n-grams in alignment-free sequence analysis

Michał Burdukiewicz

University of Wrocław, Department of Genomics, Poland

Outline

Sequence comparision methods:

- alignment-based,
- alignment-free.

Alignment-free, n-gram based methods for sequence comparison are usually faster, do not require scoring schemes, do not require gene selection.

Usual application: metagenomics (Wood & Salzberg, 2014; Ounit, Wanamaker, Close, & Lonardi, 2015).

Although alignment-free methods can be used to recover topology of a phylogeny, they cannot represent lengths of the branches, which require number of substititutions per site. Despite that, some n-gram based methods can also calculate branch lenghts (Yi & Jin, 2013).

Although alignment-free methods can be used to recover topology of a phylogeny, they cannot represent lengths of the branches, which require number of substititutions per site. Despite that, some n-gram based methods can also calculate branch lenghts (Yi & Jin, 2013).

Outline

First usage of n-gram analysis: differences of overlapping and countinous 2- or 3-grams between sets of sequences (Blaisdell, 1986).

We consider only two sequences, Q and S, that contain only symbols from the alphabet u.

$$d_{ngram}(Q,S) = \sum_{i=1}^{u^n} (q_i - s_i)^2 \tag{1}$$

- d_{ngram}: n-grams distance between sequences;
- q_i: frequency of the i-th of u^n possible substrings of length n in Q;
- s_i: frequency of the i-th of u^n possible substrings of length n in S;

The simple n-gram distance is not powerful when applied to very similar sequences (Höhl & Ragan, 2007).

The frequency of words varies between sequences independently of their relationships.
Practical corrections to remove random background:

- zero-order Markov method substracts background (Pride, Meinersmann, Wassenaar, & Blaser, 2003);
- composition vector method (Qi, Wang, & Hao, 2004);
- feature frequency profile (Sims & Kim, 2011).

The composition vector method substracts random background by normalizing counts of the n-grams (Qi et al., 2004).

$$d_{cv}(Q,S) = \frac{1 - C(Q,S)}{2} \tag{2}$$

$$C(Q,S) = \frac{\sum q_{ni} s_{ni}}{\sqrt{\sum q_{ni}^2 \sum s_{ni}^2}} \tag{3}$$

- d_{cv}: composition vector distance;
- q_i: normalized frequency of the i-th of u^n possible substrings of length n in Q;
- s_i: normalized frequency of the i-th of u^n possible substrings of length n in S.

Outline

Closely related sequences contain more exact matches than divergent sequences (Kurtz et al., 2004).
Differences between semi-grammar methods and n-gram methods:

- no longer requires specifying n;
- preserves sequence of the substrings;
- potentially much longer computation time.

Suffix tree allows quicker deconstruction of the sequence.
See:

$$S = \mathrm{TCCT\$}$$

Outline

A grammar is a set of rules for decomposing a string into its elements.

$$\mathrm{TCCT} \rightarrow \mathrm{T_1C_2T_1}$$

Differences between semi-grammar methods and grammar methods:

- even slower than grammar methods;
- calculates mutation distances.

Common methods of compression: Lempel-Ziv factorization, average common substring.

Grammar methods are the msot powerful aligment-free methods (?, ?), but are more sensitive to imperfections in data.

Outline

Blaisdell, B. E. (1986, July). A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, *83*(14), 5155–5159. Retrieved 2015-06-28, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC323909/

Höhl, M., & Ragan, M. A. (2007, April). Is Multiple-Sequence Alignment Required for Accurate Inference of Phylogeny? *Systematic Biology*, *56*(2), 206–221. Retrieved 2015-06-29, from http://sysbio.oxfordjournals.org/content/56/2/206 doi: 10.1080/10635150701294741

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004, January). Versatile and open software for comparing large genomes. *Genome Biology*, *5*(2), R12. Retrieved 2015-06-29, from http://genomebiology.com/2004/5/2/R12/abstract doi: 10.1186/gb-2004-5-2-r12

Otu, H. H., & Sayood, K. (2003, November). A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, *19*(16), 2122–2130. Retrieved 2015-06-29, from `http://bioinformatics.oxfordjournals.org/content/19/16/2122` doi: 10.1093/bioinformatics/btg295

Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015, March). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, *16*(1), 236. Retrieved 2015-06-28, from `http://www.biomedcentral.com/1471-2164/16/236/abstract` doi: 10.1186/s12864-015-1419-2

Pride, D. T., Meinersmann, R. J., Wassenaar, T. M., & Blaser, M. J. (2003, February). Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Research*, *13*(2), 145–158. doi: 10.1101/gr.335003

Qi, J., Wang, B., & Hao, B.-I. (2004, January). Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *Journal of Molecular Evolution*, *58*(1), 1–11. doi: 10.1007/s00239-003-2493-7

Sims, G. E., & Kim, S.-H. (2011, May). Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs). *Proceedings of the National Academy of Sciences*, *108*(20), 8329–8334. Retrieved 2015-06-29, from `http://www.pnas.org/content/108/20/8329` doi: 10.1073/pnas.1105168108

Wood, D. E., & Salzberg, S. L. (2014, March). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. Retrieved 2015-06-28, from `http://genomebiology.com/2014/15/3/R46/abstract` doi: 10.1186/gb-2014-15-3-r46

Yi, H., & Jin, L. (2013, April). Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Research*, *41*(7), e75. doi: 10.1093/nar/gkt003