of mutational variants that extend as “transects” across the “diameter” of sequence space. Hence, despite the enormous size of sequence space, meaningful random samples of RNA and protein sequence space can be practically constructed and evaluated. As laboratory experiments tend to be more costly than computer simulations, we suspect that maps of sequence space based on samples using computer simulations will precede, and will therefore guide, subsequent experimental efforts.

Along these lines, it will be useful to develop software tools integrating various nucleic acid and protein folding/structure prediction algorithms, with the goal to evaluate and compare the performance of existing algorithms on known sequences and on samples of sequence space. Computational sampling may in some instances require the folding of hundreds of millions of sequences in order to elucidate distinct regions of sequence space and the nature of their boundaries. Some or all of this computed data may be archived into the Sequenomics Database described below. Like any map, the scale or resolution in mapping sequence space will be chosen to present certain features of the “territory” over others. Thus, fast algorithms can be used for constructing small-scale, low-resolution maps, and more demanding algorithms for detailed analysis (e.g., neutral network analyses).

Although no folding algorithm is perfect, we can still develop useful, approximate maps of sequence space using computer simulations. By comparing the structures of evolved and unevolved sequences head-to-head, we can begin to make inferences about background sequences and how they are modified in the course of evolution, even if algorithms can make only low-resolution predictions or have some error in predicting correct folds. This is because any limitation that is idiosyncratic to a particular algorithm will be held constant over individual sequences of the sample and can therefore be handled as systematic error. In addition to the final output of an algorithm, the behavior of the algorithm may also tell us much about the space. For example, for sequences that are poised at the junction of two neutral networks, folding algorithms may demonstrate anomalous run times as the sequence is

undecided about its ultimate structure. Furthermore, the computational sampling of sequence space using different algorithms is also an opportunity to perform comparisons of various algorithms against each other, and evaluate their performance over large datasets that are not confounded by genealogical history or other factors that may be peculiar to biological samples. Such head-to-head comparisons will also indicate where factors like extreme monomer composition and sequence information complexity may expose the intrinsic limitations of a given algorithm.

Sequenomics Visualization

In our work with the RNA simplex described above, we discovered that visual representation was much more than a convenient communication tool. It was essential to conceptualizing RNA sequence space and understanding how the high-dimensionality of sequence information impacts the evolution of RNA molecules. Yet the RNA simplex is only one of many conceivable projections, and additional methods for RNA and novel methods for protein sequence spaces are needed. These visualization tools need to be interactive, taking advantage of space, motion, and parametric control of the projections themselves, in order to identify and communicate inherently complex, and sometimes high-dimensional patterns. The development of visualization software would interlock with the theoretical efforts (described above) and the database efforts (described below), resulting in interactive, low-dimensional computational windows to the high-dimensional regular-graphs of sequence space.

The Sequenomics Database

To create useful maps of sequence space, we will need a centralized system of cataloging molecular sequence and structural data that is much more general than any existing bioinformatics database. In particular, this central data source would need to archive data from both RNA and proteins. But even beyond the juxtaposition of nucleic acids and protein sequences, such a database would be unique in archiving, in a consistent format, structural information for four broad classes of sequences. These sequence classes are necessary as standards and controls when evaluating the structures of unevolved sequences.

- (1) *Reference Sequences (negative controls)*: These are homopolymers, 4 for RNA, 20 for proteins. These 24 sequences contain no information content and so their macromolecular properties reflect the physiochemical properties intrinsic to each of the nucleotide or amino acid monomers. Some homopolymers may be impossible to synthesize and/or experimentally analyze (see experimental section below), and may confound folding algorithms. Nonetheless, homopolymer sequences will still provide essential reference points if only in mapping zones forbidden to biological evolution.

- (2) *Model Sequences (positive controls)*: These are well behaved, artificial sequence designs that fold into simple, predictable structures, e.g., hairpins in RNA and the four-helix-bundle in proteins.
- (3) *Evolved Sequences*: These are sequences isolated from nature or in vitro evolution experiments. For the purposes of sequenomics, only the highest-quality, representative data for the known ranges of molecular structures, functions, source organisms, and ecological settings will be included in this database. Although datasets comprehensive for particular functional classes (e.g., 16 S rRNA used to build universal phylogenetic trees) are important when asking questions about neutral networks, it is the inclusion of sequences from a wide range of functional classes and evolutionary lineages will make this compilation unique among bioinformatics databases, and uniquely positioned to answer the core questions of sequenomics.
- (4) *Unevolved Sequences*: As samples of sequence space, the unevolved sequences are the unknowns that sequenomics wishes to understand. At first, it seems absurd to keep randomly generated sequences in a database, but the arbitrary sequences we generate and then analyze (by computational folding or by lab experiments) become valuable data that deserve to be archived for the purposes of ongoing analyses. Depending on the application, random sequences could be stored as records similar to the biological sequence data or as a seed for pseudo-random number generation.

Experimental Sequenomics

Ultimately, protein and RNA sequence space will need to be probed experimentally, using real molecules in the laboratory. As the actual synthesis, preparation and analysis of protein and RNA sequences is much more costly than its virtual analogue, the issues surrounding efficient sampling methodologies become even more important. Experimentalists, however, will be able to consult the theoretical maps of protein and RNA sequence space (and even download specific unevolved sequences from the sequenomics database) to find sets of sequences permitting specific hypotheses to be tested. There are three experimental methods by which arbitrary, unevolved sequences can be synthesized and their structures analyzed.

- (1) *The synthesis of specific arbitrary sequences*: Automated DNA synthesis has made the construction of arbitrary DNA templates routine. From these templates, the synthesis of specified RNAs and proteins sequences can be implemented. Using commercial sources of DNA synthesis, it is possible for a single lab to make and process hundreds of oligonucleotide templates per year. Using high-throughput technologies, it is not inconceivable that tens of thousands of oligonucleotide and protein sequences could ultimately be made and perhaps even characterized per year. These techniques would be used to synthesize unevolved sequences, including the rational construction of putative neutral networks.
- (2) *The synthesis of combinatorial pools*: Automated DNA synthesis can also be adapted to the synthesis of diverse pools of DNA templates. In this case, residue

positions along the length of the template are permitted to incorporate a mixture of the four nucleotides, simultaneously creating many, random-sequence DNA templates. Routine synthesis and preparative techniques can yield individual pools with 10^{15} unique sequences. Such pools have been used as the starting point for the selection of a wide variety of functional biopolymers. However, from the point of view of sequenomics, it is the sequences from these pools that are *not* selected that are of interest. Combinatorial pools can serve as a means to economically isolate large numbers of arbitrary sequences for direct analysis (although, unlike the synthesis of specific arbitrary sequences noted above, the monomer composition will fluctuate around the composition of the nucleotide mixture during pool synthesis). Combinatorial pools having biased compositions, although currently possible, have yet to be explored extensively as a means of focusing search in sequence space. Mathematical and experimental exploration of the role compositional focusing might play in optimizing combinatorial searches is an important goal of sequenomics. In contrast to this “shotgun” sampling of sequence space (where each molecule has uncorrelated sequence), these combinatorial methods can also be used to sample the local neighborhood of particular sequences. This “neighborhood” sampling method generates a pool of sequence variants that differ from the original sequence (on average) by some specified number of mutations (typically in the range of 2 to 10% of N). In this case, the 10^{15} molecules in the pool are highly correlated. Entire pools or cloned isolates from these pools, can be assayed for structure.

- (3) *Fabrication of microarrays*: Microarray technology for synthesis and display of protein and nucleic acid sequences has found wide application in genomic studies. Microarrays permit the quantitative analysis of large numbers of sequences (for binding or other biochemical tasks) in parallel using automated “chip” readers. As a cross between the synthesis of specific arbitrary sequences and the synthesis of combinatorial pools, microarrays will find extensive and novel uses in the sequenomics, especially in the exhaustive search of local mutational neighborhoods for neutral networks or intersecting neutral networks.

The characterization of folding and structure in unevolved sequences will be performed using any of various, standard techniques to be determined as a compromise between resolution and throughput. However, particularly useful techniques may include Nuclear Magnetic Resonance Spectroscopy (a technique that can yield precise measurements of the overall order-disorder in a protein or RNA) and Temperature Gradient Gel Electrophoresis (a technique that permits expedient measurement of thermodynamic and kinetic properties, e.g., [24, 25]). As noted above in the outline of the sequenomics database, positive structural controls will include sequences having trivial but well-behaved structures or sequences that code for complex structures that have been previously well characterized. Negative structural controls, or reference molecules, include sequences such as homopolymers, having no information content and in some cases no unique structure.

The dedication of expensive resources to the study of unevolved protein and RNA sequences seems, at first, anathema to both the spirit and practice of structural biology. Research proposals compete and funding is justified through the promise of

application in biotechnology and medicine. How could unevolved sequences, having by definition, no known function, find their way to the head of long queues of excellent proposals struggling for scarce resources? We believe that once a solid theory of sequence space is developed, visualization tools are in place, and the sequenomics database is constructed, a large number of well-defined, tractable experiments will present themselves, and the space of unevolved sequences will become the obvious frontier of post-genomic and post-structural proteomic research.

References

1. Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181:223–230
2. Armstrong KA, Tidor B (2008) Computationally mapping sequence space to understand evolutionary protein engineering. *Biotechnol Prog* 24:62–73
3. Beadle GW, Tatum EL (1941) Genetic control of biochemical reactions in *neurospora*. *Proc Natl Acad Sci* 27:499–506
4. Bloomfield VA, Crothers DM, Tinoco I (2000) *Nucleic acids: structures, properties, and functions*. University Science Books, Sausalito
5. Breaker RR (2004) Natural and engineered nucleic acids as tool to explore biology. *Nature* 432:838–844
6. Carothers JM, Oestreich SC, Davis JH, Szostak JW (2004) Information complexity and functional activity of RNA structure. *J Am Chem Soc* 126:5130–5137
7. Chargaff E (1950) Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia* 6:201–209
8. Chiarabellia C et al (2001) Investigation of *de novo* totally random biosequences, part II: on the folding frequency in a totally random library of *de novo* proteins obtained by phage display. *Chem Biodivers* 3:840–859
9. Creighton TE (1993) *Proteins: structures and molecular properties*, 2nd edn. Freeman, New York, pp 172–173
10. Curtis EA, Bartel DP (2005) New catalytic structures from an existing ribozyme. *Nat Struct Mol Biol* 12:994–1000
11. Davidson AR, Sauer RT (1994) Folded proteins occur frequently in libraries of random amino acid sequences. *Proc Natl Acad Sci* 91:2146–2150
12. Davidson AR, Lumb KJ, Sauer RT (1995) Cooperatively folded proteins in random sequence libraries. *Nat Struct Biol* 2:856–864
13. Doherty EA et al (2001) A universal mode of helix packing in RNA. *Nat Struct Biol* 8:339–343
14. Doi N, Kakukawa K, Oishi Y, Yanagawa H (2004) High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. *Prot Eng Des Sel* 18:279–284
15. Draper DE (1992) The RNA-folding problem. *Acc Chem Res* 25:201–207
16. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Biol* 6:197–207
17. Fontana W, Schuster P (1998) Continuity in evolution: on the nature of transitions. *Science* 280:1451–1455
18. Frauenfelder H, Wolynes PG (1994) Biomolecules: where the physics of complexity and simplicity meet. *Phys Today* 47:58–64
19. Gan HH, Pasquali S, Schlick T (2003) Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Res* 31:2926–2943
20. Green DW, Ingram VM, Perutz MF (1953) The structure of hemoglobin, IV: sign determination by isomorphous replacement method. *Proc R Soc Lond A* 255:287–307

21. Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc R Soc Lond B* 205:581–598
22. Grüner W et al (1996a) Analysis of RNA sequence structure maps by exhaustive enumeration, I: Neutral networks. *Mon Chem* 127:355–374
23. Grüner W et al (1996b) Analysis of RNA sequence structure maps by exhaustive enumeration, II: Structures of neutral networks and shape space covering. *Mon Chem* 127:375–389
24. Guo F, Cech TR (2002) Evolution of tetrahymena ribozyme mutants with increased structural stability. *Nat Struct Biol* 9:855–861
25. Hecker R et al (1988) Analysis of RNA structure by temperature-gradient gel electrophoresis: viroid replication and processing. *Gene* 72:59–74
26. Held DM et al (2003) Evolutionary landscapes for the acquisition of new ligand recognition by RNA aptamers. *J Mol Evol* 57:299–308
27. Huang Z, Szostak JW (2003) Evolution of aptamers with a new specificity and new secondary structure from ATP aptamers. *RNA* 9:1456–1463
28. Kendrew JC, Bode G, Dintzis HM, Parrish RC, Wykoff H (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* 181:660–662
29. Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217:624–626
30. King JL, Jukes TH (1969) Non-Darwinian evolution. *Science* 164:788–798
31. Kauffman SA (1993) *The origins of order: self-organization and selection in evolution*. Oxford University Press, New York
32. Kim N, Shin JS, Elmetwaly S, Gan HH, Schlick T (2007) RAGPOOLS: RNA-as-graph-pools a web server for assisting the design of structured RNA pools for in vitro selection. *Bioinformatics*. doi:10.1093/bioinformatics/btm439
33. Knight R et al (2005) Abundance of correctly folded RNA motifs in sequence space, calculated on computational grids. *Nucleic Acids Res* 33:6671–6671
34. Lambros RJ, Mortimer JR, Forsdyke DR (2003) Optimum growth temperature and the base composition of open reading frames in prokaryotes. *Extremophiles* 7:443–450
35. LaBean TH, Kayffman SA (1993) Design of synthetic gene libraries encoding random sequence proteins with desired ensemble characteristics. *Protein Sci* 2:1249–1254
36. LaBean TH, Kauffman SA, Butt TR (1995) Libraries of random-sequence polypeptides produced with high yield as carboxy-terminal fusions with ubiquitin. *Mol Divers* 1:29–38
37. LaBean TH, Schultes EA, Butt TR, Kauffman SA (2009) Protein folding absent selection (submitted)
38. Leontis N et al (2006) The RNA ontology consortium: an open invitation to the RNA community. *RNA* 12:533–541
39. Levinthal C (1968) Are there pathways for protein folding? *Extrait J Chim Phys* 65:44–45
40. Levinthal C (1969) How to fold graciously. In: DeBrunner JTP, Munck E (eds) *Mossbauer spectroscopy in biological systems: proceedings of a meeting held at Allerton House, Monticello, IL*. University of Illinois Press, Champaign, pp 22–24
41. Liu X, Fan K, Wang W (2004) The number of protein folds and their distribution over families in nature. *Proteins* 54:491–499
42. Lisacek F, Diaz Y, Michel F (1994) Automatic identification of group I introns cores in genomic DNA sequences. *J Mol Biol* 235:1206–1217
43. Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288:911–940
44. Meier S, Özbek S (2007) A biological cosmos of parallel universes: does protein structural plasticity facilitate evolution? *BioEssays* 29:1095–1104
45. Mirsky AE, Pauling L (1936) On the structure of native, denatured, and coagulated proteins. *Proc Natl Acad Sci* 22:439–447
46. Nissen P et al (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc Natl Acad Sci* 98:4899–4903
47. Pinker RJ, Lin L, Rose GD, Kallenbach NR (1993) Effects of alanine substitutions in alpha-helices of sperm whale myoglobin on protein stability. *Protein Sci* 2:1099–1105

48. Prijambada ID et al (1996) Solubility of artificial proteins with random sequences. *FEBS Lett* 382:21–25
49. Ptitsyn OB (1995) Molten globule and protein folding. *Adv Protein Chem* 47:83–229
50. Quastler H (1964) The emergence of biological organization. Yale University Press, New Haven
51. RajBhandary UL, Kohrer C (2006) Early days of tRNA research: discovery, function, purification and sequence analysis. *J Biosci* 31:439–451
52. Reidys CM, Stadler PF, Schuster P (1997) Generic properties of combinatorial maps: neural networks of RNA secondary structures. *Bull Math Biol* 59:339–397
53. Rucker AL, Creamer TP (2002) Polyproline II helical structure in protein unfolded states: lysine peptides revisited. *Protein Sci* 11:980–985
54. Salisbury FB (1969) Natural selection and the complexity of the gene. *Nature* 224:342–343
55. Sanger F (1952) The arrangement of amino acids in proteins. *Adv Protein Chem* 7:1–69
56. Schultes EA, Spasic A, Mohanty U, Bartel DP (2005) Compact and ordered collapse in randomly generated RNA sequences. *Nat Struct Mol Biol* 12:1130–1136
57. Schultes EA, Bartel DP (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* 289:448–452
58. Schultes E, Hraber PT, LaBean TH (1999a) A parameterization of RNA sequence space. *Complexity* 4:61–71
59. Schultes EA, Hraber PT, LaBean TH (1999b) Estimating the contributions of selection and self-organization in RNA secondary structures. *J Mol Evol* 49:76–83
60. Schultes E, Hraber PT, LaBean TH (1997) Global similarities in nucleotide base composition among disparate functional classes of single-stranded RNA imply adaptive evolutionary convergence. *RNA* 3:792–806
61. Smit S, Yarus MY, Knight R (2006) Natural selection is not required to explain universal compositional patterns in rRNA secondary structure categories. *RNA-A Publ RNA Soc* 12:1–14
62. Smith JM (1970) Natural selection and the concept of protein space. *Nature* 225:563–564
63. Sondek J, Shortle D (1990) Accommodation of single amino acid insertions by the native state of staphylococcal nuclease. *Proteins* 7:299–305
64. Svedberg T, Fahraeus R (1926) A new method for the determination of the molecular weights of proteins. *J Am Chem Soc* 48:430–438
65. Tompa P (2002) Intrinsically unstructured proteins. *Trends Biochem Sci* 27:527–533
66. Urfer R, Kirschner K (1992) The importance of surface loops for stabilizing an eightfold beta alpha barrel protein. *Protein Sci* 1:31–45
67. Uhlenbeck OC (1995) Keeping RNA happy. *RNA* 1:4–6
68. van Holde KE (2003) Reflections on a century of protein chemistry. *Biophys Chem* 100:71–79
69. Weissmann C (2004) The state of proin. *Nat Rev Microbiol* 2:861–871
70. Wilson DS, Szostak JW (1999) In vitro selection of functional nucleic acids. *Annu Rev Biochem* 68:611–647
71. Woese CR (2000) Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci* 97:8392–8396
72. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* 9:133–149
73. Zuker M (2003) mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31:3406–3415
74. Zukerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel H (eds) *Evolving genes are proteins*. Academic Press, New York