Cerutti et al. BMC Bioinformatics 2011, 12:58
http://www.biomedcentral.com/1471-2105/12/58

BMC Bioinformatics

RESEARCH ARTICLE Open Access

# Taxon ordering in phylogenetic trees: a workbench test

Francesco Cerutti[1,2], Luigi Bertolotti[1,2], Tony L Goldberg[3], Mario Giacobini[1,2*]

## Abstract

**Background:** Phylogenetic trees are an important tool for representing evolutionary relationships among organisms. In a phylogram or chronogram, the ordering of taxa is not considered meaningful, since complete topological information is given by the branching order and length of the branches, which are represented in the root-to-node direction. We apply a novel method based on a ($\lambda + \mu$)-Evolutionary Algorithm to give meaning to the order of taxa in a phylogeny. This method applies random swaps between two taxa connected to the same node, without changing the topology of the tree. The evaluation of a new tree is based on different distance matrices, representing non-phylogenetic information such as other types of genetic distance, geographic distance, or combinations of these. To test our method we use published trees of Vesicular stomatitis virus, West Nile virus and Rice yellow mottle virus.

**Results:** Best results were obtained when taxa were reordered using geographic information. Information supporting phylogeographic analysis was recovered in the optimized tree, as evidenced by clustering of geographically close samples. Improving the trees using a separate genetic distance matrix altered the ordering of taxa, but not topology, moving the longest branches to the extremities, as would be expected since they are the most divergent lineages. Improved representations of genetic and geographic relationships between samples were also obtained when merged matrices (genetic and geographic information in one matrix) were used.

**Conclusions:** Our innovative method makes phylogenetic trees easier to interpret, adding meaning to the taxon order and helping to prevent misinterpretations.

## Background

Phylogenetic trees are an important tool in evolutionary biology for representing the history of evolution of organisms. They are composed of nodes, representing hypothetical ancestors, and branches or edges, reflecting the relationship between nodes. Terminal nodes or taxa represent the taxa whose evolution has been investigated, and they can represent extant or extinct organisms [1]. In phylograms and chronograms, branches contain information, e.g. character changes or evolutionary time; in both cases they represent distances between nodes. Consequently, the taxon order is meaningless, and the closeness of taxa can be misleading. For example, two taxa lying adjacent to each other on a phylogenetic tree may actually be very distantly related, creating problems of interpretation.

Previous work has already accepted the challenge of ordering the taxa to add a meaning according to the genetic distance. The software Neighbor-Net [2] is shown to build a network which also minimizes the distance among taxa, given a matrix that fulfill the Kalmanson inequalities [3,4]. This software relies on phylogenetic networks, using an algorithm based on Neighbor Joining [5]. Levy and Pachter [4] showed that the algorithm, considering the problem as a "traveling salesman" problem, is robust for ordering taxa according to a distance matrix. Other studies have endeavored to minimize the distance between taxa as a minimum Hamiltonian path [6,7], either while building the tree or after having built it.

In our previous work [8,9] we introduced an innovative method to order taxa on a phylogenetic tree to make taxon order more meaningful. We used ($\lambda + \mu$) - Evolutionary

* Correspondence: mario.giacobini@unito.it
[1]Department of Animal Production, Epidemiology and Ecology, Faculty of Veterinary Medicine, University of Torino, via Leonardo da Vinci 44, 10095, Grugliasco (TO), Italy
Full list of author information is available at the end of the article

BioMed Central

© 2011 Cerutti et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Algorithms (EAs) [10,11] to reorganize taxon order by random rotation of internal nodes (thus not modifying the topology), evaluating the modified tree on the basis of genetic distances. Figure 1 reports a brief scheme of the process to generate and select the optimal tree. The main goal is to order the taxa on a phylogenetic tree according to any given distance matrix. The tree can be built with any software for phylogenetic inference and then used as input for our algorithm.

In the current study we apply this method to different data sets using different types of distances. We chose published phylogenetic trees of three RNA viruses infecting different hosts with different modes of transmission: Vesicular stomatitis virus (VSV) presented by Perez et al. [12], West Nile virus (WNV) presented by Bertolotti et al. [13] and Rice yellow mottle virus (RYMV) presented by Abubakar et al. [14]. We reorganized each tree using genetic and geographic distance matrices as well as combined distances (geographic and genetic) to improve graphical representation by optimizing the order of taxa.

## Results and discussion

### VSV case study

VSV is a negative-sense single-stranded RNA virus member of the family *Rhabdoviridae*, which causes vesicular stomatitis in horses, cattle, swine, and certain wildlife species [15]. Starting from the tree originally published in [12], we created an Euclidean distance matrix from collection sites coordinates. Then, the EA was run 50 times, each starting with the original tree plus $\lambda$-1 different initial trees (see Methods), generated from the original one (Figure 2a). The performance of the runs was comparable, and we analyzed the best trees (Figure 2b). The order of taxa follows a clear north-south progression, reflecting the geographic arrangement of collection sites, represented in the map in Figure 3. In other words, the algorithm was able to group those taxa belonging to the same state (geographically closer each other). We also tested the algorithm using a separate genetic distance matrix and a combined genetic and geographic distance matrix. Using genetic distances only, the improvement in tree representation is less evident, as might be expected considering that

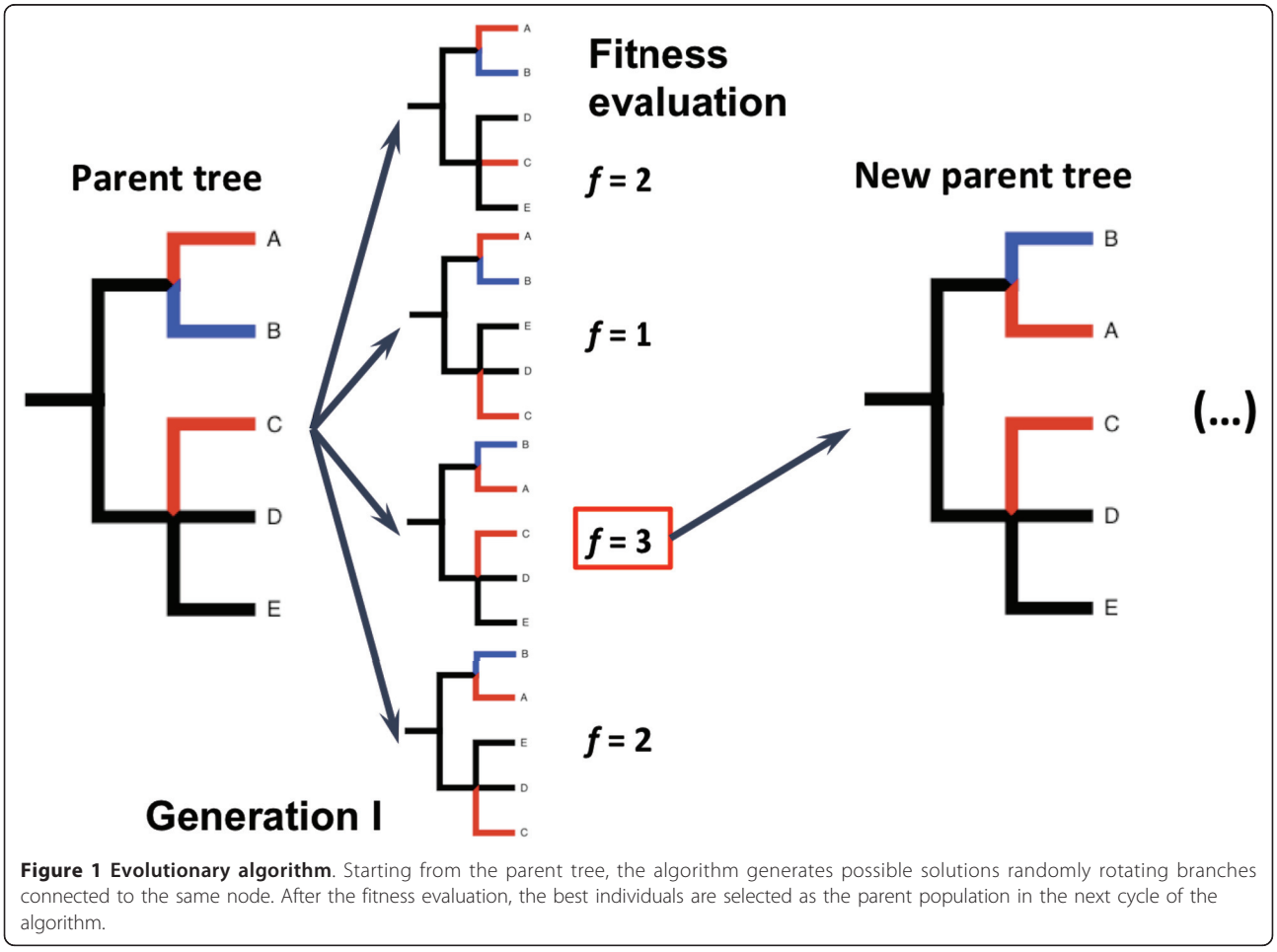


**Figure 1 Evolutionary algorithm**. Starting from the parent tree, the algorithm generates possible solutions randomly rotating branches connected to the same node. After the fitness evaluation, the best individuals are selected as the parent population in the next cycle of the algorithm.

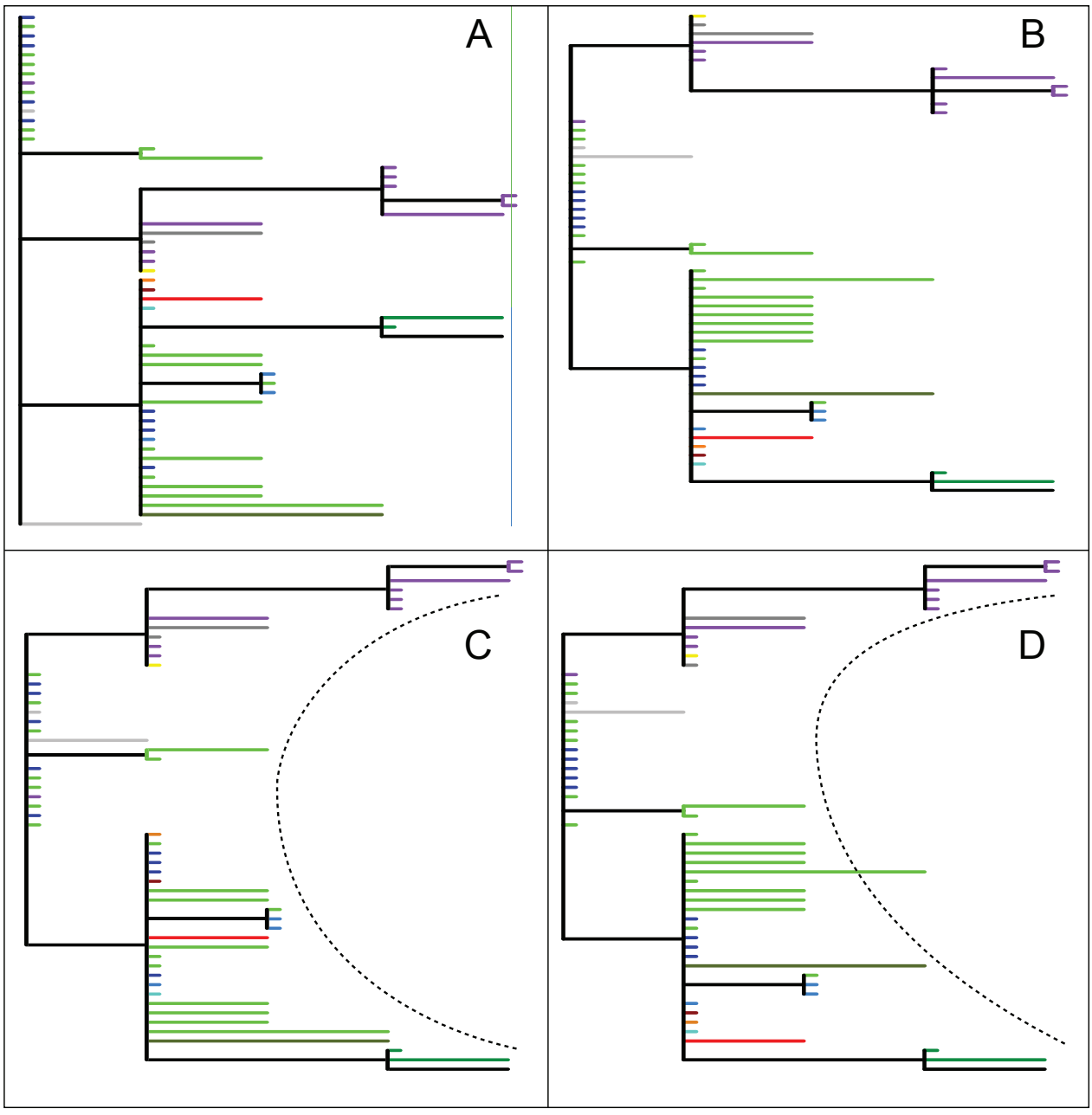


**Figure 2 VSV trees**. The original tree, as presented by Perez et al. (A). The best trees obtained using the geographic (B), genetic (C) and combined (D) distances. The dashed line in C and D highlights the "C" shape acquired by the clades.

the original tree was constructed using genetic data. However, the EA returns a tree in which the most genetically divergent clades are moved to the extremities of the phylogeny, leading to a "C"-like shape, as shown in Figure 2c. The effect of combining the two matrices is strongly evident: the tree has the same "C"-like shape arrangement as the genetically modified tree, and moreover it conserves the aggregation of taxa from same states (locations) (Figure 2d).

### WNV case study

WNV is a positive-sense single-stranded RNA virus belonging to the family *Flaviviridae*, and it is transmitted primarily through the bite of infected mosquitoes to birds. Occasionally it infects horses and humans causing West Nile febrile illness and neurologic disease [16]. The original tree, from [13], shown in Figure 4a, was rearranged using a matrix of geographic distances, as described in the Methods section. In the case of WNV,

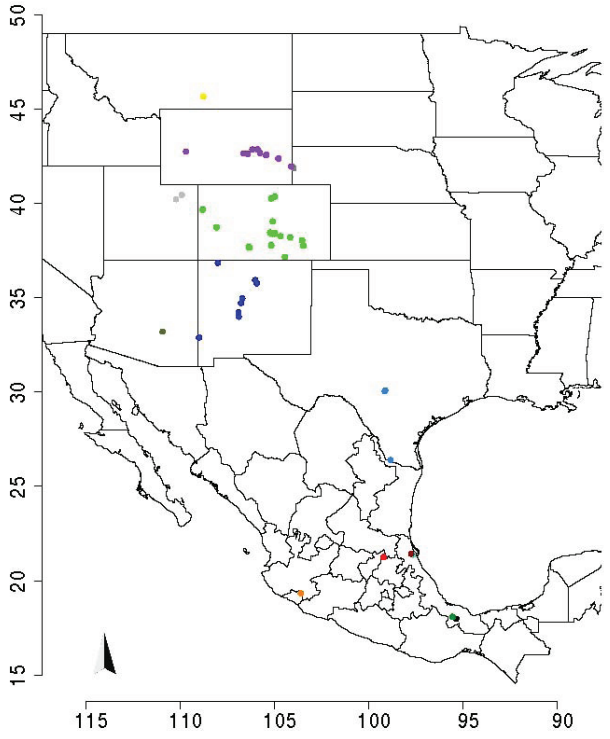


**Figure 3 Map of VSV samples**. Map representing the study area of USA and Mexico where VSV samples were collected. Each site has a color, that is the same of the tips of the samples collected in it.

the geographic arrangement of locations is not linear (Figure 5), as in the case of VSV. In the modified tree (Figure 4b), taxa nevertheless group by sampling location. Due to large sample size and constraints of tree topology, this grouping is less evident than in case of VSV.

In the original paper, the authors declared that more than the 90% of viral genetic variance was contained within sampling sites. Modifying the tree using a matrix of genetic distances yielded a tree that did not tend to group taxa by geographic location (Figure 4c); rather, the modified tree acquired the aforementioned "C"-like shape. As reported in the same paper, the authors found a significant association between viral genetic and spatial distances, with samples collected from the same site likely to be genetically similar. The modified tree obtained with merged genetic and geographic distance matrices tended to move samples collected from the same location closer, supporting the original contention that that viruses from the same location are also genetically similar (Figure 4d). Furthermore the "C"-like shape of the tree was conserved even when merged matrices were used. The new trees obtained therefore graphically support the statistical analyses in the original publication showing weak genetic-geographic association [13].

### RYMV case study

RYMV is a positive-sense single-stranded RNA virus belonging to the genus Sobemovirus and it is considered to be among the most important rice pathogens in sub-Saharan Africa [17,18]. The original tree of RYMV, published in [14], is shown in Figure 6a and has fewer taxa than trees in the previous examples. For this reason, RYMV provided a convenient system to examine the influence of the *radius* parameter in the fitness evaluation, where we previously showed [8] that $r = 8$ offers an acceptable balance between computational intensity and accuracy. In the RYMV case, where the tree had 39 terminal taxa, $r = 8$ was too large to discriminate differences among samples, such that we examined $r \in \{2, 4\}$. The best relative fitness improvement was obtained by $r = 2$, since this value was able to discriminate well among distances in the process of fitness evaluation.

The tree modified using geographic distance (shown in Figure 6b) shows two clusters of geographic locations: samples from West and East Africa (see map in Figure 7). Within clusters, the taxon order reflects the geographic relationship on the map. In addition, samples from central Africa (dots in red, dark red and dark blue on the map) were moved farther away from Tanzanian samples (dots in greens and black) than from samples in Western African countries. This result may be explained by the fact that the algorithm would be expected to optimize taxon order within clusters more efficiently than between clusters. We note that manually modifying the tree in order to improve geographic order often returned a worse fitness compared to the artificially evolved tree (data not shown).

## Conclusions

In our previous studies [8,9] we introduced an innovative method to give a meaning to the order of taxa on a phylogenetic tree using an evolutionary algorithm. Here we test the algorithm using different viral systems, trees and distances with the aim of improving the graphic representation of the tree without modifying its topology. Qualitatively, the most improved tree was obtained when we evolved the VSV phylogenetic tree using a matrix of geographic distances, since the order of taxa on the improved tree closely mirrored geography. In the case of WNV, our algorithm generated trees that support previous studies of genetic diversity and spatial correlation. In the case of RYMV, we identified a possible limitation of the algorithm: the improvement in graphical representation may be less pronounced when the tree contains small numbers of taxa. For this reason we evaluated the role of *r*, highlighting that, in case of small sample sets, small radii may be sufficient to achieve marked improvements. The strength of this method is that a phylogeny can be inferred using any available

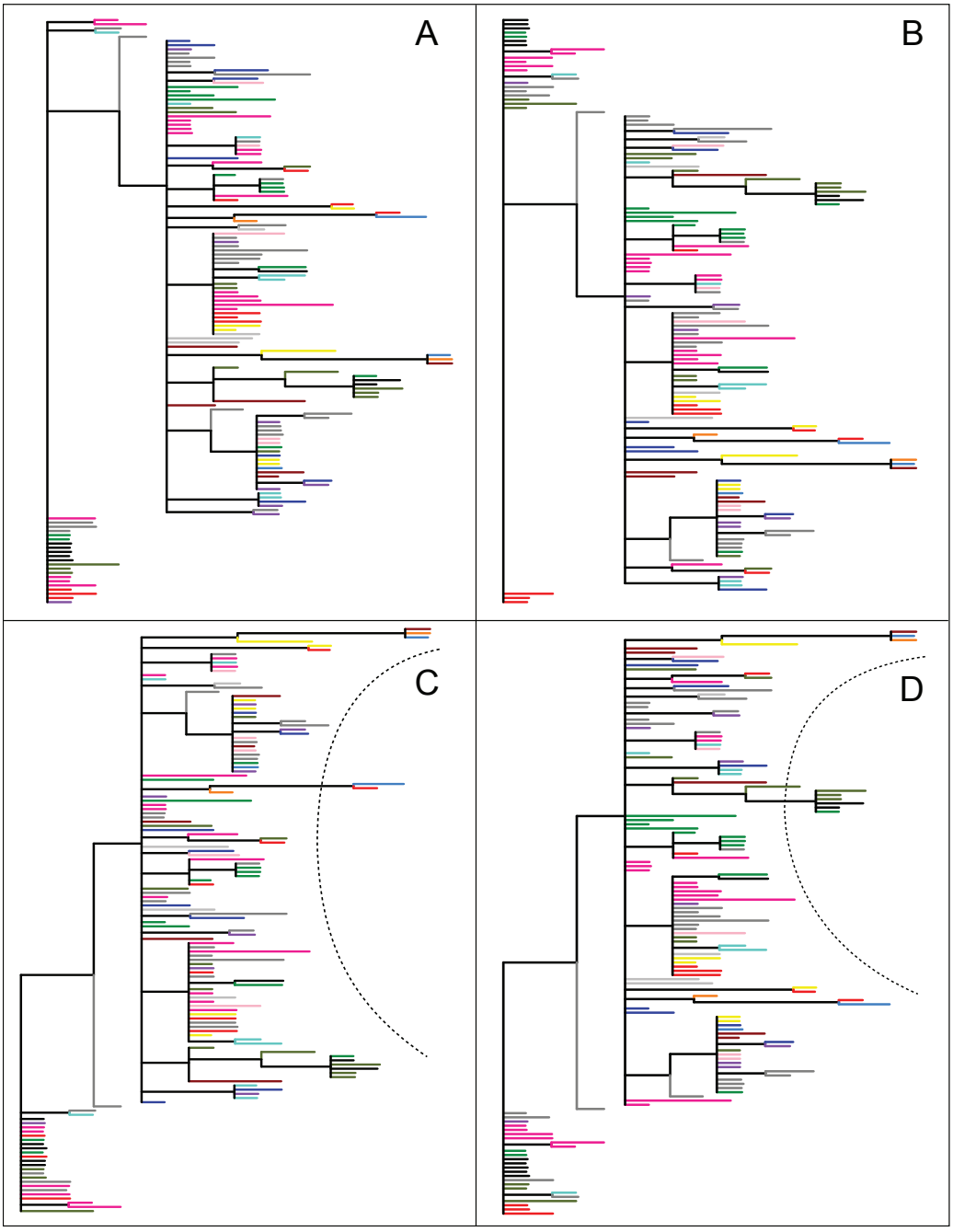


**Figure 4 WNV trees**. The original tree, as presented by Bertolotti et al. (A). The best trees obtained using the geographic (B), genetic (C) and combined (D) distances. The dashed line in C and D highlights the "C" shape acquired by the clades.

method for building trees, using distance and algorithm, from a simple approach (like Neighbor-Joining) to a more accurate Bayesian inference. The user can choose the appropriate method, using the genetic information contained in the considered sequences and, in the next step, add a second matrix, containing for example geographic information. The tests of the algorithm presented here showed promise. Very good results were achieved when the geographic distribution of the samples was linear, as in the VSV and RYMV cases. When the geographic distribution was more complex, as in the case of WNV, the grouping of taxa collected from the

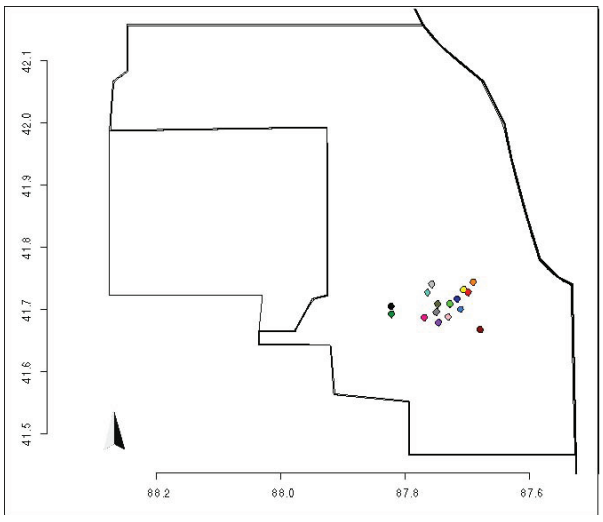


**Figure 5 Map of WNV samples**. Map representing Cook and DuPage counties, and the collection sites used in [30]. Each site has a color, that is the same of the tips of the samples collected in it.

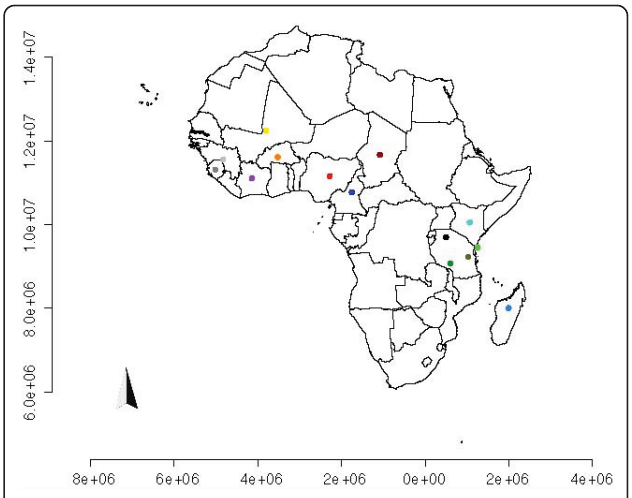


**Figure 7 Map of RYMV samples**. Map representing African states and sample origins, as reported in [14]. Each site has a color, that is the same of the tips of the samples collected in it.

same site is hard to see. In this case, geographic information alone is not enough, but much greater improvement is gained when both genetic and geographic information are incorporated, as shown in Figure 3d. In conclusion, the algorithm tested in this paper offers a customizable method that can help biologists to better represent results of their phylogenetic analyses, improving the interpretation of phylogenetic trees and making them more understandable.

## Methods

### (5 + 5)- EA

The (5 + 5)- EA used in the present study is a particular case of the large family of $(\lambda + \mu)$-EAs. As briefly reported in Figure 1, the algorithm starts from an original tree and creates. $\lambda$-1 trees by random swaps of pairs

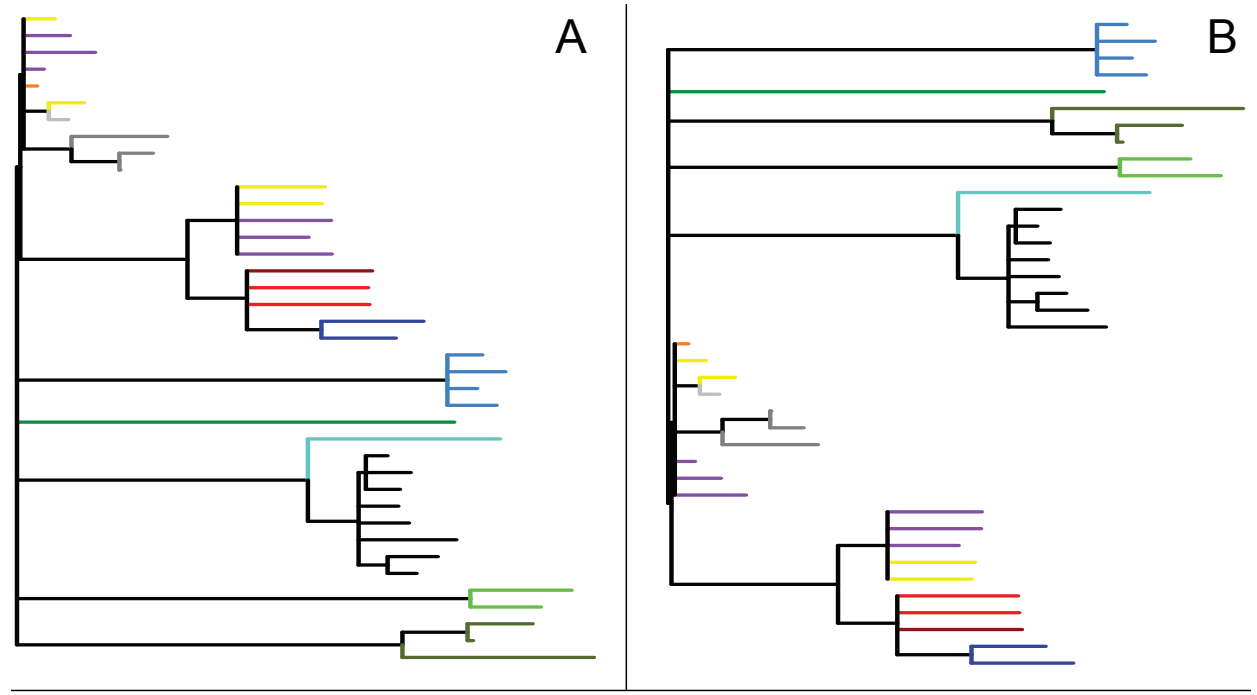


**Figure 6 RYMV trees**. The original tree, as presented by Abubakar et al. (A). The best trees obtained using the geographic (B) distances.

of internal nodes. In this way, a total of $\lambda$ starting trees, with the same topology but different order of taxa, are obtained. In each generation, $\mu$ trees are generated by random mutation of those selected with $\mu$ tournaments between couples chosen by random sampling with reintroduction among the $\lambda$ trees. The next generation parents are then the $\lambda$ fittest trees among the $\lambda + \mu$ ones.

The selection of the best trees is based on the fitness evaluated as the sum of the distances between each taxon and the next $r$ tips, where $r$ is the radius in the fitness evaluation. The distances are contained in a matrix, used as input for the algorithm. Any available data that are able to discriminate among taxa can be used to generate the distance matrix. As a starting tree, an original tree obtained by any available method of phylogenetic inference with any method (e.g. distance-based, parsimony-based, or likelihood-based) is suitable.

### Experimental validations

For the experimental validation of the algorithm, we selected three phylogenetic trees from published works, following two criteria: availability of genetic and geographic data and published phylogenetic trees and associated phylogeographic interpretations. All the distance matrices were normalized to the highest value in order to have values lying in the range [0, 1].

#### *VSV data*

We reconstructed the tree presented by Perez et al. using the maximum-likelihood optimality criterion as implemented in PAUP* version b10 [19] and the nucleotide substitution parameters as estimated using Modeltest, version 3.7 [20], as reported in [12]. Unlike the original tree, the new tree contains only 55 taxa instead of 59, because of availability of both sequences and geographic coordinates. For the genetic distance matrix, we used nucleotide-level distances corrected with the HKY substitution model [21]. For the geographic distance matrix, we used euclidean distances between collection sites in a Latitude/Longitude coordinate system. The combined matrix of distances was created averaging the cells in the genetic and geographic matrices.

#### *WNV data*

The tree is from Bertolotti et al. [13] and contains 140 samples collected from Chicago, USA. We re-evaluated the best evolution model with Modeltest, version 3.7, and inferred the phylogeny using MrBayes software [22,23]. The genetic distance matrix was created from nucleotide sequence data and an uncorrected p-distance [24]. The geographic distance matrix was created from Euclidean distances between collection sites in a Latitude/Longitude coordinate system. The combined matrix of distances was created with the same method used for the VSV example.

#### *RYMV data*

RYMV is a positive-sense single-stranded RNA virus belonging to the genus Sobemovirus and it is considered to be among the most important rice pathogens within sub-Saharan Africa [17,18]. We considered the tree published by Abubakar et al. and re-built it using neighbor-joining tree and pairwise nucleotide sequence distances with the Kimura two-parameters model, as reported in the original paper [14], but removing the out-group sequence; thus the tree has 39 taxa instead of 40. The geographic distances were computed as a Euclidean distances between the centroids of the district (for Tanzania) or of the area of the state (for the other states) in a UTM coordinate system, zone 36.

### Geographic data

Data on country borders were obtained from shapefiles available online, managed and plotted with R software [25] and the packages *maptools* [26] and *shape* [27]. In particular, USA border data were downloaded from http://www.census.gov/geo/www/cob/st2000.html in ESRI Shapefile (.shp) format for all 50 States, D.C., and Puerto Rico; Mexico shapefile was downloaded from http://www.vdstech.com/map_data.htm and Africa shapefile was downloaded from http://www.maplibrary.org/index.php. Collection site coordinates for VSV were kindly provided by L. Rodriguez, A. Perez and S. Pauszek. WNV collection site coordinates were already available. RYMV collection site coordinates were extracted from the shapefile using GRASS-GIS software [28]; specifically, coordinates of Tanzanian collection sites were selected as the centroid of the collection district (Mwanza, Mbeya, Morogoro, Pemba, as described in [14]), and coordinates of the other states were selected as the centroids of those states.

### Computational performance

Algorithms were written in R, using the package 'ape' [29]. The runs were performed on the cluster IBM-BCX available at the Supercomputing Group of the CINECA Systems & Tecnologies Department between February and June 2010.

#### Acknowledgements

The authors thank the Supercomputing Group of the CINECA Systems & Tecnologies Department for supporting in computations. MG acknowledges funding (60% grant) by the Ministero dell'Università e della Ricerca Scientifica e Tecnologica. LB gratefully acknowledges financial support by Ricerca Sanitaria Finalizzata 2008-Regione Piemonte. This work is supported by the National Science Foundation/National Institutes of Health Ecology of Infectious Diseases Program under award number EF-0840403. Furthermore the authors thank L. Rodriguez, A. Perez and S. Pauszek for kindly sharing genetic and geographic data. The authors also acknowledge Donal Bisanzio for his precious help for geographic data management and map plotting.

**Author details**
[1]Department of Animal Production, Epidemiology and Ecology, Faculty of Veterinary Medicine, University of Torino, via Leonardo da Vinci 44, 10095, Grugliasco (TO), Italy. [2]Molecular Biotechnology Center, University of Torino, via Nizza 52, 10126, Torino, Italy. [3]Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, 1656 Linden Drive, Madison, Wisconsin, 53706, USA.

**Authors' contributions**
FC implemented the computational model, carried out the simulations, and participated in the interpretation of the results. LB conceived the design of the study, implemented the computational model, and participated in the interpretation of the results. TLG conceived the design of the study and participated in the interpretation of the results. MG conceived the design of the study, participated in the interpretation of the results, and coordinated the participants' contributions. All authors participated in writing the manuscript and approved it.

Received: 8 September 2010 Accepted: 22 February 2011
Published: 22 February 2011

**References**

1. Page RDM, Holmes EC: *Molecular evolution: a phylogenetic approach* Wiley-Blackwell; 1998.
2. Bryant D, Moulton V: **Neighbor-net: an agglomerative method for the construction of phylogenetic networks.** *Molecular biology and evolution* 2004, **21(2)**:255-65.
3. Thuillard M, Fraix-Burnet D: **Phylogenetic applications of the minimum contradiction approach on continuous characters.** *Evolutionary Bioinformatics* 2009, **5**:33-46.
4. Levy D, Pachter L: **The neighbor-net algorithm.** *ArXiv Mathematics e-prints* 2007, 1-23.
5. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Molecular Biology and Evolution* 1987, **4(4)**:406-425.
6. Moscato P, Buriol L, Cotta C: **On the analysis of data derived from mitochondrial DNA distance matrices: Kolmogorov and a traveling salesman give their opinion.** In *Advances in Nature Inspired Computation: the PPSN VII Workshops.* Edited by: Corne D. PEDAL, University of Reading; 2002:37-38.
7. Cotta C, Moscato P: **A memetic-aided approach to hierarchical clustering from distance matrices: application to gene expression clustering and phylogeny.** *Biosystems* 2003, **72(1-2)**:75-97.
8. Cerutti F, Bertolotti L, Goldberg TL, Giacobini M: **Adding Vertical Meaning to Phylogenetic Trees by Artificial Evolution.** *Proceedings of the 10th European Conference on Artificial Life (ECAL 2009), Volume LNCS/LNAI 5777, 5778* Springer; 2010.
9. Cerutti F, Bertolotti L, Goldberg T, Giacobini M: **Investigating Populational Evolutionary Algorithms to Add Vertical Meaning in Phylogenetic Trees.** In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, Volume 6023 of Lecture Notes in Computer Science.* Edited by: Pizzuti C, Ritchie MD, Giacobini M. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010:240-247.
10. Eiben AE, Smith J: *Introduction to Evolutionary Computing (Natural Computing Series)* Springer; 2010.
11. Tettamanzi A, Tomassini M: *Soft Computing: Integrating Evolutionary, Neural, and Fuzzy Systems* Springer; 2010.
12. Perez AM, Pauszek SJ, Jimenez D, Kelley WN, Whedbee Z, Rodriguez LL: **Spatial and phylogenetic analysis of vesicular stomatitis virus over-wintering in the United States.** *Preventive veterinary medicine* 2010, **93(4)**:258-64.
13. Bertolotti L, Kitron UD, Walker ED, Ruiz MO, Brawn JD, Loss SR, Hamer GL, Goldberg TL: **Fine-scale genetic variation and evolution of West Nile Virus in a transmission "hot spot" in suburban Chicago, USA.** *Virology* 2008, **374(2)**:381-389.
14. Abubakar Z, Ali F, Pinel A, Traore O, N'Guessan P, Notteghem J, Kimmins F, Konate G, Fargette D: **Phylogeography of Rice yellow mottle virus in Africa.** *Journal of General Virology* 2003, **84(3)**:733-743.
15. Rodriguez LL, Nichol S: In *Vesicular stomatitis viruses.* r edition. Edited by: Webster. London: Academic Press; 1999:1910-1919.
16. Diamond MS: *West Nile Encephalitis Virus Infection: viral pathogenesis and the host immune response* Springer, New York; 2009.
17. Regenmortel M, Fauquet C, Bishop D, Carstens E, Estes M, Lemon S, Maniloff J, Mayo M, McGeoch D, Pringle C: *Virus taxonomy classification and nomenclature of viruses, seventh report of the International Committee on Taxonomy of Viruses* Academic Press; 2000.
18. Abo ME, Sy AA, Alegbejo MD: **Rice Yellow Mottle Virus (RYMV) in Africa: Evolution, Distribution, Economic Significance on Sustainable Rice Production and Management Strategies.** *Journal of Sustainable Agriculture* 1997, **11(2)**:85-111.
19. Swofford D: **PAUP: phylogenetic analysis using parsimony, version 4.0 b10.** *Sinauer Associates, Sunderland, MA* 2002.
20. Posada D, Crandall K: **MODELTEST: testing the model of DNA substitution.** *Bioinformatics* 1998, **14(9)**:817-818.
21. Hasegawa M, Kishino H, Yano Ta: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.** *Journal of Molecular Evolution* 1985, **22(2)**:160-174.
22. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP: **Bayesian inference of phylogeny and its impact on evolutionary biology.** *Science (New York, N.Y.)* 2001, **294(5550)**:2310-4.
23. Ronquist F: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19(12)**:1572-1574.
24. Nei M: *Molecular Evolutionary Genetics* Columbia University Press; 1987.
25. Team RDC: *R: A Language and Environment for Statistical Computing* 2008.
26. Lewin-Koh N, Bivand R, Pebesma E, Archer E: *S: Maptools Tools for reading and handling spatial objects* 2008.
27. Soetaert K: *shape: Functions for plotting graphical shapes, colors* 2009.
28. Neteler M, Mitasova H: In *Open source GIS: a GRASS GIS approach. Volume 9.* Springer; 2008.
29. Paradis E: **APE: Analyses of Phylogenetics and Evolution in R language.** *Bioinformatics* 2004, **20(2)**:289-290.
30. Bertolotti L, Kitron U, Goldberg TL: **Diversity and evolution of West Nile virus in Illinois and the United States, 2002-2005.** *Virology* 2007, **360**:143-9.

doi:10.1186/1471-2105-12-58
**Cite this article as:** Cerutti *et al.*: Taxon ordering in phylogenetic trees: a workbench test. *BMC Bioinformatics* 2011 **12**:58.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central