

Supporting Information

The new algorithm is tested in the alignment of 20 mitochondrial control region sequences from primates (D-loop), and of 15 genomic sequences around the *CAV2* gene from mammals and chicken (*CAV2*). In the first example (D-loop) the sequences are less than 400 bases long and contain only short indels, and we compare alignments performed with different settings in order to assess the effect of (i) our modification to the algorithm; (ii) the modelling of the substitution process; and, when the modification was enabled, (iii) the different ways of handling the insertions. As a reference, we align the same dataset using a traditional algorithm implemented in the software ClustalW v.1.83 (1). In the second example (*CAV2*), the sequence lengths vary from 5200 to 7000 bases and contain long insertion elements and non-homologous regions. Here, we test whether our progressive global aligner can infer reasonable alignments for genomic sequences and, as long insertion elements are typical in genomic sequences, compare the two different approaches for handling the inserted sites.

The software used in the analyses, data sets analysed and detailed instructions to repeat the analyses are available via <http://www.ebi.ac.uk/goldman> or upon request from A.L.

D-loop: The guide tree for the D-loop alignments is based on the tree produced with ClustalW (Figure 3, left) but we have corrected the placement of *C. aethiops* and rooted it by defining *E. fulvus* an outgroup (Figure 3, right). The sequences are aligned with ClustalW using the software default parameters (analysis ‘ClustalW’) and with our probabilistic algorithm using either the Jukes-Cantor (‘JC’) or Hasegawa-Kishino-Yano (‘HKY’) model (2, 3); either disabling (‘-’) or enabling (‘+’) the correction for insertion sites; and allowing characters to be matched to sites inferred earlier as insertion (*i.e.*, insertions may be closed, denoted ‘+’) or forcing these sites to stay as unmatched insertions (*i.e.*, insertions open forever, ‘+^F’). In all probabilistic alignments parameters are $r = 0.025$, $\varepsilon = 0.5$ and $\gamma = 0$; in the HKY model empirical base frequencies (0.342 / 0.309 / 0.104 / 0.245 for A / C / G / T) are used and κ , the transition-transversion ratio, is set to 2. The resulting alignments are shown in Figures 5–11 as follows: (5) ClustalW, (6) JC⁻, (7) JC⁺, (8) JC^{+F}, (9) HKY⁻, (10) HKY⁺ and (11) HKY^{+F}.

A comparison of the alignments performed with and without the correction for insertion sites (JC⁺ vs. JC⁻ and HKY⁺ vs. HKY⁻) shows that our method works: when enabled, the algorithm preferentially places gaps at the same sites and is more likely to create gaps that can be explained by a single insertion event. The heuristics implemented in ClustalW have a partly similar effect, and the ClustalW alignment has in total fewer gapped columns than JC⁻ and HKY⁻. However, in comparison to JC⁺ and HKY⁺, the gaps inferred by ClustalW are less consistent with the phylogeny: one should remember that two insertions are never evolutionary homologous, and that gaps at the same site in different parts of the tree require multiple independent deletions. When the gaps inferred as insertion are forced to be skipped over in all subsequent alignments (*i.e.*, JC^{+F} and HKY^{+F}; Figures 8, 11), all indel events are strictly consistent with the phylogeny. Some sequences have truncated terminal regions, however, and in intermediate alignments other (non-truncated) terminal regions are incorrectly inferred as insertions. By disabling their matching at a later stage, the strict “insertions open forever” rule spreads out the end of the alignment with multiple long gaps. The differences in the alignments inferred using different evolutionary substitution models (JC vs. HKY) suggest that

subsequent analyses (*e.g.*, phylogenetic inference) estimating the same parameters may, at least partly, depend on the initial choices made (or accepted) for the sequence alignment.

Our method can be vulnerable to wrong alignment order, however, and when the alignments were performed using the original (*i.e.*, wrong) guide tree, the algorithm attempted to explain the inconsistency in the data by inferring additional gaps (data not shown).

CAV2: For the CAV2 alignments a guide tree inferred with a maximum likelihood approach is used, although the placement of rodent and rabbit sequences in the tree is controversial (Figure 4). The sequences are aligned with ClustalW both using the software default parameters (analysis ‘ClustalW’) and, as the high penalty for long gaps seems to cause problems, using the default parameters except a gap extension penalty of 0 (‘ClustalW_0’). The probabilistic alignments, either allowing for insertions to be closed or forcing the algorithm to keep them open (‘HKY⁺’ and ‘HKY^{+F}’, respectively), are performed using parameters $r = 0.025$, $\varepsilon = 0.9$, $\gamma = 0$ and the HKY model with empirical base frequencies (0.166 / 0.314 / 0.402 / 0.117 for A / C / G / T) and $\kappa = 2$. The resulting alignments are shown in Figures 12–15, respectively.

The ClustalW alignment is very compact, only 9.7 kilobases (kb) in comparison to 12.0 kb, 13.6 kb and 14.3 kb for ClustalW_0, HKY⁺ and HKY^{+F}, respectively, and some regions are clearly over-matched by discouraging insertions in single or only a few sequences (Figure 12). The first and third exons are correctly aligned among the mammals and between mammals and chicken (sites 565–714 and 9007–9157, respectively), whereas the second exon of chicken is mis-aligned in comparison to mammals (sites 1179–1367 in mammals but 809–996 in chicken). The alignment ClustalW_0 is less compact, mostly because of the fragmentation due to many short gaps (Figure 13). The first and third exons are again correctly aligned (654–803 and 11333–11483, respectively) and the second exon not (1376–1630 in mammals, 910–1176 in chicken). The alignment HKY⁺ is even longer than ClustalW_0 but the most dramatic difference is in the structure of the gaps: the difference in sequence lengths is distributed among fewer and longer indel events, and in most cases these gaps are consistent with phylogeny, *i.e.*, they can be explained by a single insertion or deletion event and the shared history of the sequences (Figure 14). All three exons are correctly aligned among all the sequences (sites 1095–1244, 1825–2012 and 12608–12758), though the alignment of non-exonic regions of the evolutionarily very distant chicken sequence to mammals is locally incorrect and in some places the chicken sequence is matched against sites earlier inferred as insertion. The problem is largely resolved by forcing the algorithm to skip over all the insertions (HKY^{+F}): exons are correctly aligned (sites 1159–1308, 1948–2135 and 13500–13650) and, as expected, large proportions of the alignment consist of long insertions in a single or only a few sequences (Figure 15).

References

1. Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994) *Nucleic Acids Res.* **22**, 4673–4680.
2. Jukes, T. H. & Cantor, C. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. N. (Academic Press, New York), pp. 21–132.
3. Hasegawa, M., Kishino, H., & Yano, T. (1985) *J. Mol. Evol.* **22**, 160–174.