Molecular Evolution

Inken Wohlers

Abstract

If we look at living organisms we find an amazingly complex system of interacting molecules. We find that each organism's genetic information is encoded in RNA or DNA sequences out of four nucleotides and that each organism proliferates by synthesis of new and basically identical RNA or DNA sequences. A sequence's genetic information instructs the synthesis of a large number of different molecules which fulfill complex functions.
It is not known yet, how this complex molecular machinery we encounter in living organisms has developed, but two molecular prerequisites are obvious: self-reproduction and cooperation of (macro-)molecules. The possibility to store more and more information and to be part of more and more complex interactions seems to be a result of mutation and selection and thus evolution on a molecular level. Eigen and Schuster proposed two mechanisms for such a molecular evolution, quasi species and hypercycles [1]. The quasi species model explains the inner diversity of species while hypercycles explain how the genetic information stored in an RNA or DNA sequence can be increased as a result of molecular cooperation.

1 Quasi Species

Manfred Eigen and Peter Schuster proposed a model for the evolution of self-replicating molecules like RNA or DNA, the quasi species model. An outline of this model will be given in chapter 1. For more details and the mathematical formulation refer to a review by Kemena [2] or the original publication [1]. The limitations of the quasi species model will be discussed in chapter 2. To overcome these limitations Eigen and Schuster suggest hypercycles, which will be introduced in chapter 3.
A self-replicating molecule consists out of a sequence of monomers and results from a copy process of an existing molecule. It can be identical with its parent or can be mutated from it leading to a locally different molecule. A self-replicating molecule may decay into its building blocks, in the case of RNA or DNA into nucleotides. The model assumes that the building blocks are always present in sufficient amount. The selective pressure on a specific self-replicating molecule is its replication rate; molecules with higher replication rate have a selective advantage since they are able to reproduce themselves or, as a result of mutations, sequences closely related to themselves more often.
The main finding of the mathematical formulation of this model is the existence of so called quasi species, which are distributions of closely related sequences. The relatedness is usually measured by the minimal number of point mutations needed in order to transform one sequence in the other.
The selective pressure does not lead to survival of one single sequence, but to

consolidation of many different sequences which differ only slightly from each other and thus can mutate into each other during replication. The sequence with the highest number of molecules is called the master sequence and refers to what is called the wildtype. The quasi species is distributed around this master sequence where the form of the distribution may vary. If the mutation rate is small less mutations are likely to occur and the quasi species distribution gets narrow, while a high mutation rate leads to more mutations resulting in a broader distribution.

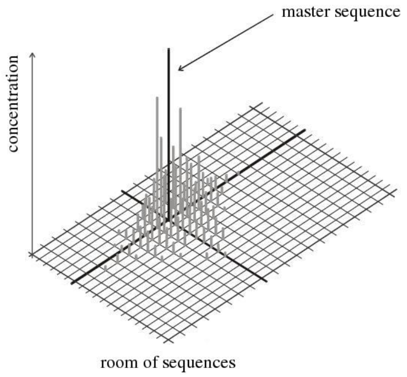

Figure 1: Quasi species in the room of sequences [3].

There are two borderline cases for the mutation rate. The first one is a mutation rate of 0, which means, that each sequence is copied without error, always leading to an identical sequence. In this case each sequence represents a quasi species. The other borderline case is that the mutation rate exceeds a specific error threshold which leads to random sequence spreading over the whole range of possible sequences. The quasi species distribution is unsteady or randomly distributed and its information therefore can not be passed on through various reproductions. As a result information can not be inherited. The next chapter describes in detail how the error threshold can be obtained and how this leads to a limit for the information which can be conserved within a quasi species.

2 Information Content and Conservation of Information

A quasi species is able to pass on information to consecutive generations if its distribution is steady within the room of sequences. This is the case if and only if the master sequence is stable. In order to derive a formula for the error rate threshold we have a look at the probability for an error free replication of the master sequence, Q_m. The probability of a mutation follows to be $1 - Q_m$. Q_m can be interpreted as a quality factor for the replication and can be written as product of the probabilities for error-free replication of the single monomers in the master sequence. If we assume as a simplification that all nucleotides have the same probability for error-free replication, namely q_m, we get

$$Q_m = q_m^{N_m}, \tag{1}$$

where N_m is the length of the master sequence. The master sequence is stable if the correct master sequence copies are able to compete with their error copies, which means that the quality factor needs to be above a certain threshold, formally

$$1 > Q_m > Q_{min}. \tag{2}$$

For the threshold Q_{min} the following equation can be derived

$$Q_m > Q_{min} = \frac{1}{\sigma_m}, \tag{3}$$

which is equivalent to

$$q_{min} = \sqrt[N_m]{\frac{1}{\sigma_m}}, \tag{4}$$

where σ_m is the superiority of the master sequence. The superiority is a weighted quotient of the master sequence's fitness divided by the rest of the population's fitness. Equation (4) shows that the probability for an error-free replication and therefore also the error rate threshold depends on two parameters, on the superiority of the master sequence and on the sequence length. From rearranging equation (4) we obtain an equation for the maximum sequence length that still permits a stable quasi species distribution and therefore inheritance of information,

$$N_{max} = \frac{\ln \sigma_m}{1 - q_m}. \tag{5}$$

Note that this limit is inversely proportional to the average error rate per monomer within the master sequence, $1-q_m$. Another way to interpret equation (5) is as expectation value of an error in the master sequence,

$$\mathbb{E}(\varepsilon_m) = N_m(1 - q_m) \Leftrightarrow \exp(\mathbb{E}(\varepsilon_m)) < \sigma_m. \tag{6}$$

The expected number of mutations during one reproduction of the master sequence has to stay below a sharply defined threshold which depends on the superiority of the master sequence. The relationship of sequence length, error rate and superiority described by equation (5) is illustrated in table 2 for different molecular mechanisms.

3 Hypercycles

The last chapter showed that the information content of quasi species is strictly limited. In table 2 sequence lengths for enzyme-free RNA replication are listed for different values of superiority. Uncatalyzed replication of RNA has never been observed to any satisfactory extent but error rates of 5×10^{-2} per nucleotide seem to be realistic. For such fairly high error rates sequence lengths up to a magnitude of 100 base pairs can be reached for high values of superiority. Such low information content is not sufficient to encode for the complexity of living systems we encounter today. But how can the sequence length and thus the information content of a system be increased? Eigen and Schuster proposed

Error	**Super-**	**Sequence**	**Molecular Mechanism**
Rate	**iority**	**Length**	**and Biological Example**
$1 - q_m$	σ_m	N_{max}	
5×10^{-2}	2	14	Enzyme-free RNA replication
	20	60	(t-RNA precursor, $N = 80$)
	200	106	
5×10^{-4}	2	1386	Single-stranded RNA
	20	5991	replication via specific replicases
	200	10597	(phage Q_β, $N = 4500$)
1×10^{-6}	2	0.7×10^6	DNA replication via polymerases
	20	3.0×10^6	including proofreading by exonuclease
	200	5.3×10^6	(E.coli, $N = 4 \times 10^6$)
1×10^{-9}	2	0.7×10^9	DNA replication and
	20	3.0×10^9	recombination in eucaryotic cells
	200	5.3×10^9	(vertebrates (man), $N = 3 \times 10^9$)

Figure 2: Error rate, superiority and sequence length for different molecular mechanisms [1].

cooperation of macromolecules within a hypercyclic linkage as an answer to this question. This chapter will discuss their model of the hypercycle in detail.
From equation (5) we know that there are two ways to increase the sequence length of a self-replicating molecule and thus the information content of its quasi species. Those are:

- Lower the error rate
- Increase the master sequence's superiority

Both can be achieved via cooperation of molecules. In the case of RNA or DNA replication for example enzymes catalyze the replication and therefore lower the error rate. Lower error rates permit for larger sequence lengths. The enlarged information content can encode more complex enzymes and therefore increase the superiority. This is visualized in figure (3).

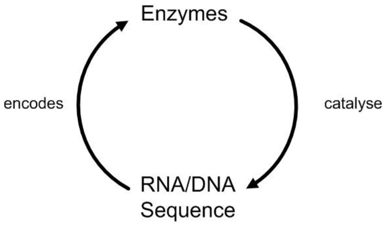

Figure 3: Relationship of RNA/DNA sequences and enzymes.

Hypercycles are chemical reaction cycles of a high level of organization. A chemical reaction cycle is a sequence of reactions where any product is identical with a reactant of a preceding step. Chemical reaction cycles are also called catalysts because they catalyze the synthesis of at least one member of the cycle. Figure (4) shows the catalytic mechanism of an enzyme.

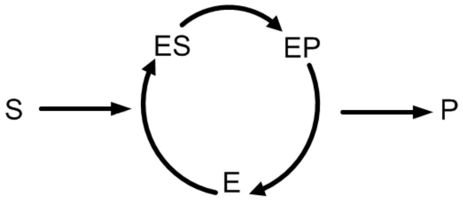

Figure 4: Catalyst: Catalytic mechanism of an enzyme (S: Substrate, P: Product, E: Enzyme).

The next higher level of organization is the catalytic cycle, a reaction cycle where one or up to all intermediates are catalysts. An example for a catalytic cycle is the replication of single stranded RNA which exhibits linear growth behavior since from one single strand one new strand is produced. The abstract catalytic cycle and the catalytic cycle for single stranded RNA replication is displayed in figure (5).

An autocatalyst is a special catalytic cycle where the product of a single reaction is identical with its substrate. Therefore it is a true self-replicating

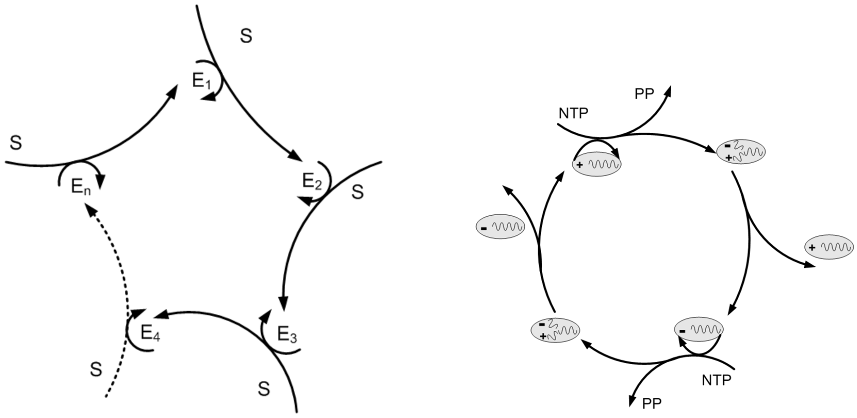

Figure 5: Catalytic cycles. Left: Schema where E_i are enzymes and S substrate. Right: Replication of single-stranded RNA.

unit. An example for an autocatalytic cycle is the process of semi-conservative DNA replication where two daughter strands are synthesized from one parent strand. Autocatalysts exhibit exponential growth behavior.

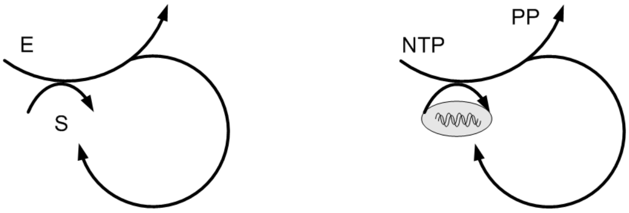

Figure 6: Autocatalyst: Left: Schema where E is the enzyme and S the substrate. Right: Semi-conservative DNA replication.

A hypercycle finally is a reaction cycle of a next higher level of organization, namely a cycle of catalytic cycles or autocatalysts. In the case of RNA or DNA molecules we find hypercyclic cooperation if the sequences are transcribed and translated to enzymes, which in return catalyze the replication of the RNA or DNA sequences and therefore lower the replication's error rate. The abstract form of a hypercycle is displayed in figure (7).

If we assume that self-organization within hypercycles was one step of molecular evolution, we might ask ourselves, why nowadays there exists only one hypercycle for the molecular machinery of the cell and why there is only one universal genetic code. This is the reason, why we finally have a look at the competition of hypercycles.

Hypercycles compete with each other if they are not locally separated from each other, if they need the same chemical building blocks and if they do not cooperate in higher order linkage. The number of molecules for each hypercycle can

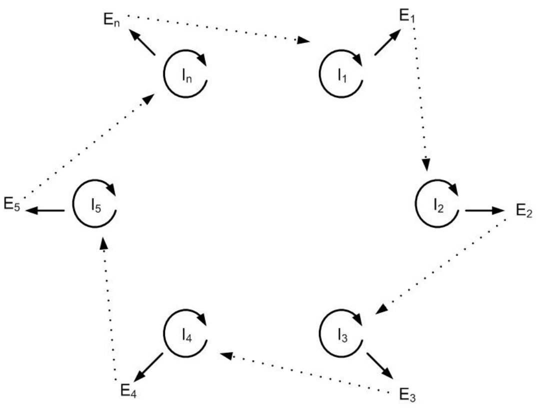

Figure 7: Hypercycle: I_i are catalytic cycles or autocatalysts, E_i are enzymes.

be mold using differential equations [4],

$$\frac{dn_{E,i}}{dt} = A_{T,i}n_{I,i} - \varphi_E(t)n_{E,i} \tag{7}$$

$$\frac{dn_{I,i}}{dt} = A_{R,i}n_{I,i}n_{E,i} - \varphi_I(t)n_{I,i}. \tag{8}$$

Here $n_{E,i}$ is the number of enzymes and $n_{I,i}$ the number of self-replicating molecules, e.g. RNA or DNA. The parameters are A_T for the rate with which a self-replicating molecule is translated to an enzyme and A_R for its replication rate. The index i denotes the hypercycle the molecules and rate constants belong to. The change of the number of enzymes depends on the number of self-replicating molecules present. The change of the number of self-replicating molecules depends on the number of enzymes catalyzing and on the number of self-replicating molecules since they are the blueprint for new molecules. The competition of hypercycles is mold using the functions $\varphi_E(t)$ and $\varphi_I(t)$ which depend on the time t. The competition therefore changes with time. As a simplification we assume that the overall number of molecules in the system is constant. This means that the rates of change, the values of the differential equations, summed up have to be equal to zero, formally

$$\sum_i \frac{dn_{E,i}}{dt} = 0 \ , \ \sum_i \frac{dn_{I,i}}{dt} = 0. \tag{9}$$

Using those equations we get for the competition functions

$$\varphi_E = \frac{\sum_i A_{T,i}n_{I,i}}{\sum_k n_{E,k}} \ , \ \varphi_I = \frac{\sum_i A_{R,i}n_{I,i}n_{E,i}}{\sum_k n_{I,k}}. \tag{10}$$

Now we are interested in the stable steady states of this dynamic system, since stable steady states are specific numbers of molecules that are reached after a certain time and that are maintained, even after small perturbations. If we consider the competition of two hypercycles we find two stable steady states,

$$(n_{E,1}, n_{I,1}, n_{E,2}, n_{I,2}) = (0, 0, C_E, C_I) \tag{11}$$

$$(n_{E,1}, n_{I,1}, n_{E,2}, n_{I,2}) = (C_E, C_I, 0, 0), \tag{12}$$

where $C_E = n_{E,1} + n_{E,2}$ is the overall number of enzymes and $C_I = n_{I,1} + n_{I,2}$ the overall number of self-replicating molecules in the system. The trajectories in the $(n_{I,1}, n_{E,1})$ plane are depicted in figure (8).

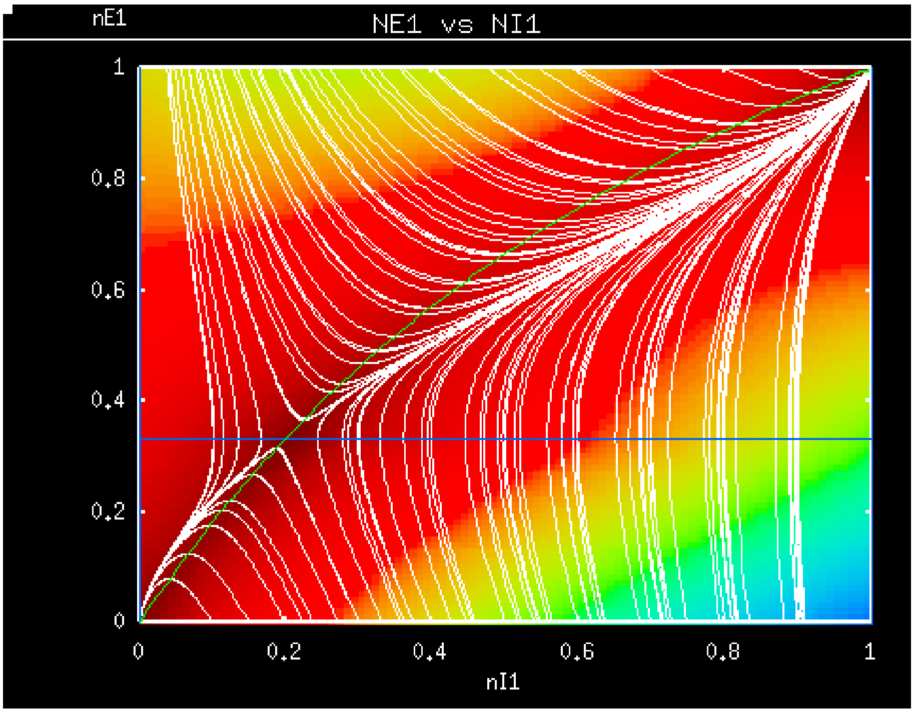

Figure 8: Competition of two hypercycles: Nullclines (blue and green lines) and trajectories for the first hypercycle (white lines) $(A_{T,1} = A_{R,1} = 1, A_{T,2} = A_{R,2} = 0.5)$

The steady states and the trajectories show that depending on the initial values only one of the two hypercycles survives. In the case which the graphic visualizes it is less likely that the first hypercycle with better replication and translation rates gets extinct, but it may happen if the initial number of molecules belonging to this hypercycle is small. Since a new evolving hypercycle usually has a low initial number of molecules it is unlikely that it can survive if another hypercycle is already present, even if the new hypercycle has a better reproduction rate. This fact can be described as "once-for-ever" selection [1]: After one hypercycle had been selected and gained many molecules, there is almost no chance for new hypercycles to prevail, even if they are more efficient. Such a "once-for-ever" selection is a possible explanation why there is only one basic molecular machinery of the cell.

References

[1] Schuster P Eigen M. The hypercycle: A principle of natural self-organization. *Naturwissenschaften*, 64:541–565, 1977.

[2] 'Carsten Kemena'. Sequence evolution. *Seminar 'Gute Ideen in der theoretischen Biologie/Systembiologie' FU Berlin*, 2007.

[3] Schuster P. Ursprung des lebens und evolution von moleklen. *http://www.tbi.univie.ac.at/ pks/PublicLectures/ursprung-1999.pdf*, 1999.

[4] Fachkurs 'Modellierung biologischer Systeme'. Competition and selection in biological systems. *HU*, 2000.