V.A. Lyubetsky, A.V. Seliverstov

Mathematical problems in biological evolution and molecular regulation

Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow

Only recently have processes in molecular biology been formalized. Despite of simpler mathematics behind them and apparent biological necessity, none has been studied with rigor. Mathematics has become superseded by computer modeling that advanced in the last years. The processes discussed in this abstract actually describe important biological phenomena at the molecular level, unlike many other publications that distantly relate to biological reality, for details see [1].

In areas 1-4 we developed original computer models that produce predictions close to real measurements, often to the precision of experimental error. In areas 5-6, on the contrary, already designing an effective modeling is a problem. The qualities discussed in areas 1-6 cannot always be measured experimentally, and are therefore estimated indirectly. Understanding biological terms below is not important for mathematical understanding of areas 1-6.

A genome is a long sequence of millions (in bacteria) or billions (in animals) of characters in the 4-letter alphabet $\{A, T, G, C\}$, and a gene is a directed (in one of the two orientations) short region within it. The number of genes may be many thousands (from 8-9 in bacteria to 50 in animals), hundreds (bacteria, plastids) or just a few (mitochondria). Shorter regions in between the genes regulate their activity, they are also directed and encode complex structures. Thus, a genome can be viewed as a set of genes and regulatory regions (”regulatory systems” or “regulations”). The rest of the genome, so called junk regions of unknown function, is not used in modeling. In many cases the character composition of the genome, gene and regulatory region is not used in modeling either.

An organism is a genome that evolves during lifetime, a species is a collection of genomes with similar characteristics, which provides for breeding compatibility and production of descendant genomes (the progeny).

All processes run in physical time, however discrete time is as yet used in models for a seeming simplicity.

1. Competing processes of binding and movement (competition of RNA polymerases). Given is a sequence in the 4-letter alphabet with directed regions of two types: genes and promoters. The mutual arrangement of promoters and genes can be arbitrary but is fixed. Each promoter, if *available,* is bound by a molecular machine (the polymerase) of a certain type out of a fixed finite set of types. A polymerase of each type has a fixed type-specific length and moves along the sequence in the corresponding direction of promoter. Many polymerases concurrently bind the sequence and move each in its direction. The promoter is *available* if none of polymerases overlaps with its sequence. The gene is “read” if a polymerase moved from its beginning to the end. The gene’s reading frequency is its *transcription level*. Each promoter, for each polymerase type, is characterized by the intensity of *binding attempts*. The polymerase concentration is assumed sufficient, thus the intensity is only a function of the promoter quality, for each polymerase type. An attempt is *successful* if at the instance of its realization the promoter is available. For certain polymerase types, binding is followed by the abort process: an alternation of movement at a fixed finite rate in the corresponding direction of promoter at an arbitrary (e.g., exponentially distributed) distance and instantaneous return to the initial position. Such alternations occur an arbitrary (e.g., geometrically distributed) number of times until the polymerase reaches at a threshold distance from the promoter. At this instance the polymerase detaches from the promoter, its size instantaneously decreases by a fixed value and movement continues in the same direction. Other polymerase types lack the abort process: the movement initiates immediately after binding, the size does not change. Binding attempts are allowed to form a Poisson process, with a polymerase moving at a predetermined rate fixed for each type until colliding with another polymerase. If two polymerases moving in the same direction collide, their rates become equal to that of the leading polymerase until it is attached to the sequence (“elongates”). In case of a front collision both polymerases detach (“terminate”). Under this model several biologically meaningful questions can be formulated. For example: inferring transcription levels of all genes given the binding attempt intensities of all promoters; inferring binding attempt intensities that best approximate given gene transcription levels; inferring binding attempt intensities that best approximate known changes of gene transcription levels under wide fluctuations of temperature and polymerase rates (described by simple combinations of affine functions). Although we obtained a rigorous description of the stochastic movement of the polymerase, such biologically preferred assumption causes large difficulties even for modeling. A computer realization of the model is available at http://lab6.iitp.ru/ru/rivals, [2]. The problem is largely simplified into a special case of the counter-flow theory with annihilation by assuming no abort processes, equal rates of polymerases, and zero sizes of polymerases and promoters. However, this case is biologically irrelevant.

Further difficulties arise if the sequence is replaced by a circle, i.e., a sequence modulo its length. The simplest case is a circle of 17.000 characters (a human mitochondrial genome), whereon competition occurs only among polymerases of the same type, and only three promoters are located at positions 407 counterclockwise, 561, and 646 clockwise. Abort processes are absent. Initially, polymerases do not complete the circle, their counter-flows from the three promoters collide and the polymerases detach. Genes distant from the promoters have nearly zero expression levels, which contradicts biological observations. This is an unstable state: one of the promoters realizes by 10-20 more bindings, the extra polymerases avoid collisions and complete the full circle including the initial promoter. It simulates the increasing number of successful bindings and increases the number of polymerases completing the circle in one direction. If another promoter also receives enough bindings, the movement in opposite direction may become more successful. The directions are rarely swapped several times, and a winning direction rapidly establishes. When the amount of polymerases moving in one direction reaches a threshold, the intensity of effective binding to one promoter and the transcription levels of its downstream genes continue to increase until polymerases occupy the entire circle and spaces between them become less than the polymerase size. Usually the circle sequence contains regions with “passing terminators”, which are sites that allow through a certain average amount of polymerases in each direction. This process dynamics including bifurcation points is to be described. Changes in characters (mutations) leading to terminators misfunction may cause severe human health disorders. The mutual arrangement of promoters and genes varies widely as well. “Passing terminators” also occur in straight sequences.

A competition of another type occurs when two promoters are overlapping or very close in the sequence, which causes spatial interference between binding polymerases in some 3-dimensional neighborhood. Many particular questions remain, such as inferring the average length of the polymerase run, asymptotic distribution of the lengths, etc.

2. Reconciliation of a set of trees (resulting in a species tree). Genes and regulatory systems are part of an organism (species), which is thus a set of genes and regulatory systems. Although genes and regulations evolve inside and with the species, their evolutionary patterns often do not coincide. Of fundamental importance is to develop a concept that describes genes, regulations and species in continuous time. Commonly, the evolutions of genes and regulations are considered independently in discrete times and reconciled “against” the evolution of species. Each evolution (gene, regulation, gene-and-regulation, species) is represented as a tree. Define the gene evolution as tree G_i («gene tree»). A set of gene trees $\{G_i\}$ is given. The sought is tree S («species tree») nearest in average to trees in $\{G_i\}$. The solution: each G_i is assigned the value $c(G_i,S)$ of its difference from unknown S, and the functional $c(\{G_i\},S) = \sum_i c(G_i, S)$ in minimized over variable S. Define $c(G_i,S)$ as the amount of difference between the evolutions of gene and species, i.e., between G_i and S. Such definition requires to determine a list of evolutionary events and to correlate discrete times along trees G_i and S. To do so one needs a mapping of vertices in G_i into vertices and edges in S (“scenario of the evolution of gene G_i along species tree S”). Our original algorithms to solve the above tasks have at maximum cubic (a very low) complexity and are available at http://lab6.iitp.ru/ru/super3gl, [3,4]. In our solution, the unknown S and gene evolutionary scenarios are built with induction as the cardinality of set V of leaves in S increases. At each induction step, trees S_1 (with leaves set V_1) and S_2 (with leaves set V_2), and their corresponding sets of scenarios f_1 and f_2 are already known. The trees are merged into a larger tree S_1+S_2 with combined scenarios f_1+f_2 such that the value $c(\{G_i\},S_1+S_2)$ is minimal against all possible partition of V into V_1 and V_2. The same principle is used to build a gene scenario along the given S, when component trees are subtrees in S, [3]. The subtrees need to be rooted within the same time slice. We developed an algorithm to impose time slices on the tree edges, [3]. Instantaneous events are allowed between edges within a time slice. Incorporating continuous time in the approach is likely to relieve uncertainties with justification of time slices.

3. Reconstruction of secondary structures along a tree (the example of attenuation regulation). Certain regulatory regions (*primary structures*) after being copied into the outside of the genome fold into *secondary structures* (SS, ref. to Figs 1-3). Each SS is a pairing of characters, $A – T$ and $G – C$, maintained by hydrogen bonds and stacking interactions of neighboring pairs. Fig. 1 shows a part of SS, the “helix”. Biological SS may contain up to several thousands of helices in a complex combination. The pairing occurs between regions (“shoulders”) of certain length (6, 3, and 4 characters in Fig. 1). A helix consists of several paired regions, “hypohelices”: two longest continuous shoulders connected by a loop (Fig. 1 shows three nested hypohelices with loops of 25, 18 and 6 characters). Certain genes are regulated by specific types of SS. One such type is attenuation regulation partly depicted in Fig. 2 (two alternative helices). Given is tree S of species or regulation factors with primary structures assigned to the leaves. Although some primary structures have experimentally known SS’, secondary structures are not given in our approach and need to be reconstructed. Known SS’ are used for verification purposes. To be found is the evolutionary inference of the distribution (*configuration*) of primary structures in internal vertices, and SS’ in all vertices of tree S. The solution realizes a Gibbs approach with energy functional $H(\sigma)$, which global minimums define the sought configurations σ'. Global minimums are found with annealing based on the Metropolis–Hastings stochastic dynamics. The functional $H(\sigma)$ is a sum of three terms. The first term defines the energy of pair interaction between the two next primary structures on each edge. More specifically, it defines the standard dynamics of the primary structure: the probability of a character substitution according to a fixed transition rate matrix, and also the probabilities of insertion/deletion of a word of any length at any position in the primary structure. At each position, the evolution rate is considered according to the gamma law. The second term defines conservativity of the secondary structure along each edge and entire paths in tree S that is specified by a sophisticated potential of non-local interaction. The third term defines the presence of other elements pertinent to the regulation of interest (e.g., the “leader peptide gene”). The first and second terms require a pairwise alignment to be found: primary structures at the ends of each edge to be aligned for the first term computation, and secondary structures – for the second term. For the latter, we developed a procedure that aligns the secondary structures of two primary structures. The algorithm is realized as a heterogeneous Markov chain, with transition probabilities being functions of current configuration $\sigma(n)$ and temperature parameter β_n. Let the chain start with any configuration $\sigma(0)$ and $\beta_n \to \infty$ such that $\lim(\log n / \beta_n) > C$. Then $\sigma(n)$ converges over probabilities to one of minimal σ', thus describing all globally minimal configurations. The algorithm is available at http://lab6.iitp.ru/ru/anneal, [5].

4. Competition of two processes (transcription and translation, the case of attenuation regulation). Two machines, a polymerase and a ribosome, move on a sequence. The ribosome recognizes and binds a specific site (like a promoter) upstream a specific gene (the “leader peptide

gene") after the polymerase had already bound to its promoter and moved forward. If the ribosome catches up with the polymerase, their rates become equal, and the polymerase is not affected. The ribosome rate is function $v(c)$ of concentration c of certain substance (amino acid) and does not exceed 45 characters/sec. The region between the two machines forms secondary structure ω with the minimal energy (by definition in our model) that decreases the polymerase rate according to function $v(\omega)$ (at no SS the rate is 42 characters/sec). If the polymerase decelerates at a *T*-rich region, its binding strength weakens and it detaches from the sequence ("transcription termination"). Given are a sequence and two functions, $v(c)$ and $v(\omega)$. The functions define the instantaneous positions of the leading polymerase and the following-up ribosome on the sequence. To be found is correlation $p(c)$ between transcription termination frequency and concentration c. Usually $v(c)$ is found according to the Michaelis-Menten law, while determining $v(\omega)$ is much more sophisticated. Our solution is available at http://lab6.iitp.ru/rnamodel/runmodel.php?lang=rus, [6].
Two special problems below are of high importance and have to be addressed.

4.1. How to determine the binding strength of a molecular machine (a polymerase, ribosome, etc.) with a sequence that it moves on; what is the effect of SS? Evidence exists that the binding strength decreases with deceleration. How the SS lowers the rate and how deceleration decreases the strength is unknown, [6].

4.2. Ample experimental observations exist but no theoretical explanation. How to categorize complex SS with many pseudoknots; how to determine energy of a given SS? Little is evident on how to classify pseudoknots and decompose a SS into elementary SS'; what is the list of elementary SS'. Consider the simplest case when the SS consists of one helix (Fig. 1). Recall that a helix consists of several hypohelices. We estimated the helix energy as the sum of the bond energy $\frac{1}{RT}\cdot\sum_i E_i$ and loop energy $\sum_i\left(1.77\cdot\ln(l_i+1)+B+\frac{C}{l_i}\right)$, where i varies over all hypohelices of the helix and E_i is the energy of *i*-th hypohelix determined from the experimentally known hydrogen bonds and stacking energies; l_i is the loop length of *i*-th hypohelix; B and C are constants, [6].
Another challenge is to decompose a huge SS space into clusters ("macrostates") and then calculate clusters energies. Such decomposition is to be effective. Consider a parentheses structure where each pair of parentheses corresponds to a hypohelix and is tagged with the number of its parental helix (Fig. 3). The parentheses are interpreted as follows: consecutive hypohelices correspond to consecutive pairs of parentheses, $(\ )_1(\ )_2$; an overlap of one hypohelix with the loop of the next is represented by the nested structure $(()_1)_2$. The notation is applicable to simple pseudoknots: $(_1(_2)_1)_2$. The macrostate is a set of all SS ("microstates") defined by a given parenthesis structure; this set must not be empty.

5. Combination of three- and one-dimensional diffusions. A promoter is very short (dozens of characters at max) comparing to a typical sequence (several million characters in bacteria), which raises a question of how the polymerase finds its specific promoter in the space of the cell. The sequence (the DNA molecule in the cell) has a peculiar spatial geometry, like the Jordan curve in the square, and this arrangement is functional. In current views, initially the polymerase binds weakly ("non-specifically") to the closest region of the sequence and moves in one of the two randomly chosen directions for a random short period of time. This is the one-dimensional diffusion along the curve. If the binding is too weak or a collision takes place, the polymerase detaches and again non-specifically binds to the next spatially close region of the sequence, which may be very distant lineally in the curve. Thus, the three- and one-dimensional diffusions switch until the polymerase finds its promoter where it binds strongly ("specifically"). To be studied is such alternation of diffusions taking into account the type or only characteristics of the curve. Ample experimental evidence exists with no sound theoretical bases.

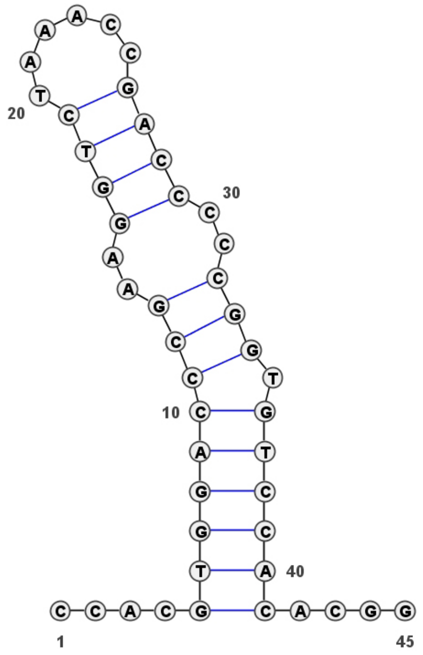
6. Origin of species (speciation). The genome is represented by sequential characteristics x, where i-th position contains number m_i of different genes, each represented by exactly i copies (copies are also genes). Numbers m_i are nonnegative, all integers or real; starting from a certain position x contains only zeros. Define $m(x) = m_1 + m_2 + ...$ as the number of all gene types and $n(x) = m_1 + 2m_2 + ...$ as the total number of genes in "genome" x. Define ***X*** as the space of all allowed sequences x and $f(x,t)$ as the density of genomes in point x at time t. Note that genes and genomes are represented in the model only via their characteristics x. The following transitions (events at the gene or genome level) are allowed in point x:
1) $<..., m_i, ...> \to <..., m_{i-1}+1, m_i-1, ...>$ loss of a gene from m_i, if $i \neq 1$ and $m_i \geq 1$, and $<..., m_i, ...> \to < m_1-1, m_2, ...>$, if $i=1$ and $m_1 \geq 1$; if $m_i = 0$ or $m_1 = 0$, then the transition is forbidden. 2) $<..., m_i, ...> \to < m_1+1, m_2, ...>$ transfer, i.e. gain of a single copy of a new gene. 3) $<..., m_i, ...> \to <..., m_i-1, m_{i+1}+1, ...>, i \neq 1$ duplication of a gene from m_i; if $m_i \geq 1$ is not true the transition is forbidden. 4) $<..., m_i, ...> \to < m_1+1, ..., m_{i-1}+1, m_i-1, ...>$ mutation of a gene from m_i, if $i \neq 1$, and $<..., m_i, ...> \to <..., m_i, ...>$, if $i=1$; if $m_i \geq 1$ is not true the transition is forbidden.
Each transition is assigned a x-dependent vector of transition rate (intensity). Let $A(x)$ be the sum of vectors that defines the vector potential. Scalar potential is denoted $-V$, where $V(x,t)$ reflects the internal congruence («survival») of a genome in x at t. Both potentials depend on parameters, including the main $m(x)$ and $n(x)$; some parameters are unknown and varied. Let $V(x,t)$ belong to class ***V*** of functions with low chaotic maxima. According to natural interpretation of V, at V-max points genomes possess certain selective advantages to survive during dynamics described by $A(x)$ but the result is still not pre-defined. Scalar potential $V(x,t)$ changes dynamically in space ***V***. The exact choice of class ***V***, as well as how $V(x,t)$ depends on time, is a matter of study. According to one possible representation, sharp perturbations occur at certain times t_i that form a Poisson distribution with parameter μ. Such changes in survival conditions correspond to transitions from $V(t_i)$ to $V(t_{i+1})$ through smooth changes of the set of local maxima in $V(t_i)$ according to a distribution with parameter λ (a noise). Are there such natural distributions and values of parameters μ and λ that would result in the formation of clusters (biologically, species) in space ***X***? More specifically, we aim at describing parameter regions, for which there exists time t_0 after that trajectories acquire the property «almost all mass $M(t) = \int f(x,t)dx$ concentrates in several disjunctive clusters in ***X***», (*). The clusters describe species. Their number can be estimated via the number of extant species in the problem statement. Then t_0 corresponds to the time of species formation (origin), i.e., speciation. Modeling provides the estimates of parameter regions, for which property (*) is true.
Our model does not incorporate a biologically more relevant genome representation as a sequence of natural numbers with repeats, where each number is the name of a gene. Such system is described by a more complex dynamics.

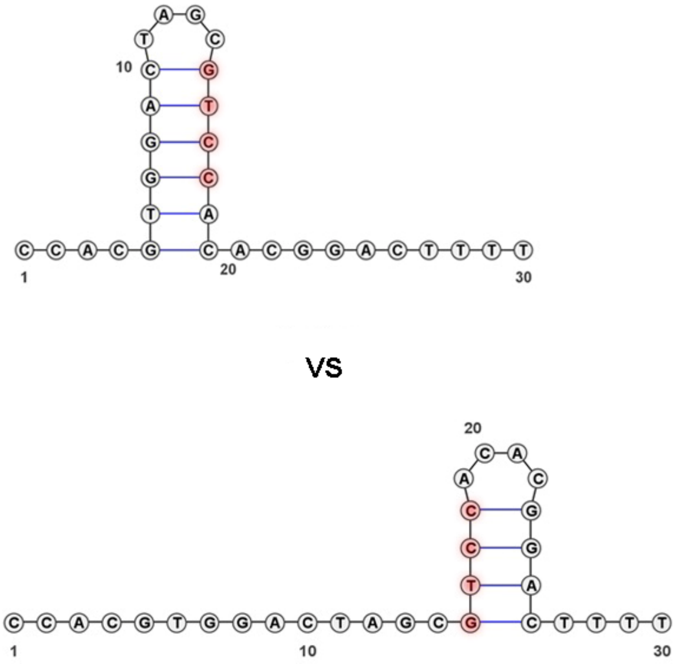
6.1. The dynamics of $x = x(t)$ can be alternatively described with equation $x' = A(x) + \varepsilon\xi$, where ξ is a noise with a certain potentials-dependent generator, and ε is a parameter. It can be assumed that there exists time t_0, after which a finite number of massive clusters exists with centers of mass $x_1, x_2, ...$, with transitions requiring exponentially long times or impossible (**). Then x_j describe the formed species, and t_0 – the time of speciation. A theory like the Ventsel-

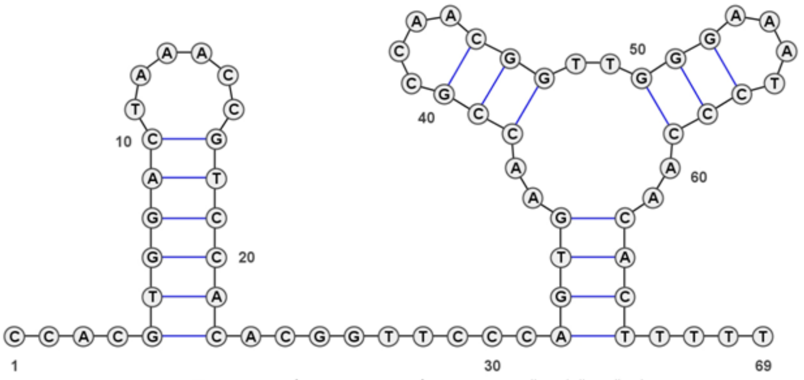
Freydlin theory allows for such function $\varphi(x)$ that $\varphi'(x)=0$ is a necessary condition of (**). Then x_j can be found from the equation.

The authors are grateful to L. Rusin, K. Gorbunov, L. Rubanov, S. Pirogov, E. Zhizhina for detailed discussions and involvement in some of the described areas of research. The authors thank L. Rusin for help with preparing the text.

The Figures below are given in the order of numbering.







Parentheses scheme: $()_1(()_2()_3)_4$

References

[1] Lyubetsky V., Gorbunov K., Rusin L., V’yugin V. Algorithms to reconstruct evolutionary events at molecular level and infer species phylogeny. A chapter in the book: Bioinformatics of Genome Regulation and Structure, II. Springer Science & Business Media, Inc. 2006, p. 189-204.

[2] V.A. Lyubetsky, O.A. Zverkov, L.I. Rubanov, A.V. Seliverstov Modeling the RNA polymerase competition: the effect of sigma-subunit knockout and temperature on gene expression, Biology Direct, 2011, 6:3, http://www.biology-direct.com/content/6/1/3.
[3] K. Yu. Gorbunov, V. A. Lyubetsky. Reconstructing the evolution of genes along the species tree. Molecular Biology, 2009, vol. 43, No. 5, pp. 881-893.
[4] K.Yu. Gorbunov, V.A. Lyubetsky, The tree nearest in average to a given set of trees, Problems of Information Transmission, 2011 (in print).
[5] V.A. Lyubetsky, E.A. Zhizhina, L.I. Rubanov, Gibbs field approach for evolutionary analysis of regulatory signal of gene expression, Problems of Information Transmission, 2008, vol.44, No.4, pp. 333-351.
[6] V. Lyubetsky, S. Pirogov, L. Rubanov, A. Seliverstov, Modeling classic attenuation regulation of gene expression in bacteria, Journal of Bioinformatics and Computational Biology, 2007, vol 5, no 1, pp. 155-180.