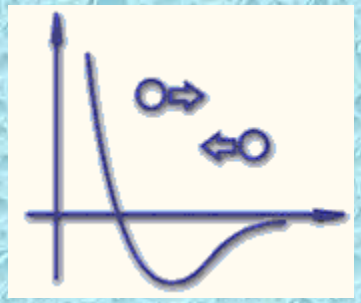
Partially-directed evolution: from the Big Bang to the problem of knowledge acquisition

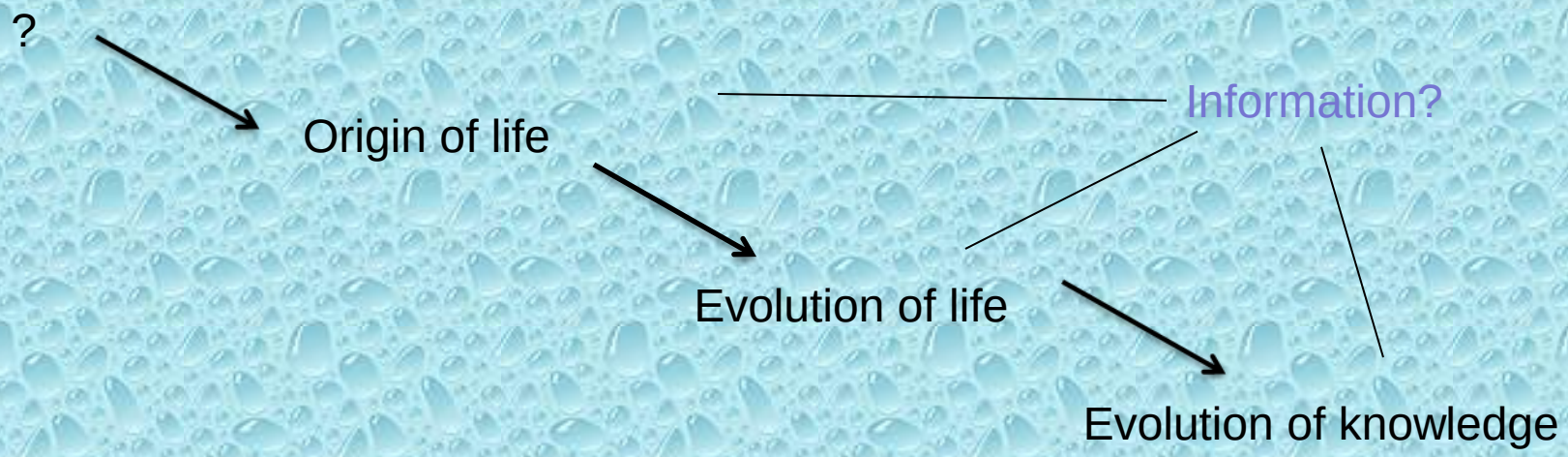
Alexey Melkikh

Ural Federal University, Mira str. 19,
620002,Yekaterinburg, Russia
melkikh2008@rambler.ru

Ural Federal University

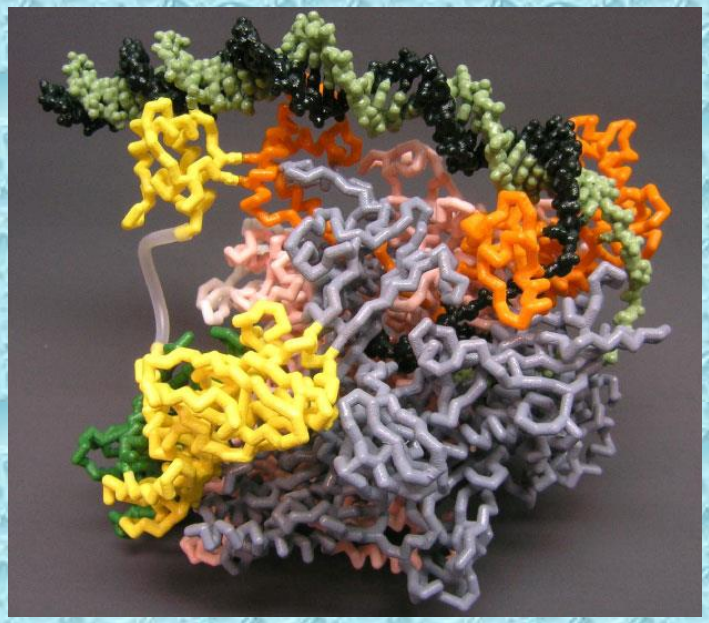


Problems of evolution of biological information



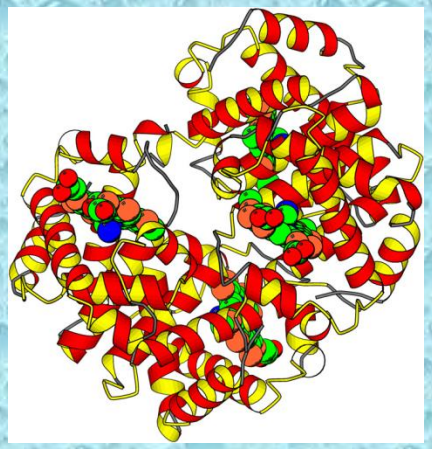
- How do complex macromolecules (replicators) function effectively, bearing in mind that the total number of possible spatial configurations of these molecules is exponentially large?
- How does selection operate on long coding sequences, given the total number of states of such a sequence is exponentially large?

1. The problem of folding and stability of replicators



RNA-world and protein-world

Protein folding – Levinthal paradox

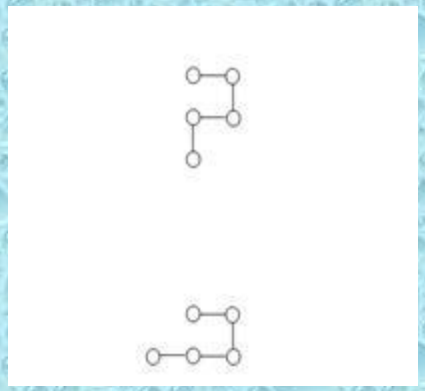


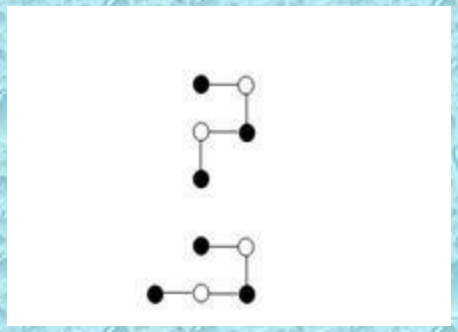
Typical number of steps required for the folding of a protein that consists of 150 domains, each of which can have three possible states, is approximately

$3^{150} \approx 10^{72}$ → 10^{48} years

How to fold a protein?

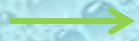
Two-dimensional folding





One- and two-component systems

energy-degenerate (but spatially different) states



funnel-like landscape is not formed and the probability of arrival of a molecule in its native conformation is low

$$P(n) = \left(\frac{1}{2}\right)^{n/2}$$

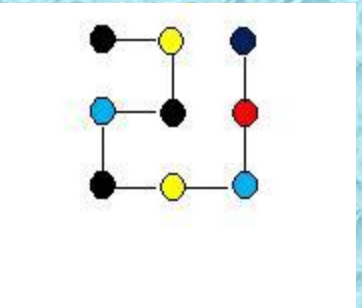
For sufficiently large *n*, this value will be exponentially small

$N << 10^3$ molecules could in principle find their native conformation through the simple enumeration of variants

Consideration of folding into a three-dimensional space will only lead to a greater number of "forks" in the way of folding, because in this case there will be additional equivalent energy states associated with the possibility of folding on an additional axis.

Energy levels

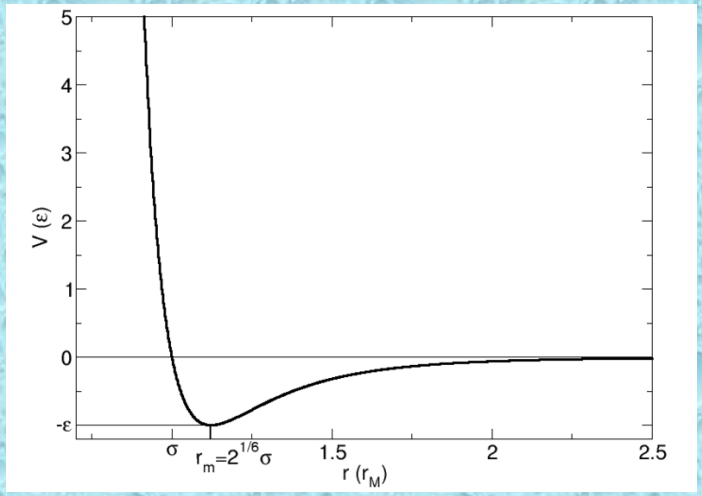
The presence of different monomers implies the presence of different energy levels (at the limited value of a full energy interval ΔE). Thus, the energy spectrum becomes almost continuous, which decreases the control of the selectivity of the folding process. This loss of control will arise because the probability that the energy of alternative conformations will slightly differ will be high, i.e., the probability that particles with strongly differing potentials are found close to each other will be small.



If the total difference in the interaction energies is approximately 5-10 kT (this value is typical for the interaction between the amino acids during protein folding), the energy difference between the alternative conformations (for the different monomers equals to 20) will be less than kT. This means that the probability of alternative conformation in such a system will be comparable with each other. This in turn will lead to a large number of configurations of the macromolecule.

Problem of short-range potentials

Lennard-Jones and Morse potentials: minimum corresponds to interatomic distance



Requirements:

- Interaction between the particles must be nonlocal; interaction based on the nearest neighbors cannot resolve the paradox;

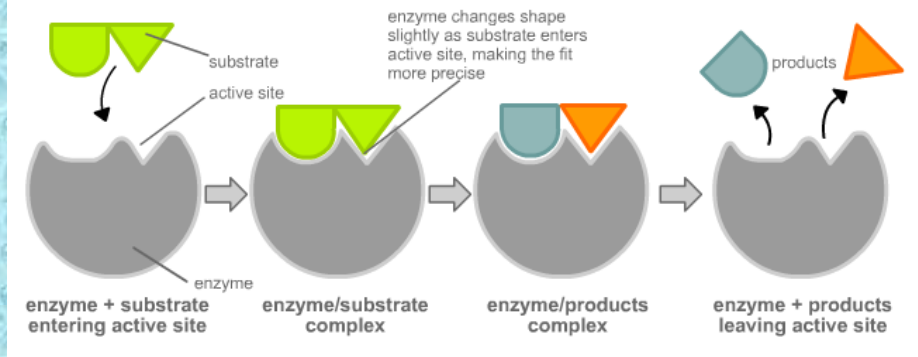
- Interaction should have a resonance (essentially structurally-dependent) character, as in most non-living systems the behavior of atoms and molecules is well explained on the basis of the known interaction potentials.

Why quantum mechanics is able to resolve the problems arising at interaction of molecules? The answer lies in the fact that classical mechanics cannot meet the requirements for proteins folding and to describe the motion of biologically important molecules. In particular, classical mechanics gives only a few well-known interaction potentials such as the Coulomb or gravitational, while others are essentially quantum ones.

The main paradox of molecular biology

interaction of two sufficiently long folded molecules – biochemical reactions

“key" and "lock” principle



Melkikh, Biosystems, 2014b

On the one hand, the number of different kinds of particles surrounding given particle must be large, otherwise it is impossible to make a clear choice in favor of a single path. But in order that the particles are different (primarily on the interaction energy) for a given energy range typical number of such particles should be small. Implementation of these contradictory requirements for known interaction potentials between particles is impossible.

Thus, the existence of a set of spatial structures of biologically important molecules, as well as the many variants of chemical reactions between them, is one of the most important obstacles to the emergence of the simplest living systems. There must be some special mechanism that significantly limits the range of possible variants for such a system.

→ Information

Problem 2: Origin of the molecular code and the problem of enumeration of variants

In the early stages of the evolution of life, biomolecules were involved in various reactions, including self-reproduction. At this stage, the problem of enumeration of variants was not important because molecules arising as a result of reactions could not perform certain types of work. However, some of the molecules further encoded other molecules. This is a crucial step, which creates a problem of enumeration of variants.

Consider a chain of nucleotides of length N. There are 4^N variants of such sequences. How large is this number? For example, for N = 1000 we receive $4^{1000}=10^{602}$.

N = 1000 corresponds to about only one modern gene

Molecular exaptation?

However, this mechanism does not work by itself, it implies the existence of a priori information in the system. Indeed, if the system has no *a priori information* about what exactly will encode a set of characters, there is no way of knowing about it, but to synthesize the molecular machines using this set and checking whether such an organism will survive or not. This is a simple enumeration of variants. If *a priori information* is available, it must have some material carrier, i.e., be recorded in some (yet unknown) intracellular structures.

A fundamental question is thus raised: is it possible to speed up the search process in the absence of *a priori information* through any other method (molecular exaptation, horizontal gene transfer, alternative splicing, which emerged later etc.)?

The answer to this question is negative: either change in the information sequence occurs randomly, or there must be a system that somehow *selects* nucleotides and performs certain operations with them

a priori information

Partially-directed evolution

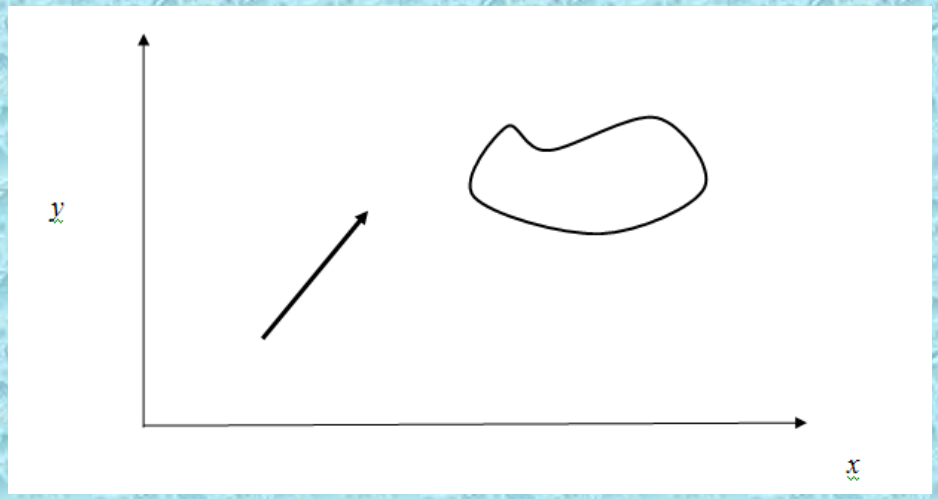
Basic equation of evolutions of replicators:

$$\frac{dp_k}{dt} = W_k p_k + \sum_l \chi^D_{kl} p_l + \sum_l \chi^*_{kl} p_l - E p_k$$

where p_k – are probabilities of existing of certain species of replicators in population

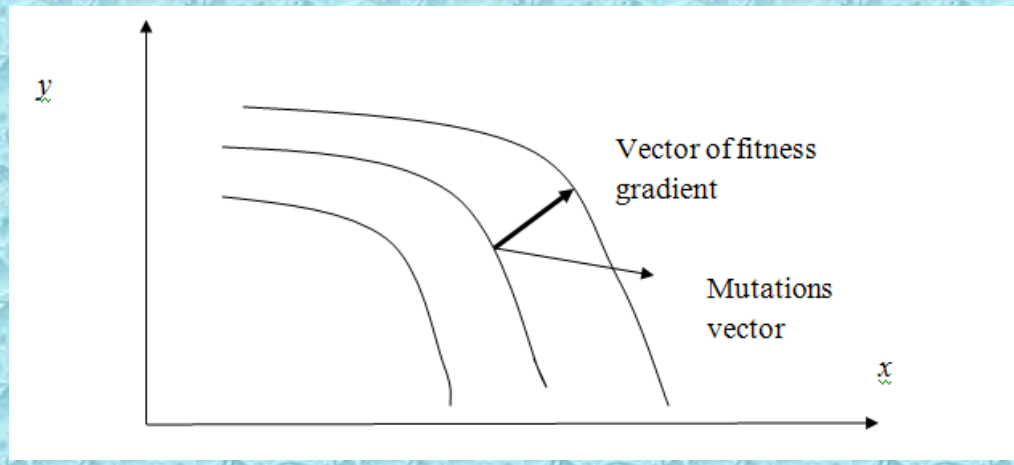
where χ^D_{kl} - are mutational fluxes not directed a priori to achieve any goal in the phase space, which corresponds to Darwinian evolution, χ^*_{kl} - mutation fluxes, directed *a priori* to achieve any «goal» in the phase space. In this case, the "goal" is a region in the phase space in which the system tends, in accordance with the equation of motion.

Phase space



terminal set (empty ecological niche)

fitness landscape

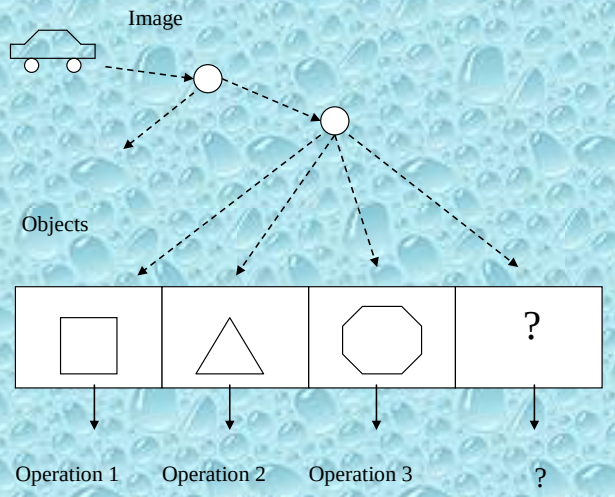


In the case of partially directed evolution, movement can be in a direction different from the local gradient

Paradox of knowledge acquisition

- If the received information is new to the organism (i.e., the images registered by the receptors are not recognized, and any innate program that responds adequately to the appearance of the image is absent), this information is not valuable (helpful),
- If the received information is valuable (i.e., the recorded images are recognized, and as a result of their recognition, certain a priori programs adequately respond to the environment and begin to work), this information is not new.

As a result, the acquisition of knowledge - receiving new valuable information from the environment - is logically contradictory



generalized behavior

	Evolution	**Behavior, learning, knowledge acquisition**	**Intracellular reactions, transport of substances, protein folding**
Time interval	Millions of years	Years, days, hours	Seconds, milliseconds
The main problem, paradox	The problem of rates and mechanisms of evolution	The problem of mechanisms of knowledge acquisition	The problem of selectivity and the rate of biological reactions
Paradox solution	Partially directed evolution	Innate programs of behavior	Biochemical reactions that are controlled at the microscopic level
Common base	Control of quantum properties of biologically important molecules		

Where *a priori information* in the early stages of evolution came from?

properties of biologically important molecules (proteins, RNA)

properties of atoms

amount of elements important for life (fraction of C, N, H, O…)

initial conditions of the Universe

Biological systems at all levels of organization require a large amount of a priori information for its operation

Where this information came from and where it was stored before the emergence of life?

Before inflation?

Up to the Planck time, 10^{-43} seconds it could be the era of the setting of the constants

Before this, constants could have a wide distribution, and as a result of evolution constants took (very accurately) quite certain values. These values together with the initial conditions contain all further information about the Universe

analogue of DNA

Evolution ↔ process of coding

Conclusions

Origin of life and its evolution require a priori information:
-Protein folding,
-Biochemical reactions,
-Evolution of information sequences,
-Partially-directed evolution,
-Paradox of knowledge acquisition

This information could be coded in initial conditions of the Universe and physical constants