Complexity as Energy, Constructive Physics and the Third Evolutionary Synthesis

S.V. Kozyrev, Steklov Mathematical Institute

10-th Biomath Conference on mathematical methods in biology,
6–8 November 2018, INM RAS, Moscow, Russia

S.V. Kozyrev, Biology is a constructive physics, p-Adic Numbers, Ultrametric Analysis and Applications, 10:4 (2018), 305–311. arXiv:1804.10518

A model of biological evolution by constructive statisitical mechanics where Hamiltonian contains contribution from logarithmic Kolmogorov complexity.

Yuri Manin's *Complexity as Energy* approach — Gibbs distribution is the Zipf's scaling law.

Eugene Koonin's *Third Evolutionary Synthesis* — scaling in genomics should be described by interacting gas of genes.

Relation to machine learning — biological fitness is described by the functional of empirical risk and Kolmogorov complexity is used as a regularization to reduce overfitting.

The idea of constructive mathematics is to add computability to properties of mathematical theory, in particular to consider theories where objects are defined constructively and maps are computable functions.

Constructive worlds. An (infinite) constructive world is a countable set X given together with a class of structural numberings: computable bijections $u : \mathbb{Z}_+ \to X$.
Moreover natural maps between constructive worlds should be given by computable functions. One can also consider constructive worlds enumerated by sequences of bits ("programs").

Example. Goedel numbering of formulae of a formal language.

Complexity as energy (Yuri Manin):

There are natural observable and measureable phenomena in the world of information that can be given a mathematical explanation, if one postulates that logarithmic Kolmogorov complexity plays a role of energy.

Y I Manin, Complexity vs energy: theory of computation and theoretical physics, Journal of Physics: Conference Series 532 (2014) 012018 arXiv:1302.6695

Relation to Zipf's law.
We apply this idea to modeling of biological evolution and discuss relation of Zipf's law and scaling in genomes.
Biological (genetic) sequences in this approach form a constructive world, i.e. *Biology is a Constructive Physics*.

Kolmogorov complexity and Kolmogorov order.

A constructive world X is generated by "programs" (sequences of bits). For $x, y \in X$ we consider the conditional entropy (complexity) as the minimal length of program p (in bits) satisfying

$$K_A(x|y) = \min_{A(p,y)=x} l(p). \tag{1}$$

i.e. program p computes x starting from y. Here A is a "way of programming".

Unconditional complexity is given by application of the above definition to some "initial" object y_0

$$K_A(x) = K_A(x|y_0).$$

Logarithmic Kolmogorov complexity of x is the length (in bits) of the shortest program which generates x.

There exists such way of programming A that for each other (semi)-computable B, some constant $c_{AB} > 0$, and all $x \in X$, one has

$$K_A(x) \leq K_B(x) + c_{AB}.$$

Here $K_A(x)$ is the logarithmic Kolmogorov complexity of x. Way of programming A is called optimal Kolmogorov numbering.

A **Kolmogorov order** of a constructive world X is a bijection $X \to \mathbb{Z}_+$ arranging elements of X in the increasing order of their complexities K_A.

Properties of Kolmogorov complexity

Any optimal numbering is only partial function, and its definition domain is not decidable.

Kolmogorov complexity K_A itself is not computable. It is the lower bound of a sequence of computable functions. Kolmogorov order is not computable as well.

Kolmogorov order of naturals cardinally differs from the natural order in the following sense: it puts in the initial segments very large numbers that are at the same time Kolmogorov simple (for example 2^k, 2^{2^k}).

This can be compared with properties of natural language, which are usually discussed as a result of historical accidents but at least partially are related to possibility to express complex meanings in short way — abundance of synonyms and senseless grammatically correct texts.

Zipf's law and Kolmogorov order. Frequencies of words of a natural language in texts. If all words w_k of a language are ranked according to decreasing frequency of their appearance in a corpus of texts, then the frequency p_k of w_k is approximately inversely proportional to its rank k: $p_k \sim k^{-1}$.

Zipf: this distribution "minimizes effort". Gibbs distribution with energy proportional to $\log k$ gives a power law.

How minimization of complexity leads to Zipf's law (Yu.I.Manin). A mathematical model of Zipf's law is based upon two postulates:
(A) Rank ordering coincides with the Kolmogorov ordering.
(B) The probability distribution producing Zipf's law is Gibbs distribution with energy equal to logarithmic Kolmogorov complexity $e^{-K(w)}$.

Zipf's law is described by the statistical sum

$$\sum_w e^{-zK(w)}$$

with the inverse temperature $z = 1$.
For natural numbers, since for majority of naturals logarithmic Kolmogorov complexity of n is close to $\log n$, the statistical sum is "similar" to zeta function

$$\sum_n e^{-zK(n)} \approx \zeta(z) = \sum_n n^{-z}. \tag{2}$$

Point $z = 1$ is the point of phase transition: for $z > 1$ the series for zeta function converge and for $z \leq 1$ diverge.

Exponent -1 in the Zipf's law — phase transition in the model where (logarithmic) Kolmogorov complexity is energy.

Scaling in genomics

Eugene V. Koonin, The Logic of Chance: The Nature and Origin of Biological Evolution, Pearson Education, 2012.

Eugene V. Koonin, Are There Laws of Genome Evolution? PLoS Comput Biol. 7(8): e1002173 (2011).

Universals of Genome Evolution:
1) log-normal distribution of the evolutionary rates between orthologous genes,
2) power law-like distributions of membership in paralogous gene families,
3) scaling of functional classes of genes with genome size.

E.Koonin: scaling in genomics should be described by some model of statistical mechanics — interacting gas of genes.
The "**third evolutionary synthesis**"
(the first is Darwinism, the second is Darwinism plus genetics, and the third should generalize Darwinism with genomics data).

Our Claim. *Scaling in genomics should be related to Zipf's law. The corresponding statistical mechanical model should contain a contribution from complexity in energy.*

Complexity as energy in biological evolution

Set of biological sequences (genes, part of genomes, total genomes). Structure of a constructive world on the set of biological sequences:

Finite set S of sequences (genes, regulatory sequences, etc.), and a finite set O of genome editing operations with contains operations of gluing together sequences and operations similar to typical evolutionary transformations (point mutations, insertions, deletions (in particular insertions and deletions of genes $s_i \in S$), duplications of parts of a sequence, etc.).

To elements $s_i \in S$ and $o_j \in O$ we put in correspondence positive numbers $w(s_i)$ and $w(o_j)$, called scores (or weights).

For a sequence s obtained from elements in S by applications of operations in O we put in correspondence score $w(s)$ equal to a sum of scores of elements $s_i \in S$ and operations $o_j \in O$: composition of o_j generates the sequence s starting from s_i. Sequence s can be obtained in this way non-uniquely. Complexity of s — minimum over possible compositions of operations $o_j \in O$ and elementary sequences $s_i \in S$ giving s

$$K_{SOW}(s) = \min_{A(s_1,\ldots,s_n)=s} \left[\sum_i w(s_i) + \sum_j w(o_j) \right] \tag{3}$$

where A is a (finite) composition of o_j applied to sequences $s_1, \ldots, s_n$.

Conditional version $K_{SOW}(s'|s)$ of complexity (3)

$$K_{SOW}(s'|s) = \min_{A(s_1,\ldots,s_n)[s]=s'} \left[\sum_i w(s_i) + \sum_j w(o_j) \right], \quad (4)$$

we generate sequence s' starting from sequence s.
$A(s_1, \ldots, s_n)[s]$ is a combination of genome editing operations containing sequences $s_1, \ldots, s_n$ applied to s.

Complexity $K_{SOW}(s)$ — weighted number of genes and edit operations generating sequence s. Logarithmic Kolmogorov complexity (approximately) — number of computational operations generating element of a constructive world.
Complexity (3), (4) gives weighted version of estimate from above for logarithmic Kolmogorov complexity.

Model of the Third Evolutionary Synthesis — constructive statistical mechanical system with states s — sequences, generated (constructed) as above, statistical sum

$$Z = \sum_s e^{-\beta H(s)}, \quad H = H_F + H_K \tag{5}$$

where β is the inverse temperature and the Hamiltonian contains two contributions:
$H_F(s)$ describes biological fitness of a sequence s,
$H_K(s)$ describes complexity of s (for the described above example $H_K = K_{SOW}$).
Here good fitness corresponds to low $H_F(s)$ (potential wells on the fitness landscape). The symbol K in H_K is for Kolmogorov (complexity).
The contribution $H_K(s)$ describes the evolutionary effort to generate the sequence s (sequences with less evolutionary effort are more advantageous).

Remark. Complexity of sequences grow sufficiently fast with addition of sequences from S and application of editing operations from O, hence for sufficiently low temperatures (large β) the constructive statistical sum (5) converges.

Scaling in genomics — Zipf's law

Using (3), (5), the scaling in genomics, in particular power law–like distributions of membership in paralogous gene families, can be discussed as a consequence of Zipf's law. Paralogous genes are genes in the same genome generated by duplication events.

Let us assume that if the genome contains a paralogous family of genes with N elements, then the Kolmogorov rank of this genome should be proportional to N (since the complexity (3) in this case will contain N contributions $w(s_i)$ for some gene s_i). Then by Zipf's law (2) contribution of this genome to the statistical sum will be proportional to N^{-z} which gives the power law.

Evolution and machine learning

A problem of biological evolution can be considered as a problem of learning where genomes learn in the process of natural selection. Contribution $H_F(s)$ in (5) (biological fitness of sequence s) can be modeled by the functional of empirical risk (number of errors on a training set)

$$H_F(s) = R_{\text{emp}}(s) = \frac{1}{l}\sum_{j=1}^{l}(y_j - f(v_j, s))^2.$$

Here genome s generates a classifier $f(v, s)$ which models biological function — it recognizes the situation v and classifies it (gives 0 or 1 for $f(v, s)$). Here $(y_1, v_1), \ldots, (y_l, v_l)$, $y_j \in 0, 1$ is the training set.

Machine learning — joint minimization of empirical risk and regularization term

$$R_{\mathrm{emp}}(\mathrm{classifier}, \mathrm{training\ set}) + \mathit{Reg}(\mathrm{classifier}).$$

Regularizing contribution describes some kind of complexity of a classifier. Regularization reduces overfitting.

Vapnik–Chervonenkis theory (or VC-theory) states that a classifier can be taught if the family of classifiers has sufficiently low VC-entropy (which is some kind of complexity).
Presence of the complexity contribution H_K in Hamiltonian (5) can be considered as a regularization by low Kolmogorov complexity in the model of learning by evolution.

Minimization of Kolmogorov complexity in learning theory with applications to low complexity art and music

J. Schmidhuber, Discovering neural nets with low Kolmogorov complexity and high generalization capability, Neural Networks 10 no 5, P. 857–873 (1997).

Summary

We have discussed the application of Yuri Manin's idea on relation of the Zipf's law and Kolmogorov order *(complexity as energy)* to biological evolution — the Hamiltonian of evolution should contain a contribution given by (weighted estimate from above for) Kolmogorov complexity — weighted number of elementary evolutionary operations (evolutionary effort).

Zipf's law in this approach should be related to scaling laws observed in genomics. The third evolutionary synthesis.

Modeling of evolution by machine learning approach — biological fitness is the functional of empirical risk and Kolmogorov complexity term is the regularization to reduce overfitting.