Evolutionary Bioinformatics

ORIGINAL RESEARCH

ColorPhylo: A Color Code to Accurately Display Taxonomic Classifications

Sylvain Lespinats[1] and Bernard Fertil[2]

[1]UMR INSERM unité U722 and Université Denis Diderot, Paris 7, Faculté de médecine, site Xavier Bichat, 16 rue Henri Huchard, 75870 Paris cedex 18, France. [2]Laboratoire LSIS (UMR CNRS 6168), Equipe I&M (ESIL) case 925, 163, avenue de Luminy, 13288 Marseille cedex 9, France. Corresponding author email: sylvain.lespinats@gmail.com

Abstract: Color may be very useful to visualise complex data. As far as taxonomy is concerned, color may help observing various species' characteristics in correlation with classification. However, choosing the number of subclasses to display is often a complex task: on the one hand, assigning a limited number of colors to taxa of interest hides the structure imbedded in the subtrees of the taxonomy; on the other hand, differentiating a high number of taxa by giving them specific colors, without considering the underlying taxonomy, may lead to unreadable results since relationships between displayed taxa would not be supported by the color code. In the present paper, an automatic color coding scheme is proposed to visualise the levels of taxonomic relationships displayed as overlay on any kind of data plot. To achieve this goal, a dimensionality reduction method allows displaying taxonomic "distances" onto a Euclidean two-dimensional space. The resulting map is projected onto a 2D color space (the Hue, Saturation, Brightness colorimetric space with brightness set to 1). Proximity in the taxonomic classification corresponds to proximity on the map and is therefore materialised by color proximity. As a result, each species is related to a color code showing its position in the taxonomic tree. The so called ColorPhylo displays taxonomic relationships intuitively and can be combined with any biological result. A Matlab version of ColorPhylo is available at http://sy.lespi.free.fr/ColorPhylo-homepage.html.
Meanwhile, an ad-hoc distance in case of taxonomy with unknown edge lengths is proposed.

Keywords: color code, dimensionality reduction, hierarchical classification, taxonomy

Evolutionary Bioinformatics 2011:7 257–270

doi: 10.4137/EBO.S7565

This article is available from http://www.la-press.com.

Introduction

Topic

Many datasets are "naturally" structured as hierarchical classifications. In particular, the Darwin's evolution theory ensures that the relationships between species can be expressed within a tree (named "phylogenic tree"/"taxonomic tree"). However, the exploration of large trees (and graphs) is not easy.[1]

Visualising taxonomy together with other pieces of information (obtained from various biological analyses for example) is often extremely instructive; some examples can be found in.[2–4] In those cases, a level of granularity of the tree is usually chosen and a color is assigned to each subclass thus defined.[4] Several drawbacks can be easily identified. The granularity level is obviously subject to arbitrariness. Moreover, proximity relationships between subclasses are ignored as well as their subdivisions.

Objective

In the following, we set up an automatic coloring method in order to address these problems. Indeed, we describe a simple method that automatically generates an intuitive color code showing proximity relationships between data in any hierarchical classification. The presented algorithm, named ColorPhylo, associates a specific color to each item so that the taxonomic relationships are shown by color proximity (the closer two items in the tree, the more similar their colors). Colors can thereafter be used in any user's analyses and figures so as to display taxonomy.

Background

The above research field is yet relatively unexplored. An automatic coloring tool for tree exploration has been proposed by Fua and co-authors:[5,6] in order to explore a given taxonomy, leaves are considered as an ordered list of items (the order can result for example from the reading direction when the tree is displayed on a phylogram). Each node can then be related to a color according to a chosen LUT (Look-Up Table). The LUT may be locally stretched or compressed in order to focus on a given area. Such an interactive control of the LUT allows exploring the taxonomy while considering various depths. This approach suffers however from a major shortcoming: the leaf ordering is not unique (trees are invariant with respect to permutation of linked branches). Very different colorings can then result from various choices. Moreover, close colors can correspond to items that belong to different branches and somewhat different colors can rely to close species. These drawbacks can be observed in Figures 6 and 8.

In the biological field, several tools may be used to color species in taxonomic tree.[4,7,8] In particular, PhyloView[4] has been specially designed to display taxonomy and other analyses (phylogenetic trees) together. Strictly speaking, however, these tools are not automatic coloring tools: the user has to define the colors for taxonomic groups that are relevant from his point of view.

Node coloring is of main concern in Self-Organising Map (SOM) framework.[9,10] SOM is a non-linear dimensionality reduction method. It is designed to map items (generally from a high dimensional Euclidean space) onto a discrete grid: data are aggregated on each neuron (vector quantisation). Because the grid is discrete, the visualisation of gaps between clusters—if they exist—is uneasy. Various coloring methods have been proposed to account for this drawback, including.[11–17] In particular, Kaski et al[14,15] display neurons in a CIELab colorimetric space;[16] Johan Himberg[17] uses a hierarchical classification in order to find clusters and use a cut-off function in the resulting tree so as to linking neurons to a LUT (thanks to a method close to the Fuas' one).[5,6]

However, the SOM context differs from the one of our present objective, despite approaches that are connected. Indeed, we focus here on taxonomies and SOM deals with multidimensional data. Moreover, if hierarchical classifications are often involved in automatic coloration of SOM, the tree is not the input of the algorithm, but a step in the method (especially in the context of cluster discovering). Such methods can be considered, however, as sources of inspiration for the present paper.

Notes concerning taxonomy

One of the examples used in the present paper (see section 3.2) is based on the Genbank taxonomy (http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy). Please note that the Genbank taxonomy can not be considered as a conclusive phylogeny, as clearly stated by Genbank owners (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=howcite). We are only looking here for a way to display any given hierarchical classification, whether it is considered as a phylogeny, a taxonomy or so.

Methods

Whole procedure overview

ColorPhylo aims at assigning a unique color to each species of a given set so that the color differences reflect the taxonomic "distances" between species (see Fig. 1):

Step 1: If not available, taxonomic distances are calculated from the taxonomic tree. Two procedures are considered depending on whether the edge lengths are known or not (see section 2.2). Thus, a distance matrix between all species is obtained.

Step 2: Species are mapped onto a 2D space while preserving the distance matrix as much as possible (section 2.3).

Step 3: The map is rescaled (and possibly rotated) (section 2.4) in order to fit a 2D colorimetric subspace (section 2.5).

Step 4: Each species located in the colorimetric subspace is consequently given a unique color. The

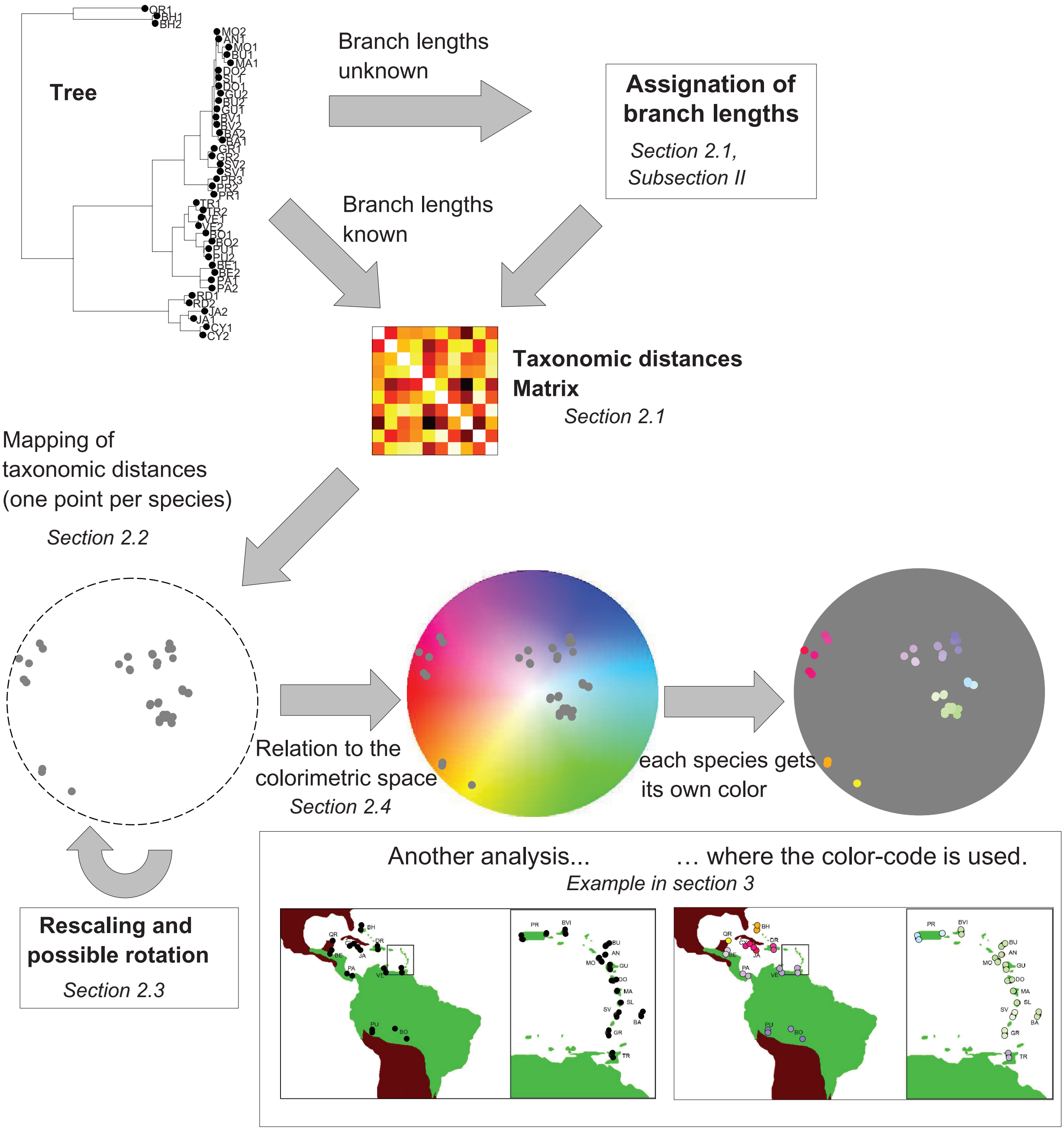


Figure 1. A typical analysis.

species coloring can therefore be used in subsequent analyses to visualise the taxonomic relationships (as done in section 3).

The generation of a color-code for hierarchically-organised data.

Taxonomic distance

The presented methodology relies on the availability of a distance based on a taxonomic tree.

Two cases are possible:

I. Edge lengths are known. Distances between species can be immediately deduced: the distance between two species corresponds to the sum of the length of edges connecting them. The "bananaquit bird" dataset offers a good illustration of that case (section 3.1).

II. Edge lengths are unknown (the GenBank taxonomy dataset: section 3.2). We then have to (arbitrary) choose the length of edges. Many solutions can be considered. One may assign the same length for all edges. However, such a choice is inappropriate when the tree depth depends on the level of refinement along branches. Such a situation is rather common in taxonomy: the Eukaryote classification is actually more refined than that of Bacteria and Achea (Fig. 2). If the length of edges is set constant along the tree, humans appear closer to some bacteria (which are microscopic prokaryotes) than to octopus (which is also an animal). The toy example presented in Figure 3, left insert (a) provides a good illustration of this bias.

In fact, as far as the tree of life is concerned, the distance between two species belonging to the same subclass (branch) must always be smaller than the distance between one of these species and any species that does not belong to the subclass. In order to ensure that this property is always true, we propose to attribute a variable length for edges: an important length is given to edges close to the tree root, and the length is successively reduced when the edge gets away from the root. A geometric progression can be implemented here. The length of edges at the tree root is 1. The length of any other edge equals half the length of the parent edge (Fig. 3, right insert (**B**)).

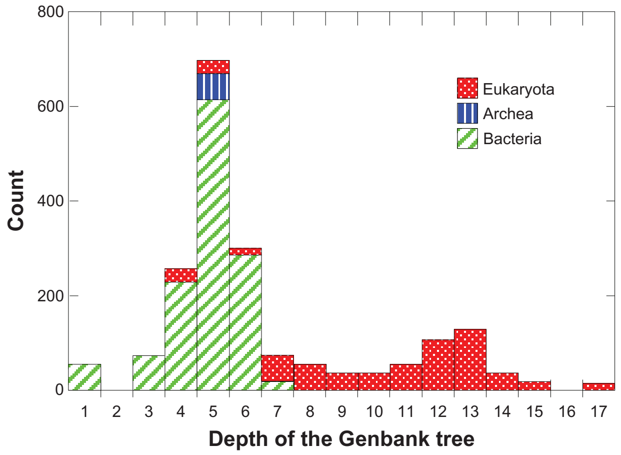


Figure 2. Depth of the Genbank taxonomic tree for 2046 species. Histograms for the three domains of life are displayed. The tree is often deeper for Eukaryotes (organisms with nuclear cells, ie, animals, plants and fungus among other species) than for Bacteria and Archea.

This arbitrary choice has a critical property:

Proposition: The distance between two species belonging to the same subclass is always smaller than the distance between one of these species and any species that does not belong to the subclass.

Proof: Let us consider three species *A*, *B* and *C* where *A* and *B* belong to a same subclass, contrary to *C*: species *A* and *B* are linked by node N_1, and species

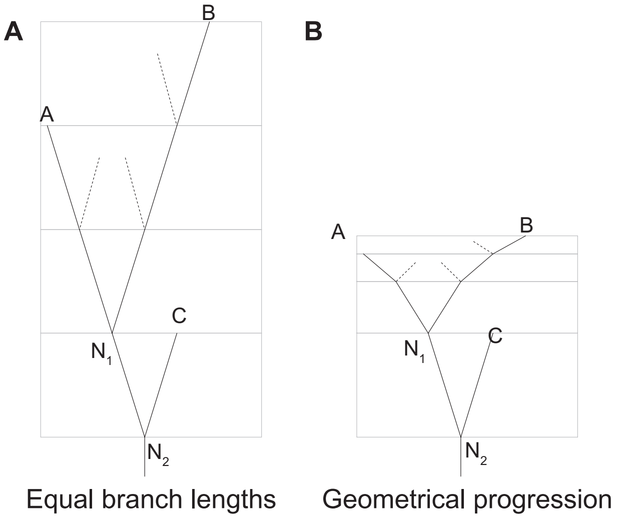


Figure 3. Representation of a hierarchical tree with various depths as a function of edge distance. *A* and *B* are deeper in the tree than *C*. When edge lengths are set to 1 -left insert (**A**)-, *d (A,B)* = 5 while *d(A,C)* = 4. *A* is then found closer to *C* than to *B* (*d(A,B)* > *d (A,C)*). However, *A* and *B* belong to the same high level branch while *C* does not. When edges lengths follow a geometrical progression -right insert (**B**)-, $d(A,B)/d(N_1,N_2) = 13/8 < d(A,C)/d(N_1,N_2) = 11/4 \Rightarrow d(A,B) < d(A,C)$.

C is linked to node N_1 by node N_2 (similarly to the example in Fig. 3). $d(x, y)$ corresponds to the distance between elements x and y in the tree. According to the properties of geometric progression with a common ratio equal to 1/2 (basically, $\Sigma_{i=2}^{\infty} 1/2^i = 1$):

$$d(B, N_1) < \sum_{i=1}^{\infty} \frac{d(N_1, N_2)}{2^i} = d(N_1, N_2)$$

As a consequence, $d(A,B) = d(A,N_1) + d(B,N_1) < d(A,N_1) + d(N_1,N_2)$ and $d(A,C) = d(A,N_1) + d(N_1,N_2) + d(C,N_2) > d(A,N_1) + d(N_1,N_2) \Rightarrow d(A,C) > \mathrm{d}(A,B)$.

Thus, the contribution of edges far from the tree root is reduced, which makes it possible to emphasize most important classes.

It must be pointed out that such a procedure should not be considered as a way to assess "true" edge lengths but rather as a heuristic method to account for classes and subclasses within the actual automatic coloring procedure.

Whatever the case, the next steps starts from the resulting matrix of taxonomic distances (noted d thereafter).

Mapping species

Non-Linear Multi-Dimensional Scaling (MDS) is a set of methods designed to show relationships between items. The explored dataset is displayed as points on a low-dimensional output space. Most of the time, the output space is a Euclidean 2D space, which is proposed as a "map" of data. These techniques are particularly used to explore high dimensional data. Indeed, MDS techniques are expected to retrieve the spatial organisation of high dimensional dataset. To achieve this goal, linear, as well as non-linear Multi-Dimensional Scaling methods preserve the distances between data "as much as possible", according to criteria depending on the method.[18] Usually, small distances are emphasized.

As far as our data are concerned, MDS methods are consequently expected to display taxonomic relationships (a distance matrix) onto small dimensional spaces. To achieve this goal, we have chosen the DD-HDS algorithm (Data-Driven High Dimensional Scaling),[19] for its efficiency whatever the distribution of input distances.

The DD-HDS projection in the two-dimensional output space is computed (code available at http://sy.lespi.free.fr/DD-HDS-homepage.html). A DD-HDS parameter (λ) can be adjusted in order to more or less emphasize smallest distances: 0 (1 respectively) for minimum (maximum respectively) consideration of longer distances. In the following examples, λ is set to 0.5. The output of DD-HDS is the set of coordinates in a 2D space (coordinates of species i are noted $X_{i,j}$ where $j \in \{1, 2\}$).

It is worth noting that additive distances (distances in a tree) cannot be perfectly preserved onto a 2D space. However, this fact is not a drawback in the present case because we are less concerned by distances preservation than subclasses segmentation. Indeed, even if the distance mapping is not perfect, it still emphasizes the various levels of the classification thanks to DD-HDS that avoids both "false neighbourhoods" (when a small distance in the output space becomes a large distance in the data space) and "tears" (when a large distance in the output space becomes a small distance in the data space).[19]

Rescaling and rotation

After being mapped, data are translated and rescaled so as to occupy a circle having a radius equal to 1. This step is required in order to ensure that the map can be related to the circle-shaped HSB colorimetric space (see section 2.5). Coordinates are subsequently centred $X'_{i,j} = X_{i,j} - mean(X_{i,j})$ (with $j \in \{1, 2\}$) and rescaled $X''_{i,j} = X'_{i,j} / \max_i \left(\sqrt{\Sigma_j X'^{2}_{i,j}}\right)$ the polar coordinates of $X''_{i,j}$ are finally calculated: the norm of item i is $\rho_i \sqrt{X''_{i,1} + X''_{i,2}}$ and its angle is $\varphi_i = \cos^{-1}(X''_{i,1} / \rho_i)$ if $X''_{i,2} \geq 0$ and $\varphi_i = -\cos^{-1}(X''_{i,1} / \rho_i)$ otherwise.

Rotation around the centre and symmetry can take place now. Rotation and symmetry does not impact MDS solutions but allow modifying the color map at will while preserving relative color differences between species (for example, it may be useful to make colors close to the ones the user is familiar with).

The map is related to the HSB colorimetric space

Each position in the mapping can now be related to a color according the HSB (Hue, Saturation, Brightness) colorimetric conic-shaped space (The HSB space is also called HSV space (V for Value) or HSL space

(L for Lightness)). The HSB space relates each position in a cone to a color:

- Hue corresponds to the angle on a circle (the base of the cone) where 0° is pure red.
- Saturation corresponds to the distance to the centre of the circle: 0 for 0% of saturation (pure white), 1 for 100% (pure color).
- Brightness corresponds to position on the axe of the cone. No brightness (black) at the cone tip to full brightness at the base.

Brightness is not used in the present framework (Brightness is set to 1).

A position in a circle is characterised by hue and saturation values (Fig. 4). Species can subsequently be given a color, norm (ρ_i) being related to saturation and angle (φ_i) being related to hue. The continuity of color in the HSB space ensures that color proximity is related to taxonomic proximity Figure 7 (section 3.2).

Angle corresponds to hue while distance to centre corresponds to saturation.

In the automatic coloring of SOM framework, Kaski et al[14,15] propose to embed data in a CIELab colorimetric space.[16] Properties of this colorimetric space also fit very well with our context: indeed distance between positions in the CIELab colorimetric space is supposed to be linked with the human perception of color difference. However, significant constraints are placed on the mapping due to the particular shape of the CIELab space. In contrast, the mapping on circle-shaped HSB sub-space is easier. Most of the times, data can be expected to fill a large part of the space after a simple rescaling procedure. Moreover, the mapping can be easily rotated and returned in the space in order to find the most convenient color code.

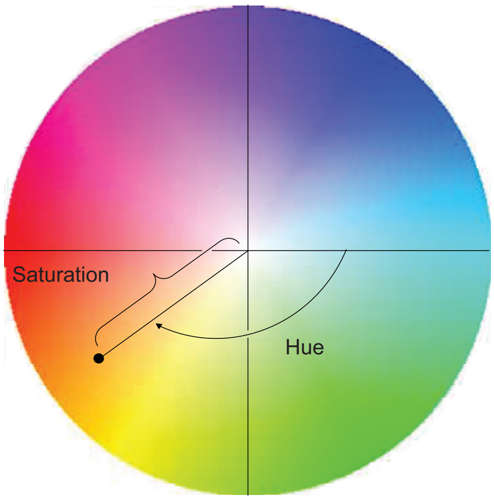


Figure 4. HSB Colorimetric 2D sub-space (Brightness is set to 1).

It is worth noting that a three-dimensional mapping (such as RGB space) may offer in some instances a better fit with the matrix of distance. However, every color would be eligible, and a species color could be close (or similar) to the background color. With the two-dimensional mapping related to HSB colorimetric space, no species can correspond to grey, which can be chosen as background color.

Results

ColorPhylo is tested on two biological applications: the first one refers to a published ornithological study,[20] the second one is related to the analysis of genomic signatures (DNA sequences characterised by oligonucleotide frequencies).

It has to be remembered that, on the one hand we study specified knowledge (or analysis results) related to a given dataset (here, birds' geographical position—first example—and oligonucleotide counts—second example—) and on the other hand, a taxonomy (the tree) of these data is available. In the present paper, we aim to display the taxonomy and the knowledge together. The procedure described in section 2 provides a taxonomic color-code used here to color geographic position of birds or genomic signatures.

Application to the work of E. Bellemain and co-authors: linking geographical distribution and phylogenetic data for bananaquit birds

Bellemain and co-authors performed a phylogenetic analysis of bananaquit birds (*Coereba flaveola*) caught in various places in Latin America and Caribbean islands.[20] Geographical positions of catching sites are displayed (in their Fig. 1) and the genetic analyses led to two phylogenetic trees (their Figs. 2 and 3). Relationships between geographical origin of birds and taxonomy are discussed, and conclusions about the evolutionary history are made. Despite that this work has been successfully achieved without the help of any visualisation tool, a coloring based on the taxonomic tree would have considerably helped these authors.

In the following, ColorPhylo is used to assign a color to each bird according to its position in the taxonomic

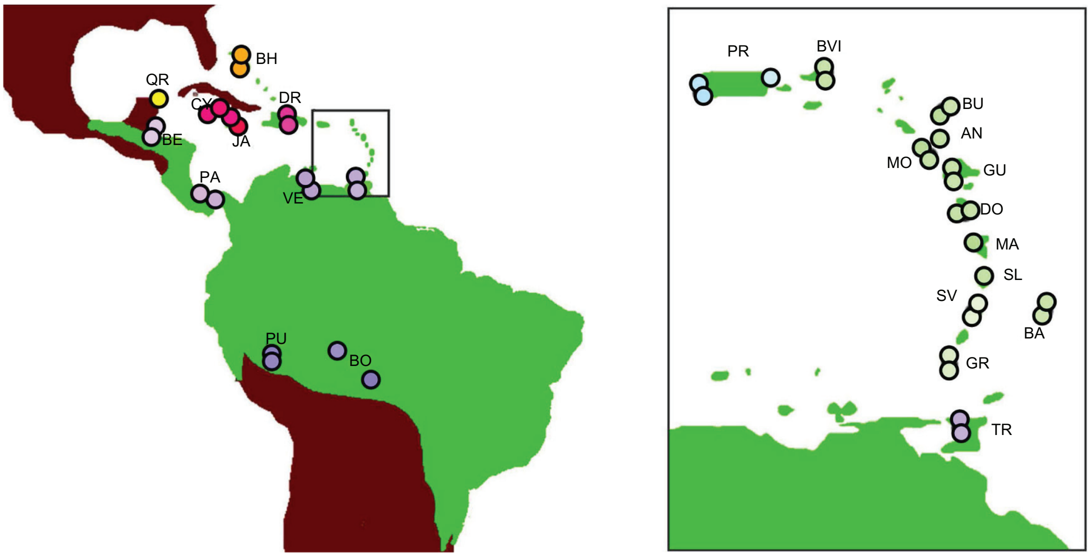


Figure 5. The evolutionary history of bananaquit birds as highlighted by ColorPhylo. The data come from.[20] A biological analysis can be found in that paper. This figure is similar to Figure 1 in Bellemain's and co-authors[20] (original publisher: Biomed central) with superimposed colored dots instead of original black dots.

tree (based on combined mitochondrial data, Figure 2 in).[20] Here, edge lengths (provided by the phylogenetic analysis) are known and used (see section 2.2, subsection I). Birds are displayed as colored dots, positioned on the map on their catching site (Fig. 5).

Dot coloring allows easy observation of the genetic groups: Lesser Antilles and Puerto Rico (green dots), Bahamas and Quintana Roo (orange dots), continent (purple dots), and Greater Antilles except Puerto Rico (red dots) as well as smaller relationships within groups.

We performed the analysis on the same dataset, according to the Fua et al[5,6] coloring method (Fig. 6). It must be pointed out that the interactive local stretching and compression of the LUT—an essential feature of the Fua et al method—cannot be implemented here (obviously impracticable on a printed document!). For that reason, a comparison between ColorPhylo and the Fua et al algorithm from the present figures exclusively would be highly unfair. However, there are noticeable differences between Figures 5 and 6. A high diversity in the population of birds from Lesser Antilles and Puerto Rico could be inferred from figure 6 right insert, whereas it is not supported in fact by the evolutionary tree (Fig. 6, left insert). Moreover, the high difference between birds

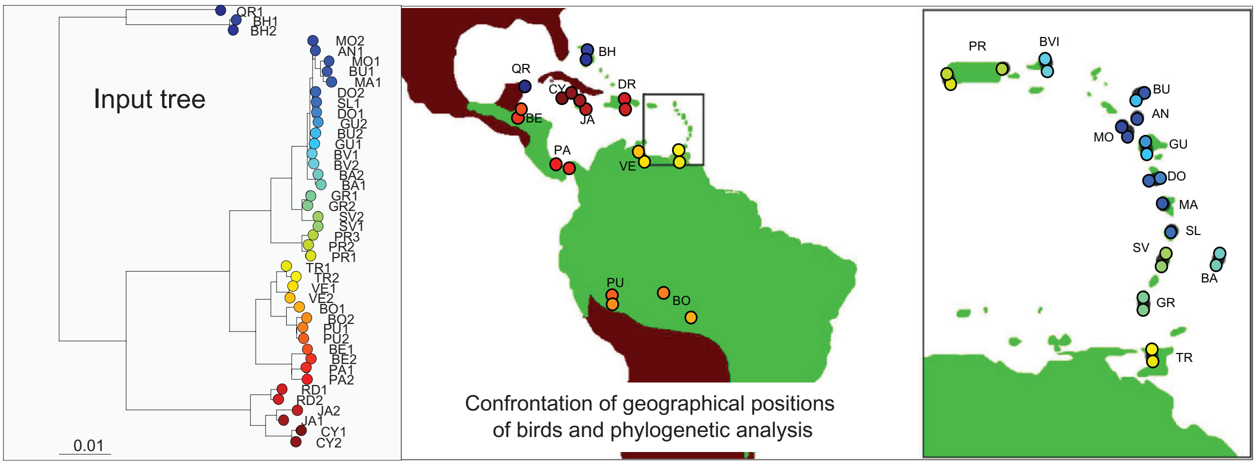


Figure 6. Analysis of bananaquit birds data, similar as the one presented in Figure 1 and Figure 5, except that the Fua et al[5,6] coloring method is used rather than ColorPhylo.

from Lesser Antilles and birds from Bahamas and Quintana Roo is clearly underestimated.

Dot coloring allows easy observation of the genetic groups: Lesser Antilles and Puerto Rico (green dots), Bahamas and Quintana Roo (orange dots), continent (purple dots), and Greater Antilles except Puerto Rico (red dots) as well as smaller relationships within groups.

Genomic signatures analysis

Coloring according to the Genbank tree: Species paths are from Genbank so that "root, cellular organism, Eukaryota, Fungi/Metazoa group, Metazoa; Eumetazoa, Bilateria, Coelomata, Deuterostomia, Chordata, Craniata, Vertebrata, Gnathostomata, Teleostomi, Euteleostomi, Sarcopterygii, Tetrapoda, Amniota, Mammalia, Theria, Eutheria, Euarchontoglires, Primates, Simiiformes, Catarrhini, Hominoidea, Hominidae, Homo/Pan/Gorilla group, and Homo Sapiens" qualifies human for example. Paths are derived from a rooted tree where edge lengths are unknown. Taxonomic distances are calculated according to the procedure described in section 2.2, subsection II. Colors are then selected according to section 2.3, 2.4 and 2.5. Data embedded in the colorimetric space are displayed in Figure 7.

Species are embedded in a two-dimensional space generated from the taxonomic distance matrix. Colors are assigned from the position in the mapping (by construction, colors are smoothly

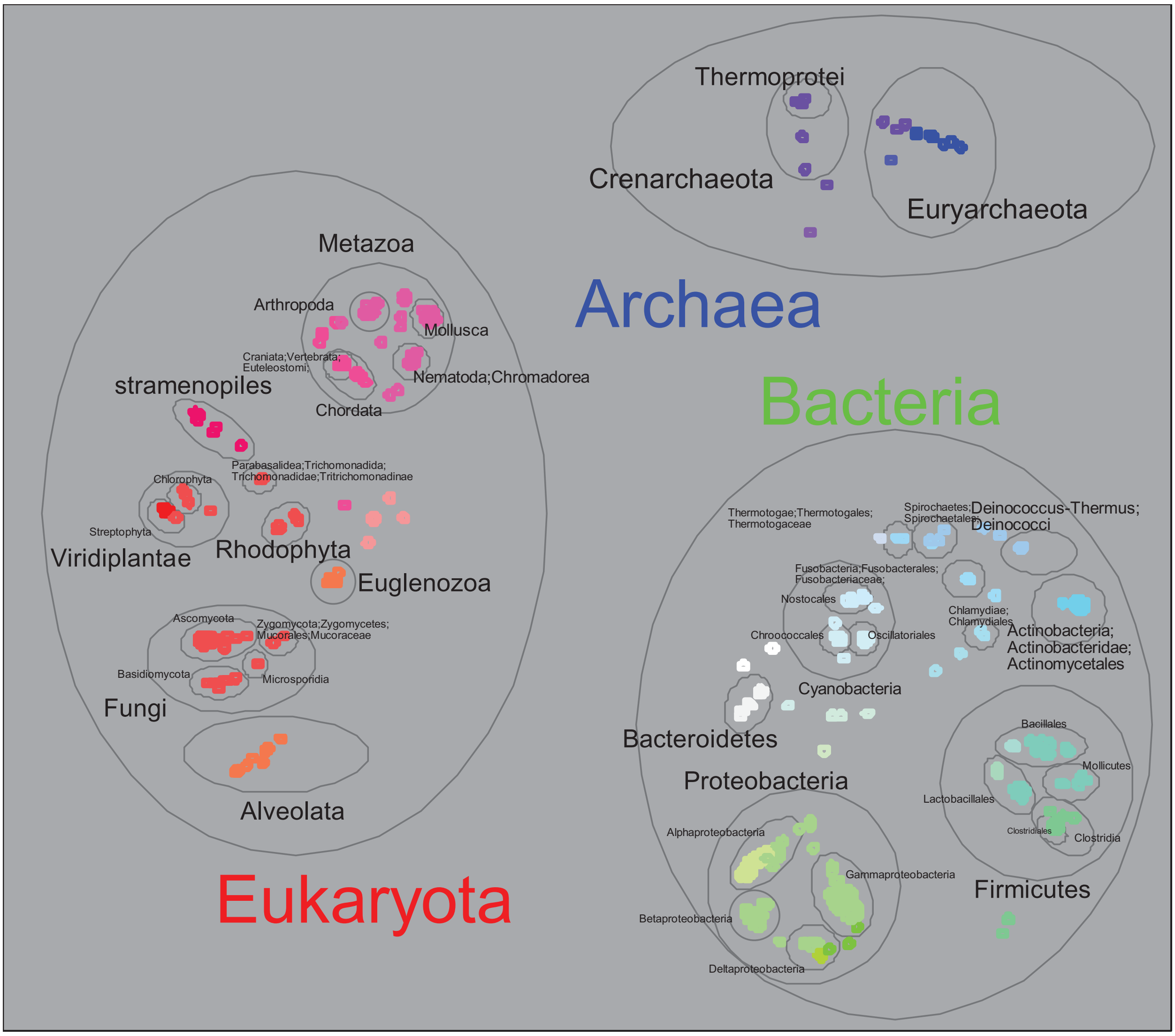


Figure 7. Embedding of 2046 species in the HSB colorimetric space, according to the Genbank taxonomy.

distributed). Gray ellipses show main classes and subclasses.

Contrasting with figure 7, the three domains of life are not clearly segmented by color in Figure 8.

An example of color code use: Study of the link between taxonomic proximity and genomic signature proximity: The whole set of short oligonucleotide frequencies observed in a DNA (DeoxyriboNucleic Acid) sequence is species-specific and is thus considered as a "genomic signature".[2,21] Moreover, a DNA segment as short as 1 Kb (kilobase) is sufficient to characterise the genomic signature of the species. As a consequence, genomic signature appears qualifying the "writing style" of the species. The genomic signature is species specific: it allows finding the species of origin of a DNA fragment with a fairly good efficiency.[2,22] Note that because the genomic signature is stable along the genome, non-homolog fragments of species can be compared (homology is required for most other methods devoted to comparative genomics). Lastly, proximity between species in terms of genomic signature is known to be linked to evolutionary proximity. Indeed many phylogenies based on hierarchical classifications of signatures have been proposed.[23–28] Another powerful approach to describe the taxonomic organisation of genomic signatures uses dimensionality reductions: data are embedded on a two-dimensional (or sometimes three-dimensional) space. These representations

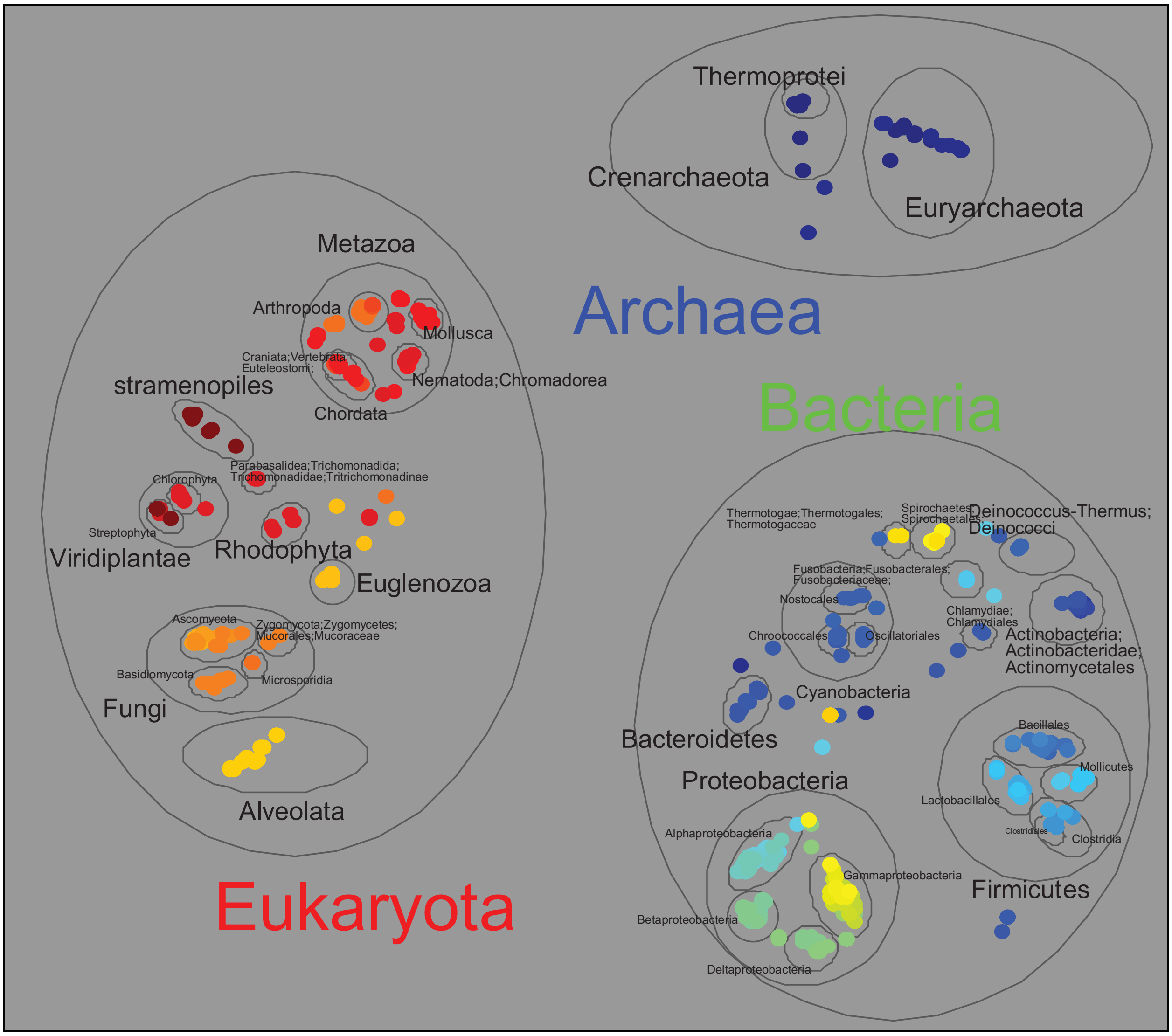


Figure 8. Same mapping than Figure 7, except that color is provided by the Fua et al method.

have been achieved by Principal Components Analysis (PCA),[2,3] by Self Organising Map (SOM)[29] or by Data-Driven High Dimensional Scaling (DD-HDS).[19] However, details of the spatial organisation of genomic signatures cannot be easily observed on a large dataset because of the limitation in term of discrete colors. In the following example we propose to visualise the taxonomic tree of species on the DD-HDS mapping of genomic signatures by means of the color code provided by ColorPhylo and Figure 7.

Prior to mapping, signatures are corrected using 1-order Markov model[30] as recommended in.[25,27] A distance matrix between signatures is then obtained, based on the Pearson's correlation coefficient as recommended by.[26] The mapping is subsequently generated: spatial proximity between items expresses proximity between species from the genomic signature point of view. Colors are provided by the ColorPhylo procedure (the map of species in the related taxonomic color-space is shown in Fig. 7): color proximity between items relies on taxonomic proximity between species. As a result, the various levels of taxonomy are simultaneously observable on a single figure (Fig. 9). The taxonomic organisation of signatures is clearly demonstrated. In particular, the patches of homogeneous colors support concluding that the similarity between genomic signatures accurately matches the taxonomic tree of the species, the signatures come from.

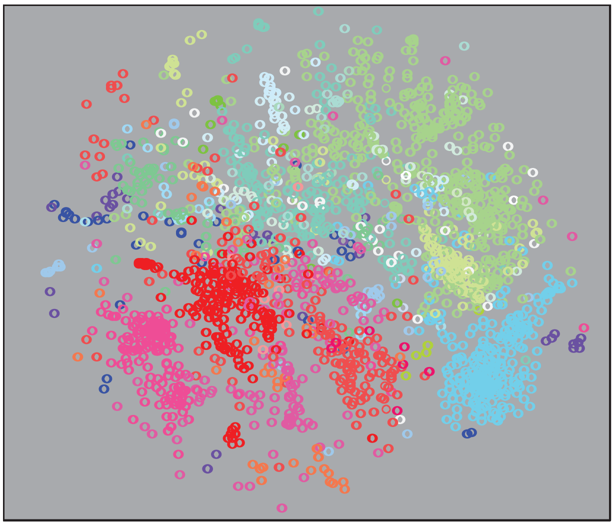


Figure 9. Genomic signature mapping (2046 items) colored according to Genbank taxonomy.

Discussion and Conclusions

Using ColorPhylo is straightforward. Hierarchical classifications can be easily displayed together with their relationships with any other organisation of the data. We have observed on real life examples that the interpretation of the resulting color code is fully intuitive. In the first example, geographical origins of bananaquit birds are related to phylogenetic data in order to analyse the evolutionary history. However, because the number of items and the complexity of the dataset was somewhat low, Bellemain and co-authors succeeded in describe the relationships between taxonomy and geographical distribution in their publication (but at the price of more irksome work). When the size of the dataset and/or the complexity of their relationships increase, ColorPhylo can provide a critical benefit, as it can be observed for the second application. In that example, (genomic signature) colors express the membership of items to one of the three domains of life, with subtle shades showing subclasses. In both cases, we have instantaneously access to the structure of classification through an attractive visualisation plot.

Although variations of color are theoretically unlimited, we rely on the perceptual discriminative power of the human eye. Surprisingly, the method gives access to a remarkable degree of detail (well above what is expected with a "manually" defined color code). In addition, a focus on a small region of the data is always possible by an ad hoc local color reallocation. Similarly, Colorphylo may be adapted to fit color-impaired users' requirements. The 2-dimensional color-space can be modified at will according to any desired effect, including of course satisfaction of the user's color perception.

In this paper, we have proposed a method to study the relationship between a given knowledge on a set of data (here, the birds' geographical position and the oligonucleotide frequencies in DNA sequences of species) and a specific organisation of these data, expressed by a taxonomic tree. Our approach can easily be adapted to other contexts. For example, if the organisation of the data results from an analysis based on a distance matrix (such as the ones performed by Neighbor joining,[31] Fitch-Margoliash,[32] …), the original distance matrix may be preferred to the taxonomic distance (such an approach may have been implemented for the bananaquit birds dataset). In fact, the procedure may be extended to the analysis of any kind of organisation of the data, given it is expressible as a distance matrix for which a 2D mapping makes sense.

A matlab version of ColorPhylo is available at http://sy.lespi.free.fr/ColorPhylo-homepage.html.

Authors' Contributions

SL and BF conceived the method together and wrote the manuscript. Both authors read and approved the final manuscript.

Acknowledgement

We thank Eva Bellemain and collaborators, as well as Biomed Central, for the permission to reproducing their figure. We also thank Mikael Cugnet for its useful comments.

Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

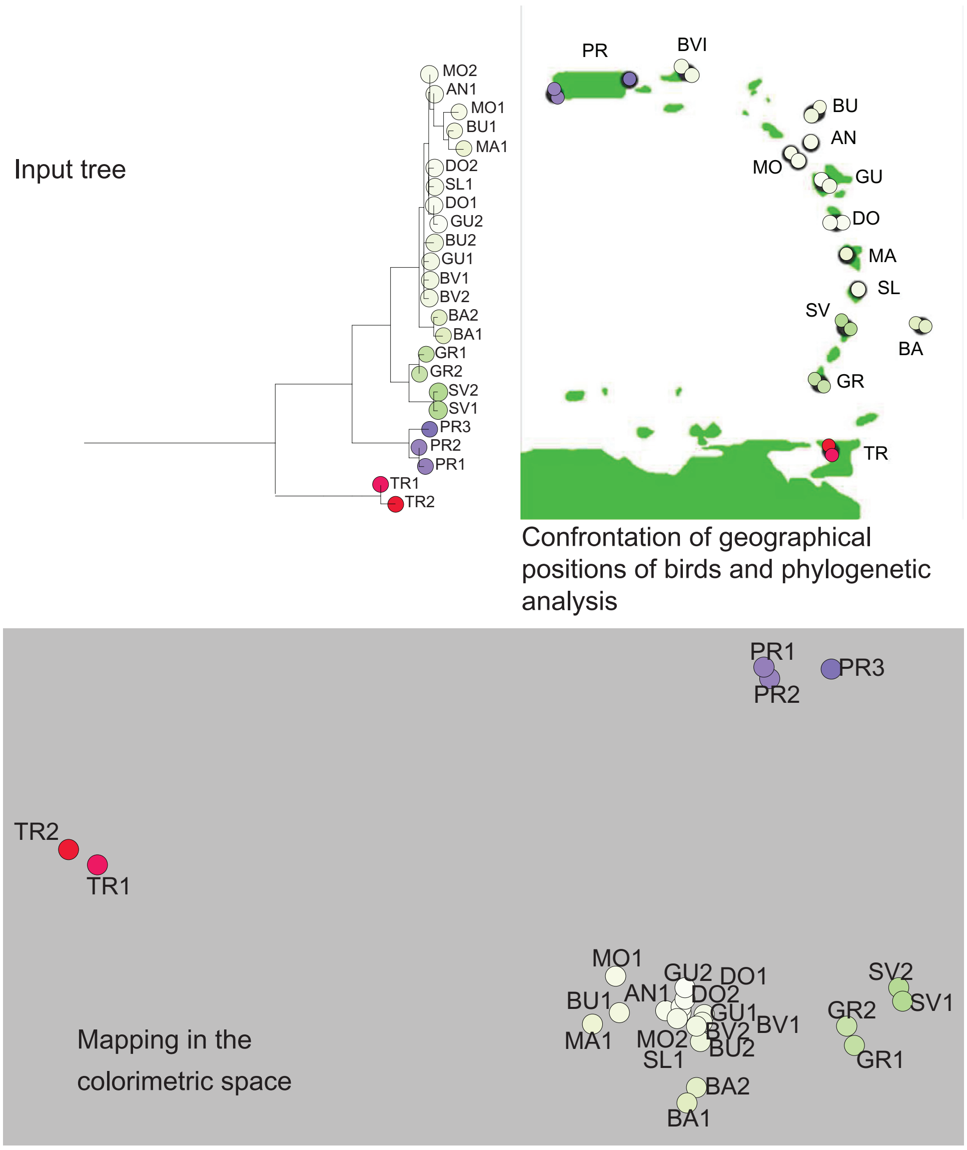
References

1. Heer J, Card SK. DOITrees revisited: scalable, space-constrained visualization of hierarchical data. Proceedings of the working conference on Advances Visual Interfaces. Gallipoli, Italy, May 2004. ACM Press, 421–4.
2. Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B. Genomic signature: characterization and classification of species assessed by Chaos Game Representation of sequences. *Mol Biol Evol*. 1999;16:1391–9.
3. Sandberg R, Winberg G, Branden CI, Kaske A, Ernberg I, Coster J. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res*. 2001;11(8):1404–9.
4. Palidwor G, Reynaud EG, Andrade-Navarro MA. Taxonomic coloring of phylogenetic trees of protein sequences. *BMC Bioinformatics*. 2006;7:79.
5. Fua Y, Ward MO, Rundensteiner EA. Navigating Hierarchies with Structure-Based Brushes. *In Proceeding of IEEE Symposium on Information Visualization, October 1999. San Francisco*. 1999:58–64.
6. Fua Y, Ward MO, Rundensteiner EA. Structure-based brushes: A mechanism for navigating hierarchically organized data and information spaces. *IEEE Trans. Visualization and Computer Graphics*. 2000;6(2):150–9.
7. Mesquite [http://mesquiteproject.org].
8. Maddison WP, Maddison DR. Interactive analysis of phylogeny and character evolution using the computer program Mac-Clade. *Folia Primatol (Basel)*. 1989;53:190–202.
9. Kohonen T. Self-organized formation of topologically correct feature maps. *Biol. Cybern*. 1982;43:59–69.
10. Kohonen T. Self-Organizing Maps. HKV Lotsch, Ed. Heidelberg, Germany: Springer-Verlag; 1997.
11. Varfis A. On the use of two traditional statistical techniques to improve the readibility of Kohonen Maps. *In Proceedings of the NATO ASI Workshop on Statistics and Neural Networks*, 1993.
12. Ainsworth EJ. Classification of Ocean Color Using Self-Organizing Feature Maps. *In Proceedings of the 5th International Conference on Soft Computing and Information/Intelligent Systems (IIZUKA98). Japan, October*. Edited by Yamakawa T, Matsumoto G. *World Scientific*. 1998;2: 996–9.
13. Villman T. Topology preservation in self-organizing maps (1999). *In Kohonen Maps*. Edited by Oja E, Kaski S. Elsevier. 1999:288–90.
14. Kaski S, Venna J, Kohonen T. Coloring that Reveals High-Dimensional Structures in Data. *In Proceedings of the 6th International Conference on Neural Information (ICONIP'99), IEEE Service Center, Piscataway, NJ*. 1999;II:729–34.
15. Kaski S, Venna J, Kohonen T. Coloring that reveals cluster structures in multivariate data. *Australian Journal of Intelligent Information Processing Systems*. 2000;6:82–8.
16. CIE (International Commission on Illumination) Technical Report. Colorimetry. 2nd ed. CIE publication No 15.2. Vienna, Austria. 1986.
17. Himberg J. A SOM based cluster visualization and its application for false coloring. *In Proceedings of International Joint Conference on Neural Networks (IJCNN2000), Como*. 2000;3:587–92.
18. Venna J. Dimensionality reduction for visual exploration of similarity structures. PhD thesis, Helsinki University of Technology; 2007.
19. Lespinats S, Verleysen M, Giron A, Fertil B. DD-HDS: a tool for visualization and exploration of highdimensional data. *IEEE Transactions on Neural Networks*. 2007;18(5):1265–79.
20. Bellemain E, Bermingham E, Ricklefs RE. The dynamic evolutionary history of the bananaquit (Coereba flaveola) in the Caribbean revealed by a multigene analysis. *BMC Evolutionary Biology*. 2008;8:240 doi:10.1186/1471-2148-8-240, available at http://www.biomedcentral.com/1471-2148/8/240.
21. Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. *Trends In Genetics*. 1995;11:283–90.
22. Deschavanne P, Giron A, Vilain J, Dufraigne C, Fertil B. Genomic signature is preserved in short DNA fragments. *In Proceedings of the IEEE international Symposium on bio-informatics & biomedical engineering (BIBE2000), 8–10 november 2000, Washington*. 2000;161–7.
23. Karlin S, Mràzek J, Campbell AM. Compositional biases of bacterial genomes and evolutionary implications. *J Bact*. 1997;179: 3 899–913.
24. Edwards SV, Fertil B, Giron A, Deschavanne PJ. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst Biol*. 2002;51:599–613.
25. Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res*. 2003;13(2):145–58.
26. Yap YL, Zhang XW, Danchin A. Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling. *BMC Bioinformatics*. 2003;4(1):43.
27. Chapus C, Dufraigne C, Edwards S, Giron A, Fertil B, Deschavanne P. Exploration of phylogenetic data using a global sequence analysis method. *BMC Evol Biol*. 2005;5:63.
28. Wang Y, Hill K, Singh S, Kari L. The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene*. 2005;346:173–85.
29. Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T. Informatics for Unveiling Hidden Genome Signatures. *Genome Res*. 2003;13: 693–702.

30. Schbath S. An efficient statistic to detect over- and under-represented words in DNA sequences. *J Comput Biol*. 1997;4(2):189–92.
31. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Evol Biol*. 1987;4(4):406–25.
32. Fitch WM, Margoliash E. Construction of phylogenetic trees. *Science*. 1967;155:279–84.

Annexe

In the Bananaquit bird example, a very similar color has been given to birds from all great Antilles Islands. We subsequently run ColorPhylo again while focusing on the Antilles Islands subclass. The result is displayed on Annexe Figure 1. The tight matching between phylogeny and geographical data is demonstrated in details by the color code.



Annexe Figure 1. Upper left insert: The subtree of birds from Great Antilles islands. Upper right insert: the map of catching sites. Lower insert: the map of species in the colorimetric taxonomic space. Colors attributed by ColorPhylo are used in the three inserts.