On the Evolution of Primitive Genetic Codes

Günter Weberndorfer, Ivo L. Hofacker and Peter F. Stadler
Institut für Theoretische Chemie und Molekulare Strukturbiologie
Universität Wien, Währingerstrasse 17, A-1090 Wien, Austria
(`{gw,ivo,studla}@tbi.univie.ac.at`*)*

2002/06/10

Abstract. The primordial genetic code probably has been a drastically simplified ancestor of the canonical code that is used by contemporary cells. In order to understand how the present-day code came about we first need to explain how the language of the building plan can change without destroying the encoded information. In this work we introduce a minimal organism model that is based on biophysically reasonable descriptions of RNA and protein, namely secondary structure folding and knowledge based potentials. The evolution of a population of such organism under competition for a common resource is simulated explicitly at the level of individual replication events. Starting with very simple codes, and hence greatly reduced amino acid alphabets, we observe a diversification of the codes in most simulation runs. The driving force behind this effect is the possibility produce fitter proteins when the repertoire of amino acids is enlarged.

1. Introduction

The evolution of the translation machinery still presents a great challenge to any theory of the Origin of Life. As far as we know, all extant life-forms use protein enzymes and they all construct them in the same way by translating an RNA message. Invariably, translation occurs in a highly complicated RNA/protein complex, the ribosome, using tRNAs that are specifically loaded with an amino acid. All organism use the same set of twenty amino acids (22 if we count selenocystein [41, 9] and the recently discovered pyrrolysine [70]). In all cases tRNA acts as an adapter that allows the transfer of an amino acid to the growing chain if and only if the three consecutive nucleotides that form the codon on the mRNA match the tree anticodon nucleotides of the tRNA. Aminoacyl-tRNA synthesis typically is performed by 20 aminoacyl-tRNA synthetases, each one specific for a single amino acid; but see [31] for an overview of an increasing collection of exception to this simple rule.

It is not hard to argue that such a complex mechanism should have developed from a much simpler one. Unfortunately, because the translation mechanism is universal, there not too much evidence left from its earlier evolutionary stages. Even the code itself, i.e., the assignment of an amino acid to a codon is almost invariant. The most direct evidence for the evolution of the genetic code is the fact that the code is not quite "universal". The first deviations from the standard code were observed in vertebrate mitochondria, soon many more were identified among different phyla, see Fig. 1.

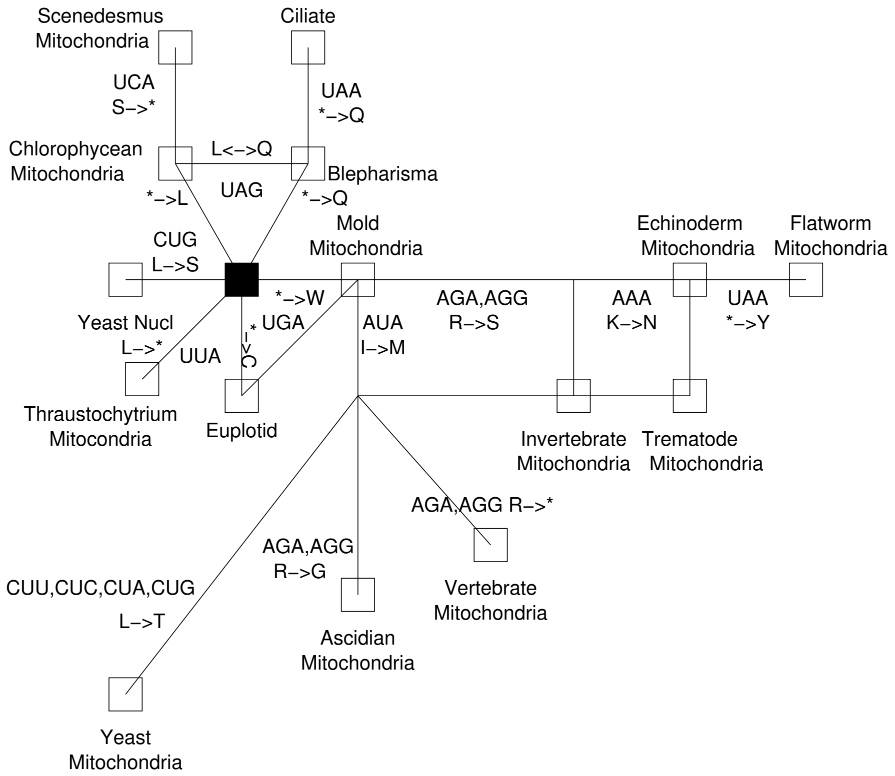


Figure 1. The genetic code shows variations among different species that can be represented as a tree-like graph. The black square marks the so-called *universal* or standard code. The definitions of the code variants were obtained from the National Center for Biotechnology Information (NCBI) website `http://www.ncbi.nlm.nih.gov/`.

All known non-standard codes, however, appear to be secondarily derived [52]. Interestingly, some changes occur independently in related linages implying multiple changes within a short period of time during evolution. Several codons seem to be more easy changeable and were assigned to different amino acids. For instance AGG has been reassigned from Arg to Ser, Gly, and STOP. In particular, STOP-codons seem to be an evolutionary degree of freedom. Their neutrality may be achieved due to their rareness (they occur once per gene) and the fact that transcriptional release factors are easy to change [53].

Another factor that may make codon reassignment evolutionary feasible is variations in codon frequencies. In fact, codon usage can vary dramatically between different species; see [16] for a recent review and [37] for a discussion the context of the genetic code.

It has been argued repeatedly by different authors that the Universal Code is optimal or near optimal in some sense. For example, Freeland *et* *al.* [21] show that the Universal Code is near optimal in terms of error minimization, adaptation for double-strand coding is discussed in [39]. In [45] a balance of robustness and changeability is advocated, the approach in [1] focuses on amino acid properties.

While the idea that the genetic code evolves towards more robust coding properties is compelling, it is by no means clear how such mutations are *accessible*. Indeed, the rewired code must be at least neutral at the level of the proteins that it produces. The selection pressures towards robustness is weak: evolution towards robustness and evolvability is a second order effect that can prevail only if the organizational changes do not cause immediate fitness losses [74, 75, 77].

Possible mechanisms of evolvability of genetic codes are reviewed in [36]: Code modifications can originate from changes in several components of the the translation apparatus, e.g.:

- Mutations of the identity elements of tRNA elements may change the specificity of aminoacylation. The tRNA may then be loaded with a different amino acid or loading may become ambiguous
- Mutation of the anticodon of the tRNA will cause the incorporation of a wrong amino acid (unless the anticodon is part of the identity elements, which is not always the case.
- Mutation of the Aminoacyl synthetase gene might lead to a change in the loading specificity.

In general, however, such changes will be deleterious because every protein that contains the modified codon will be affected.

In recent years three mechanisms of codon changes especially in mitochondria were published and each of them predicts certain codon changes that have not yet been observed.

(1) The **Codon Capture Hypothesis** [52] states that specific codons disappeared from the code by AT or GC pressure. Hence mutations in the tRNAs coding for these codons are neutral. If if the pressure is reduced the codons reappear and may now code for a different amino acid. Support for this theory comes from the mitochondrial codes, where genes are AT rich and small.

(2) **Ambiguous Intermediate Hypothesis** [83] proposes that codons undergo a period of ambiguity instead of disappearing when their meaning changes. This idea is supported by that fact that RNA in some cases mis-pairs: $\mathsf{G} \cdot \mathsf{A}$ and $\mathsf{C} \cdot \mathsf{A}$ pairs may occur at the third codon positions and $\mathsf{G} \cdot \mathsf{U}$ pairs may even occur at the first codon position. Support also comes from yeast, where a mistranslation between Ser and Leu at the CUG has been reported.

(3) The **Genome Streamlining Hypothesis** [3] assumes that the simplification of the translation apparatus is the driving force for codon reassignment in mitochondria. Reduction of the genome size has a direct selective advantage, and even the size of a single tRNA is significant for very small genomes. This is the driving force for the loss of tRNAs and hence codons.

In this contribution we describe detailed mechanistic simulations of a simplified (proto)organism that show that the genetic code can indeed evolve in the presence of strong selection on the encoded polypeptides. This approach differs from previous arguments for the adaptive nature of the code in that we need not assume a direct selection pressure on higher order properties such as evolvability. Indeed, or model is based on the reproductive success of individuals which depends only on the quality of the encoded proteins, not on the code that they use. The evolution of the encoding is therefore an emergent property in our model.

2. The Minimal Organism Model of Genetic Code Evolution

The current implementation of the Neo-Darwinian framework in the form of population genetics or quantitative genetics in essence deals with selection and is hence insufficient to describe features of phenotypic evolution such as innovation [48]. The reason is that before selection can determine the fate of a new phenotype, that phenotype must first be produced, or *accessed*, by means of variational mechanisms [17]. As far as we know, all *heritable* variations of a phenotype must occur through genetic mutation. The accessibility of a phenotype is therefore determined by the *genotype-phenotype map* which determines how phenotypes vary with genotypes [43, 78, 20, 71].

A meaningful model of evolutionary innovation, and this includes any model evolutionary model of the genetic code, must therefore make explicit assumptions on the properties of the genotype-phenotype map. In fact, the genotype-phenotype map must be modeled explicitly based on known principles of physics, chemistry, and molecular biology in order to obtain a meaningful implementation of phenotypic accessibility.

This approach was tremendously successful in the case of RNA evolution. RNA folding from sequences to secondary structures can be used as a biophysically realistic, yet extremely simplified toy-model of a genotype-phenotype map. Simulated populations of replicating and mutating sequences under selection exhibit many phenomena known from organismal evolution: neutral drift, punctuated change, plasticity, environmental and genetic canalization, and the emergence of modularity, see e.g. [18, 62, 30, 20, 2]. Laboratory experiments [69, 42, 73] have generated phenomena consistent with these patterns.

Even a minimal model for the evolution of the genetic code is necessarily much more complex. It must deal with all the key players of the translation machinery in order to provide a meaningful description of the accessibility of variant codes. In addition, it must include a biophysically reasonable fitness function.

We base our model on the assumption of an RNA World [22, 7] as a predecessor of our present DNA/RNA/Protein biology. For a recent review of the arguments for and against an RNA World Era see [82]. We emphasize, however, that we make no claim as to whether RNA was the primordial biopolymer or whether it was preceded by other, simpler molecules such as PNAs [38], that might be more plausible in terms of prebiotic synthesis [50].

The simulations presented here are motivated by a specific model organism, Fig. 2 at a (very) late stage of the RNA world, just after tRNA-based peptide synthesis has been invented and the power of protein-enzyme catalysis is utilized for replication. The main features of our hypothetical primitive cell, which we interpret as a distant ancestor of the last universal common ancestor [54, 81] are the following:

(1) *RNA genome.* It is generally believed that RNA as a molecular carrier of genomic evolution was only later replaced by by DNA genomes. A possible explanation for the advantage of DNA in larger genomes in terms of the mechanism of homologous recombination is described in [63], although it the reason may simply be the greater chemical stability of DNA.

(2) *RNA-ribosome.* Evidence form both *in vitro* studies [35, 51] and the analysis of the atomic structure [55] reveals that the ribosome is first and foremost a ribozyme. On the other hand, no isolated protein, or mixture of proteins, has ever been shown to catalyze the peptidyl-transferase reaction [25]. Furthermore, even present-day ribosomes can deal with a wide variety of amino acids, as exemplified by the incorporation of artificial amino acids by means of translation [44]. It seems reasonable, therefore, to assume that the ribosome performs its function independent of the amino acid alphabet that is used by the organism.

(3) *tRNAs* acted as crucial adaptors presumably even in the earliest versions of the translation apparatus; they are mostly likely much older than the last common ancestor [13]. Each tRNA incorporates *two* codes: the codon/anti-codon code that reads the information from the mRNA and a second *operational code* [11, 58, 61] that determines the amino acid with which the tRNA is loaded. This second code is determined by the aminoacyl-synthetases.

(4) *Ribozyme Aminoacyl synthetases.* The RNA world hypothesis implies that present-day mechanism of coded protein synthesis evolved from

ribozyme-catalyzed acyl-transfer reactions. The existence of specific aminoacyl-tRNA synthetase ribozymes has been demonstrated by means of *in vitro* evolution [40]. Furthermore, there is evidence that tRNAs predate their synthetases [56]. The present-day operational code is determined by an intricate pattern of sequence determinants that are recognized by the aminoacyl-synthetases; in the late RNA world it may have been as simple as the complementary recognition of the ribozyme designed by Lee *et al.*. There is ample evidence that amino acids may have acted as co-factors in the RNA world [59, 72]. It is plausible therefore that specific amino acid recognition and aminoacyl-transferring ribozymes have evolved long before the onset of translation.

(5) *Protein Replicase.* Ribozymes with ligase-based replication activity [46] and true replicase activity [33] were recently obtained by *in vitro* evolution, lending additional credibility to the RNA world scenario. Once replication is protein dependent all modifications of the code have an immediate impact on survival. It is therefore sufficient in our model to consider a polypeptide replicase as the only protein component.

(6) A *ribozyme based metabolism* is a convenient assumption in our setting because it need not be modeled explicitly. The wide range of chemical reactions, including carbon bond formation, that can be catalyzed by ribozymes [6, 67, 32] make this assumption even plausible.

Only a few of these components need to be modeled explicitly on the computer. We need a genomic sequence that has to be replicated, we need the tRNAs and an implementation of the operational code relating a tRNA sequence to a (set of) amino acids with which it is loaded, and we need a way of evaluating the replicase protein that is encoded on the genome. We don't have to implement the details of the replication process, the action of the ribosome, and the metabolism. This is equivalent to assuming that (i) the rate-limiting step in the "cell-cycle" of our model is the replication of the genome.

We remark that our ansatz allows an alternative interpretation as well: if we assume that replication is still RNA based and that the rate limiting step is a protein-enzyme based metabolism, we arrive at the same type of model.

3. Implementation of the Model Organism

The genome of our model organism consists of the mRNA for the replicase protein and a variable number of tRNA genes.

In order to model the structural requirements on a tRNA that are imposed by the ribosome we require that each putative tRNA must fold into

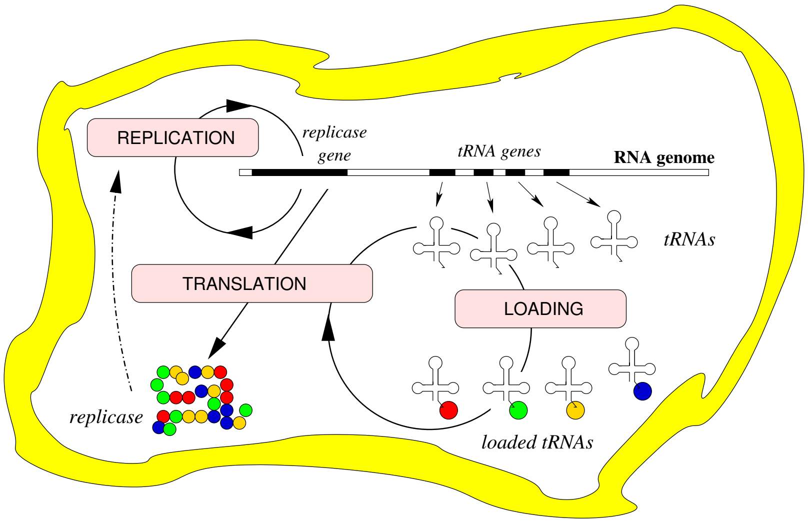


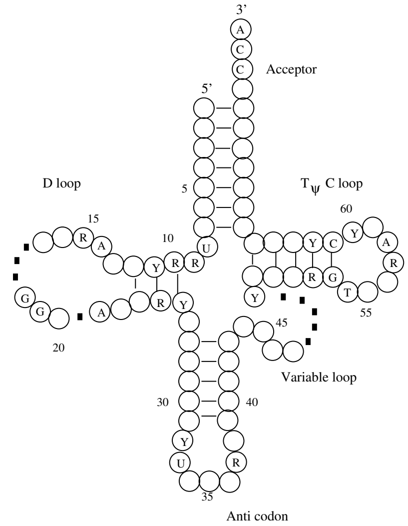
Figure 2. Model of a minimal organism with translation. It has a genome that carries genes for a protein replicase and tRNAs as well as a primitive translation apparatus and a system for loading tRNAs with amino acids. Neither the proto-ribosome nor the aminoacyl transferases are modeled in molecular detail. The protein sequence of the replicase determines rate and accuracy of replication. Translation proceeds by the usual rule of codon/anti-codon complementarity. The loading of a tRNA with a certain amino acid depends on a sequence determinants on the tRNA. The replication rate of the organism is determined by the replication rate of its genome.

the canonical cloverleaf structure that is characteristic for tRNAs, Fig. 3. RNA secondary structures can be predicted accurately and efficiently based on thermodynamic rules [85]. We use the implementation of the minimum energy folding from the `Vienna RNA Package`[1] [28]. For the purpose of our model, a functional tRNA is a sequence of length 76 whose secondary structure matches the regular expression given in Fig. 3.

There is no generally accepted model for the affinity of individual amino-acids to RNA sequences. We therefore employ a rather arbitrary table of amino acid assignments to the tRNAs that depends on the sequence of the anticodon loop and the two terminal nucleotides. The algorithm is described in the lower panel of Fig. 3.

A codon of the message is translated to the amino acid of the tRNA in the genome that has the anticodon sequence closest (in Hamming distance) to the complement of the codon. In case of equal hamming distance a match at the 1st codon position is preferred over 2nd, and 2nd over 3rd. The code may be ambiguous if two or more tRNAs match a codon equally well. In this case the assignment is done stochastically (but the assignment is then kept

[1] `http://www.tbi.univie.ac.at/RNA/`




```
# 5′ acceptor stem (5-9 pairs)
(^\({5,9}\.*
# D arm (3-5 pairs)
\({3,5}\.+\){3,5}
# variable region
\.*
# anticodon arm (3-7 pairs)
\({3,7}\.{2})(\.{3})(\.+\){3,7}
# variable region (2-7 unpaired)
\.{2,7}
# T arm (3-6 pairs)
\({3,6}\.+\){3,6}\.*
# 3′ acceptor stem
\){5,9}
# trailing bases
\.+)$
```

```
CGGGGUGGACACGCACUAGCAACGUGAUGCUUUCUACACAAGCAAUAGAACGGUCGGACCAACCGUCAUUCUGAUCA
(((((((..((((.........)))).(((((.......))))).....(((((.......))))))))))))....

CACAA => 1100110000
11001  xor  10000 = 01001 ( = 9 )
UGU => [L]
```

Figure 3. The canonical clover leave structure of a tRNA. L.h.s.: conventional drawing with the conserved nucleotides marked. The R.h.s. gives the `perl`-style regular expression that defined a tRNA for our purposes.
Given a correctly folded tRNA sequence the amino acid with which is loaded is computed by the following algorithm: (i) The determinants are the nucleotides 1, 76, and the anticodon loop. (ii) These are translated to a binary code using A=00, U=01, G=10, and C=11. (iii) The first and second five bits are combined using the "xor" operation to give a number between 0 and 31. (iv) This number is interpreted as an amino acid from the alphabet N,P, Q, A, R, S, C, T, D, E, V, F, W, G, H, Y, I, K, L, M or as a STOP signal. In this example the anti-codon is ACA, the corresponding codon is thus UGU, which is mapped to the leucine L.

fixed for the lifetime of the individual). The tRNAs that fold into the correct secondary structure together with the sequence dependent loading algorithm described in Fig.3 therefore determines the genetic code. The mRNA for the replicase is translated into its amino acid sequence according to this code.

The evaluation of the resulting protein is based on its structure. Of course we do not attempt to solve the folding problem. Instead we determine how well the amino acid sequence fits onto a target structure. We used the structure of the T7 RNA polymerase, for which an X-ray structure with a resolution of 3.3Å, PDP file `4rnpA`, is available [68], Fig. 4

Knowledge-based potential are well suited to discriminate between correctly folded and mis-folded proteins [27, 65, 66], an approach that was previously used to explore the sequence-structure map of proteins [5, 4]. For the sake of computational efficiency we do not use M. Sippl's `PROSA`-potential here. Instead we us a 4-point potential [80] that is based on Alexander Tropsha's Delauney tessellation potentials [49, 64, 84]. The idea of inverse folding [8] by means of knowledge-based potentials is to compare the energy $W(x,\psi)$ of sequence x threaded onto structure ψ with the distribution of energies obtained from threading x onto a large library of unrelated protein structures. From $W(x,\psi)$, the mean $\overline{W}(x)$ and the standard deviation $\sigma_{W(x)}$ of this distribution one computes the *z-score*

$$z(x,\psi) = \frac{W(x,\psi) - \overline{W}(x)}{\sigma_{W(x)}} \tag{1}$$

which measure how well the sequence x fits onto structure ψ. It seems natural therefore to use $z(x, \mathbf{4rnpA})$ as fitness function.

The replicase also determines the replication accuracy. Certain positions at the active site are responsible for the identification of the template base, and direct the recruitment of a nucleotide for elongation. We used the deviation of local folding energies from the values for the wild-type sequence for these 21 amino acids. For the details we refer to the PhD dissertation of the first author [79].

In summary, therefore, our model organism has a genome $\mathbf{x}$ that (via its tRNAs) defines its genetic code and (via properties of the protein resulting from this code) determines its replication rate $A_{\mathbf{x}}$ and its replication accuracy, as measured by the single-digit error rate $\mu_{\mathbf{x}}$.

4. Simulation in a Tank Reactor

The simplest experimental setup for observing a population over long periods of time is *serial transfer* [69], where at fixed time interval a tiny fraction of the population is transfered to a virgin growth medium. In chemical kinetics the *chemostat* (flow reactor) is preferred, where the population is fed a constant supply of nutrients and the total volume is kept constant. An approximate realization of an evolution reactor under constant organization is Husimi's *cellstat* [29]. From a theoretical point of view, serial transfer can be viewed as the discrete time version of the flow reactor; both lead to very similar dynamical behavior [26].

Both models are rather easily implemented on the computer. Sophisticated version are based on Gillespie's algorithm [23] that exactly simulates

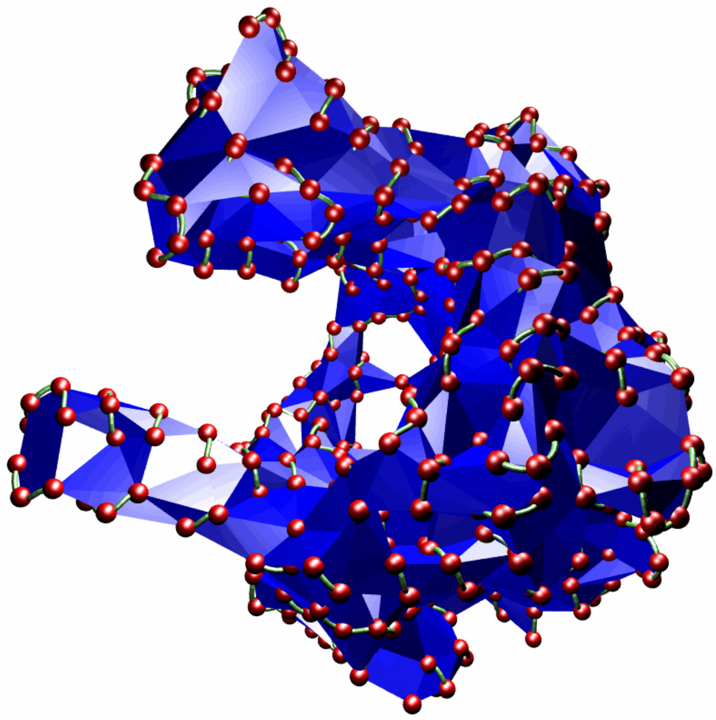


Figure 4. Delauney tessellation of the T7 RNA-polymerase structure `4rnpA`. The red balls indicate the C^{α} atoms. The energy $W(x,\mathbf{4rnpA})$ is the sum of contributions U_{ijkl} for each tetrahedron that depend on the aminoacids at corners and their relative location along the chain, and a surface term for each triangle on the surface of the molecule [80].

the stochastic reaction kinetics of mutation and fitness proportional selection [19]. In order to save computer resources we resort to a somewhat simpler approximate scheme of *tournament selection* [24] where two individuals in the population are picked at random, their fitness is compared, and the fitter one is replicated. In order to limit the population size, the child organism replaces another randomly picked individual.

This reaction scheme in essence reproduces Eigen's quasi-species model [12, 14]

$$\frac{\mathrm{d}p_{\mathbf{x}}}{\mathrm{d}t} = \sum_{\mathbf{y}} \{Q_{\mathbf{xy}}A_{\mathbf{y}}p_{\mathbf{y}} - Q_{\mathbf{yx}}A_{\mathbf{x}}p_{\mathbf{x}}\} \tag{2}$$

Here $A_{\mathbf{x}}$ is the replication rate of an organism with genome $\mathbf{x}$ and $Q_{\mathbf{xy}}$ is the mutation rate from $\mathbf{y}$ to $\mathbf{x}$. If we consider only point mutations with a mutation probability of $\mu_{\mathbf{x}}$ at each position, we get

$$Q_{\mathbf{xy}} = \left(\frac{\mu_{\mathbf{y}}}{\alpha - 1}\right)^{d(\mathbf{x},\mathbf{y})} (1-\mu_{\mathbf{y}})^{n-d(\mathbf{x},\mathbf{y})} \tag{3}$$

where $d(\mathbf{x},\mathbf{y})$ is the Hamming distance between the parent and offspring genome. Equ.(2) described replication and point mutation. In contrast to the usual quasi-species model the error rate $\mu_{\mathbf{x}}$ is an explicit function of the parental genome. Nevertheless, the model behaves dynamically just like a classical quasi-species: survival of the fittest leads to a predominant *master species* that is surrounded by a "tail" of mutants. If a mutant becomes fitter than the master, the population drifts toward this new species. The population avoids the error-threshold phenomenon by adjusting the mutation rate.

Gene duplication still is an important mechanism of genomic evolution, see e.g. [76]. Hence we include the duplication of tRNA genes as macro-mutation events. Mutation may then act on the duplicate genes and lead to diversification of the code.

We assume that a rudimentary coding system is already in place, i.e., we do not attempt to model the origin of coding itself. Thus an initial condition must be prepared consisting of a "primordial code" and a an associated gene for the replicase that leads to a non-zero replication rate.

It was shown in [5] by means of computer simulations that various small subsets of the amino acid alphabet can be used to design polypeptide sequences with native-like z-scores for known proteins. Experimental evidence is described e.g. in [10, 34, 60]. First we produce an inverse-folded protein sequence for `4rnpA` by means of adaptive walks with a restricted amino acid alphabet as described in [5], then we use the initial code to reverse-translate it into a mRNA. The tRNAs for the initial genome are produced by inverse RNA folding with prescribed nucleotides at the determinant positions using the program `RNAinverse` for the `Vienna RNA Package` [28]. The simulation is then started with the tank reactor filled with N identical copies of the "primordial organism".

5. Results

5.1. Expansion of two-amino-acid alphabets

The simplest conceivable initial alphabets distinguish only between one hydrophilic and one hydrophobic amino acid. One of these simulations is discussed in some details in Fig. 5. In some runs no new amino acid is incorporated within some 10^7 replication events. In most simulation runs, however, we find 4-7 amino acids at the end of the simulation, often with one or two additional ones that were invented and managed to spread through the population but were forgotten at later stage.

As a global indicator of evolutionary progress we consider the average fitness $\overline{F}$ of the population as a function of time. The diversity of encoded amino acids in the population is conveniently measured by the "amino acid

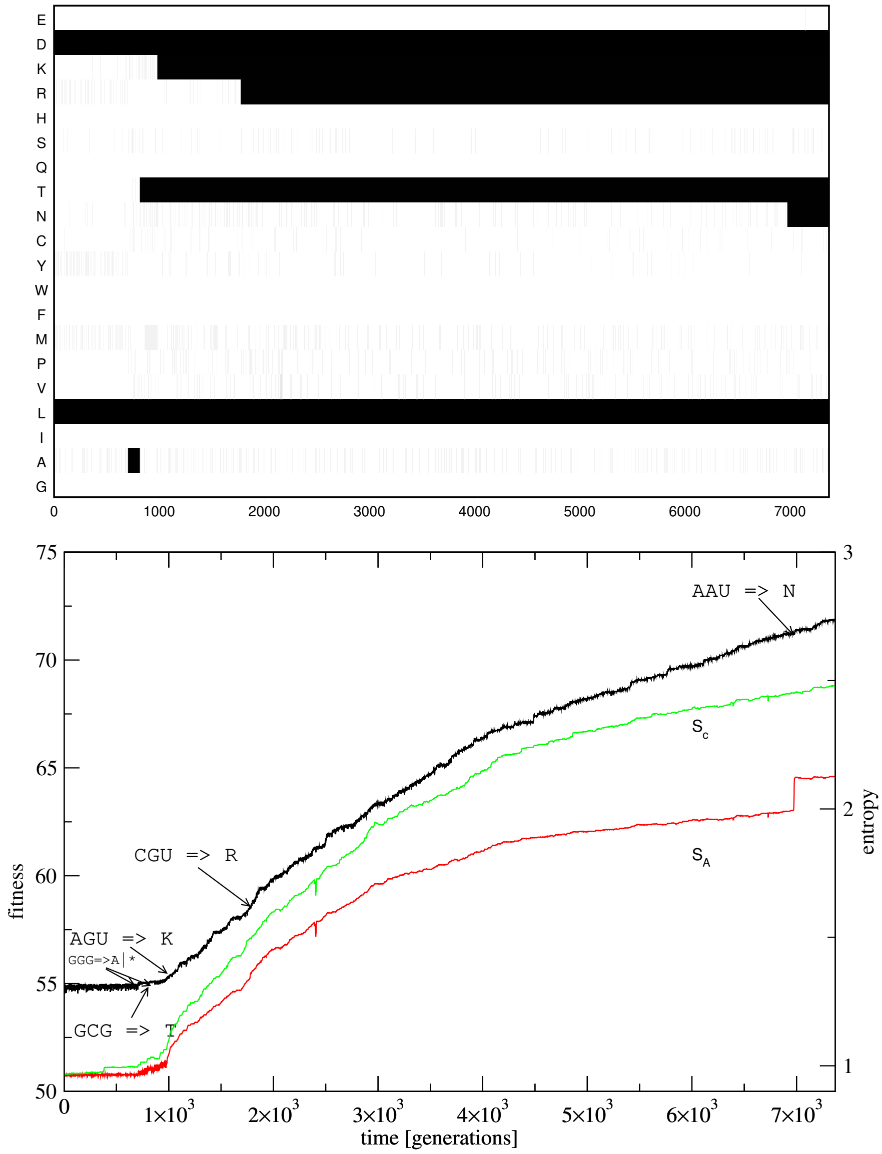


Figure 5. Extension of the LD amino acid alphabet as a function of simulation time. The upper plot shows the fraction of individuals in the population that use an amino acid (in gray scale). The lower panel displays the time evolution of (from top to bottom) the fitness, the codon usage entropy S_c, and the amino acid entropy S_A. The jump in S_A around $t = 7000$ occurs when the AAU codon is reassigned from L to N. Only 16% of the simulation run is shown, but no further innovations occured.

entropy”

$$S_A = -\sum_a f_a \log_2 f_a \tag{4}$$

where f_a is the fraction of amino acids a in an organism’s replicase. Analogously, the frequencies of codon usage can be used to compute a “codon usage entropy” S_c. Both the average fitness and the entropy measures increase with time. The increase of $\overline{F}$ is implicit in the model [12]; the increase of the entropy measures, on the other hand, describe the increase in the complexity of the evolving codes. We expect $S_A \approx S_c$ if there is only one codon in use for each amino acid. We observe, however, that $S_c > S_A$, indicating that the redundancies in the code yield to diversification in codon usage. On the other hand, the value of $S_c \approx 2.5$bit at the end of the run in Fig. 5 is much smaller than the theoretical maximum of 3×2bit for a nucleotide triplets. The slow increase in S_A and S_c shows that amino acid innovation occur via rare codons, whose usage in the genome increases as a consequence of subsequent mutations. In some cases a codon that is already commonly used for a redundantly coded amino acid is reassigned, i.e., the code is refined. Such an event can be detected form a comparison of the two entropy curves: The codon entropy S_c remains smooth while the amino acid entropy S_A sharply increases because of the novel amino acid. An example of such a refinement event can be seen in Fig. 5.

Simulations that were started with small alphabets (e.g. LD) tend in a first phase to reach “codon coverage”. By codon coverage we mean that each group of codons (ANN, UNN, GNN, and CNN) is translated unambiguously to a different amino acid. Only in a later phase further refinements of the code are observed. This is a consequence of the assignment of tRNAs to codons described in Section 3 which implies that the first codon position is more important for the matching than the second and the third.

As soon as a modification of the alphabet is fixated in the population, a further innovation becomes less likely because over the following thousands of generations fitness advantages can be drawn rather easily from spreading the usage of the novel amino acid. As the number of innovations past codon coverage is small we have not been able to extract a common pattern from the further expansion steps.

5.2. Experiments with Larger Alphabets

The amino acid alphabet AKGV, with codons of the form GNC was proposed as the primordial amino acid alphabet in [15], the alphabet ADLG is another candidate [47] for the primordial one; the restriction of inverse folding to this alphabet was studied in some detail in [5]. Computations using knowledge-based potentials suggest that this alphabet allows inverse folding of a variety of present day protein structures. A phage display experiment [57] resem-

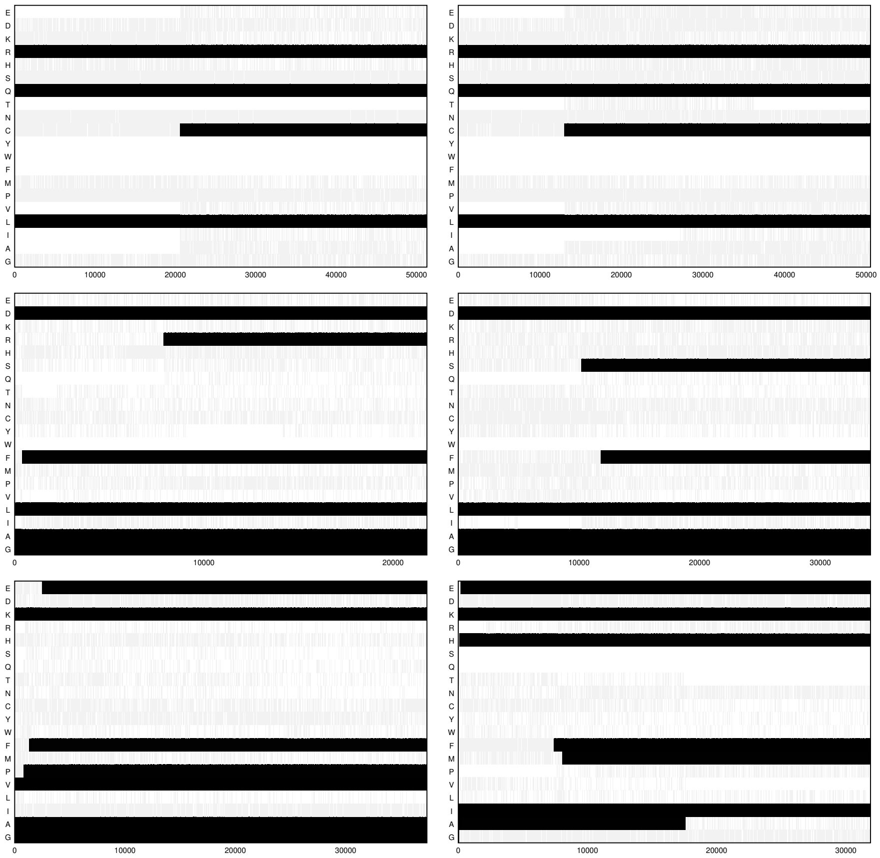


Figure 6. Coded amino acids as a function of time in six different runs that were started with three to five letter alphabets QLR (top row), ADLG (middle row), AKGV (lower left), and IKEAG (lower right).

bling the evolution of the SH3 domain (an important part of intracellular signaling) identified an alphabet consisting of two hydrophobic (I and A), two hydrophilic (K, E) and Glycine G as essentially sufficient to build the binding site.

Sauer and co-workers [10, 60] used the QLR alphabet for their work on random polypeptides. Inverse folding does not yield wild-type like z-scores for globular proteins [5]; this may not be surprising since Sauer's experimental QLR-peptides form multimeric structures. For unknown reasons it seems hard to expand the QLR alphabet in our simulation runs.

Starting from the larger alphabets yields in qualitatively the same end results as the simulations that were initiated with a two-letter alphabet: The final codes contain at most 7 coded amino-acids, Tab, I.

Table I. Summary of Simulation Runs

Run	E	D	K	R	H	S	Q	T	N	C	Y	W	F	M	P	V	L	I	A	G
ADLG_pks05		■				★							★				■		■	■
ADLG_prali		■		★									★				■		■	■
AGKV_pks04			■													■			■	■
AGKV_pks07	★		■										★		★	■			■	■
IG_pks13	★									◇			★					□		■
IKEAG_pks04	■		■															■	■	□
IKEAG_pks13	♣		■		★								★	★				■	□	□
LD_4_pks06		■	★	★				★	★								■		◇	
LD_3_pks06		■	★						★								■		◇	
LD_2_pks06		■															■			
LD_pks03		■				★											■			
QLR_pks11				■			■			★							■			
QLR_pks12				■			■			★							■			
QLR_pks00				■			■										■			

■ kept from start, ★ invented, □ lost, ◇ invented and lost again, ♣ lost and re-invented.

The model includes the possibility that the evolving organism fine-tune the mutation rate. We observe that the mutation rate decreases with with time so that the invention of additional amino acids become more and more unlikely. This can be understood by the fact that a reduction in mutation rate increases the population fitness by reducing the number of detrimental offspring. This self-adaptation of the mutation rate will require a more detailed investigation.

6. Concluding Remarks

We have described a mechanistic model of the evolution of simple genetic codes. Our simulations show that the increase in fitness that can be achieved with more diverse amino acid repertoires is sufficient to cause an increase of the alphabet size from two to about six or seven. The small size of the protein-coding part of our model genome (a single gene with only a few hundred amino acids, Fig. 4) implies that a moderate diversity of the amino acid alphabet is sufficient to produce very good sequences. We suspect that the inclusion of additional proteins in the fitness function will increase the potential fitness effects of further amino acid innovations.

In the computational setting presented in this contribution, at least, we were able to show that the genetic code can evolve. Our simulations tend to lead to codes that span the full range of polarities. We view this as an indication that the knowledge-based potentials underlying the evaluation of the protein's fitness are at least qualitatively reasonable.

In principle, simulations of the type presented here allow to test hypotheses on the origin of the genetic code, such as whether a particular property is evolved or incidental. However, even for the minimal organism presented here, the simulations require considerable computational effort. The data that we have accumulated so far are, for example, insufficient to test hypotheses about the optimality of the present-day code(s).

Further simulation with varied initial conditions may yield a realistic scenario for the expansion of the amino acid alphabet. Other questions will require extensions of the model. One might argue, for example, that the present-day code is optimized to allow rapid adaptation of proteins. But in order to optimize the code for "evolvability" our model would have to incorporate a time-dependent environment.

It will be interesting to see if extensions of the present models towards a more sophisticated protein machinery will indeed lead to a full set amino acids.

References

1. Aita, T., S. Urata, and H. Yuzuru: 2000, 'From amino acid landscape to protein landscape: analysis of genetic codes in terms of fitness landscape'. *J. Mol. Evol.* pp. 313–323.
2. Ancel, L. and W. Fontana: 2000, 'Plasticity, Evolvability and Modularity in RNA'. *J. of Exp. Zoology (Molecular and Developmental Evolution)* **288**, 242–283.
3. Andersson, S. G. and C. G. Kurland: 1995, 'Genomic evolution drives the evolution of the translation system'. *Biochem. Cell Biol.* **73**, 775–787.
4. Babajide, A., R. Farber, I. L. Hofacker, J. Inman, A. S. Lapedes, , and P. F. Stadler: 2001, 'Exploring Protein Sequence Space Using Knowledge Based Potentials'. *J. Theor. Biol.* **212**, 35–46.
5. Babajide, A., I. L. Hofacker, M. J. Sippl, and P. F. Stadler: 1997, 'Neutral Networks in Protein Space: A Computational Study Based on Knowledge-Based Potentials of Mean Force'. *Folding & Design* **2**, 261–269.
6. Bartel, D. P. and P. J. Unrau: 1999, 'Constructing an RNA world'. *Trends Biochem. Sci.* **24**, M9–M13.
7. Benner, S. A., A. D. Ellington, and A. Tauer: 1989, 'Modern Metabolism as a palimpsest of the RNA world'. *Proc. Natl. Acad. Sci. USA* **86**, 7054–7058.
8. Bowie, J. U., R. Luthy, and D. Eisenberg: 1991, 'A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure'. *Science* **253**, 164–170.
9. Commans, S. and A. Böck: 1999, 'Selenocysteine inserting tRNAs: an overview'. *FEMS Microbiology Reviews* **23**, 335–351.
10. Davidson, A. R. and R. T. Sauer: 1994, 'Folded proteins occur frequently in libraries of random amino acid sequences'. *Proc. Natl. Acad. Sci. (USA)* **91**, 2146–2150.
11. de Duve, C.: 1988, 'Transfer RNAs: the second genetic code'. *Nature* **333**, 117–118.
12. Eigen, M.: 1971, 'Selforganization of Matter and the Evolution of Macromolecules'. *Naturwiss.* **58**, 465–523.
13. Eigen, M., B. F. Lindemann, M. Tietze, R. Winkler-Oswatitsch, A. W. M. Dress, and A. von Haeseler: 1989a, 'How old is the genetic code? Statistical geometry of tRNA provides an answer'. *Science* **244**, 673–679.
14. Eigen, M., J. S. McCaskill, and P. Schuster: 1989b, 'The Molecular Quasi-Species'. *Adv. Chem. Phys.* **75**, 149–263.
15. Eigen, M. and P. Schuster: 1979, *The Hypercycle*. New York, Berlin: Springer-Verlag.

16. Ermolaeva, M. D.: 2001, ‘Synonymous Codon Usage in Bacteria’. *Curr. Issues Mol. Biol.* **3**, 91–97.
17. Fontana, W. and L. W. Buss: 1994, ‘”The Arrival of the Fittest”:Towards a Theory of Biological Organisation’. *Bull. Math. Biol.* **56**(1), 1–64.
18. Fontana, W., W. Schnabl, and P. Schuster: 1989, ‘Physical aspects of evolutionary optimization and adaption’. *Phys. Rev. A* **40**, 3301–3321.
19. Fontana, W. and P. Schuster: 1987, ‘A computer model of evolutionary optimization’. *Biophysical Chemistry* **26**, 123–147.
20. Fontana, W. and P. Schuster: 1998, ‘Continuity in Evolution: On the Nature of Transitions’. *Science* **280**, 1451–1455.
21. Freeland, S. J., R. D. Knight, L. F. Landweber, and L. D. Hurst: 2000, ‘Early Fixation of an Optimal Genetic Code’. *Mol. Biol. Evol.* **17**, 511–518.
22. Gilbert, W.: 1986, ‘The RNA World’. *Nature* **319**, 618.
23. Gillespie, D. T.: 1976, ‘A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions’. *J. Comput. Phys.* **22**, 403.
24. Goldberg, D. E. and K. Deb: 1991, ‘A Comparative Analysis of Selection Schemes Used in Genetic Algorithms’. In: G. J. E. Rawlins (ed.): *Foundations of Genetic Algorithms*. San Mateo, CA, pp. 69–93.
25. Hampl, H., H. Schulze, and K. H. Nierhaus: 1981, ‘Ribosomal components from Escherichia coli 50S subunits involved in the reconstitution of peptidyltransferase activity.’. *J. Biol. Chem.* **256**, 2284–2288.
26. Happel, R. and P. F. Stadler: 1999, ‘Autocatalytic Replication in a CSTR and Constant Organization’. *J. Math. Biol.* **38**, 422–434.
27. Hendlich, M., P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl: 1990, ‘Identification of Native Protein Folds Amongst a Large Number of Incorrect Models — The Calculation of Low Energy Conformations from Potentials of Mean Force’. *J. Mol. Biol.* **216**, 167–180.
28. Hofacker, I. L., W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster: 1994, ‘Fast Folding and Comparison of RNA Secondary Structures’. *Monatsh. Chem.* **125**, 167–188.
29. Husimi, Y.: 1989, ‘Selection and Evolution in Cellstat’. *Adv. Biophys.* **25**, 1–43.
30. Huynen, M. A., P. F.Stadler, and W. Fontana: 1996, ‘Smoothness within Ruggedness: The role of Neutrality in Adaptation’. *Proc. Natl. Acad. Sci. USA* **93**, 397–401.
31. Ibba, M. and D. Söll: 2001, ‘The renaissance of aminoacyl-tRNA synthesis’. *EMBO reports* **2**, 382–387.
32. Jäschke, A.: 2001, ‘RNA-catalyzed carbon-carbon bond formation’. *Biol. Chem.* **382**, 1321–1325.
33. Johnston, W. K., P. J. Unrau, M. J. Lawrence, M. E. Glasner, and D. P. Bartel: 2001, ‘RNA-Catalyzed RNA Polymerization: Accurate and General RNA-Templated Primer Extension’. *Science* **292**, 1319–1325.
34. Kamtekar, S., J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht: 1993, ‘Protein design by binary patterning of polar and nonpolar amino acids’. *Science* **262**, 1680–1685.
35. Khaitovich, P., A. S. Mankin, R. Green, L. Lancaster, and H. F. Noller: 1999, ‘Characterization of functionally active subribosomal particles from *Thermus aquaticus*’. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 85–90.
36. Knight, R. D., S. J. Freeland, and L. F. Landweber: 2001a, ‘Rewiring the keyboard: evolvability of the genetic code.’. *Nat. Rev. Genet* **2**, 49–58.
37. Knight, R. D., S. J. Freeland, and L. F. Landweber: 2001b, ‘A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes.’. *Genome Biology* **2**, 1–13.

38. Knight, R. D. and L. F. Landweber: 2000, 'The Early Evolution of the Genetic Code'. *Cell* **101**, 569–572.
39. Konecny, J., M. Eckert, M. Schöniger, and H. G. Ludwig: 1993, 'Neutral adaptation of the genetic code to double-strand coding'. *J. Mol. Evol.* **36**, 407–416.
40. Lee, N., Y. Bessho, K. Wei, J. W. Szostak, and H. Suga: 2000, 'Ribozyme-catalyzed tRNA aminoacylation'. *Nat. Struct. Biol* **7**, 28–33.
41. Leinfelder, W., E. Zehelein, M. A. Mandrand-Berthelot, and A. Bock: 1988, 'Gene for a novel tRNA species that accepts L-serine and cotranslationally inserts selenocysteine.'. *Nature* **331**, 723–725.
42. Lenski, R. E. and M. Travisano: 1994, 'Dynamics of Adaptation and Diversification: A 10,000-generation Experiment with Bacterial Populations'. *Proc. Natl. Acad. Sci. USA* **91**, 6808–6814.
43. Lewontin, R. C.: 1974, *The Genetic Basis of Evolutionary Change*. New York, New York: Columbia University Press.
44. Liu, D. R. and P. G. Schultz: 1999, 'Progress toward the evolution of an organism with an expanded genetic code'. *Proc. Natl. Acad. Sci. USA* **96**, 4780–4785.
45. Maeshiro, T. and M. Kimura: 1998, 'The role of robustness and changeability on the orgin and evolution of genetic codes'. *Proc. Natl. Acad. Sci. USA* **95**, 5088–5093.
46. McGinness, K. E. and G. F. Joyce: 2002, 'RNA-Catalyzed RNA Ligation on an External RNA Template'. *Chem. Biol.* **9**, 297–307.
47. Miller, S. L. and L. E. Orgel: 1974, *The Origin of Life on the Earth*. Prentice Hall.
48. Müller, G. B. and G. P. Wagner: 1991, 'Novelty in Evolution: Restructuring the Concept'. *Annu. Rev. Ecol. Syst.* **22**, 229–256.
49. Munson, P. J. and R. K. Singh: 1997, 'Statistical significance of hierarchical multi-body potentials based on Delauney tessellation and their application in sequence-structure alignment'. *Protein Sci.* **6**, 1467–1481.
50. Nelson, K. E., M. Levy, and S. L. Miller: 2000, 'Peptide nucleic acids rather than RNA may have been the first genetic molecule'. *Proc. Natl. Acad. Sci. USA.* **97**, 3868–3871.
51. Nitta, I., Y. Kamada, H. Noda, T. Ueda, and K. Watanabe: 1998, 'Reconstitution of Peptide Bond Formation with *Escherichia coli* 23S Ribosomal RNA Domains'. *Science* **281**, 666–669.
52. Osawa, S.: 1995, *Evolution of the genetic code*. Oxford: Oxford University Press.
53. Osawa, S., T. H. Jukes, K. Watanabe, and A. Muto: 1992, 'Recent evidence for evolution of the genetic code'. *Microbiol. Rev.* **56**, 229–264.
54. Penny, D. and A. Poole: 1999, 'The nature of the last common ancestor'. *Curr. Opin. Genet. Devel.* **9**, 672–699.
55. Ramakrishnan, V. and P. B. Moore: 2001, 'Atomic structures at last: the ribosome in 2000'. *Curr. Opinions Struct. Biol.* **11**, 144–154.
56. Ribas de Pouplana, L., R. J. Turner, B. A. Steer, and P. Schimmel: 1998, 'Genetic Code Origins: tRNAs older than their synthethases?'. *Proc. Natl. Acad. Sci. USA* **95**, 11295–11300.
57. Riddle, D. S., J. V. Santiago, S. T. Bray-Hall, N. Doshi, V. P. Grantcharova, Q. Yi, and D. Baker: 1997, 'Functional rapidly folding proteins from simplified amino acid sequences.'. *Nat. Struct. Biol* **10**, 805–809.
58. Rodin, S. N. and S. Ohno: 1997, 'Four primordial model of tRNA-synthase recognition, determined by the (G,C) operational code'. *Proc. Natl. Acad. Sci. USA* **94**, 5183–5188.
59. Roth, A. and R. R. B. Breaker: 1998, 'An amino acid as a cofactor for a catalytic polynucleotide'. *Proc. Natl. Acad. Sci. USA* **95**, 6027–6031.
60. Sauer, R. T.: 1996, 'Protein folding from a combinatorial perspective'. *Folding & Design* **1**, R27–R29.

61. Schimmel, P., R. Giegé, D. Moras, and S. Yokoyama: 1993, 'An operational RNA code for amino acids and possible relationship to genetic code'. *Proc. Natl. Acad. Sci. USA* **90**, 8763–8768.
62. Schuster, P., W. Fontana, P. F. Stadler, and I. L. Hofacker: 1994, 'From Sequences to Shapes and Back: A case study in RNA secondary structures'. *Proc. Roy. Soc. Lond. B* **255**, 279–284.
63. Shibata, T., T. Nishinaka, T. Mikawa, H. Aihara, H. Kurumizaka, S. Yokoyama, and Y. Ito: 2001, 'Homologous genetic recombination as an intrinsic dynamic property of a DNA structure induced by RecA/Rad51-family proteins: A possible advantage of DNA over RNA as genomic material'. *Proc. Natl. Acad. Sci. USA* **98**, 8425–8432.
64. Singh, R. K., A. Tropsha, and I. I. Vaisman: 1996, 'Delauney Tessellation of Proteins: Four Body Nearest Neighbor Propensity of Amino Acid Residues'. *J. Comp. Biol.* **3**, 213–221.
65. Sippl, M. J.: 1990, 'Calculation of Conformational Ensembles from Potentials of Mean Force — An Approach to the Knowledge-based Prediction of Local Structures in Globular Proteins'. *J. Mol. Biol.* **213**, 859–883.
66. Sippl, M. J.: 1993, 'Recognition of Errors in Three-Dimensional Structures of Proteins'. *Proteins* **17**, 355–362.
67. Soukup, G. A. and R. R. Breaker: 2000, 'Allosteric nucleic acid catalysis'. *Curr. Opin. Struct. Biol.* **10**, 318–325.
68. Sousa, R., Y. J. Chung, J. P. Rose, and B. C. Wang: 1993, 'Crystal structure of bacteriophage T7 RNA polymerase at 3.3Å resolution'. *Nature* **364**, 593–599.
69. Spiegelman, S.: 1971, 'An Approach to Experimental Analysis of Precellular Evolution'. *Quart. Rev. Biophys.* **4**, 213–253.
70. Srinivasan, G., C. M. James, and J. A. Kryzcki: 2002, 'Pyrrolysine Encoded by UAG in Archea: Charging of a UAG-decoding specialized tRNA'. *Science* **296**, 1459–1462.
71. Stadler, B. M. R., P. F. Stadler, G. Wagner, and W. Fontana: 2001, 'The topology of the possible: Formal spaces underlying patterns of evolutionary change'. *J. Theor. Biol.* **213**, 241–274.
72. Szathmáry, E.: 1999, 'The origin of the genetic code: amino acids as cofactors in the RNA world'. *Trends Genet.* **15**, 223–229.
73. Szostak, J. W. and A. D. Ellington: 1993, '*In Vitro* Selection of Functional RNA Sequences'. In: R. F. Gesteland and J. F. Atkins (eds.): *The RNA World*. Plainview, NY: Cold Spring Harbor Laboratory Press, pp. 511–533.
74. Wagner, A.: 1996, 'Does evolutionary plasticity evolve?'. *Evolution* **50**, 1008–1023.
75. Wagner, A.: 1999, 'Redundant gene functions and natural selection'. *J. Evol. Biol.* **12**, 1–16.
76. Wagner, A.: 2002, 'Selection and gene duplication: a view from the genome'. *Genome Biology* **3**, 1012.1–1012.3.
77. Wagner, A. and P. F. Stadler: 1999, 'Viral RNA and Evolved Mutational Robustness'. *J. Exp. Zool./ MDE* **285**, 119–127. Santa Fe Institute preprint 99-02-010.
78. Wagner, G. P. and L. Altenberg: 1996, 'Complex adaptations and the evolution of evolvability'. *Evolution* **50**, 967–976.
79. Weberndorfer, G.: 2002, 'Computational Models ofthe Genetic Code Evolution Based on Empirical Potentials'. Ph.D. thesis, Univ. of Vienna.
80. Weberndorfer, G., I. L. Hofacker, and P. F. Stadler: 1999, 'An Efficient Potential for Protein Sequence Design'. In: *Computer Science in Biology*. Bielefeld, D, pp. 107–112. Proceedings of the GCB'99, Hannover, D.
81. Woese, C.: 1998, 'The universal ancestor'. *Proc. Natl. Acad. Sci. USA* **95**, 6854–6859.
82. Yarus, M.: 1999, 'Boundaries for an RNA World'. *Curr. Opinions Chem. Biol.* **3**, 260–267.

83. Yarus, M. and D. Schultz: 1997, 'Toward a theory of malleability in genetic coding.'. *J. Mol. Evol.* **45**, 3–6.
84. Zheng, W., S. J. Cho, I. I. Vaisman, and A. Tropsha: 1996, 'Statistical geometry analysis of proteins: implications for inverted structure prediction'. In: L. Hunter and T. Klein (eds.): *Biocomputing: Proceedings of the 1996 Pacific Symposium*. pp. 614–623.
85. Zuker, M.: 2000, 'Calculating nucleic acid secondary structure'. *Curr. Opin. Struct. Biol.* **10**, 303–310.

Address for Offprints:
P F Stadler
Inst. f. Theoretical Chemistry and Structural Biology
University of Vienna
Währingerstrasse 17
A-1090 Vienna, Austria
Phone: +43 1 4277 52737, Fax: +43 1 4277 52793,