ISSN 0026-8933, Molecular Biology, 2007, Vol. 41, No. 5, pp. 836–842. © Pleiades Publishing, Inc., 2007.
Original Russian Text © K.Yu. Gorbunov, V.A. Lyubetsky, 2007, published in Molekulyarnaya Biologiya, 2007, Vol. 41, No. 5, pp. 918–925.

MATHEMATICAL
AND SYSTEM BIOLOGY

UDC 575.852

Reconstruction of Ancestral Regulatory Signals along a Transcription Factor Tree

K. Yu. Gorbunov and V. A. Lyubetsky

Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127994 Russia; e-mail: gorbunov@iitp.ru

Received October 19, 2006

Accepted for publication January 28, 2007

Abstract—A model and an algorithm were developed to reconstruct the ancestral regulatory signals, first and foremost, for DNA–protein interactions, at inner nodes of a transcription factor phylogenetic tree on the basis of the modern signal distribution. The algorithm simultaneously infers the evolutionary scenario as a set of tree edges along which the signal diverged to the greatest extent. The model and algorithm were tested with simulation data and biological findings on the NrdR, MntR, and LacI signals.

DOI: 10.1134/S0026893307050172

Key words: evolutionary scenario, regulatory signal, frequency matrix, evolution along a tree, transcription factor tree

FORMULATION OF THE PROBLEM

Two well-known problems are the constructing of the evolutionary trees for a protein family and a family of species and the reconstructing of the molecular events taking place during evolution of a protein family [1–4]. Since a family of transcription factors evolves together with their binding sites, a natural problem is to reconstruct the ancestral binding sites for a particular transcription factor, that is, to ascribe certain sites as ancestral to the inner nodes of a transcription factor tree *G*, based on the known sites at the terminal nodes. A similar problem is to reconstruct some significant characteristics (e.g., nucleotide frequencies) for the sites at the terminal nodes. It is assumed that the binding site abruptly changed along some edges of the tree *G*, for instance, as a result of changes in the transcription factor (what changes are regarded as abrupt depends on the threshold; the exact definition is given in the next section). Hence, part of the problem is identifying the edges corresponding to abrupt changes. These edges are termed (evolutionarily) significant and the set of edges is termed a carrier of the evolutionary scenario. In addition to the carrier, the evolutionary scenario includes the arrangement of the reconstructed (ancestral) sites or their particular characteristics (e.g., frequency matrices) through all nodes of the tree *G*. It is the pair of a carrier and an arrangement that is to be determined, as they are interdependent. The arrangement arises as total changes at all edges beyond the carrier of a scenario are minimized, reflecting the maximum parsimony principle. The principle implies that, apart from relatively few edges corresponding to significant evolutionary events, the changes along all other edges are as smooth as possible [5]. The formulation of this problem is illustrated by Examples 1 and 1a (see Results and Discussion).

Consider the simplest protein–DNA regulatory signal and, accordingly, the binding sites for an activator or repressor protein. Then, given is the evolutionary tree *G* of a certain transcription factor, with all terminal nodes having the sets of protein–DNA interaction sites found in the leader regions of homologous genes regulated by the factor. A terminal node with a multiple sequence alignment of the corresponding sites is termed a taxon, meaning that the node is actually ascribed with a set of species whose genes were found to contain these sites in the upstream regions. In our data, such sites are of the same length or are palindromic with the lengths differing by unity. When all sites are of the same length, they are written one above the other. When a set includes palindromes of an even and an odd length, one gap is added at the center of each even-length palindrome and the sites are similarly written one above the other. A trivial multiple sequence alignment arises at each terminal node in either case. When sites ascribed to one taxon considerably differ in length and have low similarity, we consider only some positions (the same number of positions for each site), assuming that the significant positions are known for each site. In some cases, a nontrivial multiple sequence alignment of such sites is obtained using one of the standard algorithms. Our algorithm works with any multiple sequence alignments assigned to the terminal nodes (taxa), but the data are usually poorly linked in the last case and the results are difficult to interpret. Thus, every taxon is ascribed with its multiple site alignment with *n* columns, where *n* is hereafter taken as constant. It is possible to consider in the same sense the evolution of a nucleotide in one particular position of a site, rather than taking the total site.

The position frequency matrix can be obtained for each terminal node and considered as a site characteristic; this variant is discussed below. The position weight matrix is a matrix of dimension $4 \times n$, where four rows correspond to the four nucleotides and *n* columns correspond to *n* positions of a site or *n* positions of a multiple sequence alignment, which is the same in our case [6, 7]. The columns containing gaps are not considered here. There are typically several evolutionary scenarios, which are ranked by scenario quality (see below). The best-quality scenarios are compared with each other and with the evolutionary specifics of the transcription factor, which are inferred from the tree *G*.

Our algorithm predicts the arrangement of position frequency matrices in inner nodes of the tree *G*, as well as the evolutionary scenarios for the tree *G*. The algorithm was tested with simulation and biological examples.

Let a position frequency matrix of dimension $4 \times n$ (see above) be computed for each terminal node (taxon) from the corresponding multiple sequence alignment. First, we will consider one *i*-th column of such a matrix for all taxa; this column characterizes the frequency distribution of the four nucleotides in the *i*-th position of the initial signal, for which the matrix has been constructed. The problem is to ascribe all ancestral nodes of the tree *G* with similar distributions (which depend on *i*, where *i* is fixed) in the way best agreeing with the maximum parsimony principle. In our model, this principle is implemented in the form of a certain function *F*, which is the sum of all changes in all distributions over the total tree *G* and should have the minimal value. The arrangement of such distributions at all inner nodes of the tree for all *i* values, ranging from 1 to *n*, suggests the best arrangement of frequency matrices at the inner nodes of the tree *G*. The edge where the two distributions ascribed to its ends display an abrupt change according to a certain threshold is included in an evolutionary scenario, which depends on *i* and, accordingly, is termed the *i*-scenario. The algorithm generates several *i*-scenarios (see below) and identifies the best of them (usually one). The edges included in the best *i*-scenario are termed significant at the *i*-th position. It is assumed that the signal abruptly changed at the *i*-th position along such edges as a result of changes in the transcription factor or the site via point mutations, etc.

The carrier of the final scenario includes the edges that belong (by weight in some cases) to the carriers of the best *i*-scenarios obtained for multiple *i* values. The final scenario is determined as a carrier and the corresponding arrangement of matrices through all nodes of the tree *G* and is inferred from the distributions observed in the best *i*-scenarios for all *i* values. The final scenario can allow for the palindromic structure of the signal. To achieve this, the above function *F* is replaced by another function, *F*, which reflects not only that the sum of all changes in distributions is minimal but also that the distributions evolve via concerted changes in pairs of associated positions *i*–*j* of a palindrome.

A position of the signal is identified as conserved when the four nucleotides do not all occur at similar frequencies (and, consequently, do not all have the same binding constant), the similarity being evaluated using a certain threshold (for the association of position frequencies with binding constants, see [8]). Among all conserved positions, the algorithm selects those conserved with respect to one nucleotide, where this nucleotide significantly prevails over the others, and, similarly, the positions conserved with respect to two or three nucleotides. The positions conserved with respect to two or three nucleotides are written in the consensus as R (R = A or G), Y (Y = C or T), etc., according to the IUPAC code [9]. Positions conserved with respect to two or three nucleotides allow many mutations within a group, which are similarly acceptable for the given position. A position can be conserved at one evolutionary period (that is, in one coherent part of the tree *G*) and nonconserved at another [4]. Note that it is important to distinguish functional and evolutionary conservation.

DESCRIPTION OF THE MODEL AND ALGORITHM

Every inner node *v* of the tree *G* is ascribed with four variables: v_A, v_C, v_G, and v_T, which are the nucleotide frequencies in the distribution corresponding to node *v*. According to the maximum parsimony principle, function *F*, obtained as a sum of the distances between the two distributions corresponding to the ends of an edge in the tree *G*, is minimized for each *i*-th position individually; the known values of the above variables are inserted only for the terminal nodes (taxa). Two obvious constraints are imposed: the sum of the frequencies is unity at each node and all frequencies are nonnegative. Other constraints are also possible in the algorithm. The *F* summand corresponding to edge *u* is designated *F*(*u*).

We will describe a general scheme of the algorithm, refining the above terms, and will indicate the two particular distances (in other words, metrics) between distributions that were used to obtain the results below. The model and algorithm are applicable to other distances as well.

Function F is minimized with the above linear constraints, and the resulting solution is termed intermediate. This solution is used to determine the *i*-scenario. The list of edges with several first $F(u)$ values are considered, where the list size is limited by a numerical parameter, *vet*, and is arranged in the order of decreasing $F(u)$. It is hypothesized that these edges include evolutionary events and, consequently, are not covered by the maximum parsimony principle. Such edges are termed significant at the *j*-th step ($j = 1$ now). According to the list size, a total of *vet* variants is generated.

The same procedure is repeated for each of the *j*-significant edges u. Namely, the changed function F is minimized: the summand $F(u)$ corresponding to the current *j*-significant edge u is excluded from the function F and one new edge is added to the *i*-scenarios as above (step $j = 2$). The process is continued until the number of steps reaches another numerical parameter, *glub*; i.e., the last step in this part of the algorithm function is $j = glub$. This yields a total of vet^{glub} sets of consecutively excluded edges; each set is considered as a disordered set of *glub* elements. Such a set is termed the carrier of the *i*-scenario (at the given values of *vet* and *glub*). The arrangement generated at the last $j = glub$ step and the corresponding carrier of the *i*-scenario together constitute the *i*-scenario.

The algorithm gradually increases the parameter *glub*, starting from zero. In the example below, we considered two variants, with fixed values of both *vet* and *glub*, which are explicitly indicated in this case, and with a fixed *vet* value and *glub* automatically changed until an algorithm-stopping criterion is reached (see below).

The power of the *i*-scenario is the number of elements in its carrier, which is equal to the parameter *glub*. Generally speaking, the lower the *glub* value, the better the *i*-scenario.

Thus, a family of *i*-scenarios is generated; the quality of each *i*-scenario is characterized by two numerical values. One is the maximal value of the summand $F(u)$ for all edges u that have not been included in the *i*-scenario: the lower the value, the better the *i*-scenario. The other one is the sum $\tilde{F}(u)$ of $\tilde{F}$ values over all edges u that have not been included in the *i*-scenario, where $\tilde{F}(u) = \sum_{d=1}^{4} \sqrt{|u(d, 0) - u(d, 1)|}$ and d runs through the four frequencies in the distribution of $u(d, 0)$, corresponding to the start of the edge u, and $u(d, 1)$, corresponding to the end (root) of the edge u. The lower this value, the better the *i*-scenario.

The second parameter is termed main and the first one is termed accessory. The corresponding terms are used for the lists of *i*-scenario carriers arranged in the order of a decreasing parameter. For each *glub* value considered during its work, the algorithm generates the main and accessory lists. When an *i*-scenario carrier falls into the upper parts of both lists, this carrier is termed best for the given *i*. In other words, when the carrier ranking first in the main list falls in the upper part of the accessory list, this carrier is identified as a carrier of the best *i*-scenario at the given values of *vet* and *glub*.

The algorithm stops when the first and second quality characteristics of the best *i*-scenario decrease abruptly at a certain *glub* value and smoothly before and after this value (in terms of the corresponding thresholds). The intersection index of two sets is defined as a ratio of the power of their intersection to the power of their pooling. The intersection index of several sets is defined as the arithmetic mean of the intersection indices obtained for all set pairs. If *glub* reaches a certain threshold (e.g., 10) but the above algorithm-stopping condition is still not met, the algorithm stops and yields the *i*-scenario carrier corresponding to the *glub* value at which the maximal intersection index of all carriers of the best *i*-scenarios has been achieved among all *glub* values examined. A high intersection index achieved at least at one *glub* value is indicative of the well-grounded initial data. A combination of the carrier of the best *i*-scenario with the nucleotide frequency distribution generated at the step *glub* yields the best *i*-scenario.

In the simplest case, the distance between distributions is determined as a sum of square differences between the corresponding nucleotide frequencies. This is a quadratic function with the two linear constraints, indicated above, and its minimization is a problem of quadratic programming. This problem has only one solution. An exception is the case where the solution is achieved at a total section or a polyhedron; we demonstrated that such a case is impossible for our problem. The problem of quadratic programming can be solved with the available rapid algorithms, which work even with many thousands of variables and constraints. Such algorithms always complete their work and yield the exact solution. When the number of variables is relatively low, it is expedient to apply the simple gradient projection algorithm, which converges especially rapidly in our case, projecting onto the standard simplex intersection. This means that the variables are not repeated, change from 0 to 1, and their sum is 1.

A drawback of the above distance is that a large number of small changes in the signal may be preferred over a single substantial change upon minimization, which complicates the search for *i*-significant edges. The simplest distance free from this drawback is the Hellinger distance, which is a sum of the square differences between the square roots of the corresponding nucleotide frequencies. This function has other advantages. For instance, the penalty for a difference between two frequencies depends not only on their difference but also on their ratio: the difference between 0.8 and 0.9 has a lower penalty as compared with the difference between 0.1 and 0.2. To eliminate the square roots from the function F with the Hellinger distance, the roots of the initial variables are taken as new variables. Then, the function F becomes quadratic again, but the equality constraints are quadratic rather than linear as with the first distance. The new constraints have a shape of spherical cylinders and the total allowable set is an intersection between the positive segments of the cylinders, as in the case of the first distance.

The loss of the linearity of the constraints complicated the minimization; in particular, the only solution is no longer guaranteed. However, the projecting onto the available set is as easy as before, preserving the high speed of the gradient projection method. The only problem is to choose the starting point. We used two approaches. First, the starting point was chosen at random and the procedure was repeated many times. Second, the problem was preliminarily solved with the first distance and the resulting solution was used as a starting point in minimization with the Hellinger distance. The two variants yielded similar solutions.

The above drawback is similarly eliminated when the third distance is taken as a sum of the roots of the absolute differences between the corresponding frequencies, i.e., as the above function $\tilde{F}$. However, this function is not always differentiable and its minimization leads to certain difficulties. In view of this, this function is indirectly employed in our algorithm, as a means to evaluate the quality of the *i*-scenario.

Collation of the best scenarios obtained for different positions of the signal. The best *i*-scenarios obtained for different *i* values with *i* changing from 1 to *n* usually have different carriers. This is quite natural, since changes in the transcription factor and the binding signal often involve only some positions of the signal according to its palindromic structure. If an event has not dramatically changed the factor–signal binding constant, the regulation is preserved with a changed binding constant. For instance, when mutations occur at two positions, a decrease in the binding constant at one of them may be compensated by a decrease at the other. The final evolutionary scenario, which no longer depends on the position, includes the edges from different best *i*-scenarios. An edge is included when its weight exceeds a certain threshold; in the simplest case, the weight is determined as the number of the best *i*-scenarios containing this edge.

A more comprehensive definition of the edge weight takes into account the quality of the best *i*-scenario, the extent of conservation of the *i*-th position, the significance of the given edge in the *i*-scenario, and the frequency of the edge simultaneously occurring at associated positions, e.g., in palindromic pairs of the signal. This weight is hereafter referred to as the complete weight.

The algorithm has a special variant for working with a palindromic signal. The list of all palindromic position pairs *i,j* is fixed and the algorithm searches for the best *i,j*-scenarios coordinated by complementarity at positions *i* and *j*. Such an *i,j*-scenario reflects the evolution, taking the structure of the transcription factor-binding site into account. A new palindromic targeted function $\bar{F}$ is determined for this purpose as a sum of the previous targeted functions F for positions *i* and *j* taken separately plus the sum of penalties at all inner nodes of the tree G. The penalty at one node v is

$$2\mu[(v_{i,A}-v_{j,T})^2+(v_{i,G}-v_{j,C})^2+(v_{i,C}-v_{j,G})^2+(v_{i,T}-v_{j,A})^2].$$

Thus, the lack of complementarity at node v is penalized, where $v_{i,A}$ is the frequency of A in position i at this node and the other variables are defined similarly. The relative importance of the complementarity of distributions for the pair of associated positions *i,j* as compared to the stability of two separate distributions at positions i and j is regulated by the parameter μ in the above equation. The results described below were obtained with $\mu = 1$.

RESULTS AND DISCUSSION

Testing the algorithm with a simulation example. As a simulation tree G, we consider a binary tree with 64 terminal nodes (taxa), each branch from the root to a taxon consisting of 6 edges. The tree “grows” downwards. The taxa are numbered left to right from 1 to 64. The edges are designated as words combining 0 and 1 and reflecting the direction from the root to the edge end, where 0 means going leftwards and 1 means going rightwards. One of the following distributions is fixed for each taxon according to the rule indicated in the table:

(1) A: 5/8, C: 2/8, G: 1/8, T: 0; (2) T: 5/8, G: 2/8, C: 1/8, A: 0; (3) G: 5/8, C: 2/8, T: 1/8, A: 0;
(4) C: 5/8, A: 2/8, G: 1/8, T: 0; (5) T: 5/8, A: 2/8, C: 1/8, G: 0; (6) G: 5/8, T: 2/8, C: 1/8, A: 0.

Frequencies accepted in the simulation example for nodes of the initial transcription factor tree

Taxon	Distribution in the taxon
1–2	(4) = C: 5/8, A: 2/8, G: 1/8, T: 0
3–16	(2) = T: 5/8, G: 2/8, C: 1/8, A: 0
17–32	(1) = A: 5/8, C: 2/8, G: 1/8, T: 0
33	(5) = T: 5/8, A: 2/8, C: 1/8, G: 0
34–40	(3) = G: 5/8, C: 2/8, T: 1/8, A: 0
41–48	(1) = A: 5/8, C: 2/8, G: 1/8, T: 0
49–50	(6) = G: 5/8, T: 2/8, C: 1/8, A: 0
51–64	(1) = A: 5/8, C: 2/8, G: 1/8, T: 0

The following scenario with five significant edges was expected for a solution. Distribution (1) occurs at the root of the tree and changes to distribution (2) at edge 00, distribution (3) at edge 100, and distribution (6) at edge 11000. Distribution (2) changes to distribution (4) at edge 00000 and distribution (3) changes to distribution (5) at edge 100000. The algorithm was tested with many trees and many tables of such data. The result was always similar to that described below.

Thus, the algorithm yielded the following result. The best scenario in the main list has the carrier {100000} at *glub* = 1 (i.e., among all scenarios with one significant edge), the carrier {00, 00000} at *glub* = 2, the carrier {00, 100, 11000} at *glub* = 3; and the carrier {00, 00000, 100, 11000} at *glub* = 4. It is seen that the same edges are mostly repeated with increasing *glub*, suggesting a well-grounded initial data set. At *glub* = 5, the algorithm yields the best-scenario carrier {00, 100, 11000, 00000, 100000}, which serves as a solution, and the corresponding arrangement. It was obtained for the first time at *glub* = 5 that one scenario ranked first in both the main and accessory lists, and it is this scenario that was the first to have the two quality characteristics equaling zero, that is, reaching their minimum. For comparison, the first and second quality characteristics were, respectively, 0.44 and 9.3 for the first scenario of the main list and 0.08 and 46.1 for the first scenario of the accessory list at *glub* = 4. The above algorithm-stopping criterion terminated a further increase in *glub* at *glub* = 5.

To test the tolerance of the algorithm, random disturbances were introduced in the frequency distributions at the terminal nodes (taxa). For each taxon of the above tree, each nucleotide frequency in each distribution was increased or decreased by 0–0.1 at equal probabilities. The algorithm applied to the disturbed data set (Table 1) yielded the same best scenario at the same *glub* = 5, while the distributions at the nodes were slightly distorted. The scenario always ranked first in the main list and, in the accessory list, ranked first in 80% of the cases and second in the other 20%.

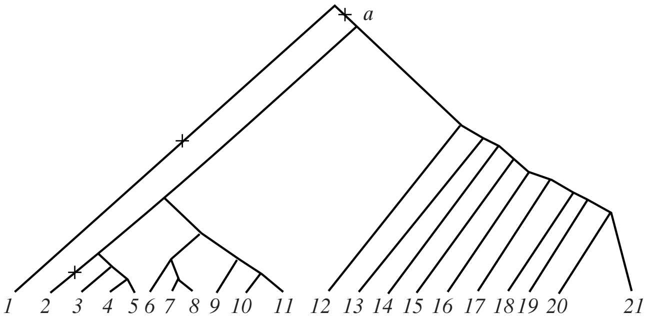

Fig. 1. Tree *G* of the species whose NrdR signals were examined. Taxa: *1* = {*Thermotoga maritima, Thermus thermophilus*}; *2* = {*Deinococcus radiodurans*}; *3* = {*Prochlorococcus marinus, Gloeobacter violaceus, Synechocystis* sp., *Synechococcus elongates, Thermosynechococcus elongates*}; *4* = {*Streptomyces coelicolor, S. avermitilis, S. scabies, Clavibacter michiganensis, Leifsonia xyli, Corynebacterium* spp., *Mycobacterium* spp.}; *5* = {*Propionibacterium acnes, Bifidobacterium longum, Thermobifida fusca*}; *6* = {*Staphylococcus aureus, S. epidermidis*}; *7* = {*Clostridium acetobutylicum, C. tetani, C. perfringens, C. botulinum, C. difficile, Thermoanaerobacter tengcongensis, Carboxydothermus hydrogenoformans, Desulfitobacterium hafniense*}; *8* = {*Bacillus subtilis, B. licheniformis, B. halodurans, B. cereus, B. stearothermophilus*}; *9* = {*Enterococcus faecalis, E. faecium*}; *10* = {*Streptococcus pyogenes, S. agalactiae, S. pneumoniae, S. mutans, Pediococcus pentosaceus*}; *11* = {*Lactobacillus* spp.}; *12* = {*Chlamydia muridarum, Chlamydophila pneumoniae, Chlamydia trachomatis, Chlamydophila abortus, C. caviae, Treponema denticola*}; *13* = {*Geobacter sulfurreducens, G. metallireducens, Desulfuromonas acetoxidans, Desulfotolea psychrophila, Bdelovibrio bacteriovorus, Bacteriovorax marinus, Myxococcus xanthus*}; *14* = {*Brucella melitensis, Mesorhizobium loti, Agrobacterium tumefaciens, Rhizobium leguminosarum, Sinorhizobium meliloti, Bradyrhizobium japonicum, Rhodopseudomonas palustris, Rhodobacter capsulatus, Caulobacter crescentus, Hyphomonas neptunium, Ehrlichia chaffeensis, Neorickettsia sennetsu*}; *15* = {*Nitrosomonas eutropha, Neisseria meningitidis, Methylobacillus flagellatus, Ralstonia solanacearum, Bordetella pertussis, B. bronchiseptica, B. avium, Burkholderia fungorum, Bu. cepacia, Bu. pseudomallei*, *Dechloromonas aromatica*}; *16* = {*Xylella fastidiosa*, *Xanthomonas axonopodis*}; *17* = {*Pseudomonas aeruginosa, P. putida, P. fluorescens, P. syringae*}; *18* = {*Vibrio cholerae, V. vulnificus, V. parahaemolyticus*}; *19* = {*Escherichia coli, Salmonella typhi, Klebsiella pneumoniae, Yersinia pestis, Y. enterocolitica, Ervinia chrysanthemi, E. carotovora, Photorhabdus luminescens*}; *20* = {*Pasteurella multocida*}; *21* = {*Haemophilus influenzae, H. ducreyi*}.

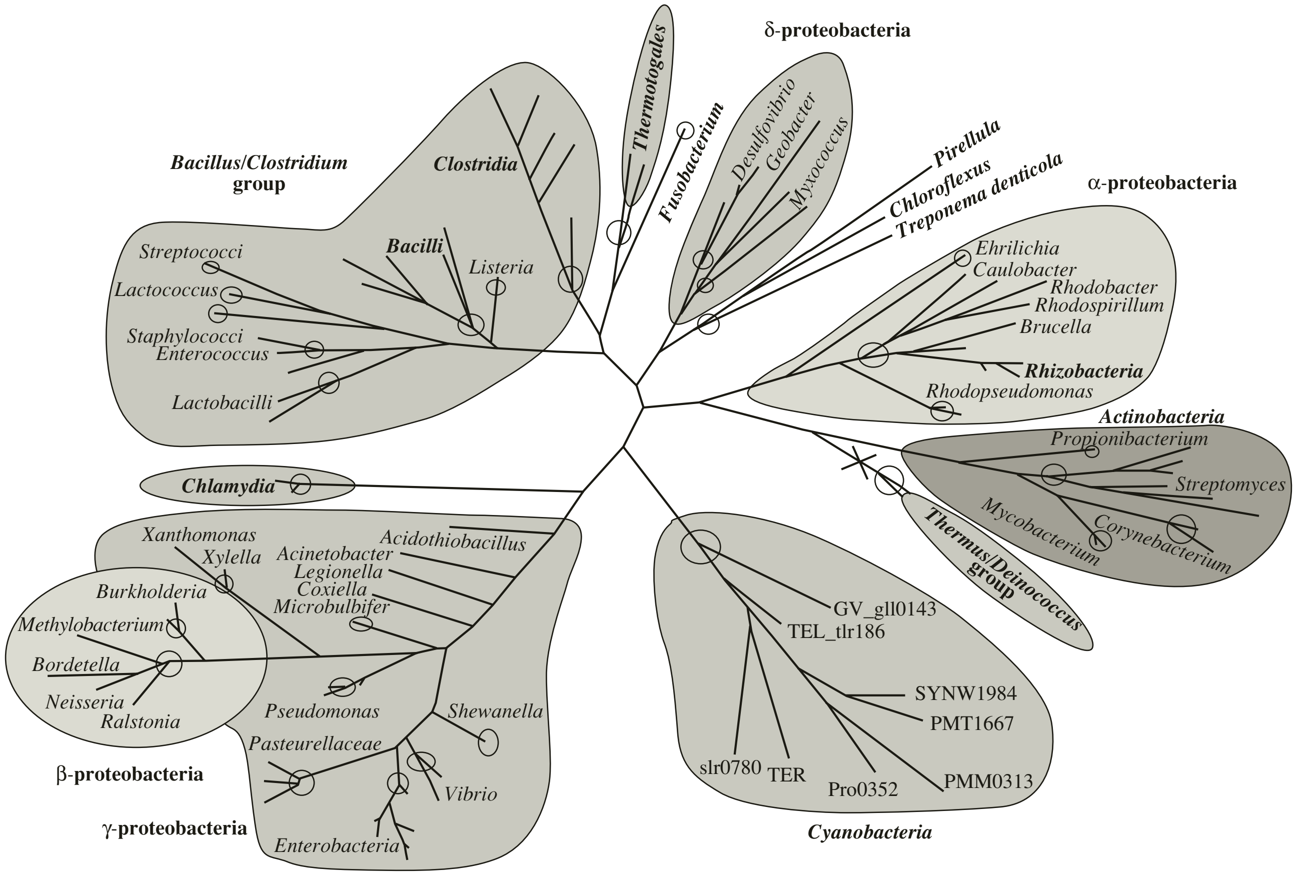

Fig. 2. Tree *G* of the taxa whose NrdR signal was examined. The taxa are indicated with circles. The number of signals considered is given after each taxon: *Fusobacterium*, 4; *Thermotogales*, 4; *Clostridia*, 25; *Listeria*, 4; Bacilli, 24; Streptococci, 52; *Lactococcus*, 6; Staphylococci, 10; *Enterococcus*, 15; Lactobacilli, 42; *Chlamidia*, 10; *Microbulbifer*, 5; *Xylella, Xanthomonas*, 4; *Burkholderia*, 10; *Methylobacterium, Bordetella, Neisseria, Ralfstonia*, 10; *Pseudomonas*, 23; Pasteurellaceae, 13; Enterobacteria, 49; *Vibrio*, 12; *Shewanella*, 6; Cyanobacteria, 13; *Thermus/Deinococcus* group, 13; *Mycobacterium*, 20; *Corynebacterium*, 16; *Streptomyces*, 12; *Propionibacterium*, 5; *Rhodopseudomonas*, 4; Rhizobacteria, *Brucella, Rhodospirillum, Rhodobacter, Caulobacter*, 26; *Ehrlichia*, 4; *Treponema denticola, Chloroflexus, Pirellula*, 4; *Myxococcus, Geobacter*, 8; *Desulfovibrio*, 14.

These findings demonstrate that our algorithm is highly tolerant of distortions in the initial data set.

Application of the algorithm to biological data. Example 1. The NrdR-binding signal of 16 nucleotides [3] regulates the production of replication proteins. As a tree *G*, we used a tree of the corresponding species with 21 taxa (Fig. 1). Example 1 and a similar result obtained for the MntR signal have been reported at the BGRS'2006 conference [10].

Three edges, each including 11 out of 16 positions, substantially prevailed in occurrence in the best *i*-scenarios for various *i* positions. Of these edges, one leads to *Deinococcus radiodurans*, another leads to the {*Thermus maritima, T. thermophilus*} taxon, and the third one goes from the root to the node indicated *a* in Fig. 1. This indicates, first, that the NrdR signal in *T. maritima* and *T. thermophilus* significantly differs from that of other species. The {*Thermus maritima, T. thermophilus*} taxon is adjacent to the root on the tree *G* and changes in NrdR and its binding site occurred at the edge leading to node *a*. This agrees with the slow evolution accepted for these species.

Second, our result indicates that the NrdR signal of *D. radiodurans* substantially differs from that in closely related species. In the tree *G, D. radiodurans* is far away from the root; it is possible to assume that the character of NrdR regulation changed during the formation of *D. radiodurans* and is associated with its rapid evolution.

Example 2. As a tree *G*, we used a tree of the NrdR proteins (Fig. 2), in which 31 taxa are shown with circles and some terminal nodes are removed.

One edge, leading to the *Thermus/Deinococcus* taxon, substantially prevailed in occurrence in the best *i*-scenarios, being found in 11 out of 16 *i*-scenarios. This edge corresponds to the three edges found in Example 1. Taking the tree (Fig. 1) unchanged, it is possible to assume horizontal transfer of the NrdR gene between *T. thermophilus* and *D. radiodurans.*

Similar results (not shown) were obtained for the 22-bp MntR signal, regulating manganese transport, and the corresponding trees of species and transcription factors, as well as for the LacI-family factors, regulating the sugar catabolism genes, and the corresponding transcription factor tree [11, 12]. The site is 20-bp in the latter case.

A transition from the consensus observed at one node of the tree *G* to the consensus at another node can occur via rapid compensatory substitutions preserving the symmetrical structure of the site; e.g., this is the case in the large LacI family. We found a situation in this family where the consensus changed via a loss of conservation at a certain step (data not shown). The algorithm was applied to other signal families as well and allowed us to identify the evolutionary significant edges, construct evolutionary scenarios for each position or a pair of associated positions of a palindrome, and to take the signal structure into account.

ACKNOWLEDGMENTS

We are grateful to M. Gelfand for fruitful discussion; O. Laikova, D. Rodionov, and A. Seliverstov for help and data provided; and to A. Vitreshchak for important comments.

REFERENCES

1. Lyubetsky V., Gorbunov K., Rusin L., V’yugin V. 2005. Algorithms to reconstruct evolutionary events at molecular level and infer species phylogeny. In: *Bioinformatics of Genome Regulation and Structure II*. Springer, 189–204.
2. Gorbunov K.Yu., Lyubetsky V.A. 2005. Identification of ancestral genes that introduce incongruence between protein and species trees. *Mol. Biol*. **39**, **5**, 847–858.
3. Rodionov D.A., Gelfand M.S. 2005. A universal regulatory system of ribonucleotide reductase genes in bacterial genomes. *Trends Genet*. **21**, 385–398
4. Kotelnikova E.A., Makeev V.J, Gelfand M.S. 2005. Evolution of transcription factor DNA binding sites. *Gene*. **347**, 255–263.
5. Fitch W.M. 1971. Towards defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool*. **20**, 406–416.
6. Schneider T.D. 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol*. **188**, 415–431.
7. Stormo G.D., Fields D.S. 1998. Specificity, energy and information in DNA–protein interactions. *Trends Biochem. Sci*. **23**, 109–113.
8. Berg O.G., von Hippel P.H. 1987. Selection of DNA binding sites by regulatory proteins. *J. Mol. Biol*. **193**, 723–750.
9. http://www.dna.affrc.go.jp/misc/MPsrch/InfoIUPAC.html.
10. Gorbunov K.Yu., Lyubetsky V.A. 2006. Inferring regulatory signal profiles and evolutionary events. *Proc. 5th Int. Conf. Bioinformatics Genome Regul. Struct.* (BGRS’2006). Novosibirsk: IC&G, vol. 3, pp. 151–154.
11. Laikova O.N. 2002. Systematic prediction of regulatory interactions in the LacI family of transcriptional regulators. *Proc. 5th Int. Conf. Bioinformatics Genome Regul. Struct.* (BGRS’2002). Novosibirsk: IC&G, vol. 2, pp. 26–28.
12. Gelfand M.S., Laikova O.N. 2003. Prolegomena to the evolution of transcriptional regulation in bacterial genomes. In: *Frontiers in Computitional Genomics*. Eds. Galperin M.Y., Koonin E.V. Wymondham, U.K.: Caiser Acad. Press, pp. 195–216.