Fast and Accurate Phylogeny Reconstruction using Filtered Spaced-Word Matches

Chris-André Leimeister [1], Salma Sohrabi-Jahromi [1,2] and Burkhard Morgenstern [1]
[1] *University of Göttingen, Institute of Microbiology and Genetics, Department of Bioinformatics, Goldschmidtstr. 1, 37077 Göttingen, Germany 37073 Göttingen* [2] *Max Planck Institute for Biophysical Chemistry, Quantitative and Computational Biology, Am Fassberg 11, 37077 Göttingen*
chris.leimeister@stud.uni-goettingen.de

Traditional approaches to phylogeny reconstruction are based on multiple sequence alignment. For the large amounts of sequence data that are now available, however, alignment-based methods are too slow. Therefore, so-called 'alignment-free' methods for sequence comparison have become popular in recent years [Vin14]. Most of these approaches represent sequences as word-frequency vectors and compare these vectors to each other instead of comparing sequences position-by-position [SJWK09, QLH04, CHL$^+$09]. Such word-based approaches are much faster than alignment-based approaches, but the distance values that they calculate are usually not very accurate.

In previous papers, we proposed to use *spaced words* in alignment-free sequence comparison. For a binary pattern P representing *match* and *don't-care* positions, a *spaced word* is a word with *nucleotide symbols* corresponding to the *match positions* and *wildcard* characters corresponding to the *don't-care positions* of P. For the pattern $P = 1100101$, for example, $AT**G*T$ would be a spaced word with respect to P. We showed that distances calculated from spaced-word frequencies are statistically more stable and lead to better phylogenetic trees than distances calculated from the frequencies of standard, contiguous words [LBH$^+$14, MZHL15]. Even so, distances between spaced-word frequency vectors are only to a rough measure of sequence dissimilarity. As with most other alignment-free methods, these distances do not represent phylogenetic distances in a statistically meaningful way.

In a more recent present paper, we proposed to estimate phylogenetic distances between DNA sequences – defined as the number of substitutions since they evolved from a common ancestor – by using matching spaced words [LSJM17]. For a given binary pattern P and a set of genomic sequences, our algorithm first identifies all spaced-word matches with respect to P. That is, we are looking for local, gap-free alignment the same length as P with matching nucleotides at the *match* positions and possible mismatches at the *don't-care* positions. Below is an example of a spaced-word match between two *DNA* sequences S_1 and S_2, with respect to the above pattern $P = 1100101$:

S_1 :	G	C	T	G	T	A	T	A	C	G	T	C
S_2 :				G	T	A	C	A	C	T	T	A	T
P :					1	1	0	0	1	0	1	

Our goal is to estimate the distance between two sequences by looking at the aligned nucleotides at the *don't-care* positions of spaced-word matches; this can be seen as a generalization of the previously proposed approaches *Co-phylog* [YJ13] and *andi* [HKP15].

If distances are to be estimated in this way, one has to keep the number of random background spaced-word matches low. In principle, this could be done by using a pattern with a sufficiently large number of *match positions*. *Co-phylog* and *andi* have taken such an approach. But this would, on the other hand, reduce the number of *homologous* spaced-word matches, to the point that no homologies are found at all. We are therefore using another approach to exclude random spaced-word matches: based on a nucleotide substitution matrix [CYM02], we calculate a *score* for each spaced-word match, by summing up the substitution scores of the nucleotide pairs aligned at the *don't-care positions*, and we discard all spaced-word matches with negative scores. Experimental results show that this way, one can easily distinguish between homologous spaced-word matches and random background matches. This allows us to use a pattern P with a small number of *match positions*; by default our patterns have 12 match positions.

In addition, ambiguous spaced-word matches are discarded: for a pair of sequences, an occurrence of a spaced word w in the first sequence is matched to at most one occurrence of w in the second sequence. If a spaced word occurs multiple times in the same sequences, a greedy algorithm is used to obtain a one-to-one matching of these occurrences. The spaced-word matches obtained in this way are then used to estimate the fraction of nucleotides *mismatches* in pairwise alignments of the input sequences. Finally, the usual *Jukes-Cantor* correction is applied to estimate for each sequence pair the number of *substitutions* that have occurred since the two sequences evolved from their last common ancestor.

In our paper, we showed that phylogenetic distance values calculated with our approach are more accurate than distances calculated with other word-based methods, even for large sequences with a low degree of similarity, and that reliable phylogenetic trees can be calculated based on these distances.

References

[CHL+09] Benny Chor, David Horn, Yaron Levy, Nick Goldman, and Tim Massingham. Genomic DNA k-mer spectra: models and modalities. *Genome Biology*, 10:R108, 2009.

[CYM02] Francesca Chiaromonte, V. B. Yap, and W. Miller. Scoring Pairwise Genomic Sequence Alignments. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, and Teri E. Klein, editors, *Pacific Symposium on Biocomputing*, pages 115–126, 2002.

[HKP15] Bernhard Haubold, Fabian Klötzl, and Peter Pfaffelhuber. andi: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, 31:1169–1175, 2015.

[LBH+14] Chris-André Leimeister, Marcus Boden, Sebastian Horwege, Sebastian Lindner, and Burkhard Morgenstern. Fast Alignment-Free sequence comparison using spaced-word frequencies. *Bioinformatics*, 30:1991–1999, 2014.

[LSJM17] Chris-André Leimeister, Salma Sohrabi-Jahromi, and Burkhard Morgenstern. Fast and Accurate Phylogeny Reconstruction using Filtered Spaced-Word Matches. *Bioinformatics*, 33:971–979, 2017.

[MZHL15] Burkhard Morgenstern, Bingyao Zhu, Sebastian Horwege, and Chris-André Leimeister. Estimating evolutionary distances between genomic sequences from spaced-word matches. *Algorithms for Molecular Biology*, 10:5, 2015.

[QLH04] Ji Qi, Hong Luo, and Bailin Hao. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Research*, 32(suppl 2):W45–W47, 2004.

[SJWK09] Gregory E. Sims, Se-Ran Jun, Guohong A. Wu, and Sung-Hou Kim. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences*, 106:2677–2682, 2009.

[Vin14] Susana Vinga. Editorial: Alignment-free methods in computational biology. *Briefings in Bioinformatics*, 15:341–342, 2014.

[YJ13] Huiguang Yi and Li Jin. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Research*, 41:e75, 2013.