Phylogenies from whole genomes: Methodological update within a distance-based framework

Alexander F. Auch(1), Stefan R. Henz(2), and Markus Göker(3)

(1) Center for Bioinformatics (ZBIT), Sand 14, Tübingen, University of Tübingen, Germany
(2) Max Planck Institute for Developmental Biology, Spemannstrasse 37-39, Tübingen, Germany
(3) Organismic Botany/Mycology, Auf der Morgenstelle 1, Tübingen, University of Tübingen, Germany
Contact: auch@informatik.uni-tuebingen.de

1 Introduction

Methods which derive pairwise distances directly from completely sequenced genomes are a potentially important and efficient tool within the growing field of phylogenomics.

We have shown in two previous studies (Henz et al. 2005; Auch et al. 2006) that the Genome BLAST Distance Phylogeny (GBDP) approach leads to reliable phylogenetic estimates if applied to prokaryotic as well as plastid and mitochondrial genomes. Basically, GBDP first invokes tools such as BLAST (Altschul et al. 1990) to identify high-scoring segment pairs (HSPs) between all pairs of genomes; afterwards, pairwise distances are estimated based on different formulae. To identify the most valuable distance formulae, Auch et al. (2006) compared quite a few modifications of GBDP (in combination with different tree reconstruction methods) with respect to topological accuracy as measured by c-scores. Additionally, δ values (Holland et al. 2002) were computed to directly estimate distance quality. This approach may be particularly useful since it does not require to specify a reference taxonomy in advance. It was demonstrated that both evaluation methods usually coincide. Thus, a framework was established to evaluate suitability of distance methods for phylogenetic inference in general.

2 Methods

Here, we examine

1. a new GBDP distance formula, based on a combination of two previously existing ones
2. use of BLAT instead of BLASTN and TBLASTX HSP search
3. an alternative measure for the agreement of a distance matrix with a predefined reference topology
4. alternative topology-independent measures of distance quality per se.

All examinations were based on an enlarged dataset compared to that used by Auch et al. (2006), additionally containing interesting key taxa (see Fig.1).

The new distance formula can be described as follows. Defining I as the sum of the number of identical characters over all HSPs between genomes X and Y, we obtain a similarity formula:

$$s_{id2} := \frac{I}{g} \tag{1}$$

using the following two denominators:

$$g_1 := |X| + |Y| \tag{2}$$
$$g_2 := 2 \cdot \min(|X|, |Y|) \tag{3}$$

To derive a dissimilarity function, subtraction from 1 or logarithmic conversion can be used, as described in Auch et al. (2006).

In addition to the c-score (Henz et al. 2005), we here measured agreement with the reference topology (the NCBI taxonomy tree) by converting it into a matrix of patristic distances and computing the Spearman correlation between it and the distance matrix put to test using CADM (Legendre 2001).

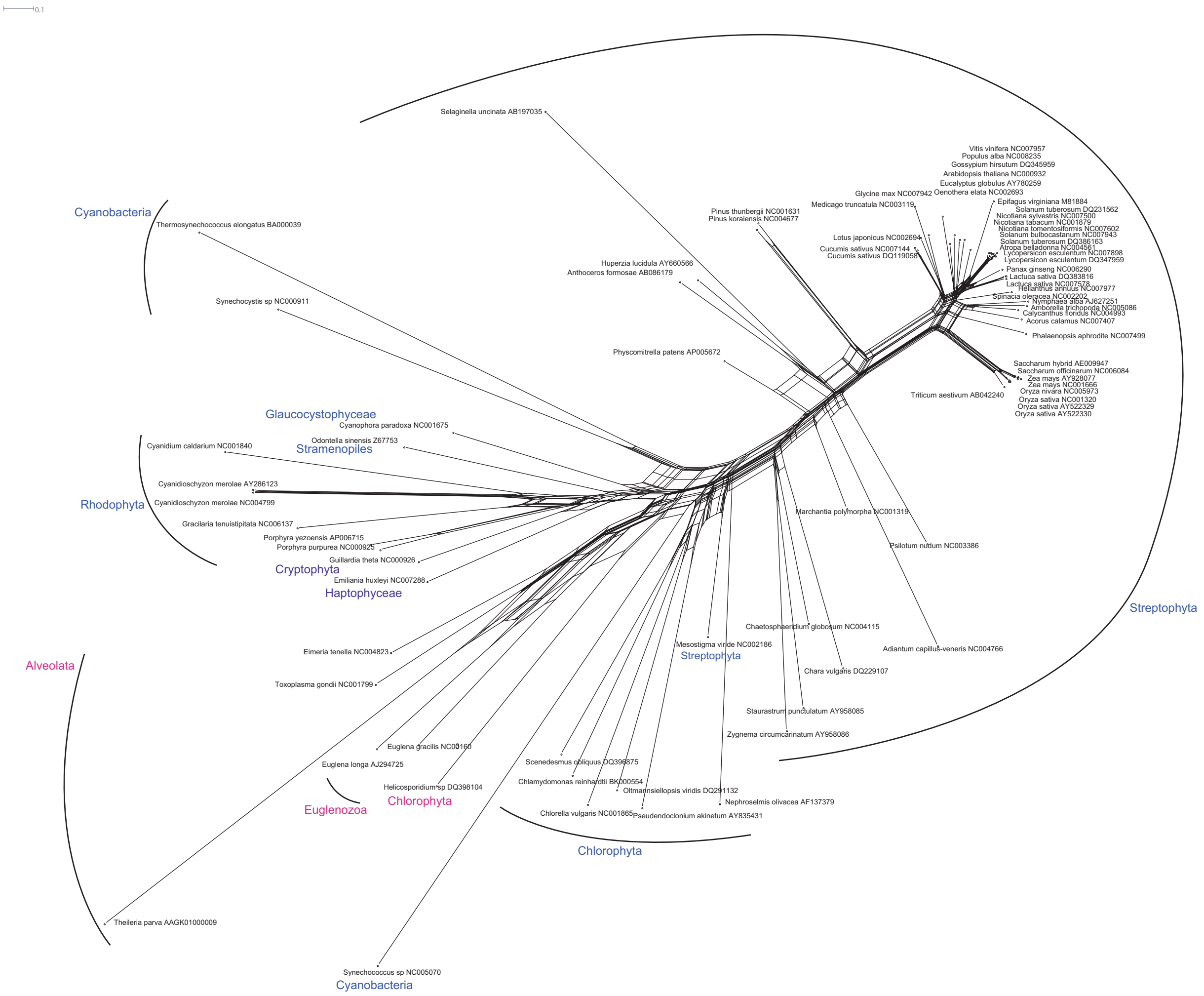
0.1
Selaginella uncinata AB197035
Vitis vinifera NC007957
Populus alba NC008235
Gossypium hirsutum DQ345959
Arabidopsis thaliana NC000932
Eucalyptus globulus AY780259
Glycine max NC007942
Oenothera elata NC002693
Medicago truncatula NC003119
Epifagus virginiana M81884
Solanum tuberosum DQ231562
Nicotiana sylvestris NC007500
Nicotiana tabacum NC001879
Nicotiana tomentosiformis NC007602
Solanum bulbocastanum NC007943
Solanum tuberosum DQ386163
Atropa belladonna NC004561
Lycopersicon esculentum NC007898
Lycopersicon esculentum DQ347959
Pinus thunbergii NC001631
Pinus koraiensis NC004677
Lotus japonicus NC002694
Cucumis sativus NC007144
Cucumis sativus DQ119058
Panax ginseng NC006290
Lactuca sativa DQ383816
Lactuca sativa NC007578
Helianthus annuus NC007977
Spinacia oleracea NC002202
Nymphaea alba AJ627251
Amborella trichopoda NC005086
Calycanthus floridus NC004993
Acorus calamus NC007407
Phalaenopsis aphrodite NC007499
Cyanobacteria
Thermosynechococcus elongatus BA000039
Huperzia lucidula AY660566
Anthoceros formosae AB086179
Synechocystis sp NC000911
Physcomitrella patens AP005672
Saccharum hybrid AE009947
Saccharum officinarum NC006084
Zea mays AY928077
Zea mays NC001666
Oryza nivara NC005973
Oryza sativa NC001320
Oryza sativa AY522329
Oryza sativa AY522330
Triticum aestivum AB042240
Glaucocystophyceae
Cyanophora paradoxa NC001675
Odontella sinensis Z67753
Stramenopiles
Cyanidium caldarium NC001840
Cyanidioschyzon merolae AY286123
Rhodophyta
Cyanidioschyzon merolae NC004799
Gracilaria tenuistipitata NC006137
Porphyra yezoensis AP006715
Porphyra purpurea NC000925
Guillardia theta NC000926
Cryptophyta
Emiliania huxleyi NC007288
Haptophyceae
Marchantia polymorpha NC001319
Psilotum nudum NC003386
Streptophyta
Chaetosphaeridium globosum NC004115
Mesostigma viride NC002186
Streptophyta
Adiantum capillus-veneris NC004766
Eimeria tenella NC004823
Alveolata
Toxoplasma gondii NC001799
Chara vulgaris DQ229107
Staurastrum punctulatum AY958085
Zygnema circumcarinatum AY958086
Euglena gracilis NC00160
Euglena longa AJ294725
Scenedesmus obliquus DQ396875
Chlamydomonas reinhardtii BK000554
Helicosporidium sp DQ398104
Oltmannsiellopsis viridis DQ291132
Chlorophyta
Nephroselmis olivacea AF137379
Euglenozoa
Chlorella vulgaris NC001865
Pseudendoclonium akinetum AY835431
Chlorophyta
Theileria parva AAGK01000009
Synechococcus sp NC005070
Cyanobacteria

Fig. 1. Plastid NeighborNet phylogeny, using the new distance formula and g_2

3 Conclusions

The GBDP distance variant introduced here seems to be quite valuable, most likely since it combines information from both within-HSP sequence similarity and total HSP length compared to total genome length (see Table 2).

Compared to BLAST, BLAT (Kent 2002) is much faster, but identifies only a subset of HSPs (at least under default values). Thus, BLAT cannot be applied if genomes are too distantly related, as in mitochondria sampled from major eukaryotic groups, and does not result as well as BLAST/TBLASTX (see Table 2). However, we observed a general trade-off between speed of HSP search and phylogenetic accuracy.

Within topology-independent distance quality measures, additivity as formulated by De Soete (De Soete 1986) as well as rescaled Q values (Guindon and Gascuel 2002) turned out to be most in accordance with c-scores (see Table 1). However, a simple modification of δ (which we called ϵ) gives even better results, particularly because it is linearly related to c-score (not shown) and, hence, displays both high parametric and non-parametric correlation coefficients (see Table 1). Albeit being a less precise measure than the c-score, agreement with a reference topology can also be measured with the simple non-parametric correlation approach as implemented in CADM which does not require to compute trees from distance matrices.

	c-score			CADM-score		
Metrics	Pearson	Kendall	Spearman	Pearson	Kendall	Spearman
R^2	-0.420	-0.074	-0.129	-0.384	-0.022	-0.049
non-additivity; Q^2 [3]	-0.638	-0.351	-0.526	-0.710	-0.393	-0.565
$R-Q$ [0]	-0.028	0.220	0.295	-0.042	0.158	0.241
R [2]	-0.604	-0.212	-0.328	-0.601	-0.226	-0.356
Q [2]	-0.680	-0.399	-0.582	-0.751	-0.448	-0.649
R [1]	-0.631	-0.125	-0.189	-0.202	0.040	0.039
Q [1]	-0.730	-0.315	-0.451	-0.375	-0.164	-0.246
$(R-Q)^2$ [3]	-0.132	0.055	0.056	0.005	0.135	0.157
ϵ; Q/R	-0.701	-0.507	-0.712	-0.789	-0.510	-0.720
non-metricity [3]	-0.543	-0.302	-0.456	-0.634	-0.296	-0.432
δ; Q/R	-0.578	-0.404	-0.601	-0.746	-0.462	-0.664
R [0]	-0.062	0.194	0.263	-0.079	0.138	0.217
Q [0]	-0.147	-0.046	-0.079	-0.170	-0.106	-0.184
non-ultrametricity [3]	-0.459	-0.241	-0.379	-0.588	-0.268	-0.390
$R-Q$ [2]	-0.442	-0.132	-0.226	-0.370	-0.117	-0.201

Table 1. Correlation of Distance Quality Metrics

We define non-ultrametricity and non-additivity as in the minimization formulae of Makarenkov and Legendre (1999) and De Soete (1986). Non-metricity is analogously defined as the square root of the sum of $(d_{ij}-d_{ik}-d_{jk})^2$ for all triplets of taxa i, j, and k in which $d_{ij} > d_{ik} + d_{jk}$, divided by the sum of all squared distances (SSD). Q is defined as in Guindon and Gascuel (2002) as the sum of q (see Fig. 3) for all quartets of taxa; R analogously sums up r, $R-Q$ sums up $r-q$, and $(R-Q)^2$ sums up $(r-q)^2$. δ (Holland et al. 2002) sums up Q/R if $R \neq 0$ and 0 if $R = 0$; ϵ adds 1 if $R = 0$. Scaling formulae are [0], division by the total number of quartets; [1], division by the total number of quartets and the largest distance value in the matrix; [2] division by the square root of SSD; [3] taking the square root after division by SSD.

With respect to phylogenetic outcome, the analyses corroborate our earlier findings that GBDP groups Apicomplexa organelle genomes with the ”green lineage” of plastids, i.e. Euglenozoa. This outcome was not affected by the inclusion of the highly derived ”apicoplast” genome of *Theileria parva* (see Fig. 2). As emphasized by Auch et al. (2006), this

	c-score (adj. $R^2 = 0.636$)		ϵ value (adj. $R^2 = 0.859$)		CADM-score (adj. $R^2 = 0.549$)	
explanatory var.	coefficient	$P(x > \|t\|)$	coefficient	$P(x > \|t\|)$	coefficient	$P(x > \|t\|)$
Intercept	0.4510	$< 2 \cdot 10^{-16}$	0.3390	$< 2 \cdot 10^{-16}$	0.3366	$< 2 \cdot 10^{-16}$
UPGMA	-0.0540	0.0002				
Plastids	0.1371	$< 2 \cdot 10^{-16}$	-0.1744	$< 2 \cdot 10^{-16}$	0.1794	$1.59 \cdot 10^{-12}$
BLAT	-0.0303	0.0008	eliminated from model		eliminated from model	
translated	0.0843	$< 2 \cdot 10^{-16}$	-0.0347	0.0002	eliminated from model	
log	not significant		0.0273	0.0027	eliminated from model	
eq.4 (Auch et. al 2006)	-0.2068	$< 2 \cdot 10^{-16}$	0.1003	$3.43 \cdot 10^{-12}$	-0.1506	$1.82 \cdot 10^{-6}$
g_2	-0.0214	0.0324	0.0386	0.0002	not significant	

Table 2. Regression Analysis (insignificant parameters omitted)

result is in disagreement with the "Chromalveolata hypothesis" which states that plastids of recent Alveolata and Stramenopiles are derived from the plastids of the common ancestor of the group.

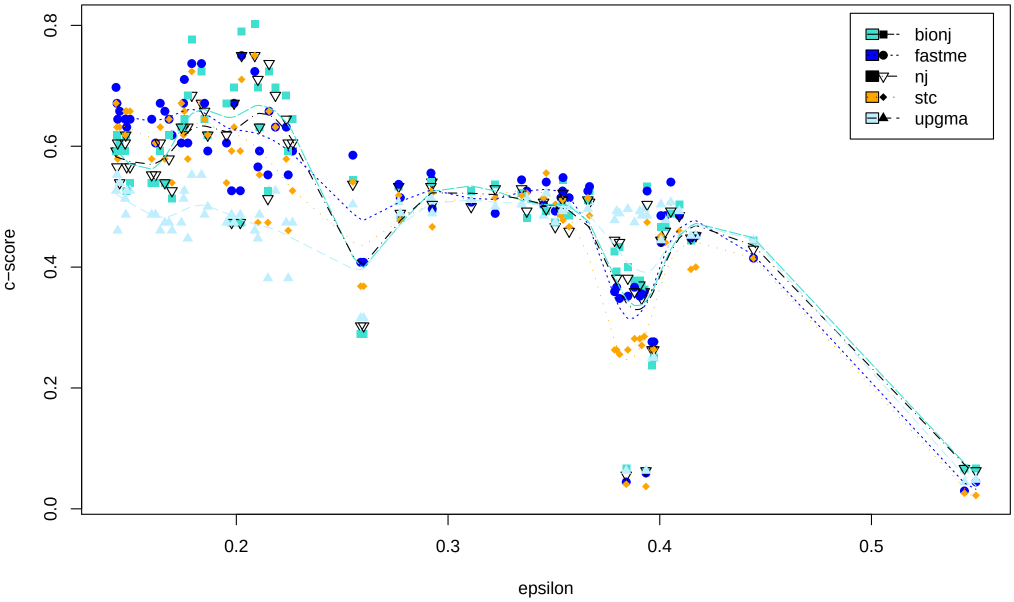

Fig. 2. Comparison of phylogenetic reconstruction methods

Strikingly, *Helicosporidium*, which has been considered as a highly derived, non-photosynthetic green alga, also clusters with this group. This placement additionally supports a green algal ancestry both of the plastids of Euglenozoa as well as the "apicoplasts". However, *Helicosporidium* placement was not consistent between all the trees computed (not shown).

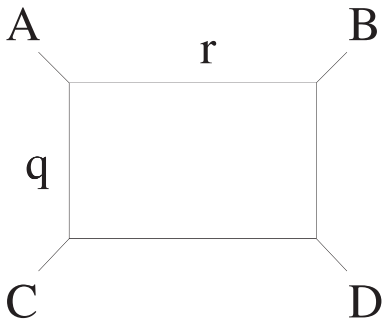

Fig. 3. Quartet of taxa and their distances (see Holland et al. (2002) for further details)

Bibliography

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. J. Mol. Biol., 215:403–410.

Auch AF, Henz SR, Holland BR, Göker M, 2006. Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. BMC Bioinformatics, 7:350.

De Soete G, 1986. Optimal variable weighting for ultrametrix and additive tree clustering. Quality&Quantity, 20:169–180.

Guindon S, Gascuel O, 2002. Efficient biased estimation of evolutionary distances when substitution rates vary across sites. Mol. Biol. Evol., 19(4):534–543.

Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC, 2005. Whole Genome-based Prokaryotic Phylogeny. Bioinformatics, 21:2329–2335.

Holland BR, Huber KT, Dress A, Moulton V, 2002. δ Plots: A Tool for Analyzing Phylogenetic Distance Data. Mol. Biol. Evol., 19(12):2051–2059.

Kent WJ, 2002. BLAT–the BLAST-like alignment tool. Genome Res., 12(4):656–664.

Legendre P, 2001. Congruence among distance matrices: Program CADM users guide. Dèpartement de sciences biologiques, Université de Montréal.

Makarenkov V, Legendre P, 1999. Optimal variable weighting for ultrametric and additive tree clustering. Dèpartement de sciences biologiques, Université de Montréal, 5 pp.