INFERRING REGULATIORY SIGNAL PROFILES AND EVOLUTIONARY EVENTS

Gorbunov K.Yu.**, *Lyubetsky V.A.
Institute for Information Transmission Problems, RAS, Moscow, 127994, Russia
* Corresponding author: e-mail: gorbunov@iitp.ru

Key words: regulatory signal reconstruction, evolutionary scenario, species tree, frequency matrix, NrdR repression signal, MntR repression signal

SUMMARY

Motivation: Inferring evolution of regulatory signals on the species tree is a timely task. The signal is known at the tips of the tree and is to be reconstructed at its internal nodes.

Results: An algorithm is proposed to reconstruct frequency of every nucleotide and infer evolutionary important edges in the given species tree. Its performance is tested on artificial and biological data including NrdR and MntR regulatory signals. Evolutionary scenarios are inferred for these signals.

THE TASK

Let the species tree be given with its tip taxa assigned multiple alignments of regulatory signals homogeneous across species in each taxon. NrdR is one of such signals (for an example of such types to Rodionov, Gelfand, 2005). Each alignment of n columns produces a corresponding $4 \times n$ frequency matrix of 4 nucleotide letters and n columns. The task is to reconstruct corresponding frequency matrices at internal nodes, as well as, for each column i of the signal, to infer edges in the species tree containing important events in the signal evolution. Note that i also designates i-columns of frequency matrices generated at the tips. The edge is considered *evolutionary important* if maximum parsimony requirement for it is violated in the sense that frequency matrices reconstructed at its nodes (and their i-columns) differ considerably. Evolutionary patterns at i- and j-positions of the signal can be different. Therefore, evolutionary scenarios are reconstructed for each position separately and then juxtaposed. Under fixed i-position in each leaf, i-column in the corresponding frequency matrix is defined (it is called the signal profile or frequency *distribution* at i-position). The aim is, to reconstruct such distributions in each inner node under fixed i-position and infer edges containing evolutionary important events for i-position.

Such scenarios for a position are defined as *individual scenarios* and are further joined in a *resulting scenario*, which combines the edges contributing the most into evolution of the entire signal. Also, *substantial positions* are inferred as those having relatively robust evolutionary scenarios with respect to the signal structure.

ALGORITHM

Maximum parsimony is used to solve this task. Namely we minimize function F, which is the sum of pairwise distances between distributions at adjacent nodes. Two

conditions are imposed: the sum of fractions at each node gives 1, all fractions are positive. Although our algorithm allows for other functions F and other distribution conditions as well. For edge u the corresponding sum of four items in function F is denoted $F(u)$. An iterative step is as follows: F is minimized, and edges with highest values $F(u)$ are considered in the number determined by parameter *vet*. For *each such u*, item $F(u)$ is independently subtracted from F, thus hypothesizing edge *u to contain an evolutionary event* and therefore not be maximally parsimonious, after which resulting F is again minimized.

The procedure iterates until the number of excluded edges exceeds the value, determined by the *glub* parameter. Each succession of excluded edges is called the evolutionary *i*-scenario under given *vet* and *glub* (*branching* and *depth*, respectively). The settings in test run were *vet* = 15 and *glub* = 4. A *robust scenario* is defined with a combination of three requirements: lower number of its contained edges, lower $F(u)$ value for all non-excluded edges, lesser sum of pairwise distances between distributions at non-excluded edges and higher – at excluded ones. The algorithm also implements comparative analysis of different *i*-scenarios to reveal their consistency (i.e. presence of an evolutionary event at the same edge in several corresponding positions of the signal) and robustness in individual signal positions (those are called evolutionary important positions). Evolutionary not important positions are excluded and the algorithm is applied to the rest of signal. The *resulting evolutionary scenario* includes edges from different individual scenarios, especially those shared by *i*-scenarios of coevolving positions (e.g., direct or inverted repeats) and which are robust for many positions. The edge contributes in the resulting scenario if contained in several robust *i*-scenarios, with weight of corresponding *i*-positions being high.

RESULTS AND DISCUSSION

Here we describe the algorithm's performance on two biological datasets. NrdR-box of length 16 (Rodionov and Gelfand, 2005) is involved in biosynthesis regulation of replication-associated molecules. The species tree used in the study is shown in Fig. 1 (edges designated with numbers of their corresponding descendant nodes). For 16 signal positions our algorithm found the following robust scenarios: (1) 40, 6, 12, 16; (2) 40, 30, 26, 23; (3) 2, 17, 29, 25; (4) 2, 4, 11, 13; (5) 40, 2, 12, 16; (6) 2, 3, 6, 16; (7) 30, 26, 7, 37; (8) 40, 2, 4, 3; (9) 40, 2, 3, 15; (10) 40, 30, 26, 16; (11) 40, 2, 17, 18; (12) 26, 31, 39, 16; (13) 40, 2, 3, 13; (14) 40, 2, 3, 17; (15) 40, 2, 12, 16; (16) 40, 2, 5, 32. The resulting scenario is combined from edges 40 and 2, which suggests considerable changes in NrdR signal to have happened during its evolution in this part of the tree.

The recently discovered MntR-box serves as the second example (Mn transport regulation, Rodionov D.A., Gelfand M.S., 2006, personal communication). The species tree is given in Fig. 2. For 22 signal positions the algorithm output 22 robust scenarios: (1) 20, 24, 18, 14; (2) 1, 6, 11, 12; (3) 1, 6, 7, 9; (4) 2, 4, 24, 21; (5) 2, 24, 6, 21; (6) 1, 3, 6, 10; (7) 1, 4, 6, 10; (8) 1, 2, 4, 14; (9) 1, 2, 4, 24; (10) 1, 24, 10, 13; (11) 2, 24, 21, 18; (12) 6, 9, 22, 13; (13) 1, 21, 10, 11; (14) 4, 24, 21, 14; (15) 1, 4, 6, 14; (16) 1, 4, 6, 10; (17) 1, 6, 7, 10; (18) 1, 2, 24, 21; (19) 1, 2, 4, 13; (20) 1, 6, 17, 14; (21) 1, 3, 9, 11; (22) 2, 4, 21, 22. The resulting scenario is combined from edges 1 and 6, thus suggesting phylogenetic localization of the signal change over time.

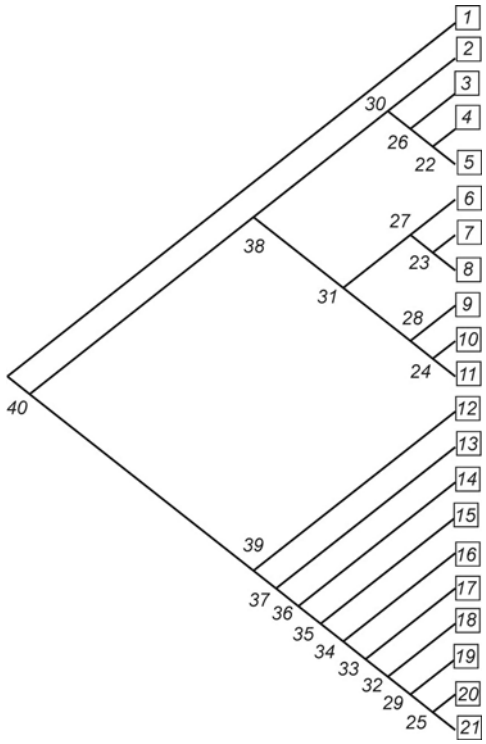

Figure 1. Species tree for the case of NrdR-box with consensus acaC(a/t)AtATaT(a/t)Gtg.

Taxa designations in Fig. 1 are as follows: 1 = {*T. maritima*, *T. thermophilus*}; 2 = {*D. radiodurans*}; 3 = {*P. marinus*, *G. violaceus*, *Synechocystis sp.*, *S.elongates*, *T. elongates*}; 4 = {*S. coelicolor*, *S. avermitilis*, *S. scabies*, *C. michiganensis, L. xyli*, *Corynebacterium* spp., *Mycobacterium* spp.}; 5 = {*P. acnes*, *B. longum*, *T. fusca*}; 6 = {*S. aureus*}; 7 = {*C. acetobutylicum*, *C. tetani*, *C. perfringens*, *C. botulinum*, *C. difficile*, *T. tengcongensis*, *C. hydrogenoformans*, *D. hafniense*}; 8 = {*B. subtilis*, *B. licheniformis*, *B. halodurans*, *B. cereus*, *B. stearothermophilus*}; 9 = {*E. faecalis*, *E. faecium*}; 10 = {*S. epidermidis*, *S. pyogenes*, *S. agalactiae*, *S. pneumoniae*, *S. mutans*, *P. pentosaceus*}; 11 = {*Lactobacillus* spp.}; 12 = {*C. muridarum*, *C. pneumoniae*, *C. trachomatis*, *C. abortus*, *C. caviae*, *T. denticola*}; 13 = {*G. sulfurreducens*, *G. metallireducens*, *D. acetoxidans*, *D. psychrophila*, *B. bacteriovorans*, *B. marinus*, *M. xanthus*}; 14 = {*B. melitensis*, *M. loti*, *A. tumefaciens*, *R. leguminosarum*, *S. meliloti*, *B. japonicum*, *R. palustris*, *R. capsulatus*, *C. crescentus*, *H. neptunium*, *E. chaffeensis*, *N. sennetsu*}; 15 = {*N. europaea*, *N. meningitides*, *M. flagelatus*, *R. solanacearum*, *B. pertussis*, *B. bronchseptica*, *B. avium*, *B. fungorum*, *B. cepacia*, *B. pseudomallei*, *D. aromatica*}; 16 = {*X. fastidiosa*, *X. axonopodis*}; 17 = {*P. aeruginosa*, *P. putida*, *P. fluorescens*, *P. syringae*}; 18 = {*V. cholerae*, *V. vulnificus*, *V. parahaemolyticus*}; 19 = {*E. coli*, *S. typhi*, *K. pneumoniae*, *Y. pestis*, *Y. enterocolitica*, *E. chrysanthemi*, *E. carotovora*, *P. luminescens*}; 20 = {*P. multocida*}; 21 = {*H. influenzae*, *H. ducreyi*}.

Taxa designations in Fig. 2 are as follows: 1 = {*T. fusca*, *R. xylanophilus*, *C. diphtheria*, *C. efficiens*, *C. glutamicum*}; 2 = {*Streptococcus* spp., *L. lactis*, *P. filamentus*, *C. hutchinsonii*}; 3 = {*E. faecalis*}; 4 = {*B. subtilis*, *B. cereus*, *B. halodurans*, *O. iheyensis*, *L. monocytogenes*}; 5 = {*Staphylococcus* spp.}; 6 = {*Treponema* spp.}; 7 = {*M. magnetotacticum*, *R. capsulatus*, *Mesorhizobium* spp.}; 8 = {*E. coli*, *S. typhi*, *K. pneumoniae*}; 9 = {*Xanthomonas* spp., *X. fastodiosa*}; 10 = {*Methanosarcina* spp.}; 11 = {*A. fulgidus*}; 12 = {*M. thermophila*}; 13 = {*Pyrococcus* spp.}; 14 = {*M. jannaschii*, *M. maripaludis*}.

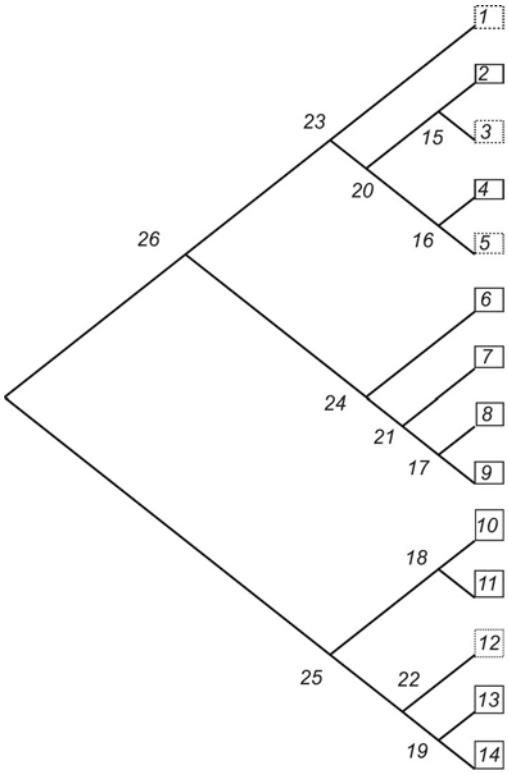

Figure 2. Species tree for the case of MntR-box with consensus a(a/t)(a/t)TTTAG(c/g)nnnn(g/c) ctA Aa(a/t)(a/t)n.

ACKNOWLEDGEMENTS

The authors are greatly indebted to prof. M.S. Gelfand for fruitful discussions, D.A. Radionov for providing test data and discussions, to L.Y. Rusin and A.V. Seliverstov for valuable help and discussions.

REFERENCES

Rodionov D.A., Gelfand M.S. (2005) A universal regulatory system of ribonucleotide reductase genes in bacterial genomes. *Trends in Genet.*, **21,** 385–398.