Fast alignment-free sequence comparison using spaced-word frequencies

Chris-André Leimeister, Marcus Boden,
Sebastian Lindner, Sebastian Horwege, Burkhard Morgenstern
University of Göttingen, Institute of Microbiology and Genetics,
Department of Bioinformatics, Goldschmidtstr. 1, 37073 Göttingen, Germany
bmorgen@gwdg.de

Sequence alignment is traditionally the first step in DNA and protein sequence analysis. With the amount of sequence data that are now available, however, pairwise or multiple alignment has become too slow in many applications. Therefore, alignment-free methods are increasingly used for genome comparison and phylogeny reconstruction, and the development of such methods has become a very active area of bioinformatics [Vin14]. While alignment-free methods are generally less accurate than alignment-based approaches, they are much faster since they run in linear time. Most alignment-free algorithms work by comparing the *word composition* of sequences. Sequences are represented by *word-frequency vectors*, and standard distance measures on vector spaces can be applied to calculate a pairwise distance matrix for a set of input sequences [HRR06, CHL$^+$09, VCF$^+$12, SJWK09]. Phylogenetic trees can then be calculated from these distance matrices with the usual distance-based methods for phylogeny reconstruction. A certain drawback of these word-based methods is the fact that word occurrences at adjacent sequence positions are far from independent.

Database search programs such as *BLAST* [AGM$^+$90] originally used *word matches* of a fixed length k as *seeds* to search for local homologies. Here, the seed length k is a trade-off between *sensitivity* and *speed*. It has been shown that the sensitivity and speed of these programs can be substantially improved if *spaced seeds* – *i.e.* word matches with possible mismatches at certain pre-defined *mismatch positions* – are used instead of *contiguous* word matches as used in the original version of *BLAST* [MTL02]. Considerable efforts have been made since then, to find suitable patterns for this *spaced-seed* approach, see *e.g.* [BBV04, KNR06, Bro08, IIB11].

Inspired by these approaches, we previously proposed to use *spaced words* for alignment-free sequence comparison, *i.e.* words containing *wildcard* characters at fixed positions, according to an underlying *pattern* P of *match* and *don't care* positions [BSH$^+$13]. The first version of our approach used one single pattern P: for a given set of input DNA or protein sequences and a pattern P, we calculated pairwise distances based on the spaced-word frequency vectors of the sequences with respect to P. A certain draw-back of this original *single-pattern* approach was the necessity to select one specific pattern P of *match* and *don't care* positions, since the results of this method strongly depend on the selected pattern.

In a subsequent paper [LBH$^+$14], we used a *hashing* algorithm to compare the spaced-word composition of sequences that was much more efficient than the tree-based algorithm that we used in the previous implementation. This way, we were able to extend our approach to using sets $\mathcal{P} = \{P_1, \ldots, P_m\}$ of randomly generated patterns P_i of a fixed length and number of *match* positions, instead of a single pattern P. (Multiple patterns of *match* and *don't care* positions have also been proposed to generate *spaced seeds* for database searching [LMKT03].) In this *multiple-pattern* version of our approach, spaced-word frequencies are then calculated and compared with respect to *all* patterns in the set $\mathcal{P}$; we define the distance between two sequences as the *average* distance over all distance values obtained with the individual patterns $P_i \in \mathcal{P}$ that are calculated as in our previous *single-pattern* approach.

As in our previous paper, we evaluated this *multiple-pattern approach* by applying it to phylogeny analysis. We tested two different approaches to calculate pairwise distances between the input sequences based on their (multiple) spaced-word-frequencies, namely the *Euclidean* distance and the *Jensen-Shannon* distance [Lin91]. The resulting distance matrices were used as input for *Neighbour Joining* [SN87] to generate trees, and we compared the resulting tree topologies to trusted reference topologies using the *Robinson-Foulds* distance [RF81]. As benchmark data sets, we used simulated and real-world

DNA and protein sequences.

In our first paper, we had shown that the *single-pattern* version of our *spaced-words* leads to slightly better trees than the same approach used with *contiguous words* [BSH+13]. In [LBH+14], we could show that our new *multiple-pattern* approach produces much better phylogenies than the previously implemented *single-pattern* approach and is also superior to established alignment-free methods that are based on *contiguous* words. On some data sets, the quality of our results was even comparable to trees that were obtained with traditional alignment-based approaches.

Also, we showed empirically that distance values calculated with our *multiple-pattern* program are statistically more stable than distances based on the previous *single-pattern* approach which were, again, more stable than distances based on the frequencies of *contiguous* words. In a more recent paper [MZHL15], we studied the statistical behaviour of our spaced-word-based distance functions in detail and showed analytically why spaced-word-based distances are statistically more stable than distances calculated from contiguous words and why, in turn, the new *multiple-pattern* version of *spaced words* is more stable than the previous *single-pattern* approach.

Our software is freely available as source code. In addition, we provide a user-friendly WWW interface that is described in [HLB+14]. Source code and WWW interface are available through *Göttingen Bioinformatics Compute Server (GOBICS)* at

`http://spaced.gobics.de/`

References

[AGM+90] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene M. Myers, and David J. Lipman. Basic Local Alignment Search Tool. *J. Mol. Biol.*, 215:403–410, 1990.

[BBV04] Brona Brejova, Daniel G. Brown, and Tomas Vinar. Optimal Spaced Seeds for Homologous Coding Regions. *Journal of Bioinformatics and Computational Biology*, 1:595–610, 2004. Early version appeared in CPM 2003.

[Bro08] Daniel G. Brown. *Bioinformatics Algorithms: Techniques and Applications*, chapter A survey of seeding for sequence alignment, pages 126–152. Wiley-Interscience, New York, Fev. 2008.

[BSH+13] Marcus Boden, Martin Schöneich, Sebastian Horwege, Sebastian Lindner, Chris-André Leimeister, and Burkhard Morgenstern. Alignment-free sequence comparison with spaced k-mers. In Tim Beißbarth, Martin Kollmar, Andreas Leha, Burkhard Morgenstern, Anne-Kathrin Schultz, Stephan Waack, and Edgar Wingender, editors, *German Conference on Bioinformatics 2013*, volume 34 of *OpenAccess Series in Informatics (OASIcs)*, pages 24–34, Dagstuhl, Germany, 2013. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[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.

[HLB+14] Sebastian Horwege, Sebastian Lindner, Marcus Boden, Klaus Hatje, Martin Kollmar, Chris-André Leimeister, and Burkhard Morgenstern. *Spaced words* and *kmacs*: fast alignment-free sequence comparison based on inexact word matches. *Nucleic Acids Research*, 42:W7–W11, 2014.

[HRR06] Michael Höhl, Isidore Rigoutsos, and Mark A. Ragan. Pattern-Based Phylogenetic Distance Estimation and Tree Reconstruction. *Evolutionary Bioinformatics Online*, 2:359–375, 2006.

[IIB11] Lucian Ilie, Silvana Ilie, and Anahita M. Bigvand. SpEED: fast computation of sensitive spaced seeds. *Bioinformatics*, 27:2433–2434, 2011.

[KNR06] Gregory Kucherov, Laurent Noé, and Mikhail Roytberg. A unifying framework for seed sensitivity and its application to subset seeds. *Journal of Bioinformatics and Computational Biology*, 4:553–569, 2006.

[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.

[Lin91] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37:145–151, 1991.

[LMKT03] Ming Li, Bin Ma, Derek Kisman, and John Tromp. PatternHunter II: Highly Sensitive and Fast Homology Search. *Genome Informatics*, 14:164–175, 2003.

[MTL02] Bin Ma, John Tromp, and Ming Li. PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18:440–445, 2002.

[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.

[RF81] DF Robinson and LR Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[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.

[SN87] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

[VCF+12] Susana Vinga, Alexandra M. Carvalho, Alexandre P. Francisco, Luís M. S. Russo, and Jonas S. Almeida. Pattern matching through Chaos Game Representation: bridging numerical and discrete data structures for biological sequence analysis. *Algorithms for Molecular Biology*, 7:10, 2012.

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