A Computer-Glimpse of the Origin of Life

CHRISTOPH KUHN

Institute für Molekularbiologie und Biophysik, Gruppe Biophysik, ETH–Hönggerberg,
CH 8093 Zürich, Switzerland
(e-mail: c-k@gmx.ch)

Abstract. Evolution is assumed to begin in a very particular compartmentalized location with periodic conditions. A highly diversified world is the driving force for the continuous increase in complexity by colonizing increasingly less favourable regions. Modeling the "origin-of-life" a Darwinian cyclic process is simulated (multiplication with sporadic errors followed by a construction and selection).

Starting from a RNA-world (R-strands of R_1 and R_2 monomers building Hairpin-Assembler devices) and introducing another kind of monomers (A_1 and A_2 which interlink to the Hairpin-Assembler devices such that they become bound and form an A-oligomer) it is shown that a simple translation apparatus evolves producing enzymes (specific sequences of A_1 and A_2 monomers given by the sequences of R_1 and R_2 monomers on the assembler-strands). Later on D-strands are introduced, which are not capable of participating in the synthesis of A-oligomers. These D-strands become carriers of the genetic information and induce the formation of increasingly complex entities of functionally interplaying components.

Key words: origin of life, evolution, genetic apparatus, genetic code, emergence and storage of information, self organisation, fundamental conditions, assembler, hairpin, RNA world, assembler-hairpin-enzyme device, RNA-protein world, DNA-RNA-protein world, computer simulation

1. Introduction

What motivates life's origin ? Let us assume (Figure 1) that prebiotic chemistry was capable of forming the building blocks of life. How can the genetic apparatus then have developed from this starting point? The explosion of all the forms of life follows from the existence of a genetic apparatus.

To investigate the principles [1–4] we search for a gapless sequence of physico-chemical steps (Figure 2) that are as simple as possible and detailed. But we do not attempt to find historical reality. Thus starting with artificial monomers which build up replica-forming oligomers in a simple artificial environment we try to find such a path to a simple artificial genetic apparatus.

Darwin's main idea on the origin of species was that an entity which is selected for in a given ecological niche does not survive in a neighbouring, less favorable region. A mutant may occasionally form, which is selected for in the less favorable region. In the present view this mechanism is also the basis to understand the very first steps in the origin of life. This is in contrast to the view that life originates by self organisation in a uniform environment.
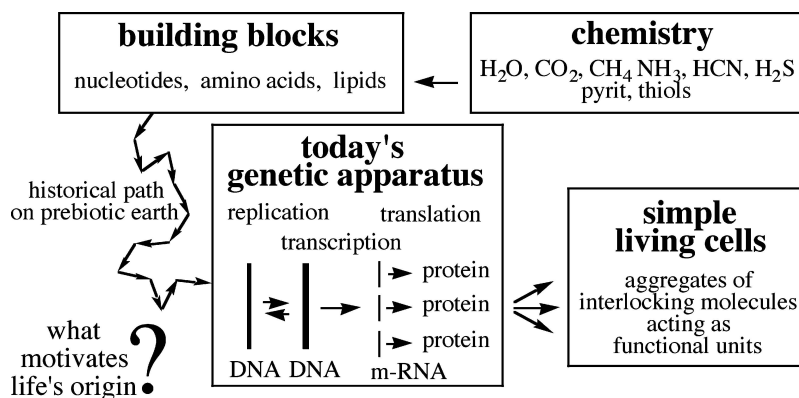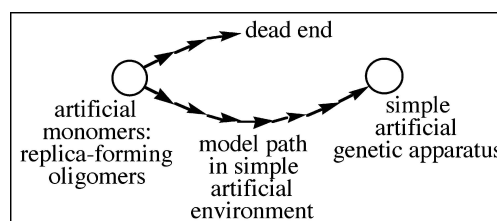
Figure 1. What motivates life's origin?.

Figure 2. A model path of a gapless sequence of physico-chemical steps.

2. Ideas on Which the Computer Simulation is Based

We assume distinct conditions to be given that should drive an evolution in a distinct direction as sketched in 2.2 and 2.3. Does this process actually take place in a computer simulation, can the conceptions be substantiated? The result of such a computer simulation is given in 3 for the critical steps leading to a simple genetic device (HAE-device). A rough draft is given in 4.1 and 4.2 of the next hypothetical steps toward the modern genetic apparatus not yet treated with the computer.

2.1. FUNDAMENTAL CONDITIONS

In the view of the present author and others [1–4] the origin of life requires the following fundamental conditions to allow the evolution of aggregates of ever increasing complexity and functional intricacy (Figure 3):
– spontaneously formed short strand and energy-rich monomers are present in every particular environment that allows formation of a replica of this strand acting as a template. Errors in the copying process can again lead to forms that replicate.
– periodic course of environmental conditions such as temperature (day and night), where at lower temperature aggregation of folding structures and selection of
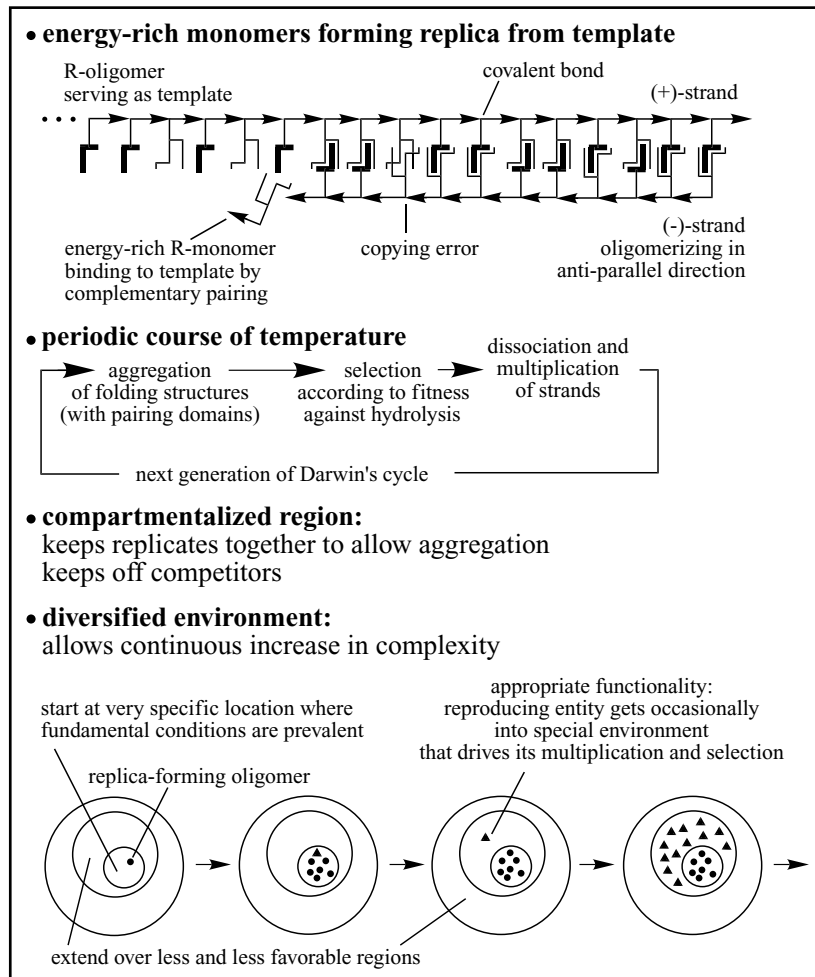
• energy-rich monomers forming replica from template

R-oligomer
serving as template
covalent bond
(+)-strand

energy-rich R-monomer
binding to template by
complementary pairing
copying error
(-)-strand
oligomerizing in
anti-parallel direction

• periodic course of temperature

aggregation ⟶ selection ⟶ dissociation and
of folding structures   according to fitness   multiplication
(with pairing domains)   against hydrolysis   of strands

next generation of Darwin's cycle

• compartmentalized region:
keeps replicates together to allow aggregation
keeps off competitors

• diversified environment:
allows continuous increase in complexity

start at very specific location where
fundamental conditions are prevalent
appropriate functionality:
reproducing entity gets occasionally
into special environment
that drives its multiplication and selection

replica-forming oligomer

extend over less and less favorable regions

Figure 3. The present view: Origin of life requires distinct fundamental conditions to allow
the evolution of aggregates of increasing complexity.

the entities occurs and at higher temperature dissociation and multiplication of
strands. In this way a cyclical Darwinian process is generated.
– a compartmentalized region (e.g. porous rock) which keeps replicates together
to allow aggregation and which keeps off competition.
– a diversified environment which allows continuous increase in complexity.

The increasing independence of these highly specific initial conditions by gradually
populating a diversified area can be realized as follows:
– populating regions of compartments with increasing size of pores,
– evolving devices that produce envelopes as self made compartments,
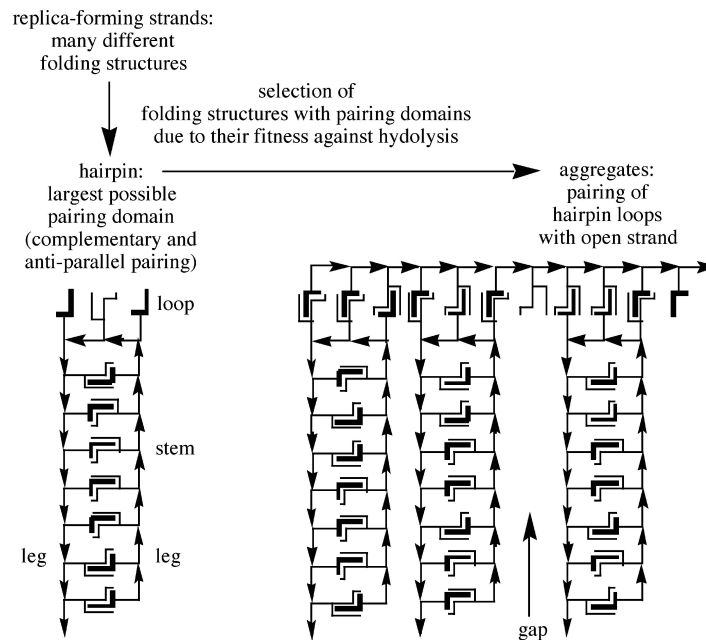– evolving metabolism (increasing intricacy of the living machinery).

Figure 4. Hairpin: largest possible pairing domain. The formation of hairpins requires a replication where template strand and replica are complementary and anti-parallel.

2.2. THE RNA WORLD

We consider the case where only one kind of complementary R-monomers, R_1 and R_2, are present and we assume that oligomers appear that are capable of replicating (Figure 3, above). The initial sequence is random and by the errors in the copying process there is variation of this sequence. If by chance a strand is found in the population that has a sequence such that it folds onto itself to form a hairpin (Figure 4) the hairpin will be selected for because it is more protected against hydrolysis and has higher probability of being selected.

The loop of hairpins is random, thus all eight possibilities of sequences are equal probable in the population and the hairpin finds an open strand to bind by the loop with complementary and anti-parallel means. If hairpins are adjacent the entity will be selected for because it is more protected against hydrolysis and has higher probability of being selected. A reading frame only exists if all hairpins in the population have loops of one kind (Figure 5, upper framed pair of loops) and all the hairpins bound to the assembler with the corresponding sequence are then adjacent (Figure 6).

2.3. THE RNA-PROTEIN WORLD

In our model A-monomers A_1 and A_2 are present and by an error in the copying process the two R-monomers at the lower end of a hairpin are not complementary,
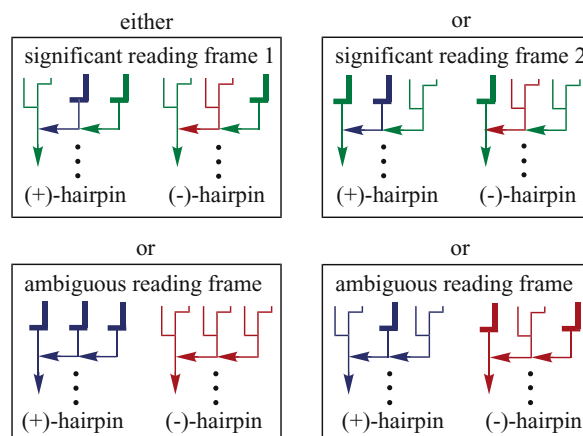
Figure 5. Loop sequences: initially the eight possibilities are equally probable in the population. A reading frame only exists if all hairpins in the population have loops of one kind (upper framed pair of loops).
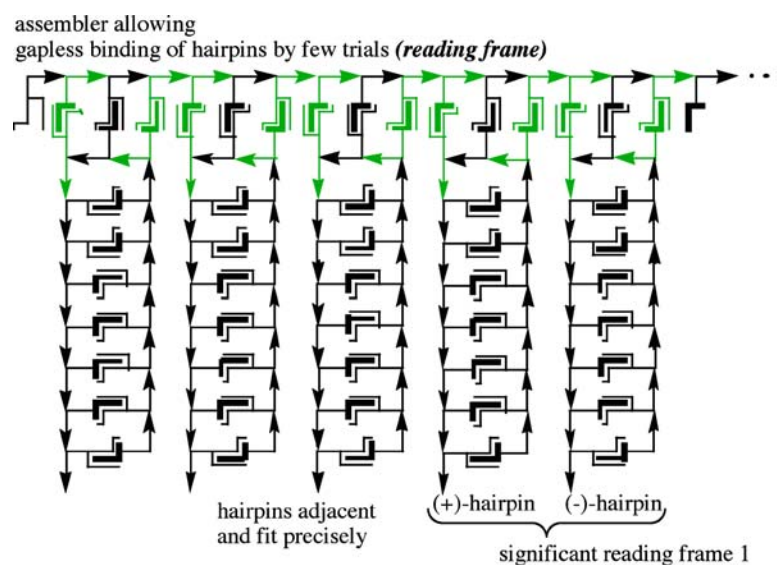

Figure 6. HA-device: Picket-fence-like aggregate. All hairpins in the population have loops of one kind and all the hairpins on the assembler are adjacent (corresponding sequence on the assembler).

then the hairpin carries an A-monomer such that R_1-R_1 carries A_1 and R_2-R_2 carries A_2 (Figure 7). If all hairpins on an assembler carry A-monomers these can polymerize (Figure 7), and the product, an A-oligomer E_0 with a random sequence has no enzymatic power, but functions as an agglutinate-forming envelope (Figure 10)
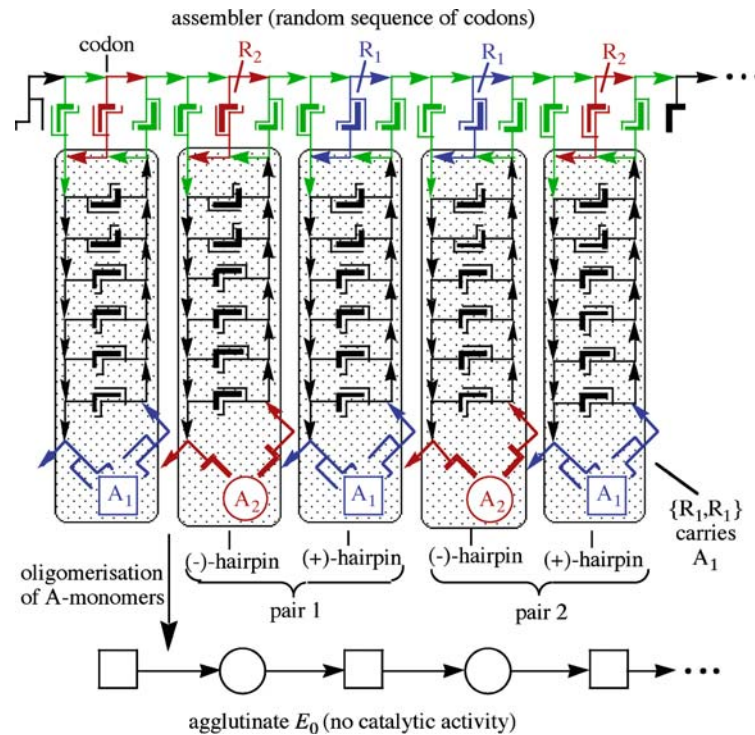
Figure 7. The HAE$_0$-device produces agglutinate forming envelope (A-oligomer with random sequence). Note that at the lower legs of the hairpins the two R-monomers are not complementary, but R$_1$-R$_1$ carries A$_1$ and R$_2$-R$_2$ carries A$_2$. The A-monomers polymerize as indicated.
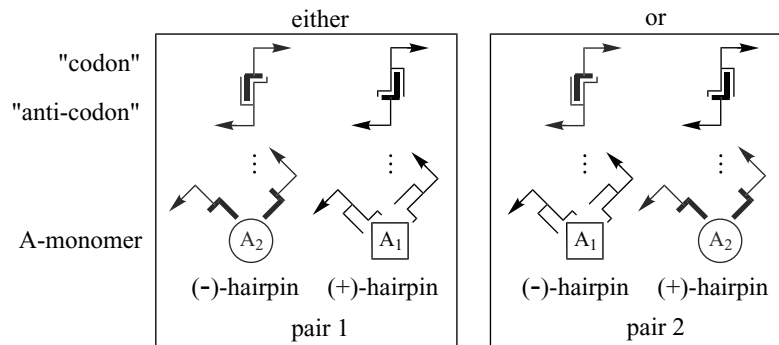

Figure 8. All four possibilities between the anti-codon and the A-monomer carried by the hairpin are equally probable in the population. A translation would exist if all hairpins in the population belong to one pair (framed).

with higher probability of being elected. The sequence for the middle R-monomer in the hairpin-loops and assembler-triplets respectively is random, and there is no correspondence between the anti-codon and the A-monomer carried by the hairpin (Figure 8), thus all four possibilities are equally probable in the population and
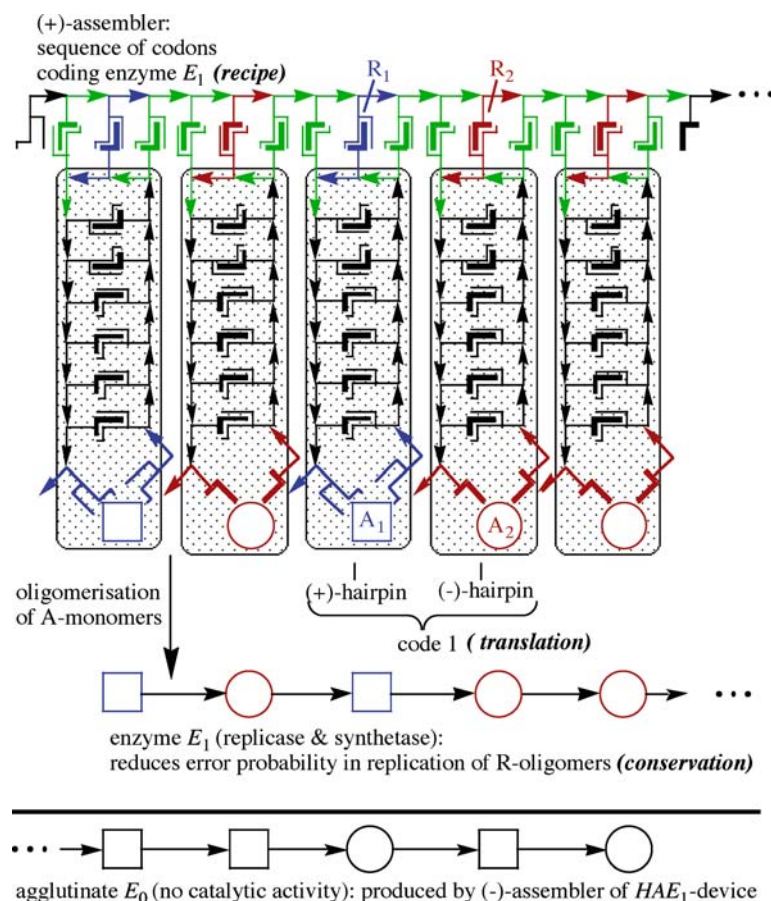
Figure 9. The HAE$_1$-device. Its action as a translation apparatus is a by-product. An enzyme (an A-polymer with a specific sequence that has the function of reducing the error probability in the replication of R-oligomers) is only selected for if there is the recipe on the (+)-assembler for it and if there is a translation by the hairpins. The (−)-assembler is the anti-parallel complementary copy of an (+)-assembler and its product is non sensical (HAE$_0$-device).

there is no translation. A translation would exist if all hairpins in the population belong to one pair (Figure 8, framed).

The HAE$_1$-device is a translation apparatus as by-product (Figure 9). An enzyme (an A-polymer with a specific sequence that has the function of reducing the error probability in the replication of R-oligomers) is only selected for if there is the recipe on the (+)-assembler for it and if there is a translation by the hairpins. The (−)-assembler is the anti-parallel complementary copy of an (+)-assembler and its product is non sensical (HAE$_0$-device). Other enzymes E_2, E_3, ... evolve (their sequences are conserved due to presence of E_1) and cooperative entities with increased functional complexity come into being (Figure 10).
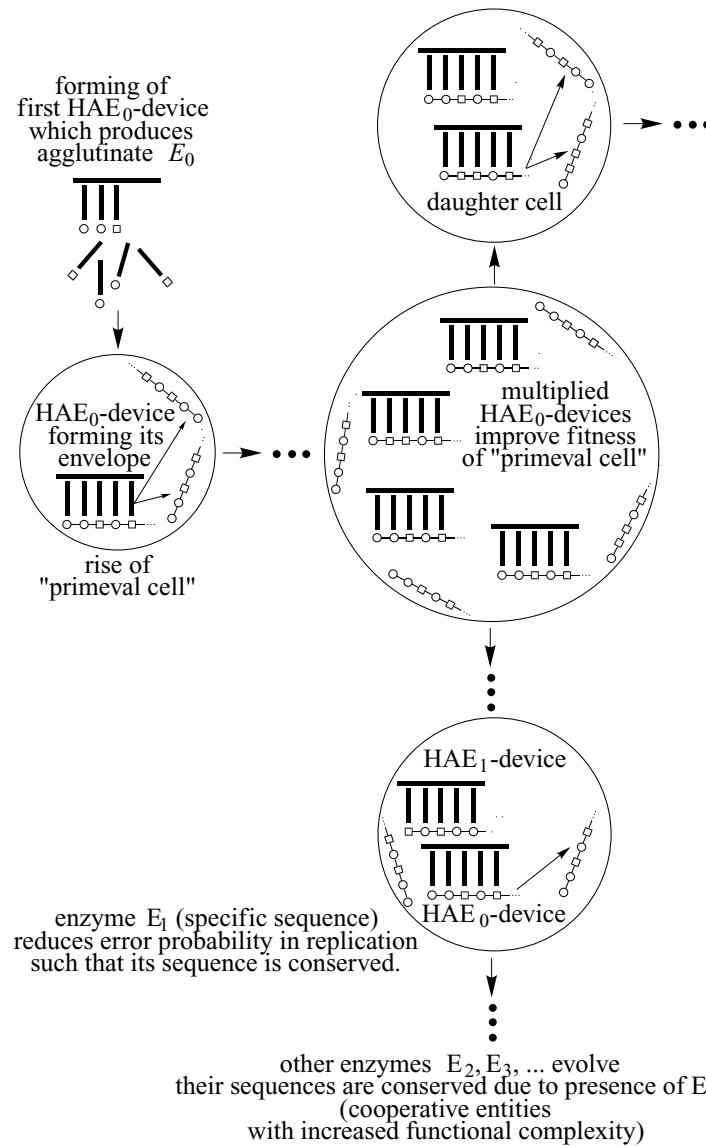
Figure 10. HAE-devices that produce envelopes as self made compartments. Division when amount larger than a critical value. Emergence of a translation apparatus as by-product.

3. Computer Implementation and its Results

In Figure 11 the computer implementation is shown in an overview-like flow-chart. There is a construction phase where aggregates are formed. According to the fitness of the formed devices, parameters are assigned to these entities such as the survival chance, the multiplication probability and the probability of occurrence of an error in the multiplication. In the selection phase only a fraction of the

start (generation g = 1)

- **initial strand** of *random* sequence in singled out compartment.

next generation

$g \rightarrow g+1$

construction phase:
formation of devices
by diffusion and interlocking

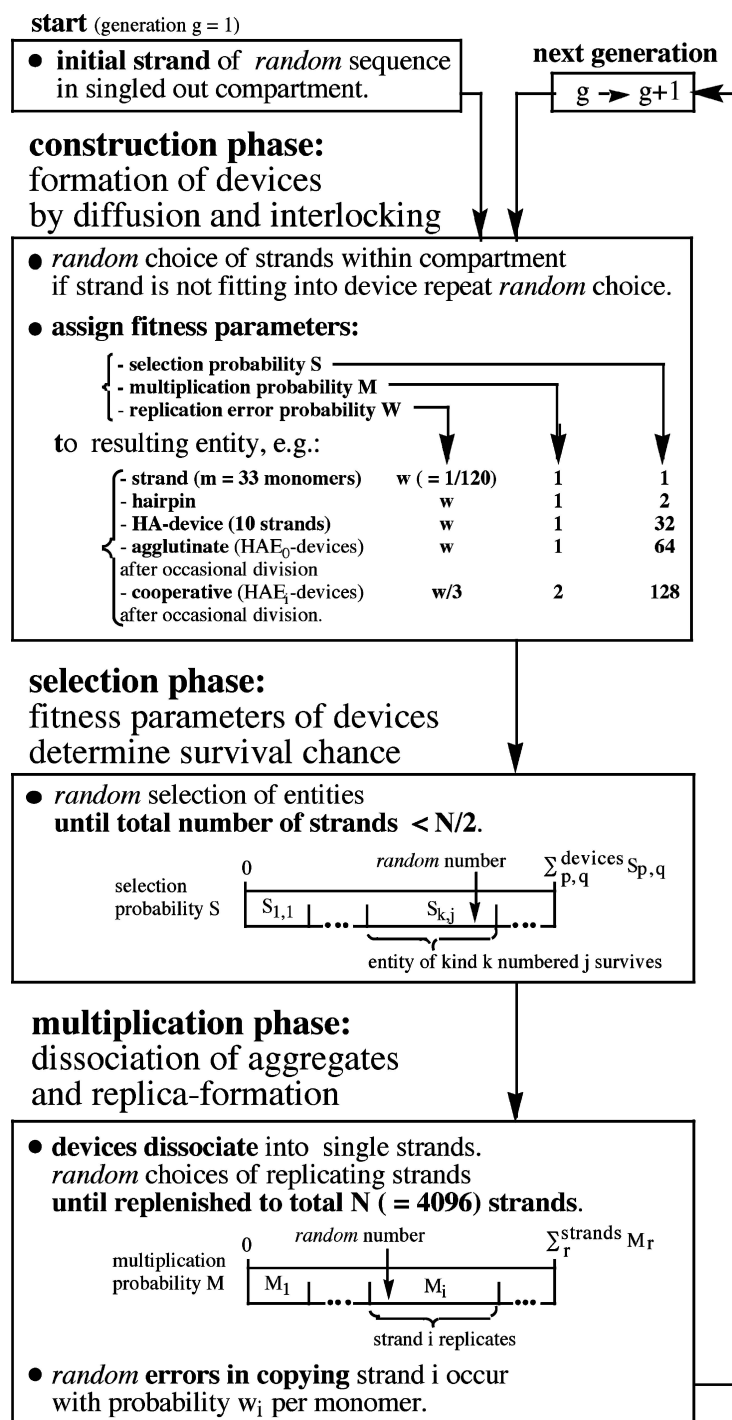- *random* choice of strands within compartment if strand is not fitting into device repeat *random* choice.
- **assign fitness parameters:**
  - **- selection probability S**
  - **- multiplication probability M**
  - **- replication error probability W**

to resulting entity, e.g.:

- strand (m = 33 monomers)	**w (= 1/120)**	**1**	**1**
- hairpin	**w**	**1**	**2**
- HA-device (10 strands)	**w**	**1**	**32**
- agglutinate (HAE_0-devices) after occasional division	**w**	**1**	**64**
- cooperative (HAE_i-devices) after occasional division.	**w/3**	**2**	**128**

selection phase:
fitness parameters of devices
determine survival chance

- *random* selection of entities **until total number of strands < N/2.**

selection probability S

0   *random* number $\sum_{p,q}^{devices} S_{p,q}$

$S_{1,1}$ ... $S_{k,j}$ ...

entity of kind k numbered j survives

multiplication phase:
dissociation of aggregates
and replica-formation

- **devices dissociate** into single strands. *random* choices of replicating strands **until replenished to total N (= 4096) strands.**

multiplication probability M

0   *random* number $\sum_{r}^{strands} M_r$

M_1 ... M_i ...

strand i replicates

- *random* **errors in copying** strand i occur with probability w_i per monomer.
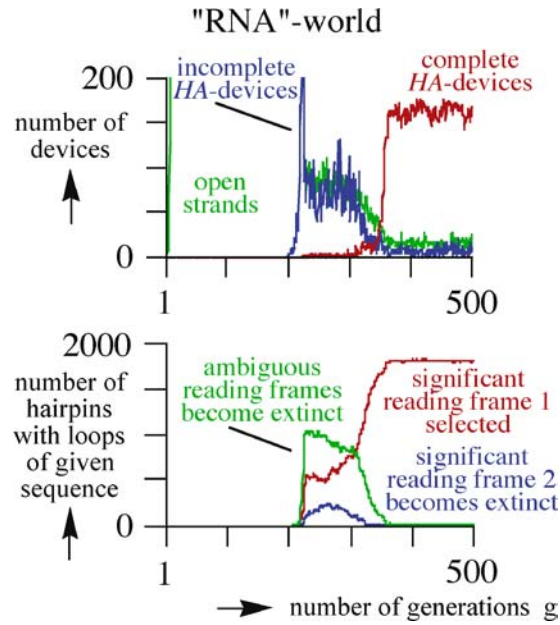
Figure 11. Computer implementation.

Figure 12. Computer results. The RNA-world. Open strands, hairpins, HA-devices and the reading frame.
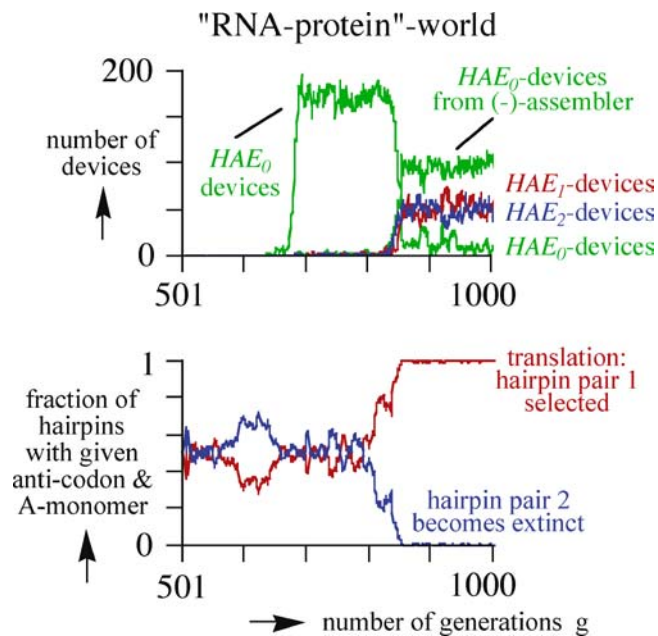

Figure 13. Computer results. RNA-protein world. HAE_0-devices, HAE_i-devices and the translation.

population survives (corresponding to half of the total number of strands). In the multiplication phase the aggregates dissociate and the strands are copied by chance until the total number of strands is replenished. Starting with a single strand of random sequence a typical simulation with a positive outcome is shown in Figures 12 and 13.

The RNA-world: as soon as the hairpins are formed, they bind to open strands (assembler function) forming incomplete HA-devices (blue in upper graph, Figure 12). Only with the occurrence of complete HA-devices (red in upper graph, Figure 12) the reading frame 1 (red in lower graph, Figure 12) becomes selected.

The RNA-protein-world: as soon as all the hairpins bound to an assembler carry A-monomers an A-polymer (agglutinate E_0) is formed, thus HAE_0-devices (green in upper graph, Figure 13) come into being. Only with the occurrence of HAE_1-devices and HAE_2-devices (red and blue in upper graph respectively) together with the agglutinate producing HAE_0-devices from the $(-)$assembler (green in upper graph) the translation 1 (red in lower graph, Figure 13) becomes selected.

4. Considerations on the Evolution of HAE_i-Devices to Today's Genetic Apparatus

4.1. THE DNA-RNA-PROTEIN WORLD

The complementary and anti-parallel copy of a $(+)$-assembler (producing an enzyme E_i) is the $(-)$-assembler (producing an agglutinate E_0). Only the complementary copy of the complementary copy is the true copy (Figure 14a). In each generation there is always a mixture of $(+)$-assembler and $(-)$-assembler. This leads to a catastrophe when the number of cooperating enzymes reaches a critical value by building up too many agglutinate producing HAE_0-devices from the $(-)$-assembler.

We introduce D-strands which are not capable of HA-device forming. The D-strands are formed from energy-rich D-monomers by polymerizing at R-strands templates. The mixture of $(+)$-assembler and $(-)$-assembler is unstable eliminating $(-)$-assembler: $(+)$-R-strands (by transcription having $(-)$-D-strands as template) and $(-)$-D-strands (by reverse transcription having $(+)$-R-strands as template) will be selected (Figure 14b).

There is again a catastrophe when too many different enzymes E_i are in the entity: a HAE_i-device is then easily lost. A $(-)$-D-superstrand is constructed (by reverse transcriptase with $(+)$-R-strands and $(+)$-R-bridges (Figure 14c) as templates. The $(+)$-R-strands are formed individually from the $(-)$-D-superstrand by transcriptase with $(-)$-D-bridges as recognition sites (Figure 14d).

A replicase is assumed to appear that directly copies the $(-)$-D-superstrand to a $(+)$-D-superstrand and vice versa (Figure 14e). The $(+)$-D-bridges are not recognition sites and the $(+)$-D-superstrand is not transcribed. The reverse transcription can

be eliminated (Figure 14e). Thus a separation of replication (information carrying system and information forwarding system in the line of descent) from transcription and translation (product producing system) has occurred, the basic structure of the modern genetic machine is reached (Figure 1).
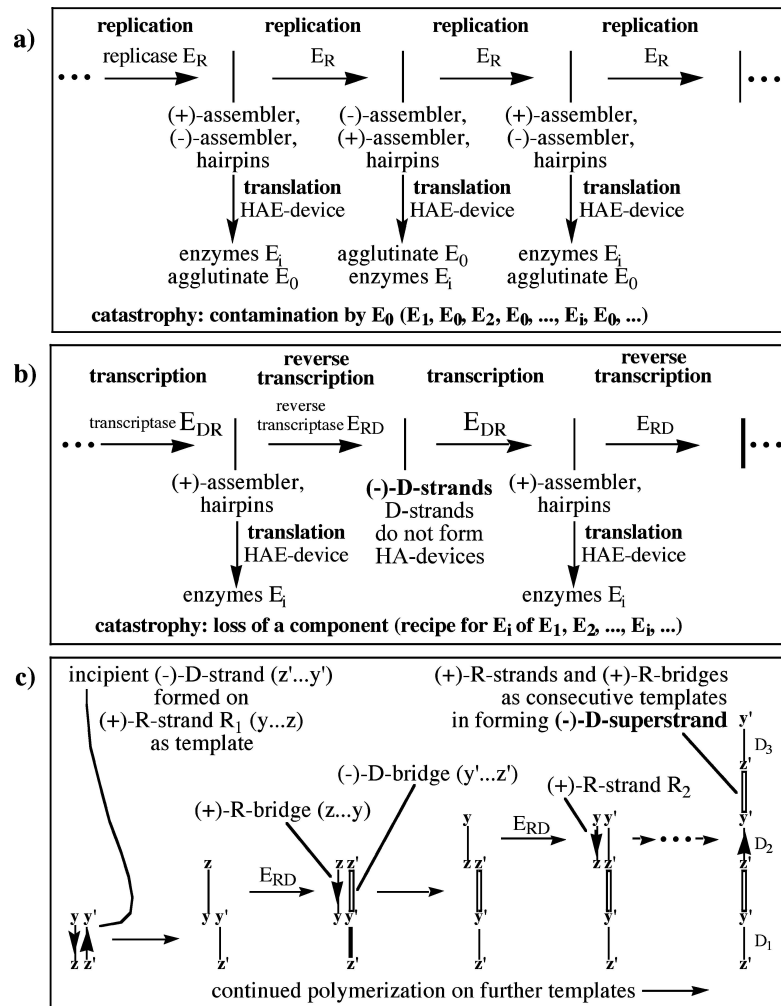

Figure 14. Separation of replication (information carrying and forwarding system) from translation (product producing system). (a) Replication and translation by $(+)$-R-strands and $(-)$-R-strands in succession. (b) Replication by $(-)$-D-strands and $(+)$-R-strands in succession (transcription and reverse transcription). Translation only by $(+)$-R-strands. (c) and (d) Formation of $(-)$-D-superstrand by $(+)$-R-strands where (y, y') and (z, z') are complementary end groups. (e) Replication only by D-superstrands. Transcription only from $(-)$-D-superstrands to $(+)$-R-strands and translation only by $(+)$-R-strands. Reverse transcription declines. (Today's genetic apparatus, see Figure 1).
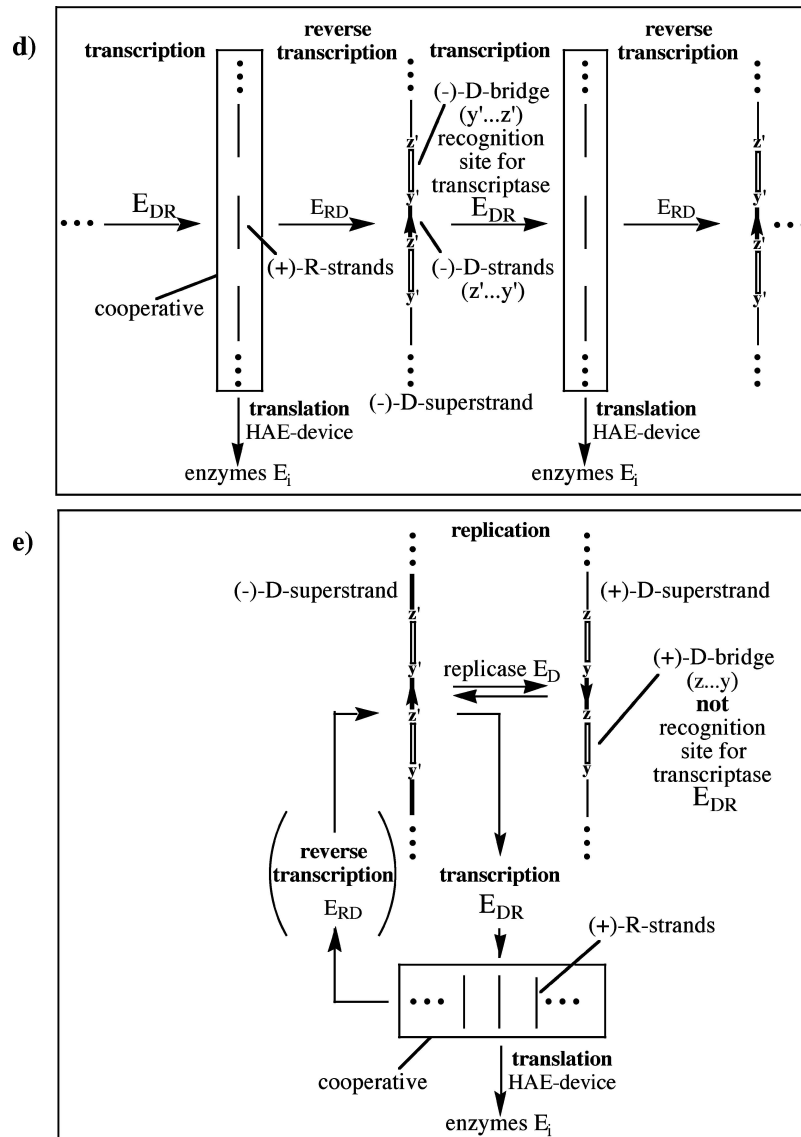
(*Continued on next page*)

Figure 14.   (*Continued.*)

4.2. STAGES LEADING TO TODAY'S GENETIC CODE

The earliest phase among candidates for the monomers of the first variety (R-monomers) are the nucleosides guanosine (G), cytidine (C), adenosine (A), and uridine (U). We favour G and C for the beginning, because they base pair with three rather than two hydrogen bonds. Polymer strands have the essential features required for HA-devices. For the A-monomers A_1 and A_2 we choose the most

polarity*		A	B	C	D	E
		GNC	PuNN	3rd position opens for coding	1st position opens for Py	modern genetic code
++	Gly	GGC	GGN			GGN
0	Ala	GCC	GCN			GCN
+++	Asp	GAC	GAN	GAPy		GAPy
+++	Glu	GAC	GAN	GAPu		GAPu
0	Val	GUC	GUN			GUN
		stop (best suppressing-binding triplet):				
		UPuPu	UPuN		UGA	UAPu
0	Leu		AUN			
0	Ile		AUN		AUPy	AUA
++	Ser		AGN	AGPy		AGPy
+	Thr		ACN			ACN
+++	Lys		AAN	AAPu		AAPu
++	Arg			from Ser AGPu		AGPu
+++	Gln			from Lys AAPy		
+++	Asn			from Lys AAPy		AAPy
0	Pro				CCN	CCN
					Arg CGN	CGN
					Gln CAN	CAPu
					Leu CUN	CUN
					Leu UUN	UUPu
					Ser UCN	UCN
0	Phe					from Leu UUPy
+++	Tyr		N = A, G, C, U			from stop UAPy
+++	His		Pu = A, G			from Gln CAPy
++	Cys		Py = C, U			from stop UGPy
0	Met					from Ile AUG
0	Trp					from stop UGG
++	Sec					from stop UGA

decreasing availability of amino acids

* polarity hypothesis

2nd position	polarity of amino acid	
A →	more polar	+++
G →	less polar	++
C →	weakly polar	+
U →	non-polar	0

Figure 15. Sketched stages in development to today's genetic code. Stage A: code GNC for the five most abundant aminoacids, Stage B: code PuNN for the ten most abundant aminoacids, Stage C: third position opens for coding, Stage D: first position opens for Py, stage E: emergence of modern code.

prevalent amino acids alanine (Ala) and glycine (Gly). The earliest "code" in our model involves just these two nucleotides and two amino acids (Figure 15).

The translation device becomes increasingly sophisticated by incorporation of increasingly complex translation products stabilizing the attachment of hairpins to the assembler strand. Strength in base pairing may be no longer beneficial. The reading frame gradually loses in importance. Triplet position 1 is also pressed into coding service, and finally all three positions are used for this.

Each step of relaxation of constraints on the code by improvement of the translation machinery confers a selective advantage, increasing the number of amino acids

that can be coded for. In this way the modern code can be reconstructed, when it is assumed

(1) that the order in which amino acids are introduced into the code is in the order of their prebiotic availability;

(2) that the code, first restricting to GNC (N = G, C, A, U) changes over to GNN (first position sufficient for docking), then the code changes to PuNN (Pu = G, A; a purine is sufficient for docking); later the third position is used for coding and the first position is opened for pyrimidines C and U (Py = C, U;)

(3) that A in position 2 is associated with the most polar amino acid to be introduced, G, C, U are associated with amino acids in the order of decreasing polarity. This is a generalisation of the initial codes GAC for Glu, Asp; GGC for Gly; GCC for Ala; GUC for Val.

The modern code results from these postulates when taking into consideration that important amino acids are provided with an additional codon at the stage when position 1 is opened for C and U. The insight thus afforded strongly suggests that the ancestral self-reproducing system resembled the HA-device.

References

1. Kuhn, H. and Waser, J.: Self-Organization of Matter and Early Evolution of Life, in W. Hoppe, W. Lohmann, H. Markl and H. Ziegler (eds.) *Biophysics*, Springer-Verlag, Berlin, 1983, pp. 830–874.
2. Kuhn, H. and Waser, J.: A model of the origin of life and perspectives in supramolecular engineering in J.-P. Behr (ed.) *Lock-and-Key Principle*, Wiley, Chichester, 1994, pp. 247–306.
3. Kuhn, C.: Computer-Modelling Origin of a Simple Genetic Apparatus, *Proceedings of the National Academy of Sciences USA*, **98** (*2001*), 8620–8626.
4. Kuhn, H. and Kuhn, C.: Diversified World: Drive of Life's Origin?! *Angew. Chemical International Edition* **42** (*2003*), 262–266.