A TREE-BASED METHOD OF SEQUENCE ALIGNMENT [1]

V. Lyubetsky [2], L. Rubanov, A. Seliverstov

Institute for Information Transmission Problems of the RAS (Kharkevich Institute)
Bolshoy Karetny lane 19, 127994 Moscow, Russia
[2] lyubetsk@iitp.ru

We describe an original fast algorithm of sequence alignment and its computer realization. Here in examples the aligned are regions upstream the same gene in different genomes. An alignment is constructed with the algorithm, which uses a binary tree representation of distances between any pair of sequences from corresponding genomes (organisms). If the binary tree is unknown, it is inferred by resolving all non-binary nodes in a given non-binary tree, which is usually known. Thus, the algorithm realizes fast generation of binary trees compatible with a given non-binary tree and produces the best alignment by sampling the generated tree space. The algorithm was tested with biological data and simulations.

The problematic

Sequence alignment is among major areas of bioinformatic research. The alignment is referred to as "multiple" if more than two sequences are to be aligned (e.g., homologous DNA regions from different organisms), and as "pairwise" if there are only two sequences. We developed an original fast algorithm to construct sequence alignments on the basis of a binary tree representation of distances between corresponding organisms (a species tree). Such binary tree is obtained by resolving polytomies (non-binary nodes) in a polytomic species tree, which is usually known, with an original procedure of generating all non-redundant binary trees topologically compatible with the given polytomic tree. During sampling the generated tree space, the algorithm infers the best tree-based alignment.

The given is a set of finite sequences of arbitrary length in a defined alphabet, e.g. of four characters: A, C, T, G. The goal is to stack the sequences such that they become of equal length and the resulting text matrix contains the maximum possible number of "conserved" (i.e. least heterogeneous in character state content) columns by inserting the arbitrary number of gaps "=" in initial sequences. This matrix corresponds to the resulting alignment. Maximizing the number of conserved columns requires defining and optimization of a fixed functional of the alignment. Obviously, exhaustive sampling of the variant space is intractable, even under constraining the overall amount of gap insertions (i.e. the amount of columns in the matrix).

Approach and techniques

Let a binary tree be given with tips marked with sequences from the initial set. The sequences close on the tree are expected to share more conserved positions in the alignment, and more distant ones – to share less. Here again, a functional is required that is dependent on the alignment and the binary tree. In a simplistic case, the "distance along the tree" is the number of edges separating two sequences. Such a tree can either be specified to our algorithm by the user or inferred. If knowledge is insufficient, sequence relatedness can be specified in form of a polytomic tree, e.g. sequences from 1st to 5th are regarded equidistant from their common

ancestor according to available evidence. The algorithm, however, requires each ancestral node to produce exactly two descendants and resolves polytomies by constructing all possible combinations of binary nodes compatible with the current polytomic node in the given non-binary tree. These combinations must not include redundant topologies to not exceed theoretical size of the binary tree space: each polytomy with n descendants allows for

$$R(n) = \frac{(2n-3)!}{2^{n-2} \cdot (n-2)!} \qquad (1)$$

topological alternatives to resolve it. As polytomies are resolved independently, the overall tree space has the number of elements equal to the product of (1) for all non-binary nodes in the tree.

The algorithm constructs the alignment along the current binary tree as follows. A tree node is assigned sequences of character frequency distributions calculated for the subtree rooted in the node; in the tree tips frequency distributions are defined trivially to describe actual sequences: one for the actual character state at the given position and zeros for other states. Along the branches toward the ancestral node, the two distribution sequences in descendants are aligned using dynamic programming, with reward a_j for matching frequencies x_i and y_i in the descendants not fixed for character i but computed anew at each position j taking into account the distance between the descendent nodes. The algorithm uses scalar square of differences of non-zero frequencies

$$a_j = 1 - \sum_{i=1}^{4} w_i (x_i - y_i)^2 , \qquad (2)$$

where w_i are character weights summing up to one. (The program also realizes other metrics, e.g. L_1 and L_2 in the distribution space). Gap penalty a'_j (i.e. for the alignment of zero and non-zero distributions) is a decreasing function of the gap string length: the longer the string, the lower the penalty. In the pairwise procedure, alignment of zero distributions is forbidden.

At each alignment position, the ancestral distribution is defined as half-sum of the distributions aligned in the descendants:

$$\mathbf{Z} = \frac{1}{2}(\mathbf{X} + \mathbf{Y}) \qquad (3)$$

When the ancestral sequence is reconstructed at the root of the tree, its gaps are progressively inserted in corresponding positions of the sequences in descendent nodes. Sequences thus obtained in the tree tips constitute the resulting multiple alignment along the current binary tree.

The above mentioned function is sum $\sum_j a_j + a'_j$ of rewards or penalties over all positions j of the pairwise alignment of two descendants of the root sequence.

The described procedure has a complexity linearly proportional to the number of tips. Particularly, for a balanced tree with $m = 2^k$ tips it consists of m-1 pairwise alignments formed progressively from tips to the root. The pairwise alignment procedure is very fast.

Multiple alignments obtained for the multitude of binary resolutions of the polytomic tree are compared using the "conservativity index" defined as follows:

$$I = (N_a + N_s)b + \sum_{i=1}^{N_s} (b + s)(l_i - 1) + N_b c , \qquad (4)$$

where N_a is the number of individual perfectly conserved (i.e. containing exactly one character state) alignment columns, N_s is the number of contiguous perfectly conserved regions with the length of two or more columns (l_i is the length of i-region), N_b is the number of highly conserved (i.e. containing exactly one mismatch) columns; b, c and s are program parameters defining different reward values.

Upon sampling the binary tree space, the algorithm outputs the best alignment with the highest function value or, if several such found, the one with highest conservativity index score (4).

Results

The originally developed effective algorithm of tree-based sequence alignment was implemented in a computer program (available currently upon request to the authors, web interface will be implemented by the beginning of the conference at http://lab6.iitp.ru). On typical hardware (Pentium-4 3 GHz PC), computing a multiple alignment of 16 sequences with 120-223 nt length takes less than one second.

The algorithm has been extensively tested. Here we only present results that suggest novel hypotheses of gene regulation (details reported in the Russian journal "Molecular Biology" and during the BGRS'2008 international meeting). Fig. 1 shows the multiple alignment of regions upstream gene and pseudogene *ycf24* in *Th. annulata*; Fig. 2 – alignment of the same regions in *T. gondii*; Fig. 3 – alignment of regions upstream gene *rpoB*. Here and in the Figures species names are set in italics. In Figs. 1-3 the regions are adjacent to the gene sequence suggesting their role as potential binding sites of regulatory proteins to RNA. Fig. 4 shows the multiple alignment of the region upstream gene *rps20*; here the hypothesized regulation is related to transcriptional competition between the two DNA strands.

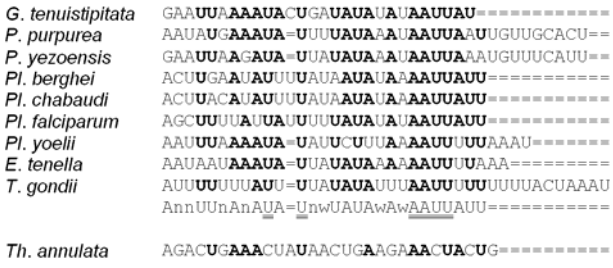

Fig. 1. Alignment of the region upstream gene *ycf24* in *Th. annulata*. Perfectly conserved columns are underlined, highly conserved set in bold.

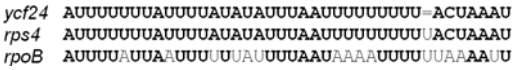

Fig. 2. Alignment of the region upstream the same gene in *T. gondii*. Designations as in Fig. 1.

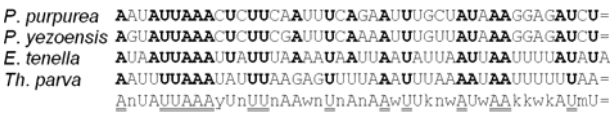

Fig. 3. Alignment of the region upstream gene *rpoB*. Designations as in Fig. 1.

Cyanidioschyzon merolae	ACTC**TTGCTT**TTTGCCATCTGCT=ATTT**TATCTT**TATGTAGACT	-33
Cyanidium caldarium	AAAT**TTGTTT**ATTTTACTTTAAT=A**TG**A**TACAGT**AATTTATAAC	-32
Porphyra purpurea	GCTA**TTGCCT**ATTCTTTTTTTTTAA**TG**T**TATAAT**ACGGCGCATA	-78
Porphyra yezoensis	ACTA**TTGCCT**ATTGTTTTTCTTTAA**TG**T**TATAAT**ACGCCGCATA	-78
Gracilaria tenuistipitata	GTTC**TTGTCT**ATTTTAATGTATTAA**TG**A**TATAAT**CCAATTAGAT	-63
Guillardia theta	TTAA**TTTATT**CCATTATTTCTTATA**TG**T**TATAAT**CTTTTATTAC	-59
Rhodomonas salina	TCTT**CTTATT**C=ATAATTTGTTCTA**TG**T**TATAAT**CACTAATCGT	-55

Fig. 4. Alignment of the region upstream gene *rps20*. Conserved positions are set in bold, putative factor binding sites double underlined and shaded.

Conclusion

The original algorithm of multiple sequence alignment along the known or dynamically inferred binary tree has the advantages:

- high performance with large alignments: computational complexity is linearly proportional to the number of sequences;
- ease of parallel implementation;
- all binary resolutions of polytomies are investigated when a fully resolved tree is unknown;
- the method can be naturally extended to deal with degenerate sequences containing more character states (e.g., in IUPAC-IUB nomenclature);
- flexibility to weigh character states;
- possibility to specify any user-defined functional of reasonable complexity for optimization purposes.

The authors are grateful to L. Rusin for fruitful discussions and help.