GENERIC PROPERTIES OF THE SEQUENCE-STRUCTURE RELATIONS OF BIOPOLYMERS

PETER F. STADLER
Institut für Theoretische Chemie und Strahlenchemie
Universität Wien, Währingerstr. 17, A-1090 Wien, Austria, &
Santa Fe Institute
1399 Hyde Park Rd., Santa Fe NM 87501, U.S.A.
Email: `studla@tbi.univie.ac.at` *or* `stadler@santafe.edu`

1. Modeling Molecular Evolution

Biological evolution is a highly sophisticated dynamical phenomenon, and its complexity is often confusing. For the purpose of analysis it may be partitioned into the four partial processes depicted in figure 1 [25]: population dynamics, (population) support dynamics, genotype-phenotype mapping, and fitness evaluation. These components are properly visualized as mappings between abstract metric spaces:

– the *concentration space* of biochemical reaction kinetics,
– the *sequence space* of polynucleotide sequences,
– the *shape space* of biopolymer structures, and
– the *parameter space* of kinetic rate constants.

Kinetics of chemical reactions or changes in populations are recorded in concentration space, the conventional space of chemical reaction kinetics. Variables count numbers of particles, molecules, virions, cells, or organisms, for example, and display their changes over time. Population dynamics is commonly described by differential equations, by difference equations in case of synchronized generations, or by a stochastic process in case of small particle numbers [12]. Concentration space deals only with the classes of genotypes actually present.

Sequence space (Q^n_α, d) is a metric point space containing all α^n possible genotypes of given chain length n, where α denotes the size of the polynucleotide or amino acid alphabet. Distances in sequence space are expressed by the minimum number of single nucleotide exchanges or point mutations

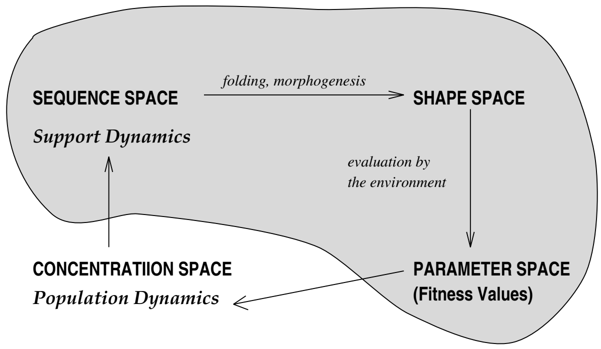

Figure 1. Evolutionary dynamics. *In vitro* evolution of biomolecules is partitioned into four components: (i) population dynamics, (ii) population support dynamics, (iii) genotype-phenotype mapping, and (iv) evaluation of the phenotype. Wright's notion of a *fitness landscape* is the composite of genotype-phenotype mapping and phenotype evaluation (shaded in gray). Population dynamics is tantamount to chemical reaction kinetics of replication, mutation and selection. Population support dynamics describes the migration of populations in the space of genotypes. Genotype-phenotype mapping unfolds biological information stored in polynucleotide sequences, which in turn is translated into kinetic constants for population dynamics.

converting one sequence into another: the Hamming distance d. The notion of a sequence space encapsulates a simple modeling of mutation frequencies because genotypes are grouped according to the minimum number of point mutation events that separate them.

The genotypes present in a population form a collection of points in sequence space. When a mutant is produced, an new point in sequence space is occupied and a new variable "appears" in concentration space; when a variant dies out, the corresponding point in sequence space becomes vacant and a variable "disappears" from the population dynamics picture. The cloud of occupied points in sequence space, the so-called *support* of a populations, therefore moves. Molecular phylogenies therefore can be viewed as reconstructions of the time evolution of the population support (or rather its center of mass, the consensus sequence) from the "snapshot" of the present-day sequences.

A complete model of evolutionary dynamics has to explain also the parameters (rate constants) entering the kinetic equations. These *fitness values* are the consequence of the species' *phenotypes*. The *shape space* $\mathcal{S}$ contains all phenotypes that are formed by processing genotypes in a given context. The genotype-phenotype mapping encapsulates all of morphogenesis in the case of higher organisms, in the context of molecular evolution, however, phenotypes are tantamount to the spatial structures of biopolymers. Sewall Wright [30] introduced the term *fitness landscape* for the combined mapping from genotypes to fitness values.

The dynamics of evolution is oftentimes simplified to adaptive walks on fitness landscapes [16], thereby disregarding concentrations. This is an acceptable shortcut as long as different phenotypes do not interact directly but only compete for common resources. Many simulation studies in this field use *ad hoc* substitutes for biological fitness landscapes such as spin glass Hamiltonians. It can be shown, however, that the properties of a fitness landscape are determined almost exclusively by its underlying sequence-structure map [28]. In the case of the most important biopolymers the landscapes deviate significantly from simple disordered systems.

2. RNA Secondary Structures

Meaningful notions of structure are of course context-depend: successful descriptions of active sites of enzymes (or ribozymes) or specific binding sites for regulatory elements on DNA, for example, require high precision at atomic resolution as provided by X-ray crystallography whereas studies on phylogenetic conservation of structures can be done much better on the coarse-grained level of protein *ribbon* diagrams or RNA secondary structures. Some levels of coarse-graining are not only of physical relevance but also suitable for mathematical modeling. RNA is a particularly fruitful system because the structure prediction problem can be solved efficiently at least at the approximate level of secondary structures.

Experimental energy parameters are available for the individual contributions as functions of type (stacked pair, interior loop, bulge, multi-stem loop), of size, of the type of the delimiting base pairs, and partly of the sequence of the unpaired subsequences, see e.g. [8]. Additivity of the energy contributions allows us to calculated minimum energy structure of an RNA sequence by means of dynamic programming [31]. An efficient implementation of this algorithm is part of the `Vienna RNA Package` [11] (available from `www.tbi.univie.ac.at`). The statistical properties of RNA secondary structures reported in this contribution depend very little on choices of algorithms and variations of the parameter set [29].

The number of acceptable secondary structures can be computed exactly. An asymptotic expansion shows that it grows like 1.8488^n, assuming that (i) the minimum stack length is two base pairs and (ii) the minimum size hairpin loop has three unpaired nucleotides. The number of different structures that are actually formed as minimum energy structure of some sequence is even smaller. Our best estimate is 1.65^n. In contrast, there are 4^n possible RNA sequences composed from the natural **AUGC** alphabet; thus many sequences must fold into the same structure. We call the set $S(\psi)$ of all sequences folding into the same structure ψ the *neutral set* of ψ.

The logic of RNA base pairing implies *a priori* restrictions on $S(\psi)$.

Only the combinations **GC**, **CG**, **AU**, **UA**, **GU**, and **UG** form stable base pairs. Therefore, if positions i and j are paired in a structure ψ, then only one of these six combinations of nucleotides may occur in the corresponding sequence positions. The sequences satisfying this condition form the set of *compatible sequences* $C(\psi)$. Clearly we have $S(\psi) \subset C(\psi)$. Proteins sequences apparently are not subject to a simple restriction of this form.

3. Protein Structures

Computational surveys of protein space similar to the explorations of the RNA world are precluded by the notorious complexity of the protein folding problem, and by the fact that there is no obvious biophysically meaningful and computationally simple coarse resolution of protein structures. For our purposes, however, a technique is sufficient that allows us to decide whether a sequence x folds into a particular structure ψ. This *inverse folding problem* is less demanding than predicting the unknown protein structure of a given amino acid sequence [3].

Our starting point is a potential function $W(x, \psi)$ evaluating the energy of a sequence x when folded into a structure ψ, the latter being defined defined by the spatial coordinates of its C^{α} and C^{β} atoms, respectively [1]. Recent studies using so-called knowledge based potentials, see e.g. [27], demonstrated that the energy of the native fold of a sequence x can be estimated from the distribution of the energy values of x in its conformation space. As a measure for the quality of fit of sequence x and structure ψ the *z-score* [4] $z(x, \psi) = [W(x, \psi) - \overline{W}(x)]/\sigma_W(x)$ is used. Here $\overline{W}(x)$ is the average energy of sequence x in all conformations in a database and $\sigma_W(x)$ is the standard deviation of the corresponding distribution. Empirically, native folds have z-scores in a narrow characteristic range. Hence one may assume that x is a member of the neutral set $S(\psi)$ if and only if the z-score of x in conformation ψ is in the native range. Of course, only native structures ψ that are already in a structure database can be explored by this method.

The number of possible protein sequences is enormous: for $n = 100$ residues there are 20^{100} sequences. On the other hand, the repertoire of stable native folds seems to be highly restricted or even vanishingly small [5]. For example, it has been frequently observed that seemingly unrelated sequences have essentially the same fold [13]. Again, there seem to be by far more sequences than structures.

4. The Topology of Neutral Sets

Data from both large samples of long RNA sequences [26] and from exhaustive folding of all short RNAs [9] suggest a distinction between *common* structures, which are formed by more than the average number of sequences, and *rare* structures. Common structures represent only a small fraction of all structures, but almost all RNA sequences fold into one of these few secondary structures. Lattice protein models exhibit the same behavior [18]. Only the common structures play a role in natural evolution and in evolutionary biotechnology [24].

For an understanding of molecular evolution we need to know how sequences folding into the same common structure are distributed in sequence space, because it is the topology of neutral sets that determines whether similar structures with distant sequences (a phenomenon commonly observed with both RNA and proteins) may have originated from a common ancestor, or whether they must be the result of convergent evolution. Three approaches have been applied so far to studying the topology of neutral sets: a mathematical model of genotype-phenotype mapping based on random graph theory [23], extensive sample statistics [26, 1] and exhaustive folding of all sequences with given chain length n [10]. It was noticed already in early work on RNA secondary structures [7] that a substantial fraction of point mutations are neutral, i.e., that many sequences differing only in a single position fold into the same secondary structure. The same seems to be true for proteins. On the other hand, sequences folding into the same structure are almost randomly distributed in sequence space (more precisely in the subspace of compatible sequences $C(\psi)$ in the case of RNA) [26]. The mathematical model therefore assumes that sequences forming the same structure are distributed randomly using the fraction λ of neutral neighbors as (the only) input parameter. If λ is large enough the model makes three rather surprising predictions [23, 21]:

• The connectivity of networks changes drastically when λ passes the threshold value $\lambda_{cr}(\alpha) = 1 - \sqrt[\alpha-1]{1/\alpha}$ where α is the size of the alphabet. Neutral sets consist of a single component that spans sequence space if $\lambda > \lambda_{cr}$. Below threshold, $\lambda < \lambda_{cr}$, the network is partitioned into a large number of components, in general, a giant component and many small ones. (For RNA it is necessary to split the random graph into two factors corresponding to unpaired bases and base pairs, respectively, and to use a different value of λ for each factor.) In the first case we refer to $S(\psi)$ as the *neutral network* of ψ. For both RNA and proteins λ seems to be well above the threshold. Extensive neutral networks in protein space discovered using knowledge-based potentials [1] meet a claim raised by Maynard Smith [20] for protein spaces that are suitable for efficient evolution. Lattice protein model also show neutral networks [19, 2]. For RNA secondary structures their existence is demonstrated in [10].

• There is *shape space covering* for $\lambda > \lambda_{cr}$, that is, in a moderate size ball centered at an arbitrary position in sequence space there is a sequence x

TABLE 1. Shape space covering radii and closest approach distances of neutral network for RNA secondary structures.

n	r^*	r_{cov}	$r_{cov} - r^*$	r_{ca}
30	3.1	5.2	2.0	
50	6.5	9.2	2.7	2.1
70	10.0	13.7	3.7	3.4
100	15.2	20.5	5.3	5.6

that folds into a an arbitrary prescribed secondary structure ψ. The radius $r_{cov} = \min\{h \mid B(h) \geq S_n\}$ of such a sphere, called the *covering radius*, can be estimated from simple probability arguments [24] in terms of the number of structures S_n and the number of sequences $B(h)$ contained in a ball of radius h. It is much smaller than the radius n of sequence space, see [10]. The average distance r^* of a random sequence to set of compatible sequences $C(\psi)$ of the target structure ψ is clearly a lower bound on the covering radius. Best upper bounds on r_{cov} listed in table 1. The covering sphere contains all common structures and therefore forms an evolutionarily representative part of shape space.

• The neutral networks of any two common structure ψ and ψ' "touch" each other if $\lambda > \lambda_{cr}$. The intersection theorem [22] guarantees that one can always find sequences that are compatible with both ψ and ψ' [22], i.e., $C(\psi) \cap C(\psi') \neq \emptyset$. Random graph theory then predicts that the distance between the neutral networks of two arbitrary common structures is $r_{ca} = 2$. This prediction has been tested computationally for RNA secondary structures [10]. One expects $r_{ca} \approx r_{cov} - r^*$ for long chains. Note that computational studies can only provide upper bounds for r_{cov} and r_{ca}. Preliminary results (Babajide, pers. comm.) suggest that there is shape space covering also for proteins.

5. Discussion

Algorithms for folding RNA sequences into secondary structures as well as inverse folding procedures for proteins based on knowledge-based potentials predict extended connected networks of sequences with identical structure and shape space covering. These observations have striking consequences for adaptation, based on a fairly realistic model of test-tube evolution [15]: (1) Acceptable starting points for evolution can be found everywhere in sequence space. (2) Finite populations diffuse along neutral networks, where their dynamics conforms the predictions of Kimura's neutral theory [17]. After a sufficiently long period of time all sequence information is lost, yet the phenotype is conserved. It is the maintenance of a phenotype, not of a genotype, which defines the mutation threshold beyond which adaptation breaks down. (3) On a single neutral network the population splits into well separated clusters. A population is not a single localized quasi-species [6] in sequence space, but rather a collection of different, independently diffusing quasi-species that share the same phenotype. (4) Neutral networks of different structures are interwoven. While drifting on a neutral network a population produces a fraction of mutants off the network and thereby explores new phenotypes [14]. A selection-induced transition between two structures, i.e., a phase of Darwinian selection, occurs in regions of sequence space where their networks come close to one another.

Neutral evolution, therefore, is not a dispensable addendum to evolutionary theory as it has often been suggested. On the contrary, neutral networks, arising as a consequence of the redundancy of sequence-structure (and possibly also structure-function) relationships of biopolymers, provide a powerful mechanism through which evolution becomes truely efficient.

Acknowledgments. The research on the sequence structure maps of biopolymers is an on-going joint effort with Peter Schuster, Walter Fontana, and Ivo Hofacker.

References

1. A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. *Folding & Design*, 2:261–269, 1997.
2. E. G. Bornberg-Bauer. How are model protein structures distributed in sequence space? *Biophys. J.*, 73, 1997. in press.
3. J. U. Bowie, R. Luthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170, 1991.
4. G. Casari and M. J. Sippl. Structure-derived hydrophobic potentials — hydrophobic potentials derived from X-ray structures of globular proteins is able to indentify native folds. *J. Mol. Biol.*, 224:725–732, 1992.
5. C. Chothia. Proteins. One thousand families for the molecular biologist. *Nature*, 357:543–544, 1992.
6. M. Eigen, J. McCaskill, and P. Schuster. The molecular Quasispecies. *Adv. Chem. Phys.*, 75:149–263, 1989.
7. W. Fontana and P. Schuster. A computer model of evolutionary optimization. *Biophys. Chem.*, 26:123–147, 1987.
8. S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. (USA)*, 83:9373–9377, 1986.
9. W. Grüner, R. Giegerich, D. Strothmann, C. M. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. *Monath. Chem.*, 127:355–374, 1996.
10. W. Grüner, R. Giegerich, D. Strothmann, C. M. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by ex-

haustive enumeration. II. Structures of neutral networks and shape space covering. *Monath. Chem.*, 127:375–389, 1996.
11. I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chemie*, 125:167–188, 1994.
12. J. Hofbauer and K. Sigmund. *Dynamical Systems and the Theory of Evolution.* Cambridge University Press, Cambridge U.K., 1988.
13. L. Holm and C. Sander. Dali/FSSP classification of three-dimensional protein folds. *Nucl. Acids Res.*, 25:231–234, 1997.
14. M. A. Huynen. Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, 43:165–169, 1996.
15. M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. (USA)*, 93:397–401, 1996.
16. S. Kauffman. *The Origin of Order.* Oxford University Press, New York, Oxford, 1993.
17. M. Kimura. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK, 1983.
18. H. Li, R. Helling, C. Tang, and N. Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273:666–669, 1996.
19. D. J. Lipman and W. J. Wilbur. Modelling neutral and selective evolution of protein folding. *Proc. R. Soc. London B*, 245:7–11, 1991.
20. J. Maynard-Smith. Natural selection and the concept of a protein space. *Nature*, 225:563–564, 1970.
21. C. M. Reidys. Random induced subgraphs of generalized n-cubes. *Adv. Appl. Math.*, 1997. in press.
22. C. M. Reidys and P. F. Stadler. Bio-molecular shapes and algebraic structures. *Comp. & Chem.*, 20:85–94, 1996.
23. C. M. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps: Neural networks of RNA secondary structures. *Bull. Math. Biol.*, 59:339–397, 1997.
24. P. Schuster. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *J. Biotechnology*, 41:239–257, 1995.
25. P. Schuster. Landscapes and molecular evolution. *Physica D*, 107:351–365, 1997.
26. P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Roy. Soc. London B*, 255:279–284, 1994.
27. M. J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. an approach to the computational determination of protein structures. *J. Computer-Aided Molec. Design*, 7:473–501, 1993.
28. P. F. Stadler. Fitness landscapes arising from the sequence-structure maps of biopolymers. *J. Mol. Struct. (THEOCHEM)*, 1997. submitted.
29. M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. Algorithm independent properties of RNA secondary structure prediction. *Eur. Biophys. J.*, 25:115–130, 1996.
30. S. Wright. The roles of mutation, inbreeding, crossbreeeding and selection in evolution. In D. F. Jones, editor, *Int. Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366, 1932.
31. M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591–621, 1984.