Evgeniy A. Gladyshev

On Gaps

Just as any other changes in nucleotide or protein sequences, gaps, or indels, result from mutations, and thus, contain important information on the evolution, structure, and function of sequences. However, unlike the other changes in the primary structure, in most cases indels are regarded merely as place-holders, or as a nuisance, interfering with subsequent sequent comparisons. This attitude must be changed and appropriate methods must be introduced to allow the incorporation of information on gaps into sequence analysis.

Clustering represents a very powerful instrument in understanding the structure and function of molecular sequences. In most cases it includes two semi-independent stages: 1) aligning sequences to establish provisional homology among them, and 2) subsequent grouping of aligned sequences based on some measure of sequence similarity or overall "goodness" of a resultant cluster. In this traditional approach, gaps are generated during the first stage, via dynamic programming (Needleman & Wunsch, 1970), and passed to the clustering algorithm, which can be either distance-based, or likelihood-based (maximum likelihood, ML), or parsimony-based (maximum parsimony, MP). I regard Bayesian approach as an extension of the ML method, as the posterior probability of the cluster given the alignment is essentially its maximum likelihood value scaled by its prior probability and normalized by the probability of observing a given multiple alignment. During the alignment step, a single multiple alignment can be generated based on certain gap-insertion and gap-extension costs. This approach is implemented in the majority of multiple alignment programs, like ClustalW (Thompson et al., 1994). It has many caveats, for instance, the arbitrary nature of gap costs that results in an arbitrary alignment of rapidly evolving sequence regions, and strong dependence of the clustering outcome on correct alignment information. In short, specifying different [and arbitrary] gap penalties can potentially lead to profound differences in cluster topologies.

An alternative approach was taken in SOAP, which generates the distribution of multiple alignments as the function of two variables, gap opening and gap extension penalties (Loyynoia & Milinkovitch, 2001). Thus, SOAP provides a novel opportunity to explore the indel parameter space, while using a standard multiple alignment algorithm, ClustalW. All multiple alignments are then subsequently compared and the regions of instability, i.e., showing considerable variation of indels, are excluded from further analysis. Although this approach takes one step further towards statistically robust sequence clustering, it deals with the problem in the most straightforward way, decreasing uncertainty by throwing out information on gaps, and, therefore, decreasing the resolving power of the alignment. As a result, clustering would produce a more reliable cluster that may lack sufficient resolution to support a fully bifurcating topology.

The question of utilizing information on gaps may also be addressed in the second phase, i.e., by clustering algorithms. Clustering algorithms treat indels in three principal ways: they may exclude them from the analysis, they may ignore them, or they may try to make some use of them. Gaps can be treated as an additional nucleotide or amino acid state, but at this moment logistical difficulties arise, as gaps do not represent observable character states. Furthermore, this approach violates the principal assumption of phylogenetic clustering, i.e., character independence: an indel of k characters long, where all characters are altered at the same time, is treated as a sequence of k independent characters.

In many clustering programs, for instance in PAUP, gaps are treated as missing information by default, and the majority of phylogenetic studies follow this setting (Giribet & Wheeler, 1999).

An alternative technique to the two-stage clustering was developed to avoid the problem of misusing information on gaps: clusters are built from unaligned sequences using *direct optimization* technique, proposed by Ward Wheeler (1996) and implemented in POY (ftp://ftp.amnh.org/pub/molecular/poy). Direct optimization approach treats each indel not as a set of independent characters, but as a single event that contains phylogenetic signal. Only one gap penalty has to be specified, and the total indel cost does not depend on indel size. This is equivalent to specifying no gap extension cost in the Needleman-Wunsch algorithm. A cluster is constructed from unaligned sequences, as the latter are aligned dynamically during the process of clustering. Along with a traditional measure of statistical support, bootstrap values, Wheeler proposes "congruence" criterion, by which a set of cluster parameters receives support if topologies produced from different molecular datasets for the same set of taxa are congruent. Of course, for molecular data such comparisons are possible only if the evolutionary histories of sequences are known to be congruent. POY was originally based on MP methods, but now includes ML techniques as well. However, in this case it is also not know a priory, which gap cost should be chosen, and the dataset should be explored with sensitivity analysis to find the optimal parameters (Giribet & Wheeler, 1999). I did sensitivity analysis in POY for the set of 1.7 kbp18S rRNA fragments of marine mussels, Mytilidae (Mollusca: Bivalvia), and found that the most congruent topologies were produced with the gap-to-nucleotide-change ratio of 2-3:1.

I believe that direct optimization has the promise to succeed the traditional two-stage phylogenetic clustering methods, however, statistical model of indel dynamics is yet to be developed. Such model must accommodate the probability of indel event as the function of several parameters, including indel length, sequence function, rate of nucleotide substitutions around the site of indel, etc. This would allow incorporation of indel likelihood functions into existing ML models. In order to develop the indel model, it would be necessary to perform direct-optimization clustering of nucleotide and protein sequences in the database, comparing different clusters of sequences for the same groups of organisms and looking for the congruence of resultant topologies, provided the probability of lateral gene transfer for these sequences is ruled out. A set of frequency distributions of indel event will be created and analytical function will be fitted into these distributions to enable the estimation of indel probabilities, an approach that is similar to deriving EVD from all-to-all pairwise sequence alignment (Levitt & Gerstein, 1998).

REFERENCES

Giribet G. & W.C. Wheeler. 1999. On Gaps. Mol. Phyl. Evol. 13: 132-143.

Levitt M & M. Gerstein. 1998. A Unified Statistical Framework for Sequence Comparison and Structure Comparison. PNAS 95: 5913-5920

Loyynoia A. & M.C. Milinkovitch. 2001. SOAP, cleaning multiple alignments from unstable blocks. Bioinformatics 17: 573-574.

Needleman S.B. & C.D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48: 443-453.

Thompson J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penaties, and weight matrix choice. Nucleic Acid Res. 22: 4673-4680.

Wheeler W.C. 1996. Optimization alignment: The end of multiple sequence alignment in phylogenetics? Cladistics 12: 1-9.