Bioinformatics, 2017, 1–3
doi: 10.1093/bioinformatics/btx367
Advance Access Publication Date: 10 June 2017
Applications Note

OXFORD

Genome analysis

MoDMaps3D: an interactive webtool for the quantification and 3D visualization of interrelationships in a dataset of DNA sequences

Rallis Karamichalis[1] and Lila Kari[1,2,*]

[1]Department of Computer Science, University of Western Ontario, London, ON N6A 5B7, Canada and [2]School of Computing Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on March 20, 2017; revised on May 3, 2017; editorial decision on June 2, 2017; accepted on June 5, 2017

Abstract

Summary: MoDMaps3D (Molecular Distance Maps 3D) is an alignment-free, fast, computationally lightweight webtool for computing and visualizing the interrelationships within any dataset of DNA sequences, based on pairwise comparisons between their oligomer compositions. MoDMaps3D is a general-purpose interactive webtool that is free of any requirements on sequence composition, position of the sequences in their respective genomes, presence or absence of similarity or homology, sequence length, or even sequence origin (biological or computer-generated).
Availability and implementation: MoDMaps3D is open source, cross-platform compatible, and is available under the MIT license at http://moleculardistancemaps.github.io/MoDMaps3D/. The source code is available at https://github.com/moleculardistancemaps/MoDMaps3D/.
Contact: lila@uwaterloo.ca
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Phylogenetic trees have been the traditional means to represent evolutionary history and species classification, but there is a growing realization, see Gusfield (2014), that some type of graphs or networks rather than trees are often needed, e.g. to take into account phenomena such as recombination, hybridization, horizontal gene transfer, and convergent evolution. At the same time, alignment-free methods, see Bonham-Carter *et al.* (2014), have been proposed to complement conventional morphological or sequence-alignment-based methods for phylogenetic analysis. Combining features of both these approaches, we propose MoDMaps3D (Molecular Distance Maps 3D, hereafter MoDMaps), an alignment-free webtool for computing and displaying sequence and species' relatedness. MoDMaps uses approximated information distance (AID) to quantify the pairwise differences in oligomer composition for all input genomic sequences, and visualizes the interrelationships thus obtained as an interactive map in three-dimensional Euclidean space.

As oligomer composition of genomic sequences has been shown to be a source of taxonomic information, see e.g. Deschavanne *et al.* (1999), MoDMaps could complement other alignment-based or alignment-free methods for species identification and classification. The advantage of the distance computation module of the MoDMaps webtool is that it is general-purpose and can be applied to any dataset of genomic sequences. In particular, this module is free of any requirements on: sequence composition, position of the sequences in their respective genomes, presence or absence of similarity or homology, and sequence length (same or different, several kbp-long or complete genomes). The advantages of the visualization module are that the output map is easy to explore, as well as easy to interpret visually. In particular, the spatial distances between a sequence-representing point and all the other points in the map are meaningful, in terms of their interrelatedness.

Note that, although this webtool can compute pairwise distances and visualize the resulting relationships among any DNA genomic

sequences (and, implicitly, among their originating organisms and species), it is alignment-free and, as such, it does not invoke the concept of phylogeny. Nevertheless, all prebuilt ModMaps are remarkably consistent with known taxonomies, which confirms that oligomer composition can be a source of taxonomic information.

2 Materials and methods

Creating a MoDMap involves three computational modules: Chaos Game Representation (CGR), Approximated Information Distance, and Multidimensional Scaling (MDS).

The CGR of a DNA sequence was defined in Jeffrey (1990) as a graphical representation of a DNA sequence, where the patterns correspond to the frequencies of k-mers in the sequence. CGR was proposed by Deschavanne *et al.* (1999)—and later confirmed—as a candidate for the role of *genomic signature*, defined by Karlin and Burge (1995) as any quantitative characteristic of a DNA sequence that is pervasive along the genome, while being dissimilar for genomic sequences originating from organisms of different species. The CGR of a sequence s is stored as a $2^k \times 2^k$ matrix, where each entry represents the number of occurrences of one of the possible k-mers in the sequence s.

The AID between two DNA sequences s and t, introduced in Li *et al.* (2004) and slightly modified in Karamichalis *et al.* (2015), is defined as:

$$d^k(s,t) = \frac{|M_k(s) \backslash M_k(t)| + |M_k(t) \backslash M_k(s)|}{|M_k(\{s,t\})|};\ 0 \leq d^k(s,t) \leq 1,$$

where for a sequence s, the set $M_k(s)$ comprises all distinct k-mers that occur in s, and $M_k(\{s,t\})$ comprises all distinct k-mers that occur in s or t. For two sets X and Y, the set $X \backslash Y$ comprises all elements that belong to X but not to Y, while $|X|$ is the number of elements of X. Informally, $d^k(s,t)$ is the ratio of the number of k-mers that occur in s but not in t or vice versa, to the total number of k-mers that occur in s or in t. In MoDMaps, $M_k(s)$ is computed as the number of non-zero entries in the CGR matrix of the sequence s, while $M_k(\{s,t\})$ is the number of non-zero entries of the sum of the CGRs of the sequences s and t.

MDS is an information visualization technique, see Kruskal (1964), that takes as input a distance matrix of pairwise distances among a set of items, and outputs a spatial representation of the items in a common Euclidean space. Each item is represented as a point, and the spatial distance between any two points approximates the distance between the items in the distance matrix.

Given an input set of n DNA sequences, MoDMaps first computes the CGR of each DNA sequence. Secondly, it computes all pairwise approximation information distances between CGRs, and stores the distance values in a distance matrix. The third step is to use this distance matrix as input for MDS, which then outputs a visualization of the input DNA sequences as points in a 3D space. The overall time complexity of the algorithm is $O(nm) + O(n^2) + O(n^3)$, for the computation of CGR, distance matrix, and MDS, respectively, where m is the maximum length of a sequence in the dataset.

3 Software description

Figure 1, a MoDMap for subphylum Vertebrata, illustrates one of the pre-built maps that can be explored interactively.

MoDMaps not only allows an intuitive and clear static representation of multi-dimensional data, but also encourages a viewer's involvement through dynamic interactions: One may zoom in and zoom out, rotate, move through the map, query the underlying genomic sequence details, visualize the structural composition of a selected DNA sequence, and search and highlight subsets of DNA sequences of interest.

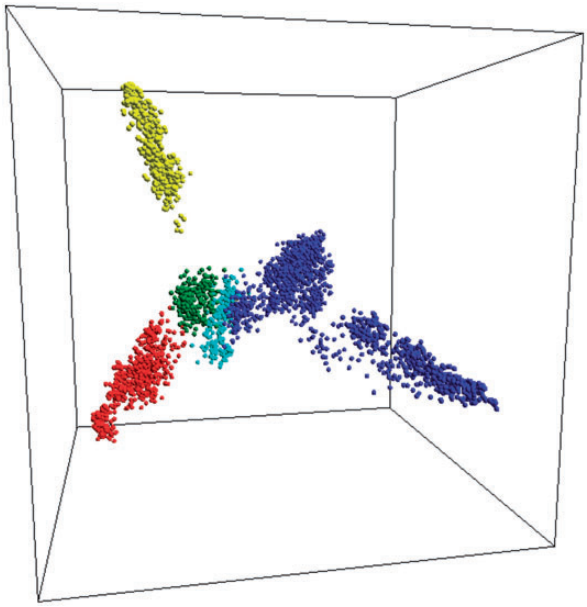

Fig. 1. MoDMap of subphylum Vertebrata: all 4, 322 complete reference vertebrate mtDNA genomes on NCBI (11 January, 2017). Blue, turquoise, green, red and yellow, represent 2, 319 fish, 291 amphibian, 285 reptile, 874 mammal and 553 bird mtDNA genomes

MoDMaps can be used in different ways: to explore a prebuilt map, to build a map *ab initio* for a new set of sequences, or add new DNA sequences to an existing map. All pre-built maps use $k = 9$, as this value empirically produced the best results while at the same time being computationally inexpensive. For *ab initio* or extended maps, the user can choose other values of k, from $k = 3$ (for short, or for dissimilar sequences) to $k = 12$ (for whole genomes, or for highly similar sequences). MoDMaps also provides the option to compute separately the AID between any pair of sequences, entered as NCBI numbers. See Supplementary Material for additional information.

MoDMaps is written in Javascript, and uses jQuery (a free open-source cross-platform Javascript library), Bootstrap (a free open-source collection of user interface elements), and Three.js (a cross-browser JavaScript library using WebGL for displaying animated 3D computer graphics). In addition the Parallel.js library is used for parallel computation when applicable.

Acknowledgements

We thank K. Hill, A. Poon, D. Smith and S. Solis-Reyes for discussions.

Funding

This research was supported by NSERC (Natural Sciences and Engineering Research Council of Canada) Discovery Grants R2824A01 (L.K.).

Conflict of Interest: none declared.

References

Bonham-Carter,O. *et al.* (2014) Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Brief. Bioinformatics*, **15**, 890.

Deschavanne,P.J. *et al*. (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol*., **16**, 1391–1399.

Gusfield,D. (2014) *ReCombinatorics: The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks*. MIT Press, Cambridge MA, USA.

Jeffrey,H.J. (1990) Chaos game representation of gene structure. *Nucleic Acids Res*., **18**, 2163–2170.

Karamichalis,R. *et al*. (2015) An investigation into inter- and intragenomic variations of graphic genomic signatures. *BMC Bioinformatics*, **16**, 246.

Karlin,S., and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*., **11**, 283–290.

Kruskal,J. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1–27.

Li,M. *et al*. (2004) The similarity metric. *Inform. Theory IEEE Trans*. **50**, 3250–3264.