Notes of synNMap

Mingwei Li

2017-1-22

Contents

1 Introduction

In genomic data analysis, one of the visual methods that helps studying the strutral differences and realtions between genomes is called a syntenic dot-plots. It is a scatter plot that points out all of the matching regions between two genomes, infering regions that originated from the same ancestor.

An interactive tool that enables different levels of inspections is SynMap. [https://genomevolution.org/wiki/index.php/SynMap] It is a tool that compare two genomes. [fig: 2d synMap]

In a more general case, one may want to compare multiple genomes, trying to study their relations in a big picture. One simple expansion to the current tools is comparing 3 genomes in a 3D syntenic plot. This works for simple cases. For example, three genome being very similar pairwise, resulting a 3D plot with a diagnal and several segments else where[fig 3d synmap] However, this may not work in general, especially in the case such as one very different genome 'breaks' the alignment of the other 2 similar genomes in a 3D plot.

Another way of generalization with current visual tools is doing pairwise syntenic plots, with the believe that pairwise comparison of all genomes of interest gives full information about their relations. This method may work for up to 4 or 5 genomes, with the hope that one is able to see the whole picture with up to 10 syntenic plots.[pic]

We propose a method that generates a 'genome map' that depicts their relations by pointing out their locations in map. With a general notion of a map that two points being closer infers a close relationship between the two entities, in this case genomes, one can easily see a big picture of multiple genomes that he/she is studying.[fig synNMap]

2 Problem setting

In the process of drawing a synmap of two genomes, the program computes a measure called synonymous mutation rate (ks) describing a matching score for each pair of aligned genes between two genomes. This score ranges from 0 to infinity, with 0 means perfect alignment and infinity indicates no alignment or an uncaught error. We want to utilize these measures to build a notion of distance between two genomes. Once we have the pairwise distances among multiple genomes, we want to construct a space infered by these distances and project it onto 2D screen.

In particular, when comparing two genomes, for example, human and chimp, we are given gene pairs and their synonnymous nutation rate ks as a measure of similarity. (see table)

Table 1: Problem sample input

Gene labels		
Human	Chimpanzee	ks
h1	c1	0.0056
h2	c2	0.0188
h3	c3	72.6574
...	...	...

We want to build a notion of distances between two genomes.

$$dist(Human, Chimp) = f(ks1, ks2...)$$

Generally the gene indices in such table may be duplicated, since there may be a match of a single gene to multiple genes. To simplify the notation in matrix form, define

$$KS = \begin{bmatrix} ks_{1,1} & ks_{1,2} & ks_{1,3} & \dots & ks_{1,c} \\ ks_{2,1} & ks_{2,2} & ks_{2,3} & \dots & ks_{2,c} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ ks_{h,1} & ks_{h,2} & ks_{h,3} & \dots & ks_{h,c} \end{bmatrix}$$

where h, c are gene count of human and chimpanzee. Now

$$dist(human, chimp) = f(KS)$$

(TBC)

3 TODOs

3.1 math

from KS to dist measure, why it works,
note loss of order in genome/chromosomes
dot product/euclidean dist =¿ PCA

3.2 viz

basic static viz
propose indication of lost distance
interaction scheme

extension from genomes comparison to gene/chromosome/population

Some paragraph asdasda

3.3 Section 2

3.4 Section 3

aaa.aaa.aaa