Cladistics

Cladistics 24 (2008) 1070

10.1111/j.1096-0031.2008.00208.x

Application note: on extension gap in POY version 3

Andrés Varón[a,b,*] and Ward C. Wheeler[a]

[a]*Division of Invertebrate Zoology, American Museum of Natural History Central Park West at 79th Street, New York, NY 100245192, USA;*
[b]*Computer Science Department, The Graduate School and University Center, The City University of New York, USA*

Accepted 7 December 2007

Direct Optimization (DO; Wheeler, 1996) is a heuristic tree cost calculation procedure for the tree alignment problem. In its original description, DO was defined for insertion–deletion distance functions of the form $\Sigma_{1<i<|s|}\ cm(s_i)$, where cm($x$) is the cost of inserting or deleting the base x and the total indel cost for inserting or deleting sequence s is the sum of single nucleotide indels over its length ($|s|$). That is, cost functions where no gap opening parameter is defined, as oposed to those cases where it is (Gotoh, 1982).

POY version 3 (Wheeler et al., 1996–2005) implemented an extension to the original DO and Gotoh (1982) algorithms for tree searches under extension gap cost calculations. However, when extension gaps are used, POY3 can underestimate the overall tree cost, yielding trees with unattainably low scores. The error does not occur when the blocks of insertions and deletions form ‘‘nicely’’ delimited blocks, but when those blocks are ragged, with partial overlaps. The simplest (and most dramatic) example is four sequences containing A, AA, AAA and AAAA. Although at least three independent indel blocks are required on any tree for these four sequences, POY3 will always estimate a single block (e.g., for the same data set, the command poy –gap 6 –extensiongap 1 would yield a tree with cost 8 (one initial and two extension gaps), which is clearly unachievable; see Fig 1). The problem is somewhat limited in real data sets, but its degree depends only on the gap structure of the input sequences.

Although the algorithms in POY 4 (Varón et al., 2008) were corrected (the equivalent command is transform (tcm:(1,1),gap opening:5)), a bug in the extension gap cost implementation left some inaccuracies in the released builds 1665, and 1724. This was detected during the beta testing process, and reported by various users. Builds 1824 and later provide fully verified tree cost for the extension gap cost case.

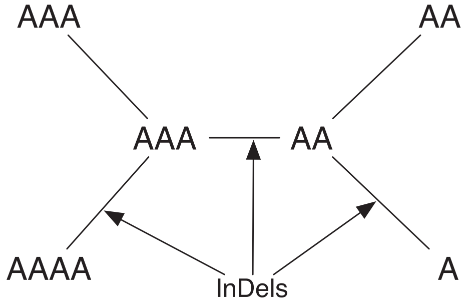

Fig. 1. Cladogram linking four simple sequences requiring a minimum of three indels. POY3 reports a length of 8 (one initial and two extension indels), POY4 the correct length of 18 (three initial indels).

References

Gotoh, O., 1982. An improved algorithm for matching biological sequences. J. Mol. Biol. 162, 705–708.

Varón, A., Vinh, L.S., Bomash, I., Wheeler, W.C., 2008. POY 4.0.2911. American Museum of Natural History, New York. http://research.amnh.org/scicomp/projects/poy.php.

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

Wheeler, W.C., Gladstein, D.S., De Laet, J., 1996–2005. POY, Version 3.0. Program and documentation available at. http://research.amnh.org/scicomp/projects/poy.php (current, Version 3.0.11. documentation by D. Janies and W.C. Wheeler. commandline documentation by J. De Laet and W.C. Wheeler. American Museum of Natural History, New York.

[*] Corresponding author:
E-mail address: avaron@amnh.org