SynMapN: Interactive Visual Comparison for Multiple Genomes

Mingwei Li* Andreina Castillo Siri† Asher Keith Haug-Baltzell‡ Eric Lyons§ Carlos Scheidegger¶

University of Arizona

ABSTRACT

Interactive visualization has become a powerful means to explore syntenic relationships among two genomes, with a variety of available tools for domain scientists to employ. However, these tools do not tend to scale well in the case where *many* genomes are compared against one another. This poster describes ongoing efforts to build techniques and tools to help genomicists understand sets of genomes and their syntenic relationships. Our main contribution is a mechanism that defines *set distances*: this can be used to compare entire genomes to one another, as well as *sets of genomes* to each other. Currently, we use this mechanism to generate dimensionality reduction visualizations. We discuss limitations of this approach, as well as future directions.

Index Terms:

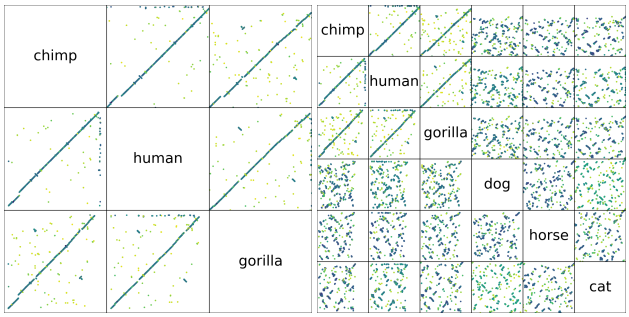

Figure 1: Two synteny matrix plots of three and six species respectively. Note that it becomes harder to inspect the relationship between genomes as the number of genomes involved increases.

1 INTRODUCTION

In comparative genomics, one of the visual methods that helps studying the structural relations between two genomes is called the syntenic dotplot [2, 3]. It is a scatterplot that depicts matched genes between two genomes; these matched genes imply *syntenic regions*: those that likely originated from the same ancestor. During genomic analysis, a measure called synonymous mutation rate between two genes (*ks*) is computed. Because many codons translate to the same amino acid (for example, TCT and TCC both translate to serine), it is possible that a mutation does not change the amino acid that is encoded in the gene: this is a "synonymous" mutation. The *ks* value is a rate of such mutations, normalized by gene sizes. Because those mutations are mostly harmless, they tend to not change selection pressure on the genomes. At the same time, they are transmitted hereditarily, and so they can be seen as "biological clocks", and thus can be used to estimate evolutionary relations. The larger the *ks* value is, the longer it has been since these two genomes diverged. In the dotplot, two axes represent gene locations of two genomes respectively. Each dot on the plot represents a match between two genes, and the dot colors usually encode *ks* values.

In some ideal cases, a perfect alignment of two genomes can be observed by seeing a line along the diagonal of a plot, similar to a plot of function $y = x$. This means that the two entities in concern have same genes presented in order. More interesting events like duplication and inversion of genes can be easily spotted from the syntenic dotplot as well (**??**). In other cases, it is also common to see only sparse dots, indicating no significant alignment in the two genomes.

Although tools such as SynMap [2] (our collaborator's tool for drawing syntenic dotplot) and MizBee [6] are widely useful, especially in comparing two genomes, there are not many tools designed to explore more than two genomes at once. Because of projects such as the thousand genomes project [1], there is a demand for understanding the relationships among many genomes. Adding one more axis of genome to make a 3D scatter plot, for example, may work in some (very limited) cases. We would see a line in diagonal in a comparison among human, chimpanzee and gorilla genomes [4]. However, it is not visually intuitive to find out complicated relations in three species in general. In fact, Tory et al. [7] showed that a 3D landscape works not as well as a 2D map in specific tasks such as search and point estimation. This casts doubt on whether three-dimensional techniques are useful for other tasks such as synteny visualiatization. Moreover, a 3D scatterplot clearly does not generalize to more than three genomes.

Traditional scatterplot matrix displays can be readily adapted for comparing multiple genomes, with each subplot being a syntenic dotplot of two of the genomes. However, scatterplot matrices do not scale up well beyond a moderate number of plots []. As the number of genomes increases, it becomes harder to tell the overall relation between the genomes.

In this poster, we describe an ongoing collaboration with domain scientists using multiple-genome comparisons in two use cases. The first use case involves a comparison of about one hundred *arabidopsis thaliana* genomes. In this case, the differences between the genomes are encoded by a set of SNPs (single-nucleotide polymorphisms). The second use case involves 17 genomes, each of a different species of the plasmodium genus. In this case, the genome differences are more complex, often architectural (that is, involving inversions and duplications of entire portions of the genome), and need to be encoded by the entire synteny map.

2 METHOD

In the process of generating a SynMap of two genomes, the synonymous mutation rate (*ks* as described above) is computed for each pair of aligned genes between two genomes. This score ranges from 0 to $+\infty$, or arbitrary large number in data, with 0 means perfect alignment and infinity indicates no alignment or an error. We want to utilize these measures to build a notion of distance between the two genomes (that is, between two *sets* of genes).

Now we want to find a function of all the *ks* values that describe

*e-mail: mwli@email.arizona.edu
†email: aicasti1@email.arizona.edu
‡email: ahaug@email.arizona.edu
§email: elyons.uoa@gmail.com
¶e-mail: cscheid@cs.arizona.edu

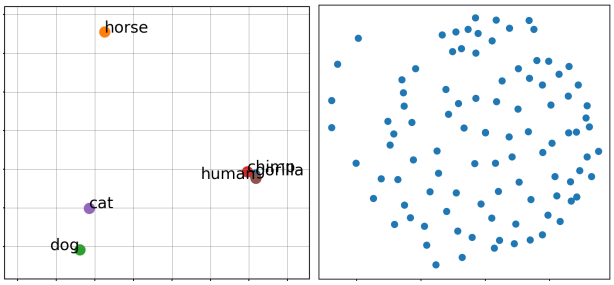

Figure 2: With the notion of similarity between sets of genes we define, we can create a single plot that summarizes the similarities across genomes. Note that although this is helpful for the six-species plot on the left, this plot is not particularly helpful for the similarity comparison between the *arabidopsis* genomes (on the right). We discuss this further in the text.

the distance between two genomes or, equivalently, a notion of similarity (or a "kernel") between the genomes. Generally, for any gene in one genome, there might be syntenic matches to more than one gene on the other genome. Computationally, this means we encode the similarity between the two genomes by arranging the *ks* values in a matrix:

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

where $ks_{1,1}$ stores the *ks* value between the first gene of human and the first gene of chimpanzee, and so on. The indices of rows and columns are in the order of gene locations in ordered chromosomes. h and c are gene counts of human and chimpanzee respectively. One can think of K as an image data of the SymMap plot. Now we want a function of K that outputs a scalar value to describe the similarity between two genomes.

To define a function so that comparison of a genome to itself gives a similarity measure close to 1, we used f to compute kernels between any two genomes, which are later used to plot Fig. 2

$$similarity(Human, Chimp) = f(K) = \sum_{i,j} e^{-\lambda K_{ij}} / \sqrt{c \times h} \qquad (1)$$

Where λ is a scalar constant related to sensitivity of the similarity measure. c and h are the gene counts defined before.

Once we have the pairwise distances among multiple genomes (stored in matrix M), we have an implied space, which we can then project it onto a 2D screen using (for example) Kernel PCA. One example of this technique being used can be seen in Fig. 2.

The definition above works for general comparisons between genomes whose differences are encoded with syntenic matches. In the case of genome differences encoded through sets of SNPs, we use a simpler definition of distances, based entirely on comparing the nucleotide polymorphisms and counting their differences.

3 Limitations and Ongoing Work

SynMapN gives an overview of genomes, but the detailed patterns within two genomes such as inversions and duplications are not visible. As can be clearly seen on the right side of Fig. 2, the dimensionality reduction plots break down for large numbers of genomes. In addition (and in contrast to the SynMap plots) they provide little insight about *what* causes the genomes to be different. In our ongoing collaboration, we are developing tools that will provide the level of detail present in syntenic dotplots, with the visual scalability of dimensionality reduction plots. Consider, for

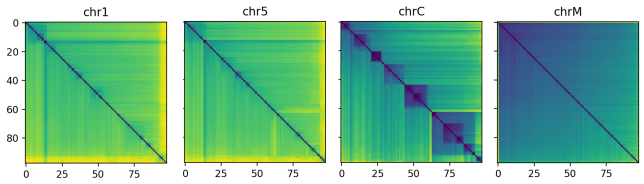

Figure 3: A matrix plot of the genomic differences broken down by chromosome. Note the difference in variability across different chromosomes (specifically the genome in the chloroplast, encoded as "chromosome C"). We are currently developing tools to help genomicists to better understand these differences.

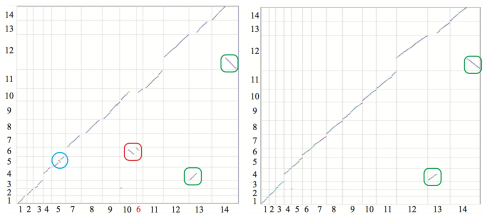

Figure 4: In cases where the task involves complex genomic differences such as the ones in our *plasmodium* genome comparison use case shown here, currently-available techniques are not sufficient. Even though in this figure we show only two syntenic dot plots, the use case involves a total of 136 different comparisons, and that requires the development of novel tools and techniques.

example, a breakdown of the *arabidopsis* genome differences with respect to different chromosomes, as shown in Fig. 3. Especially in settings where genomic differences are complex and structural in nature (Fig. 4), the current dimensionality reduction plot will be not be sufficient.

In a complete visualization system, we expect to enable users to navigate through various visualizations for different levels of details, from high level SynMapN to very detailed comparison of two specific genomes in SynMap. To further increase the number of levels, we can split genomes into individual chromosomes and compare them in SymMapN. We are also exploring the possibilities of capturing certain features in a SynMap plot, for example, the diagonal alignments or anti-diagonal alignments, using image processing techniques. More improvements on interaction can be done through studying observation level interactions [5], to enable users to specify desired clustering criteria through dragging points in the plot.

References

[1] 1000 genomes project. `http://www.internationalgenome.org/about`. Accessed: 2017-06-14.

[2] CoGepedia synmap. `https://genomevolution.org/wiki/index.php/SynMap`. Accessed: 2017-06-14.

[3] CoGepedia syntenic dotplot. `https://genomevolution.org/wiki/index.php/Syntenic_dotplot`. Accessed: 2017-06-14.

[4] Synmap 3d. `https://genomevolution.org/r/lf1i`. Accessed: 2017-06-14.

[5] A. Endert, C. Han, D. Maiti, L. House, and C. North. Observation-level interaction with statistical models for visual analytics. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pp. 121–130. IEEE, 2011.

[6] M. Meyer, T. Munzner, and H. Pfister. Mizbee: a multiscale synteny browser. *IEEE transactions on visualization and computer graphics*, 15(6):897–904, 2009.

[7] M. Tory, D. Sprague, F. Wu, W. Y. So, and T. Munzner. Spatialization design: Comparing points and landscapes. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1262–1269, 2007.