Extended abstract

Fast alignment-free phylogeny reconstruction using spaced words

Chris-André Leimeister[1]*, Lars Hahn[1]*, Burkhard Morgenstern[1]

[1]*Department of Bioinformatics, University of Göttingen, Germany*
***Corresponding author**: bmorgen@gwdg.de

Abstract

In our presentation, we give an overview about the *spaced-words* approach that we developed during the last years. Spaced words are words with wild-chard characters at certain positions defined by a binary pattern P of *match* and *don't-care* positions. We show that standard word-based approaches to sequence comparison can be improved if spaced words are used instead of contiguous words. To find suitable patterns, we implemented a program called *rasbhari* that optimizes patterns for alignment-free sequence comparison and database searching. Most recently, we developed a *Filtered Spaced Words* approach that accurately estimates phylogenetic distances between genomic sequences. In this approach, we identify matching spaced words between a pair of genomic sequences and estimate distances based on the nucleotides aligned to each other at the *don't-care* positions.

Keywords

Alignment-free — Spaced Words — Spaced Seeds — Phylogeny Reconstruction

1. Introduction

Sequence alignment is usually the first step in DNA and protein sequence analysis. With the huge amount of sequence data that are now available, however, alignment programs are often too slow. Therefore, fast alignment-free methods are increasingly used for genome comparison and phylogeny reconstruction; developing alignment-free methods has become an active area of research in bioinformatics [1]. While alignment-free methods are usually less accurate than traditional, alignment-based methods, they are much faster as they essentially run in linear time. Most alignment-free algorithms represent sequences by *word-frequency vectors* and apply standard distance measures on vector spaces to calculate a pairwise distances between sequences [2, 3, 4, 5]. Phylogenetic trees can then be calculated from these distance matrices by applying the usual distance-based methods for phylogeny reconstruction.

2. Spaced Words

Database search programs such as *BLAST* [6] could be substantially improved by using *spaced seeds* – *i.e.* word matches with possible mismatches at certain pre-defined *mismatch positions* – instead of *contiguous* word matches that have been originally used [7]. Inspired by these approaches, we proposed to use *spaced words* for alignment-free sequence comparison and phylogeny reconstruction. That is, we consider words

containing *wildcard* characters at fixed positions, according to an underlying *pattern* P of *match* and *don't care* positions [8]. In a first implementation of our approach, we used one single pattern P. For a given set of input DNA or protein sequences, we calculated pairwise distances based on the spaced-word frequency vectors of the sequences with respect to the selected pattern P. In a follow-up paper [9], we used a more efficient algorithm to compare the spaced-word composition of sequences and extended 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. 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 calculated from individual patterns $P_i \in \mathcal{P}$ that are calculated as in our previous *single-pattern* approach.

We evaluated this *multiple-pattern approach* by applying it to phylogeny analysis. For a given set $\mathcal{P}$ of patterns, 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 [10]. The resulting distance matrices were used as input for *Neighbour Joining* [11] to generate trees, and we compared the resulting tree topologies to trusted reference topologies using the *Robinson-Foulds* distance [12]. As benchmark data sets, we used simulated and real-world DNA and protein sequences.

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 could show 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 subsequent paper [13], 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.

3. rasbhari

The performance of pattern-based approaches such as *spaced words* depends on the underlying patterns. In a recent paper [14], we showed that the *overlap complexity* of a pattern set that has been introduced by Ilie and Ilie [15] is closely related to the *variance* of the number of matches between two evolutionarily related sequences with respect to this pattern set. We proposed a modified hill-climbing algorithm to optimize pattern sets for database searching, read mapping and alignment-free sequence comparison of nucleic-acid sequences; the implementation of this algorithm is called *rasbhari*. Depending on the application at hand, *rasbhari* can either minimize the *overlap complexity* of pattern sets, maximize their *sensitivity* in database searching or minimize the *variance* of the number of pattern-based matches in alignment-free

sequence comparison. We showed that, for database searching, *rasbhari* generates pattern sets with slightly higher sensitivity than existing approaches. In *spaced words*, pattern sets calculated with *rasbhari* led to more accurate estimates of phylogenetic distances than the randomly generated pattern sets that we previously used. Finally, we used *rasbhari* to generate patterns for short read classification with *CLARK-S* [16]. Here too, the sensitivity of the results could be improved, compared to the default patterns of the program.

4. Filtered Spaced-Words Matches

Most recently, we applied the *spaced-words* concept in an approach called *Filtered Spaced Word Matches (FSWM)* to estimate phylogenetic distances between large genomic sequences. For a binary pattern P, *FSWM* rapidly identifies *spaced word-matches* between input sequences, *i.e.* gap-free local alignments with matching nucleotides at the *match* positions and with mismatches allowed at the *don't-care* positions. The program then estimates the number of nucleotide substitutions per site by considering the nucleotides aligned at the *don't-care* positions of the identified spaced-word matches. To reduce the noise from spurious random matches, *FSWM* uses a filtering procedure where spaced-word matches are discarded if the overall similarity between the aligned segments is below a threshold. We showed that this approach can accurately estimate substitution frequencies even for large, distantly related sequences that cannot be analyzed with existing alignment-free methods; phylogenetic trees constructed with *FSWM* distances are of high quality. A program run on a pair of eukaryotic genomes of a few hundred *Mb* each takes a few minutes.

Acknowledgments

Salma Sohrabi-Jahromi wrote a program to simulate genomic sequences that we used in our study.

References

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

[2] Michael Höhl, Isidore Rigoutsos, and Mark A. Ragan. Pattern-based phylogenetic distance estimation and tree reconstruction. *Evolutionary Bioinformatics Online*, 2:359–375, 2006.

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

[4] 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.

[5] 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.

[6] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene M. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

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

[8] 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.

[9] 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.

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

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

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

[13] 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.

[14] Lars Hahn, Chris-André Leimeister, Rachid Ounit, Stefano Lonardi, and Burkhard Morgenstern. *rasbhari*: optimizing spaced seeds for database searching, read mapping and alignment-free sequence comparison. *PLOS Computational Biology*, 12(10):e1005107, 2016.

[15] Lucian Ilie and Silvana Ilie. Multiple spaced seeds for homology search. *Bioinformatics*, 23:2969–2977, 2007.

[16] Rachid Ounit and Stefano Lonardi. *Algorithms in Bioinformatics: 15th International Workshop, WABI 2015, Atlanta, GA, USA, September 10-12, 2015, Proceedings*, chapter Higher Classification Accuracy of Short Metagenomic Reads by Discriminative Spaced k-mers, pages 286–295. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.