Genome Visualization in Space

Leandro S. Marcolino[1], Bráulio R. G. M. Couto[2] and Marcos A. dos Santos[1]

Departamento de Ciência da Computação, Universidade Federal de Minas Gerais / UFMG[1]; Programa de Doutorado em Bioinformática, UFMG and Curso de Ciência da Computação, Centro Universitário de Belo Horizonte / UNIBH[2], Av. Antonio Carlos 6627, Belo Horizonte, Minas Gerais, 31270-010, Brasil. Email addresses: LSM: soriano@dcc.ufmg.br; BRGMC: braulio.couto@unibh.br; MAS: marcos@dcc.ufmg.br.

Abstract Phylogeny is an important field to understand evolution and the organization of life. However, most methods depend highly on manual study and analysis, making the construction of phylogeny error prone. Linear Algebra methods are known to be efficient to deal with the semantic relationships between a large number of elements in spaces of high dimensionality. Therefore, they can be useful to help the construction of phylogenetic trees. The ability to visualize the relationships between genomes is crucial in this process. In this paper, a linear algebra method, followed by optimization, is used to generate a visualization of a set of complete genomes. Using the proposed method we were able to visualize the relationships of 64 complete mitochondrial genomes, organized as six different groups, and of 31 complete mitochondrial genomes of mammals, organized as nine different groups. The prespecified groups could be seen clustered together in the visualization, and similar species were represented close together. Besides, there seems to be an evolutionary influence in the organization of the graph.

1. Introduction

Phylogeny is a very important field to understand evolution and the organization of life. However, many molecular phylogenies are built using sequences sampled from only a few genes. Besides, most methods depend highly on manual study and analysis, making the construction of phylogeny based on whole genomes difficult and error prone. The problem of analyzing genomes, however, is very similar to information retrieval from a large set of documents. In both problems, it is necessary to deal with an enormous amount of information, and to find semantic links between data. Fortunately, there are very good algorithms to deal with information retrieval. Singular value decomposition (SVD), for example, is used with great success (Berry *et al.* 1994). For example, linear algebra methods are used even by Google, enabling a better comprehension of a system as complex as the Internet (Eldén 2006; Stuart *et al.* 2002) presents a method to build phylogeny trees using SVD to analyze genomes. The method is demonstrated with verte-

brate mitochondrial genomes, and is later used to analyze whole bacterial genomes and whole eukaryotic genomes (Stuart and Berry 2004). Linear algebra methods are also used to study the different genotypes in the human population (Huggins *et al.* 2007).

Visualization techniques are essential to better analyze complex systems and can be very helpful to categorize species. There are a number of visualization tools to study a single genome (Lewis *et al*. 2002; Engels *et al*. 2006; Rutherford *et al*. 2000; Stothard and Wishart 2005; Gibson and Smith 2003; Ghai *et al*. 2004). However it is desirable to visualize the relationships between a set of genomes, in order to better comprehend the species. In Xie and Schlick (2000) is presented a visualization technique using SVD to analyze chemical databases. In this paper, we used that technique as a basis to develop a method for using genomes to visualize relationships among species in space (2D and 3D). This can facilitate the construction of phylogeny trees, enabling the analyzer to quickly have insights in the similarities between the different species. We are going to show the results of our approach using 832 mitochondrial proteins obtained from 64 whole mitochondrial genomes of vertebrates.

2. Material and methods

2.1 Sequence data

We used the same set of proteins as Stuart *et al*. (2002), 64 whole mitochondrial genomes from the NCBI genome database, each one with 13 genes, totaling 832 proteins in the data set. The following species were used in this paper: *Alligator mississippiensis, Artibeus jamaicensis, Aythya americana, Balaenoptera musculus, Balaenoptera physalus, Bos taurus, Canis familiaris, Carassius auratus, Cavia porcellus, Ceratotherium simum, Chelonia mydas, Chrysemys picta, Ciconia boyciana, Ciconia ciconia, Corvus frugilegus, Crossostoma lacustre, Cyprinus carpio, Danio rerio, Dasypus novemcinctus, Didelphis virginiana, Dinodon semicarinatus, Equus asinus, Equus caballus, Erinaceus europaeus, Eumeces egregius, Falco peregrinus, Felis catus, Gadus morhua, Gallus gallus, Halichoerus grypus, Hippopotamus amphibius, Homo sapiens, Latimeria chalumnae, Loxodonta africana, Macropus robustus, Mus musculus, Mustelus manazo, Myoxus glis, Oncorhynchus mykiss, Ornithorhynchus anatinus, Orycteropus afer, Oryctolagus cuniculus, Ovis aries, Paralichthys olivaceus, Pelomedusa subrufa, Phoca vitulina, Polypterus ornatipinnis, Pongo pygmaeus abelii, Protopterus dolloi, Raja radiata, Rattus norvegicus, Rhea americana, Rhinoceros unicornis, Salmo salar, Salvelinus alpinus, Salvelinus fontinalis, Scyliorhinus canicula, Smithornis sharpei, Squalus acanthias, Struthio camelus, Sus scrofa, Sciurus vulgaris, Talpa europaea, and Vidua chalybeata*.

2.2 Representation method

In order to visualize the genomes, we must represent each one as a point in space. The distance between the points should represent the differences in the genomes as a whole. Therefore, we might expect similar species to be close together in space. The genome proteins were represented as vectors of frequencies of groups of amino acids. In this paper, a sliding window of size 3 was used to measure the frequency. To represent the genome we used the vector sum of all its proteins. We are going to evaluate the appropriateness of this representation in the sequence. Therefore, we can obtain a database of genomes, S, as a rectangular matrix, X, where each line corresponds to one of the *n* genomes:

$$X = (X_1, X_2, \ldots X_n)^T = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{211} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

As can be seen, the representation cannot be visualized in this high-dimensional space. With 20 amino acids, and considering that unknown amino acids are represented as a separated letter of the alphabet, each genome vector has m = 2^{13} = 9, 261 dimensions. Therefore, to generate a suitable visualization, it is necessary to reduce the dimensionality of the space, with the minimum loss of information. When a representation in reduced space, Y, is generated for the database matrix X , we can calculate an error function E as following:

$$E = \sum_i \sum_j \left(\delta_{ij} - \gamma_{ij}\right)^2$$

where δ_{ij} is the *euclidean distance* between genome i and j in the original space, represented in the matrix X, and γ_{ij} is the *euclidean distance* between genome i and j in the reduced space, represented in the matrix Y . The best representation of S in the reduced space will be the Y with the minimal associated error function. Therefore, we must solve an unconstrained optimization problem. Many methods can be used to solve this problem. In Xie and Schlick (2000), the truncated-newton minimization method is used. In this paper, we used a technique based on the interior-reflective Newton method. Singular value decomposition (SVD) is a popular method to reduce the dimensionality of a space, keeping the fundamental semantic association among the vectors in that space. Therefore, a good initial solution for the optimization problem can be obtained using the singular value decomposition (SVD) of X . The matrix is represented as $X = U\Sigma V^T$, where $U = [u_1\ u_2\ \ldots\ u_p]$, $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_p)$, $V = [v_1\ v_2 \ldots v_p]$. An approximation of X in reduced space (X_k) is given by:

$$X_k = \sum_{i=1}^{p} u_i \sigma_i v_i^T ; k \le p.$$

In this paper, we generated both two and three dimensional representations. We used a rank 2 approximation of X as the initial solution for the former, and a rank 3 approximation as the initial solution for the latter. After the optimization procedure, we have the best representation of the genomes to be visualized in a reduced space.

3. Results and discussion

We used the proposed approach to generate two and three dimensional visualizations of 64 whole mitochondrial genomes with 832 proteins. First, we are going to evaluate if the *euclidean distance* of genomes using the chosen representation is suitable to evaluate the similarities between them. Couto *et al*. (2007) showed that the similarity of genome sequences can be measured by the *euclidean distance* in a reduced dimensional space of tripeptides descriptors. They found a correlation between the euclidean distance and global distance sequence alignment of +0.70. To perform a similar analysis we created 64 supersequences by concatenating the 13 genes from each organism. These supersequences were compared by using global edit distance between each pair of sequences and euclidean distance in the high-dimensional space. As in Couto *et al*. (2007), the correlation between the edit distance and *euclidean distance* was +0.70, but this time in a cubic model ($P < 0.01$; Figure 1). We can see, therefore, that the *euclidean distance* of genome sequences using the chosen representation can be used as a measure of similarity.

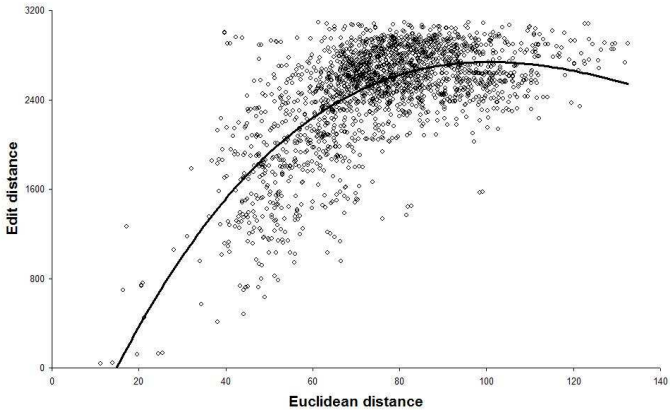

Fig. 1. Scatter plot of euclidean distance and global edit distance.

We classified the species according to the class. Therefore, the following groups were used: *Aves*, *Mammalia*, *Reptilia*, *Actinopterygii*, *Sarcopterygii*, *Chondrichtyes*. In Figure 2 we can see the 2D and 3D results. As can be observed, the different class had a tendency to form groups in space. In the 2D graph we can see that mammals (*mammalia*) are in the bottom, birds (*aves*) are in the upper left, reptiles (*reptilia*) are generally in the middle left, and fishes (*actinopterygii*,

sarcopterygii, *chondirchthyes*) are in the upper right. It is notorious how the birds are close together in a single cluster. In the results in 3D the classes are even better clustered. This time, reptiles, birds and fishes are in distinctly separated groups. Only the class of the fishes are somewhat mixed.

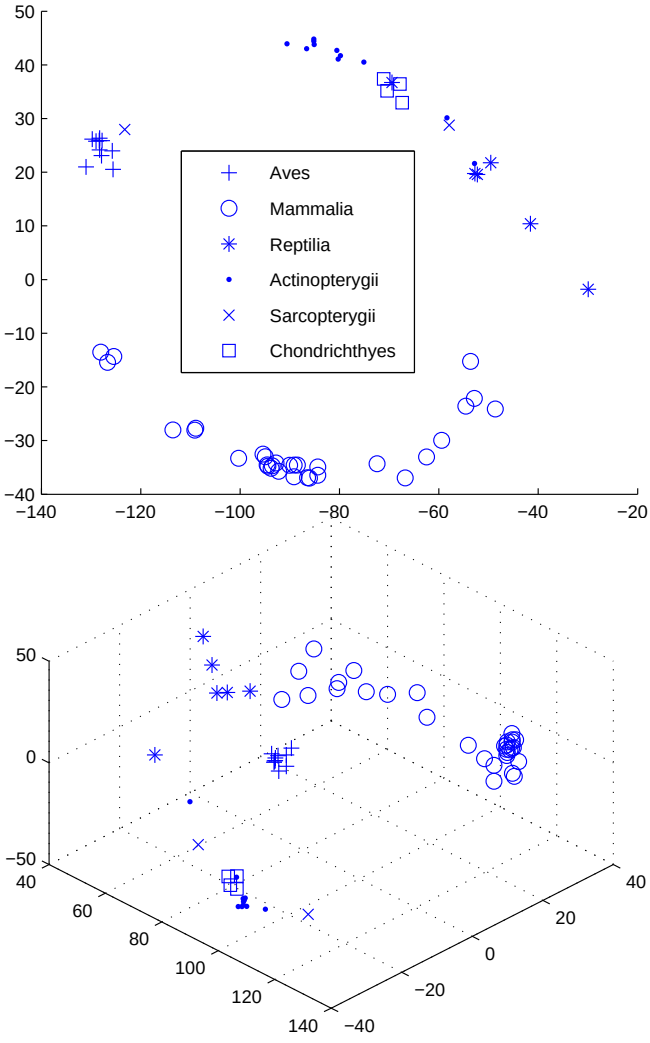

Fig. 2. Visualization of genomes in 2D and in 3D.

It is interesting to observe the relationships between the classes, as similar groups tend to be near in space. The position of the class in the graphs seems to be related to the evolutionary scale. Considering the 2D graph as an ellipse, we can see that the reptiles are between the mammals and the fishes. In 3D this can be ob-

served a second time. However, the evolutionary relationship between reptiles and birds is more clear in 3D, as there is no group between them.

Both in 2D and in 3D, mammals form a clearly distinct group from all other classes. They occupy a vast area, which might indicate more extensive diversity. We can also note that some mammals form clusters, what might be interesting to analyze. In order to better explore how the mammals are organized we separated this class in nine different groups: (i) *Prototheria*, corresponding to species in this subclass; (ii) *Marsupialia*, corresponding to species in this infraclass; (iii) *Chiroptera*, corresponding to species in this *ordo*; (iv) *Cetartiodactyla*, corresponding to species in this *superordo*; (v) *Carnivora*, corresponding to species in this *ordo*; (vi) *Perissodactyla*, corresponding to species in this *ordo*; (vii) *Primates*, corresponding to individuals in this *ordo*; (viii) *Rodentia*, corresponding to individuals in this *ordo*; (ix) *Placentalia*, corresponding to all other individuals that are in this infraclass, but were not classified in any other group. In Figure 3 we can see an approximation of the region of the mammals with this new classification. Similar species appeared close together, as was expected. This shows another advantage of the proposed method: as each genome is represented as points in space, we can easily select a region to better explore, zooming in and out in the graph as appropriate for the analysis.

The proposed method, however, allows another way to visualize a selected group of genomes. We can reduce the original set and run the algorithm a second time. Therefore, in order to better visualize the mammals, we executed the algorithm with only this class in the database. The result can be seen in Figure 4. It is interesting to note that the 2D graph has a similar elliptic format as in Figure 2. Clusters that were difficult to observe in Figure 3 are very clear in this graph. Similar species are again near to each other, showing visually the proximities of the genomes. In 3D the only group that mixed with the others is the Placentalia, but this was expected, as this group is very general, holding greatly different individuals. All other groups occupy distinct positions in space. We can see, therefore, that the proposed method allows many interesting observations and analysis of a group of genomes. Prespecified groups could be seen as clusters in the resulting graphs and the positions of the species seem to be related to their evolutionary stage. We also showed how approximating a region of the graph or running the algorithm a second time with a reduced data set allows a better insight of the relationships among selected groups of genomes. The resulting graphs can be generated both in two and in three dimensions for visualization.

4. Conclusion

In this paper, we used a linear algebra method, followed by optimization, to visualize genomes in two and in three dimensional spaces. A set of complete mitochondrial genomes were used to test the algorithm. Graphs were generated to visualize the complete set and a reduced set of similar species. We noted that the

method was able to automatically cluster some of the predefined groups and biologically similar species were represented as near points in space. We also noted that the position of the genomes in space seems to be related to the evolutionary stage of the species. Our future work is directed towards using this mechanism to visualize a large set of proteins. In this way, relationships between them can be easily observed and quickly explored, facilitating new discoveries. It would also be interesting to use this technique to explore a vast number of genomes, and further explore how it can be used to gain insight in evolution and in the phylogenetic relationships between the species.

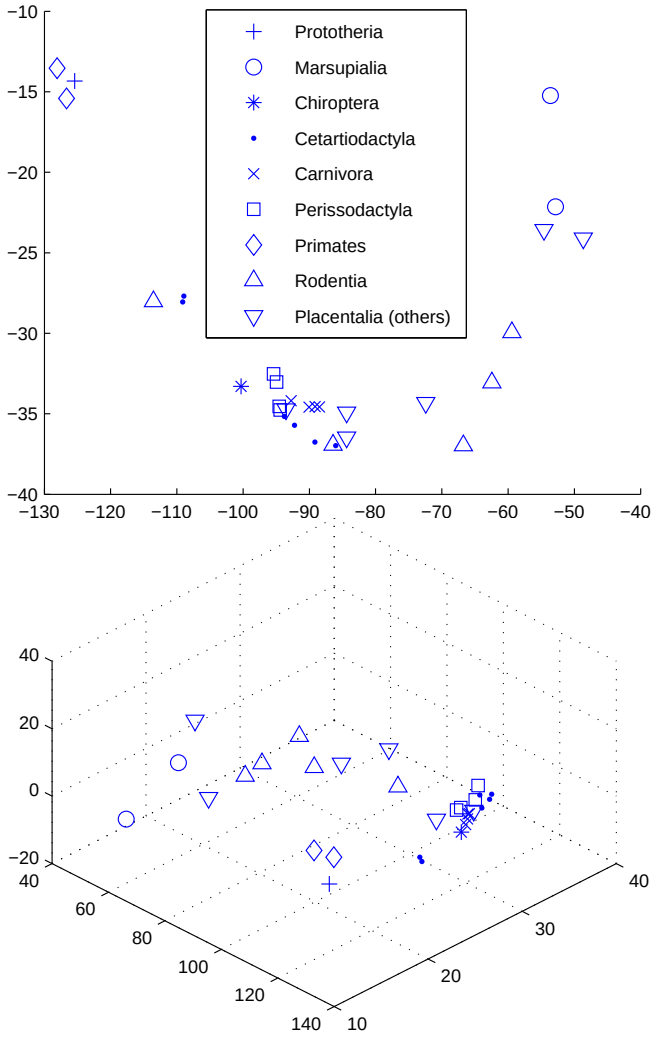

Fig. 3. Approximation of the region of the mammals in 2D and in 3D.

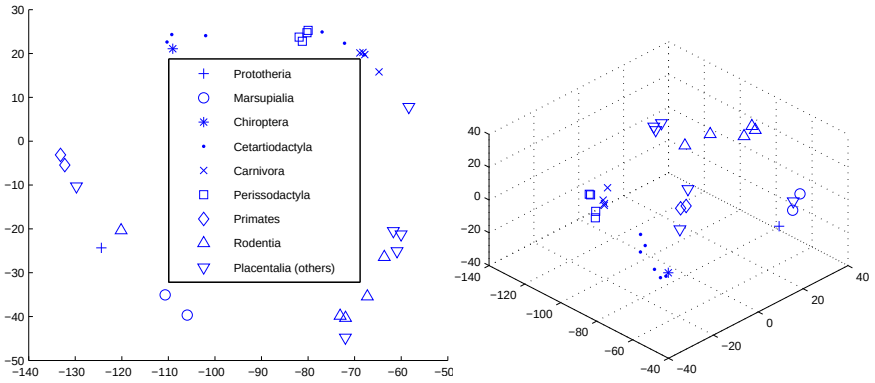

Fig. 4. Visualization of a reduced set in 2D and 3D.

5. References

Berry MW, Dumais ST, O'Brien GW (1994) Using linear algebra for intelligent information retrieval. Technical Report UT-CS-94-270, University of Tennessee, Knoxville, TN, USA

Couto BRGM, Ladeira AP, Dos Santos MA (2007) Application of latent semantic indexing (LSI) to evaluate the similarity of sets of sequences without multiples alignments character-by-character. Genetics and Molecular Research 6:983–999

Eldén L (2006) Numerical linear algebra in data mining. Acta Numerica 15:327–384

Engels R, Yu T, Burge C *et al* (2006) Combo: a whole genome comparative browser. Bioinformatics 22(14):1782–1783.

Ghai R, Hain T, Chakraborty T (2004) Genomeviz: visualizing microbial genomes. BMC Bioinformatics 5:198

Gibson R, Smith DR (2003) Genome visualization made fast and simple. Bioinformatics, 19(11):1449–1450

Huggins P, Pachter L, Sturmfels B (2007) Toward the human genotope. Bulletin of mathematical biology 69(8):2723–2735

Lewis S, Searle S, Harris N *et al* (2002) Apollo: a sequence annotation editor. Genome Biology. doi:10.1186/gb-2002-3-12-research0082

Rutherford K, Parkhill J, Crook J *et al* (2000) Artemis: sequence visualization and annotation. Bioinformatics 16(10):944–945

Stothard P, Wishart DS (2005) Circular genome visualization and exploration using cgview. Bioinformatics 21(4):537–539

Stuart GW, Berry MW (2004) An svd-based comparison of nine whole eukaryotic genomes supports a coelomate rather than ecdysozoan lineage. BMC Bioinformatics 5: 204

Stuart GW, Moffett K, Leader JJ (2002) A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. Molecular Biology and Evolution 19(4): 554–562

Xie D, Schlick T (2000) Visualization of chemical databases using the singular value decomposition and truncated-Newton minimization. In: Floudas CA, Pardalos PM (eds) Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches, vol. 40, Kluwer Academic Publishers, Dororecht/Boston/London