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.

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

- d_{ngram}: distance between n-grams;
- 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 zero-order Markov method substracts bias background (Pride, Meinersmann, Wassenaar, & Blaser, 2003).

$$E(W) = n \prod_{j \in u} f_j^j \tag{2}$$

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

$$d_{cv}(Q,S) = \frac{1 - C(Q,S)}{2} \tag{3}$$

$$C(Q,S) = \frac{\sum q_i s_i}{\sqrt{\sum q_i^2 \sum s_i^2}} \tag{4}$$

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

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