Visual Exploration of Genomic Data

Michail Vlachos[1], Bahar Taneri[2], Eamonn Keogh[3], and Philip S. Yu[1]

[1] IBM T.J. Watson Research Center, Hawthorne, NY, USA
[2] Scripps Institute of Oceanography, UCSD, CA, USA
[3] University of California Riverside, CA, USA

Abstract. In this study, we present methods for comparative visualization of DNA sequences in two dimensions. First, we illustrate a transformation of gene sequences into numerical trajectories. The trajectory visually captures the nucleotide content of each sequence, allowing for fast and easy visualization of long DNA sequences. Then, we project the relative placement of the trajectories on the 2D plane using a spanning-tree arrangement method, which allows the efficient comparison of multiple sequences. We demonstrate with various examples the applicability of our technique in evolutionary biology and specifically in capturing and visualizing the molecular phylogeny between species.

1 Introduction

Identification of evolutionary distances among species has always been a topic of interest to researchers. Several different methods have been used to identify the evolutionary relationships between species, including taxonomic, phylogenetic analyses, geometric morphometric data analysis. In the post-genome era, more accurate evolutionary views have been reached using DNA sequence analyses of species [6, 9].

In this work, we also provide a molecular vision of evolution through comparison and visualization of DNA sequences. Using comparative mitochondrial DNA analysis, we illustrate the evolutionary distances among various mammalian species. Mitochondrial DNA (mtDNA) analyses have been proven useful in establishing phylogeny among a wide range of species [2, 7, 8]. We achieve our goal by mapping the DNA nucleotide sequences into 2-dimensional trajectories. The purpose of this conversion is to facilitate the quick visual comparison between long DNA sequences. We evaluate the affinity between the resulting DNA trajectories by employing an elastic warping distance function. Our empirical results on mitochondrial DNA from various species, suggest that the utilized distance measure can reflect with great accuracy the divergence point between species. Finally, for visually comparing the evolutionary distance between the DNA trajectories we present a spanning-tree-based mapping technique. The technique arranges the objects on the 2-dimensional space, while retaining as much of the original structure as possible. We depict the enhanced visualization power that can be induced from the proposed mapping technique. All our results

are validated with freely available genomic data obtained from Genbank [4], and corroborate the current prevalent views on evolutional biology.

Previous work on DNA visualization has appeared in [3, 4], but the techniques pose limitations regarding the visual comparison between multiple DNA sequences. A technique that allows the comparison between different sequences in terms of their common subsequences has been presented in [1]. Our method is unique in that, it not only provides a visual representation of the nucleotide sequences, but also it deciphers the comparative phylogenetic distances among different species.

In the sections that follow we present a DNA conversion technique into trajectories and later on we demonstrate the spanning-tree mapping technique. The final section contains the empirical evaluation of both methods using mammalian DNA sequences.

2 Converting DNA to Trajectories

Visual comparison of DNA symbol strings can be particularly troublesome to perform, because typical DNA datasets contains thousands of symbols. Humans cannot easily compare or visually represent bulk of text; our brains are much more efficient at comparing lines or shapes. Therefore, a technique for converting a DNA string into a low dimensional shape, can significantly enhance our ability of interpreting and comparing very long DNA sequence data. Given a string of length n drawn from the alphabet `A,T,C,G`, which we will denote as *DNA*, we wish to convert it to a two-dimensional trajectory of length $n + 1$, which we denote as T. We can use the following rule to build the trajectory vector: $T(i) = T(i-1) + \mathbf{V}$, where $\mathbf{V}$ is a basis vector constructed as follows:

$$\mathbf{V} = \begin{cases} [0\ 1], \text{ if } DNA(i) = \texttt{A} \\ [1\ 0], \text{ if } DNA(i) = \texttt{T} \\ [0\ \text{-}1], \text{ if } DNA(i) = \texttt{C} \\ [\text{-}1\ 0], \text{ if } DNA(i) = \texttt{G}. \end{cases}$$

That is, starting from an initial reference point we will direct the trajectory on the relevant direction (up, down, left or right) based on the currently examined symbol. For example, if the sequence contains many `A` symbols then it will demonstrate a predominantly upward movement. Below, we demonstrate an arithmetic example of the trajectory construction.

Example: Suppose that the starting position T(1) = [0 0]. Then, for the DNA string `AATCG`, we get the trajectory vector {[0 0],[0 1],[0 2],[1 2],[1 1],[0 1]}.

2.1 Comparing Trajectories

In order to quantify the similarity between the resulting trajectories we utilize a warping distance, which can allow for a flexible matching between the DNA

[4] `http://www.ncbi.nlm.nih.gov/Genbank/`

trajectories, supporting local compressions and decompressions. The warping distance can be seen as a real-valued counterpart of the *Edit Distance*, which is customarily used for comparing DNA transcripts.

Suppose that Q and T are the trajectories that we wish to compare. If $Q = (Q_1, Q_2, \ldots, Q_n)$ and $Head(Q) = (Q_1, Q_2, \ldots, Q_{n-1})$ (and similarly for a sequence T) then the recursive equation to provide then warping distance between Q and T is:

$$DTW(Q,T) = D(Q_n, T_n) + min \begin{cases} DTW(Head(Q), Head(T)) \\ DTW(Head(Q), T) \\ DTW(Q, Head(T)) \end{cases}$$

where $D(\cdot,\cdot)$ is the distance between two points of the sequence. Typically, D is the Euclidean distance. The warping distance can be computed using a well known dynamic programming [12]. In Fig. 1 we can see the flexibility of matching that can be achieved between trajectories when utilizing the warping distance. On the left side we demonstrate the mapping between the human and the chimpanzee trajectories, which were derived from their respective mitochondrial DNA. On the right side, the matching between the human and the bear mtDNA is illustrated.

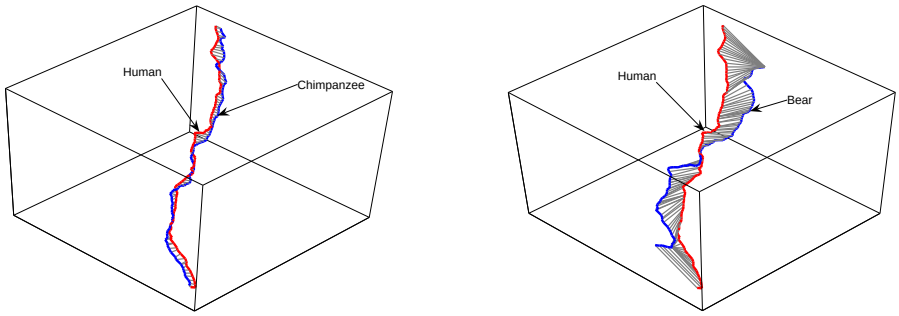

Fig. 1. Matching of DNA trajectories using DTW. Left: Human vs Chimpanzee, Right: Humans vs Bear

Even though the Warping distance can accommodate a flexible matching between the resulting DNA trajectories, it does not obey the triangle inequality (unlike the Euclidean distance). We will utilize this fact to motivate extensions on the triangulation mapping technique that is presented subsequently.

3 Spanning-Tree Visualization

Given a set of pairwise distances between objects we are seeking a way of visualizing their relationship on two dimensions, while retaining as much of the

original structure as possible. We revisit a mapping technique proposed by Lee, *et al.*, in [5], which utilizes the Minimum-Spanning-Tree (MST) and a triangulation method for preserving 2 distances per object on the two-dimensional space. The first distance preserved is the distance to the nearest neighbor of every object. The second distance can either be different for every object (e.g. its 2NN), or it can be the distance to a reference point. The latter option is the one that we adapt, which creates a powerful visualization technique that not only allows for preservation of Nearest Neighbors distances (local structure), but additionally retains distances with respect a single reference point, giving the option for global data view using that object as a pivot.

Once the MST is calculated the mapping on the 2D space can commence from any point/object that the user designates and the MST tree is mapped either in a breadth-first-search (BFS) or depth-first-search (DFS) manner. In this work we utilize a BFS mapping. We illustrate how the mapping works with a running example.

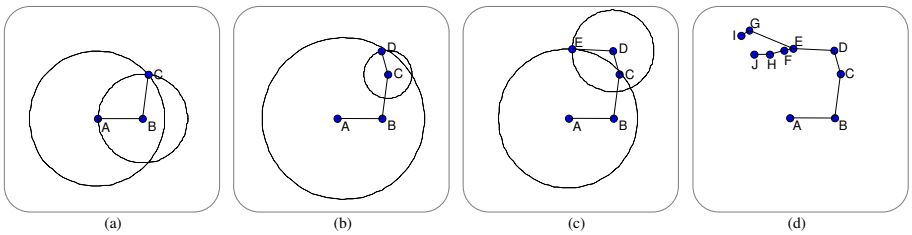

Fig. 2. 2D mapping of objects using spanning-tree and triangulation

Suppose the first two points (A and B) of the MST are already mapped, as shown in Fig. 2 (a). Let's assume that the second distance preserved per object is the distance with respect to a reference point which in our case is the first point. The third point is mapped at the intersection of circles centered at the reference points. The circles are centered at A and B with radii of $D(A, C)$ - the distance between points A and C- and $D(B, C)$, respectively. Due to the triangle inequality, the circles either intersect at 2 positions or at tangent. Any position on the circles' intersection will retain the original distances towards the two reference points. The position of point C is shown in Fig. 2 (a). The fourth point is mapped at the intersection of circles centered at A and C (Fig. 2 (b)) and the fifth point is mapped similarly (Fig. 2 (c)). The process continues until all the points of the MST at positioned on the 2D plane and the final result is shown in Fig. 2 (d).

3.1 Extensions for Non-metric Distances

The triangulation method proposed by Lee, *et al.*, is only applicable for metric distances when the circles around the reference points are guaranteed to intersect. Recall that, the Warping Distance used to quantify the distance between

the DNA trajectories does no obey the triangle inequality. This means that the reference circles may not necessarily intersect. We highlight necessary extensions to the triangulation method that allows its proper usage under non-metric distances.

We can identify two cases for the non-intersecting circles:

1. **Case 1:** One circle encloses each other,
2. **Case 2:** The two circles are disjoint and not enclosed within one another

For each of these cases we need to identify the position where to position an object with respect to the two circles, so that the object is mapped as close as possible to the circumference of both circles. In order words, we need to identify the locus of points that minimize the sum of distances to the perimeters of two circles.

One can show that the desired locus always lies on the line connecting the centers of the two circles. Case 1 is shown in Figure 3, and we can identify two sub-cases.

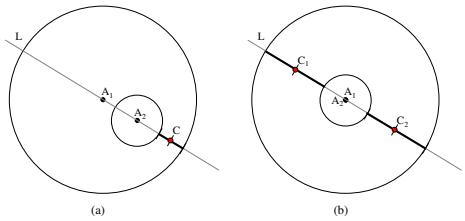

Fig. 3. Circles enclosed within one another

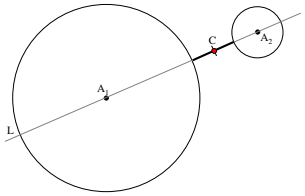

Fig. 4. Circles that are disjoint

- When the two circles have disjoint centers, then the point that minimizes the sum of distances to both perimeters, is point C on Fig. 3 (a), which lies on the line L connecting the two centers, and midway on the line segment with vertices the intersection of L with the circles' perimeters.
- In the case when the two circles have common centers, then there exist two points that satisfy the distance minimization property as shown in Fig. 3 (b).

Case 2 can be resolved in a similar way, which is shown in Fig. 4.

With the addition of the above rules, we can now discover the mapping positions of the objects on the two-dimensional plane, so that the original pairwise distances are satisfied as well as possible using the spanning-tree triangulation method.

4 Experimental Results

We demonstrate the usefulness of the proposed techniques on comparative molecular phylogenetic studies via visualization of mitochondrial DNA sequences.

We utilize publicly available mtDNA obtained from Genbank (see Table 1). All datasets used in this paper along with supplementary material can be found at the project website [5].

Mitochondrial DNA is passed on only from the mother during sexual reproduction, making the mitochondria clones. This means that there are minor changes in the mtDNA from generation to generation, unlike nuclear DNA which changes by 50% each generation. Therefore, mtDNA has a long *memory*. Each mtDNA string consists of approximately 16000 symbols (with mtDNA of humans being 16,571 nucleotides long, and all other mammals mtDNA are within plus or minus 1-3% of this).

Name	Species	mtDNA bps
Indian Elephant	Elephas Manimus Indicus	16800
African Elephant	Loxodonta Africana	16859
Blue Whale	Balaenoptera Musculus	16402
Finback Whale	Balaenoptera Physalus	16398
Hippopotamus	Hippopotamus Amphibius	16407
Human	Homo Sapiens	16571
Chimpanzee	Pan Troglodytes	16554
Pygmy chimpanzee	Pan Paniscus	16563
Dog	Canis familiaris	16727
American Bear	Ursus americanus	16841
Polar Bear	Ursus maritimus	17017

Table 1. Example from subset of the mitochondrial DNA data used for our visualizations

For our first experiment we utilize mtDNA from *Homo sapiens* and other primates. Figure 5 illustrates the spanning-tree mapping for 8 species. Our results are in general agreement with current evolutionary views. We also observe that not only the mapping is very accurate with regard to the evolutionary distance of the species, but the mapping preserves the clustering between the original groups that the various primates belong to. Specifically, Human, Pygmy chimpanzee, Chimpanzee and Orangutan belong to the *hominidae* group, the Gibbon to the *hylobatae* group and the Baboon and the Macaque to the *cercopithicae*.

Adjacent to this mapping, we provide a spanning-tree visualization that utilizes the most commonly used Euclidean distance instead of the Warping distance. One can observe that the use of the Euclidean distance introduces errors, such as mapping the gibbon closer to the human rather than to the orangutan, which is incorrect. Human and orangutan divergence took place approximately 11 million years ago. Whereas, gibbon and human divergence occurred approximately 15 million years ago [10]. According to the same source, gorilla divergence occurred about 6.5 million years ago and chimpanzee divergence took place about 5.5 million years ago.

[5] `http://www.cs.ucr.edu/~mvlachos/VizDNA/`

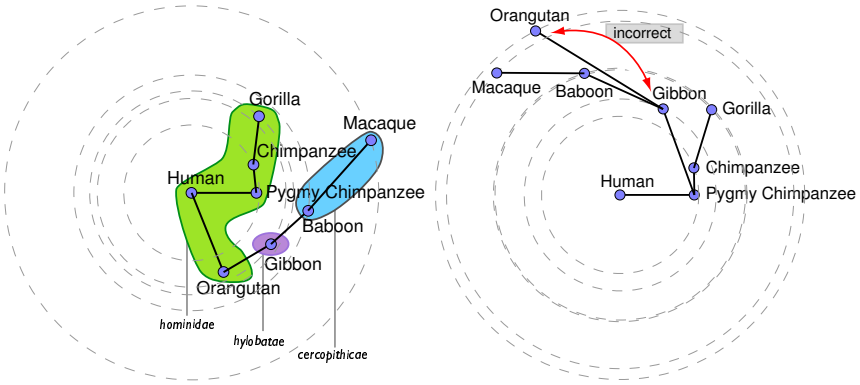

Fig. 5. Visualization of humans and other primates. Left: Using the Warping distance, Right: Using Euclidean distance to compare the DNA trajectories. Various mapping errors are indicated on the figure.

For our second example in Figure 6 we utilize a larger mammalian dataset and again take the human as the referential point. On this plot we use the formal names of the species (instead of their common names) and we also overlay on the figure the DNA trajectory of the respective mtDNA sequence. Again, the spanning-tree technique exhibits a very strong visualization capacity, particularly in unveiling the similarities and connections between the different species. For example, one can notice the great similarity of the hippopotamus with the whales. The hippopotami are indeed closely related to whales than to any other mammals. Whales and hippopotami diverged 54 million years ago, whereas the whale/hippopotamus group parted from the elephants about 105 million years ago. The group that includes hippopotami and whales/dolphins, but excludes all other mammals above is called Cetartiodactyla [11].

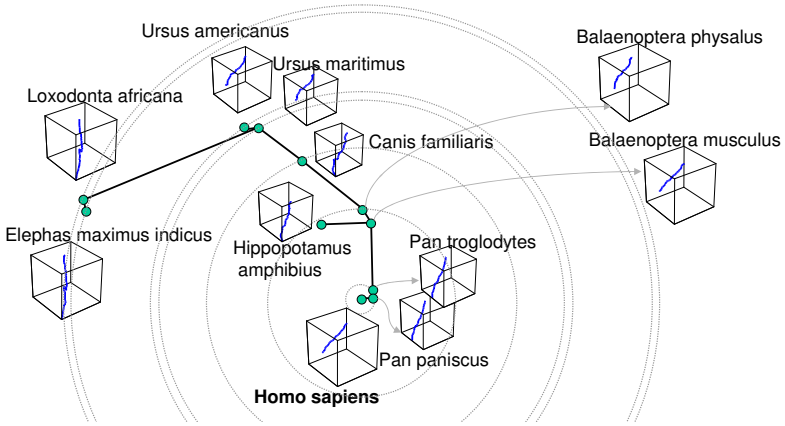

Fig. 6. Evolutionary visualization of mammalian species with respect to the human

5 Conclusions

We presented techniques that allow the effective visualization and comparison between DNA sequences, by transforming them into trajectories and mapping them on the two-dimensional plane. The mapping technique can have many biomedical applications, including advancement of diagnostic techniques for cancer data. This technology could both be applied for distinguishing cancer transcripts from normal ones, and for the identification of different cancer stages. Future direction of this work, includes expansion of our technique to transcriptome-wide screens of cancer transcripts in human and mouse transcriptomes.

References

1. A. Apostolico, F. Gong, and S. Lonardi. Verbumculus and the discovery of unusual words. In *Journal of Computer Science and Technology, Vol 19, No1: 22-41*, 2003.
2. A. Auch, S. Henz, B. Holland, and M. Goker. Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. In *BMC Bioinformatics 7:350*, 2006.
3. H. T. Chang, N.-W. Lo, W. C. Lu, and C. J. Kuo. Visualization and Comparison of DNA Sequences by Use of Three-Dimensional Trajectories. In *First Asia-Pacific Bioinformatics Conference, (APBC)*, 2003.
4. J. Herisson, N. Ferey, P. Gros, and O. M. R. Gherbi. 3D visualization and virtual exploration of genomic sequences. In *Data Science Journal 4: 82-91*, 2005.
5. R. Lee, J. Slagle, and H. Blum. A Triangulation Method for the Sequential Mapping of Points from N-Space to Two-Space. In *IEEE Transactions on Computers, Volume: C-26, Issue: 3*, pages 288–292, 1977.
6. C. Lockwood, W. Kimbel, and J. Lynch. Morphometrics and hominoid phylogeny: Support for a chimpanzee-human clade and differentiation among great ape subspecies. In *Proc. Natl. Acad. Sci. USA, 101(13), 4356-4360*, 2004.
7. T. Orrell and K. Carpenter. A phylogeny of the fish family Sparidae (porgies) inferred from mitochondrial sequence data. In *Mol Phylogenet Evol. 32(2): 425-434*, 2004.
8. A. K. Royyuru, G. Alexe, D. Platt1, R. Vijaya-Satya, L. Parida, S. Rosset, and G. Bhanot. Inferring Common Origins from mtDNA. In *Research in Computational Molecular Biology: 246-247*, 2006.
9. M. Ruvolo. Comparative primate genomics: the year of chimpanzee. In *Curr. Opin. Genet Dev. 14(6): 650-656*, 2004.
10. R. Stauffer, A. Walker, O. Ryder, M. Lyons-Weiler, and S. Hedges. Human and Ape Molecular Clocks and Constraints on Paleontological Hypotheses. In *The Journal of Heredity, 92(6): 469-474*, 2001.
11. B. M. Ursing and U. Arnason. Analyses of mitochondrial genomes strongly support a hippopotamus-whale clade. In *Proc. of the Royal Society of London, Series B, vol 265: 2251-2255*, 1998.
12. M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh. Indexing Multi-Dimensional Time-Series with Support for Multiple Distance Measures. In *Proc. of SIGKDD*, 2003.