Available online at www.sciencedirect.com

SCIENCE DIRECT®

C. R. Biologies 328 (2005) 11–22

COMPTES RENDUS BIOLOGIES

http://france.elsevier.com/direct/CRASS3/

Genetics / Génétique

Hierarchical clustering based upon contextual alignment of proteins: a different way to approach phylogeny

Anna Gambin [a,*], Piotr P. Slonimski [b,c]

[a] *Institute of Informatics, Warsaw University, Warsaw, Poland*
[b] *Centre de génétique moléculaire, CNRS, 91198 Gif-sur-Yvette, France*
[c] *Institute of Biochemistry and Biophysics, P.A.S. Warsaw, Poland*

Received 25 July 2003; accepted after revision 18 October 2004

Available online 8 December 2004

Presented by Jean Rosa

Abstract

We perform a computational study using a new approach to the analysis of protein sequences. The contextual alignment model, proposed recently by Gambin et al. (2002), is based on the assumption that, while constructing an alignment, the score of a substitution of one residue by another depends on the surrounding residues. The contextual alignment scores calculated in this model were used to hierarchical clustering of several protein families from the database of Clusters of Orthologous Groups (COG). The clustering has been also constructed based on the standard approach. The comparative analysis shows that the contextual model results in more consistent clustering trees. The difference, although small, is with no exception in favour of the contextual model. The consistency of the family of trees is measured by several consensus and agreement methods, as well as by the inter-tree distance approach. ***To cite this article: A. Gambin, P.P. Slonimski, C. R. Biologies 328 (2005).***

Résumé

Classification hiérarchique fondée sur alignement contextuel des protéines : une nouvelle maniere d'aborder la phylogénie. Nous avons utilisé dans notre étude un nouveau modèle d'alignement des séquences protéiques, le modèle contextuel proposé par Gambin et al. (2002). Il postule que, lors de la construction d'alignements, la substitution d'un résidu par un autre dépend de la nature des résidus adjacents. Plusieurs familles protéiques de la base de données COG ont été examinées selon ce nouveau procédé. Il en résulte une classification hiérarchique des taxa microbiens. Les arbres phylogénétiques ainsi obtenus ont été comparés à ceux dérivés de procédés standards. Nous montrons que le modèle contextuel conduit à des hiérarchies qui sont plus cohérentes entre elles et plus conformes à la phylogénie. La différence, bien que petite, est systématique : le modèle contextuel, sans exception, améliore la cohérence entre les arbres phylogénétiques. ***Pour citer cet article : A. Gambin, P.P. Slonimski, C. R. Biologies 328 (2005).***

[*] Corresponding author.
E-mail address: aniag@mimuw.edu.pl (A. Gambin).

1631-0691/$ – see front matter © 2004 Académie des sciences. Published by Elsevier SAS. All rights reserved.
doi:10.1016/j.crvi.2004.11.001

Keywords: Contextual sequence alignment; Hierarchical clustering; Phylogenetic tree; Consensus measures

Mots-clés : Alignement contextuel des séquences ; *Clustering* hiérarchique ; Arbre phylogénétique ; Mesures consensuelles

1. Introduction

One of the fundamental problems in biological classification is the question how to interpret the phylogenetic information contained in a collection of different phylogenetic trees that classify the same set of taxa. One reason for the uncertainty about the true phylogenetic tree is that different choices for molecular sequences often point to different trees, called *gene trees* or *protein family trees* (see e.g. [1–9]). Finding the best way of combining the information contained in numerous different gene trees for the same set of species remains an open problem in contemporary biology.

It is textbook knowledge that a range of methods has been proposed to construct trees from genetic sequences. At one end of the spectrum lie the *parametric* models, such as the maximum-likelihood method. Many researchers believe that as more data become available, the mutation rates will be known with better accuracy and these models will be better justified. At the other end of the spectrum of tree building methods lie the *non-parametric* approaches, such as the parsimony. The distance-based methods lie in between these two extremes. In this approach, the mutation model is parametric (with very few parameters considered) and the tree-building procedure is non-parametric. Distance-based methods are very popular, because the problem of computing the best tree for two other mentioned methods (maximum-likelihood and parsimony) is computationally difficult.

We have decided to enrich the parameter-space of distance-based methods by applying the new approach to the protein-sequence alignment, which takes into account the context-dependence of the amino acid substitution pattern. Several trees are reconstructed for protein families based on contextual similarity data. This set of trees is compared with the one obtained by standard methods. Our main goal was to verify the following conjecture:

> *The contextual model should yield a more consistent set of trees.*

A number of authors (e.g., [5,8]) have proposed different methods to investigate the consistency of the set of trees. Most of them aim at the construction of one tree (so-called *supertree*), which represents the set of source trees. Because existing supertree methods suffer from serious limitations (see, e.g., [9]), we have decided to work with several different gene trees.

In order to estimate the consistency of a set of trees, we have tested some mathematical properties, like pairwise distances between trees or common homeomorphic subgraphs. The results obtained are analysed for both contextual and non-contextual trees. All computational experiments, justified in some cases by a theoretical analysis, show more consistent results for contextual trees, which is in agreement with our conjecture.

This paper is organised as follows. At the beginning we present briefly the main ideas behind the contextual alignment model, then we describe the methods used in our analysis. The results of several computational experiments are analysed in Section 3 and followed by conclusions and further research.

1.1. The model of contextual sequences alignment

It is well known that the role an amino acid plays at a site in a protein depends on its environment. The evidence of this context-dependency contrasts with widely-used sequence comparison models, which assume the independence of the evolution for different sites. Recently, some research was done in the field of non-simplified models of DNA sequence evolution [10,11]. The authors consider a probabilistic model, in which a molecular sequence undergoes random changes due to substitutions, whose probability is context-dependent. This leads to a Markov chain model of quite complicated structure.

The contextual alignment model defined in [12] can be viewed as an algorithmic counterpart of these works, which is also suitable to analyse the protein sequences. It extends the classical alignment model, with the intention to bring it a step closer to the biological reality without sacrificing its algorithmic properties.

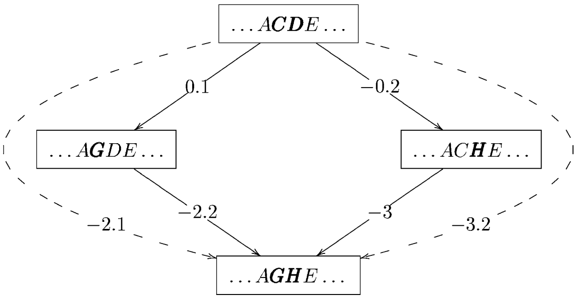
The set of operations is the same as that of the classical non-contextual alignment model, but the score function of a substitution changes. In our model, the score of a substitution *depends on the surrounding letters in the sequence*, too. The score for insertions and deletions is inherited from the classical model. As it is easy to see, in the contextual-alignment model, the score of a set of operations depends on the chronology of operations. On the other hand, the operations performed at distant fragments of the sequence are independent, in the sense that neither of them changes the context of the other. Particularly, it is sufficient to have two identical and adjacent columns in the alignment; they constitute a 'wall' separating two independent regions. Operations in independent regions can be performed in any order. Therefore, there are typically many orders that give the maximal score. Thus, the algorithms find not only an optimal set of operations, but also reconstruct a precise characterisation of the set of all possible orders (we call them admissible orders), in which the operations may be performed to yield the maximal score. More detailed study of the structure of optimal alignments and the description of efficient algorithms constructing them are included in [12,13].

1.2. Contextual substitutions tables

The contextual alignment algorithm assumes, as an important part of its input data, a contextual scoring table, providing the score for every possible substitution in every possible context. In [14], the procedure for constructing the family of contextual matrices has been proposed. It is based on the methodology introduced by [15]. The entries in the matrices are log-odds of the observed and expected mutation rates between given pairs of amino acids in a given context. For readers interested in this topic, the matrices parameterised by different clustering constants can be found at: http://www.mimuw.edu.pl/~aniag/ALIGN/TABLES.

1.3. Example

Consider the following example, that explains how the relative order of two substitutions applied to the same sequence affects the score, if the contextual substitution table from [14] is used. On the left path, the substitution $C \mapsto G$ is followed by the substitution $D \mapsto H$. In the second scenario, the order of these two substitutions is inverted: $D \mapsto H$ is followed by $C \mapsto G$. The summarized score for the left path is -2.1, while for the right path it is -3.2. This difference is caused by different contexts for the considered substitutions (e.g., the substitution $C \mapsto G$ is performed in the context A–D on the left path, while for the right path the context for this substitution is A–H, i.e. the left context is changed from D to H).



2. Methods

2.1. The dataset

We decided to use the database of Clusters of Orthologous Groups of proteins [16,17], COG in short. It consists currently of 3307 COGs, including 74 059 proteins from 43 genomes of bacteria, archaea and the yeast *Saccharomyces cerevisiae*. COG database represents an attempt to a phylogenetic classification of the proteins encoded in complete genomes. Each COG includes proteins that are thought to be orthologous, i.e. connected through vertical evolutionary descent.

Two groups of COGs are considered. The first one consists of 12 gene families of different tRNA synthetases. They are functionally related in contrast to the second group investigated here which is functionally more diverse (8 COGs). It includes DNA polymerases as well as ribosomal proteins and CDP-diglyceride synthetases.

These two sets are selected from the list of 85 COGs, in which all organisms are represented (i.e. each COG contains at least one protein from each genome). Our dataset is listed in Table 1.

Table 1
Two groups of COGs considered

Group I		Group II	
No.	COG name	No.	COG name
0013	Alanyl-tRNA synthetase	0013	Alanyl-tRNA synthetase
0442	Prolyl-tRNA synthetase	0185	Ribosomal protein S19
0016	Phenylalanyl-tRNA synthetase α	0201	Preprotein translocase subunit
0072	Phenylalanyl-tRNA synthetase β	0202	DNA-directed RNA polymerase
0162	Tyrosyl-tRNA synthetase	0361	Translation initiation factor
0018	Arginyl-tRNA synthetase	0575	CDP-diglyceride synthetase
0124	Histidyl-tRNA synthetase	0592	DNA polymerase IIIβ subunit
0143	Methionyl-tRNA synthetase	0636	FOF1-type ATP synthetase
0495	Leucyl-tRNA synthetase		
0525	Valyl-tRNA synthetase		
0172	Seryl-tRNA		
0441	Threonyl-tRNA synthetase		

2.2. *Pairwise alignments inside COGs*

In the first phase of the experiment, the protein sequences from each COG have been pairwise locally aligned using the standard Smith–Waterman algorithm and the contextual alignment procedure from [12]. The statistical significance of the obtained alignments have been computed using the method from [18], which has been also adapted to the contextual setting.

2.3. *Hierarchical clustering*

Several methods to derive the pairwise evolutionary distance (sometimes called difference score) from alignment scores are proposed (see, e.g., [19]). Being aware of the drawbacks of all these approaches (see [20] for a detailed discussion) we decided to use two independent methods.

The first one is based on the notion of statistical significance considered for local pairwise alignments in [18]. Roughly speaking, for a given pair of sequences, this value corresponds to the probability that two random sequences of the same length and amino acids composition have local similarity score higher then our pair. The statistical significance for two sequences u and v with length n and m having local similarity score $s(u,v)=S$ is given by the formula:

$$\mathit{Significance}(S) \approx 1 - \exp(-\gamma m n p^{S})$$

where p and γ are two parameters, which have to be estimated with respect to assumed alignment model (i.e. the alignment algorithm, amino acids substitution matrix, gap penalties, amino acid composition of the dataset). In our computation, we assume a typical distribution of amino acids [21], which has been verified to be very close to the distribution of amino acids inside the COG database. This method gives us the evolutionary distances that are not directly related to the genetic divergence. Hence, we treat the trees built from these distance data as cladograms taking into account the topologies of branching but not the lengths of the branches.

A second transformation between scores and distances is proposed in [22]. Assuming that $s(u,w)$ is the local similarity score between the sequences u and w, then their distance is defined via:

$$\begin{aligned} d(u,v) &= d(v,u) \\ &= s(u,u) + s(v,v) - s(u,v) - s(v,u) \end{aligned}$$

In the non-contextual case, $s(u,v)$ is given by the Smith–Waterman algorithm and, in the contextual setting, $s(u,v)$ is computed by the contextual procedure from [12]. One may observe that the measure defined in such a way may fail to satisfy the triangle inequality. However, as it was noticed in [22], such failures occur with frequency below 10^{-7}, and hence presumably hardly affect our results.

Inside the 20 orthologous gene families considered in our analysis, majority contains only one gene per species. In a few cases the family contains more than one gene (paralogous genes) per taxon. Only one of the paralogs is retained, while the more distant one is excluded from our analysis (e.g., APE0809 is excluded, while APE0117 is retained in the COG0441).

In all cases, the paralogs coding in yeast for mitochondrial t-RNA synthetases have been eliminated. All trees are reconstructed by a Neighbour-Joining algorithm implemented in PHYLIP package [23].

3. Results

We have analysed the groups of trees constructed from contextual and non-contextual data. Our main goal has been to examine whether the use of contextual model of sequence alignment has an influence on the phylogenetic clustering.

Adopting the widely-accepted assumption that the vertical descents dominates horizontal gene transfer (see, e.g., [24] for a recent discussion) we can formulate the following conjecture:

> *Set of trees reconstructed in contextual model should be more consistent (i.e. the trees should share more common structure).*

There exist several methods to investigate the consistency of the set of trees. Most of them aim at the construction of one tree (so-called *supertree* or *consensus tree*), which captures all non-conflicting information contained in the set of trees.

Supertree and consensus tree methods suffer however from inherent mathematical limitations (see [9]). More precisely, one can prove the non-existence of the method that possesses simultaneously the desirable properties, like:

- the method is independent of the order of the input trees;
- the renaming of all the species in the input trees can be reversed by the appropriate renaming of all species in the output tree;
- if the set of input trees is compatible, the output tree displays all of them.

The existing methods to combine trees are rather heuristic. The widely-used method is to re-code trees by characters and apply some standard tree reconstruction algorithm like maximum parsimony or Neighbour-Joining [2,5]. The verification of already constructed supertree is often based on some 'biological feeling' (especially in the case of bacterial phylogeny, when no different molecular data are available).

Having all mentioned limitation in mind, we have decided to examine also the whole family of species trees, not only a single super- or consensus tree.

To estimate the consistency of the set of trees, we have calculated all pairwise distances between trees. These are compared for both contextual and non-contextual trees. All computational experiments (parameterised by different tree metrics) yield better results for contextual trees than for non-contextual ones, in agreement with our conjecture.

The single tree derived from the set of trees has been also considered in two settings: Adams consensus and maximum agreement subtree. The outcomes of experiments are supported by probabilistic analysis, which justifies the significance of the superiority of the contextual approach.

In all the computational experiments mentioned above, we do not test any biological hypothesis, but only some mathematical properties (like common topologies or pairwise distances) for the set of trees. In contrast to this approach, we start the presentation of our results with a short example, which deals with the evolution of proteobacteria.

3.1. A biological example

In order to illustrate the rationale used to compare various types of alignments and the derived phenograms, an example of application to bacterial phylogeny may be of interest. We have analysed a set of trees (phenograms) constructed for several COGs in the contextual and non-contextual setting. We have been interested in the evolution of two groups of bacteria: α proteobacteria (CauCr = *Caulobacter crescentus*, MesLo = *Mesorhizobium loti*, RicPr = *Rickettsia prowazekii*) and β, γ proteobacteria (HaeIn = *Haemophilus influenzae*, PasMu = *Pasteurella multocida*, Ecoli = *Escherichia coli* K12, VibCh = *Vibrio cholerae*, PseAe = *Pseudomonas aeruginosa*, XylFa = *Xylella fastidiosa*, NeiMa = *Neisseria meningitidis* MC58, NeiMb = *Neisseria meningitidis* Z2491) together with the Buchnera (Buchn) species. It is generally believed that both groups are well clustered and most importantly Buchnera should be monophyletic with the β, γ proteobacteria family (see [25] for a recent discussion). To define a measure of evolutionary closeness, we consider the subtree rooted at the most recent common ancestor

Table 2
The evolution of proteobacteria

β, γ proteobacteria			
COG family	COG multiple alignment	non-contextual pairwise alignment	contextual pairwise alignment
0072: Phenylalanyl-tRNA synthetase β	14	26	0
0016: Phenylalanyl-tRNA synthetase α	13	17	0
0592: DNA polymerase IIIβ subunit	2	11	0
0636: FOF1-type ATP synthetase	0	0	0
Σ	29	54	0
α proteobacteria			
0072: Phenylalanyl-tRNA synthetase β	0	33	0
0016: Phenylalanyl-tRNA synthetase α	0	0	0
0592: DNA polymerase IIIβ subunit	9	23	25
0636: FOF1-type ATP synthetase	24	0	0
Σ	33	56	25

(MRCA) of the considered family (MRCA subtree). The MRCA subtree contains as leaves all members of our family and, in the ideal case, nothing more. Now, to distinguish between the quality of trees that describe the evolution of proteobacteria, we count the number of leaves that have to be pruned from MRCA subtree, because they do not belong to the considered family. The smaller is the number of leaves to be pruned, the better is the fit between hypothesis and results.

Table 2 summarizes our results. The entries correspond to the numbers of leaves that have to be pruned. In the first column, the number of COG families is given, then in the consecutive columns: the results obtained for the non-contextual tree based on multiple alignment (as presented at COG web pages), the results for tree based on non-contextual pairwise alignment data and the results for tree based on contextual pairwise alignment data. In all protein families, but one, the contextual data give monophyletic results (0 leaves pruned), while in non-contextual and and in multiple alignment, the majority (5/8) of families is inconsistent with the monophyletic origin of proteobacteria. In conclusion, the evolution of these protein families, as judged by the contextual approach, is more consistent with the rRNA phylogenetic tree [25].

3.2. Pairwise distances inside a set of trees

Several distance models for evolutionary trees have been proposed in the literature (see, e.g., [4]). From the computational point of view they fall into two categories: those model, in which the distance between trees can by computed efficiently (i.e. in the polynomial time) and the second group of models for which the approximation approach is necessary (because computing the distance in such a model is NP-hard).

For our analysis, we have chosen several methods to measure the degree of dissimilarity for a set of trees. These methods are:

- the *partition metric* treats trees as a set of clusters, it measures the amount of different clusters between trees. It is easy to compute, but its resolution is rather poor (two trees differing solely in the position of one taxon can be maximally different);
- the *Nearest-Neighbour Interchange distance* (NNI) is defined in terms of transforming one tree into another. It counts the minimum number of operations (called nearest neighbor interchanges) required for such a transformation. The main disadvantage of this approach is that no exact, efficient algorithm for NNI distance exists. In our experimental study we use several approximations proposed in the COMPONENT package [26];
- the *Maximum Agreement Subtree* (MAST) of two or more trees is an identical subtree of maximal size that can be obtained from all considered trees by pruning leaves with the same label. There exists an efficient algorithm finding MAST for two trees [7] and for more trees of bounded maximal degree. We can consider the distance between two

trees as the number of leaves removed to obtain MAST.

3.2.1. *The partition distance*

The Tables 3–6 present the outcome of computing all pairwise partition distances for all trees inside each considered set. The results for cladograms (i.e. trees built from significances of scores) are in agreement with the same experiment performed on trees with Linial's distance transformation (values in parentheses).

In both cases, the contextual models yield to more consistent set of trees. For example, the average over all pairwise distances are smaller. The differences are not very big, but, more importantly, contextual data give always better results.

Table 3

	8 COGs		12 t-RNA synthetases	
	context	non-context	context	non-context
min	24 (24)	22 (28)	24 (32)	34 (34)
max	46 (50)	46 (56)	56 (58)	62 (62)
ave.	34 (39)	35 (44)	44 (45)	50 (48)
st. dev.	6 (7)	6 (8)	6 (6)	6 (6)

3.2.2. *The NNI distance*

See Table 4.

Table 4

	8 COGs		12 t-RNA synthetases	
	context	non-context	context	non-context
min	17 (18)	18 (26)	15 (24)	28 (27)
max	46 (55)	49 (66)	82 (82)	83 (88)
ave.	30 (36)	32 (44)	49 (51)	55 (56)
st. dev.	8 (12)	7 (13)	11 (13)	11 (13)

3.2.3. *MAST distance*

See Table 5.

Table 5

	8 COGs		12 t-RNA synthetases	
	context	non-context	context	non-context
min	10 (13)	14 (14)	17 (14)	18 (14)
max	26 (26)	25 (28)	28 (29)	30 (30)
ave.	18 (20)	20 (22)	23 (22)	24 (24)
st. dev.	4 (4)	3 (4)	3 (4)	3 (3)

3.2.4. *Distances in a set of random trees*

The results from above can be compared with simulations, which have been done for a set of randomly generated trees of the same cardinality and with the same number of leaves. The outcomes in Table 6 are obtained as an average from 300 simulations. It can be seen that NNI and partition metrics discriminate better cognate trees from random ones, while the MAST metric is less informative. In Section 3.4, a more resolving application of MAST is described.

Table 6

	8 COGs			12 COGs		
	Partition	NNI	MAST	Partition	NNI	MAST
ave.	75.8	132	30.6	75.7	135.5	30.9
st. dev.	0.6	7.7	0.9	0.6	8.1	1.05

3.3. *Consensus methods*

To express the degree of agreement between cladograms, it may be sometimes useful to combine the phylogenetic information from two or more trees into one 'consensus' tree. Such a tree is a summary of how well the original trees agree. A number of different types of consensus trees has been proposed; each is calculated differently to answer different kinds of questions. Each summarizes common or average relationships among the original set of trees.

Unfortunately, consensus methods are of limited value: large disagreement among trees results in completely unresolved consensus tree. In our study, we decided to compare Adams consensus trees [1] calculated for both contextual and non-contextual groups of trees.

Adams consensus tree is characterized by the notion of *nesting*. For A and B being the subsets of the set of leaves of some phylogenetic tree, we say that A *nests in* B if the most recent common ancestor of A is a proper descendant of the most recent common ancestor of B. For a family of trees $\{T_1, T_2, \ldots, T_k\}$, sharing the same set of leaves, Adams consensus tree $\boldsymbol{T}_A$ is defined as a unique phylogenetic tree on the same set of leaves that satisfies the following:

(A1) let A and B be subsets of the set of leaves. If A nests in B in the tree T_i for all $i \in \{1, 2, \ldots, k\}$, then A nests in B in $\boldsymbol{T}_A$;

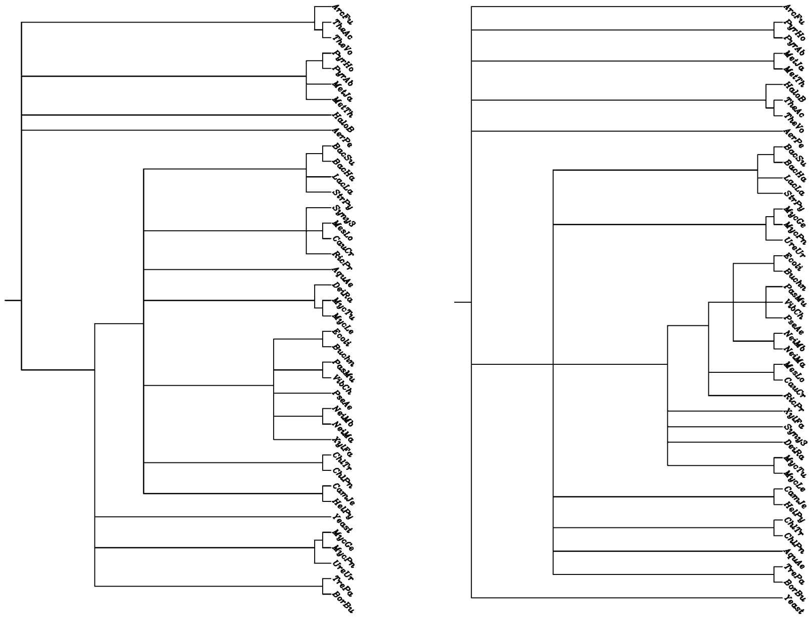


Fig. 1. Contextual (left) vs non-contextual (right) Adams consensus for 8 COGs cladograms.

(A2) let C and D be clusters of $\boldsymbol{T}_A$, such that C nests in D in $\boldsymbol{T}_A$, then C nests in D in each T_i for all $i \in \{1, 2, \ldots, k\}$.

Adams' consensus tree is particularly useful for identifying common tree structure, when one or more taxa have very different positions in the set of trees. In Fig. 1, we present as an example Adams' consensus trees for two sets of 8 COGs (contextual vs non-contextual) based on statistical significance [18]. Notice that the phylogeny of β, γ proteobacteria is more consistent in the contextual alignment tree than in the non-contextual one. Similar consensus trees are also constructed for groups of trees based on Linial's distances.

We propose a new approach to measure the quality of consensus trees. The idea of consensus tree is to capture as much common structural information of considered trees as possible. Hence better (more informative) trees should differ significantly from the 'bush' or null tree (star tree).

Consider the following characteristic of a rooted tree with n leaves: the integer vector $(i_1, i_2, \ldots, i_k)$ is called the *level density vector* if $\sum_{j=1}^{k} i_j = n$ and i_j is equal to the number of leaves on the jth level of the tree (the root level is counted as level 0). For example, for Adams' consensus trees of 8 COGs build from contextual and non-contextual data (Fig. 1), the level-density vectors are the following:

	$\rho_{\mathrm{C}} = (2, 4, 8, 13, 12)$	$\rho_{\mathrm{NC}} = (3, 6, 14, 7, 2, 7)$
level 1	● ●	● ● ●
2	● ● ● ●	● ● ● ● ● ●
3	● ● ● ● ● ● ● ●	● ● ● ● ● ● ● ● ● ● ● ● ● ●
4	● ● ● ● ● ● ● ● ● ● ● ● ●	● ● ● ● ● ● ●
5	● ● ● ● ● ● ● ● ● ● ● ●	● ●
6		● ● ● ● ● ● ●

The level-density vectors can be represented as a diagram similar to *Ferrers' diagram* [27], which is the pictorial representation of numerical partition of an integer n. In contrast to Ferrers' diagram, our vector corresponds to ordered partition of the set of leaves. A star tree (completely unresolved) with n leaves has the level-density vector $(n, 0, 0, \ldots)$. More resolved trees correspond to vectors with more non-empty levels, where those levels, which are close to root, have smaller cardinality.

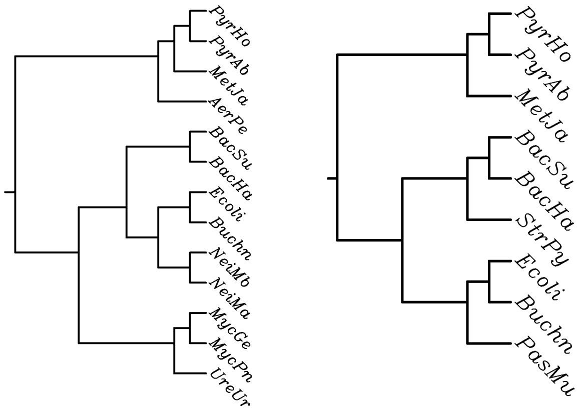


Fig. 2. MAST for contextual (left) and non-contextual (right) cladograms.

Consider the measure Ψ associated with the level-density vector $\rho = (i_1, i_2, \ldots, i_k)$:

$$\omega(\rho) = \sum_{i_j} i_j \cdot \frac{1}{2^j}$$

$$\Psi(\rho) = \frac{\omega_{\text{star}} - \omega(\rho)}{\omega_{\text{star}} - \omega_{\text{bin}}}$$

where $\omega_{\text{star}} = \max_\rho \omega(\rho) = \frac{n}{2}$ is the weighted level-density vector sum for a star tree with n leaves and $\omega_{\text{bin}} = 1$ states for this sum for the completely resolved binary tree. This measure satisfies several useful properties (for general discussion of tree information measures, see [28]):

- it is not sensitive to the tree balance, i.e. all completely resolved trees are equally informative ($\Psi = 1$), contrary to all measures whose calculation is based on summing the size of clusters;
- it takes into account the size and the height of the split;
- it is monotonous, i.e. while considering Adams consensus for several source trees, the measure of consensus tree cannot exceed the maximum over the source trees.

In the case of our example of 8 COGs trees, the Ψ measure for Adams' consensus tree is $\Psi(\rho_{\text{C}}) = 0.83$ for the contextual case and $\Psi(\rho_{\text{NC}}) = 0.72$ for the non-contextual data. We conclude that Ψ can be used as an efficiently computable alternative for the tree information measures proposed in [28].

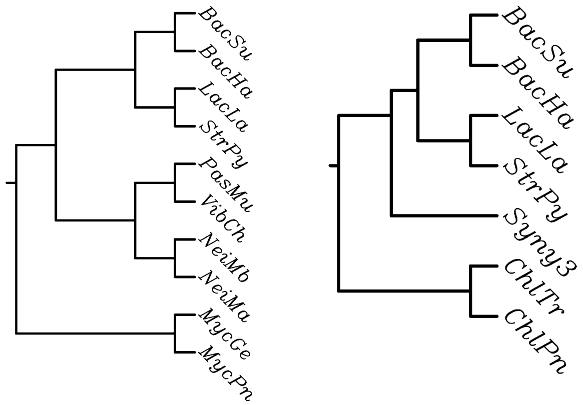


Fig. 3. MAST for contextual (left) and non-contextual (right) phenograms.

3.4. MAST for the set of trees

The algorithm described in [7], implemented in PAUP phylogeny software package [29], computes the MAST of a set of leaf-labelled trees. The comparison of these trees obtained for the contextual and the non-contextual model shows significant differences. Let us consider two pairs of trees being MASTs for the set of 8 cladograms (Fig. 2) and phenograms, i.e. trees based on Linial's distance (Fig. 3). The contextual MAST tree has more leaves than its non-contextual counterpart (13 leaves vs 9 leaves in the case of cladograms, and 10 vs 7 in the case of phenograms). The probabilistic analysis below shows that these differences are indeed significant, when compared with the expected size of MAST for the set of given numbers of trees. Moreover, the contextual approach results in consistent evolutionary classification of β-proteobacteria (Neisserias) and mycoplasmas, which are absent in the right-hand tree.

3.5. The significance of the size of MAST

In this section, we give an estimation of the expected size of MAST for a given number of random trees. We consider the uniform model, in which each labelled-rooted tree with n leaves is assigned an equal probability: $P_n(T) = \frac{1}{N(1,n)}$, where $N(1,n) = \cdot 3 \cdot 5 \cdot \cdots \cdot (2n-5) \cdot (2n-3)(2n-3)!!$ is the number of rooted trees with n labelled leaves. Denote by:

$$N(k,l) = (2k-1)(2k+1)\ldots(2l-3)$$

($l - k$ multipliers) for $1 \leqslant k \leqslant l \leqslant n$.

Table 7
The expected size of MAST for random trees

# of trees / # of leaves	3	4	5	6	7	8	9	10	11	12
10	3.91	3.06	3.01	2.52	2.16	2.05	2.01	2.00	2.00	2.00
20	5.04	4.01	3.09	3.01	3.00	2.52	2.17	2.05	2.02	2.01
30	5.68	4.12	3.54	3.04	3.00	3.00	2.62	2.21	2.07	2.02
40	6.18	4.57	4.00	3.12	3.01	3.00	3.00	2.50	2.17	2.05
50	6.95	5.02	4.02	3.30	3.02	3.00	3.00	2.99	2.33	2.11
100	8.48	6.01	4.62	4.00	3.34	3.02	3.00	3.00	3.00	2.91
150	10.02	6.26	5.01	4.05	4.00	3.12	3.01	3.00	3.00	3.00
200	10.94	7.02	5.10	4.20	4.00	3.38	3.03	3.00	3.00	3.00

The number of pairs of n-leaf trees having agreement subtree (common homomorphic subtree) of size at least k (i.e. with k leaves) is given by the formula:

$$L_2(k) = \binom{n}{k} N(2,k) \cdot N(k,n)^2$$

For a fixed subset of k leaves, there are $N(2,k)$ leaf-labelled trees. Any such a tree can by extended to the n-leaf tree in $N(k,n)$ ways. Notice that the above number is overestimated, as some pairs of trees are counted several times. This equation can be easily generalized for the set of trees of cardinality greater than 2:

$$L_r(k) = \binom{n}{k} N(2,k) \cdot N(k,n)^r$$

Now, the probability that the r-tuple of random trees has an agreement subtree with at least k leaves can be estimated as follows:

$$\begin{aligned} \mathsf{P}_r(k) &\leqslant \frac{L_r(k)}{N(2,n)^r} \\ &= \binom{n}{k} \frac{N(2,n) \cdot N(k,n)^{r-1}}{N(2,n)^r} \\ &= \binom{n}{k} \frac{N(k,n)^{r-1}}{N(2,n)^{r-1}} \\ &= \frac{\binom{n}{k}}{N(2,k)^{r-1}} \end{aligned}$$

The expected size of the maximum agreement subtree for the r-tuple of trees is calculated by the following formula:

$$\sum_k \mathsf{P}_r(k) \leqslant \sum_k \frac{\binom{n}{k}}{N(2,k)^{r-1}} \qquad (1)$$

This simple bound yields surprisingly tight estimation, especially for a bigger number of trees – the values calculated from Eq. (1) are summarized in Table 7.

The expected size of a MAST for two random trees was experimentally estimated in [30]. The authors also cite there some values obtained from the analytical estimations, which are not given. In contrast to them, analytical bound derived here work for several trees and are very close to the values obtained from simulations.

The analysis above readily confirms the significance of our results for contextual trees compared with non-contextual ones. The difference of 4 leaves in the case of MAST for 8 cladograms appears really large, when we look at the expected size of a MAST for 8 random trees which is less than 3.

4. Conclusions and further developments

It is clear that the experimental analysis described in this work is just a beginning and cannot be treated as a definitive answer. Various improvements and another experiments can be envisaged. Particularly, more COGs can be considered, different distances studied, supertree approaches proposed in [5] or [8] can be examined.

It would be also very interesting to check whether the Gap Alignment approach, described e.g. in [31], can be applied in the contextual setting. In this approach, phylogenies are reconstructed based only on the presence and evolution of gap-containing regions in the sequences. The analysis of gap-trees derived from contextual alignments seems to be an interesting extension of our work.

However, in view of the results presented in this work, we conclude that the concept of contextual approach, which *improves albeit modestly but nevertheless systematically the consistency of evolutionary changes in protein sequences, should be fruitful in phylogenetic studies*.

Some possible extensions of our analysis are:

4.1. Duplication distance

Widely studied approach to explain the discrepancies among differents gene trees is based on the notion of *reconciliation* [32]. In this formulation, one considers appropriate *tree-mapping*, which recovers all duplication events. More ambitious models take into account the phenomena of *horizontal transfer*. The problem is in general NP-hard for several gene trees, however promising approximate approaches are under study. It would be interesting to build and then to compare the reconciled *species trees* resulting from our families of gene trees.

4.2. Contextual multiple alignment

In [33], the relaxation of the contextual model was proposed, which gives the possibility to consider the multiple alignments. The effective progressive multiple alignment algorithm has been developed. Preliminary results obtained for the BaLIBASE benchmark alignments database are very promising. We plan to continue our analysis for families of trees build from contextual multiple alignment data (by parsimony and maximum-likelihood methods).

Acknowledgements

This work was partially supported by E.C. contract No. ICA1-CT-2000-70010 (EMBEU) of the IBB PAS Centre of Excellence in Molecular Biotechnology and by the Polish Research Council KBN grant 7 T11 F016 21. A. Gambin was also supported by the NATO science fellowship and the postdoctoral fellowship from the French CNRS within the framework of the 'Centre franco-polonais de biotechnologie des plantes'.

References

[1] E. Adams, N-trees as nestings: complexity, similarity, and consensus, J. Classif. 3 (1986) 299–317.

[2] B. Baum, Combining trees as a way of combining data sets for phylogenetic inference, Taxon 41 (1992) 3–10.

[3] L. Billera, S. Holmes, K. Vogtmann, Geometry of the space of phylogenetic trees, Adv. Appl. Math. 27 (4) (2001) 733–767.

[4] B. DasGupta, H. Xin, M. Li, J. Tromp, L. Wang, L. Zhang, Computing distances between evolutionary trees, in: D.-Z. Du, P. Pardalos (Eds.), Handbook of Combinatorial Optimization, Kluwer Academic Publishers, 1998, pp. 35–76.

[5] V. Daubin, M. Gouy, G. Perrière, Bacterial molecular phylogeny using supertree approach, Genome Inf. 12 (2001) 155–164.

[6] J. Doyle, Gene trees and species trees: molecular systematics as one-character taxonomy, Syst. Bot. 17 (1992) 144–163.

[7] M. Farach, M. Przytycka, M. Thorup, On the agreement of many trees, Inf. Process. Lett. 55 (1995) 297–301.

[8] C. Semple, M. Steel, A supertree method for rooted trees, Discrete Appl. Math. 105 (2000) 147–158.

[9] M. Steel, S. Böcker, A. Dress, Some simple but fundamental limits for supertree and consensus tree methods, Syst. Biol. 42 (2) (2000) 363–368.

[10] M. Schöniger, A. von Haeseler, A stochastic model for the evolution of autocorrelated DNA sequences, Mol. Phylogenet. Evol. 3 (1994) 240–247.

[11] J.L. Jensen, A.-M. Krabbe Pedersen, Probabilistic models of DNA sequence evolution with context dependent rates of substitution, Adv. Appl. Probab. 32 (2000) 499–517.

[12] A. Gambin, S. Lasota, R. Szklarczyk, J. Tiuryn, J. Tyszkiewicz, Contextual alignment of biological sequences, Proc. Eur. Conf. Comput. Biol. (ECCB special issue of Bioinformatics) 18 (2002) 116–127.

[13] A. Gambin, J. Tiuryn, J. Tyszkiewicz, Alignment with context-dependent scoring function, J. Comput. Biol. (in press).

[14] A. Gambin, J. Tyszkiewicz, Substitution matrices for contextual alignment, in: Proc. JOBIM, 2002, pp. 227–238.

[15] S. Henikoff, J. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl Acad. Sci. USA 89 (1992) 10915–10919.

[16] R. Tatusov, E. Koonin, D. Lipman, Genomic perspective on protein families, Science 278 (1997) 631–637.

[17] R. Tatusov, D. Natale, I. Garkavtsev, T. Tatusova, U. Shankavaram, B. Rao, B. Kiryutin, M. Galperin, N. Fedorova, E. Koonin, The COG database: new developments in phylogenetic classification of proteins from complete genomes, Nucleic Acids Res. 29 (1) (2001) 22–28.

[18] M. Vingron, M. Waterman, Statistical significance of local alignments with gaps, in: Bioinformatics: from Nucleid Acids and Proteins to Cell Metabolism, 1995, pp. 75–84.

[19] D.-F. Feng, R. Doolittle, Progressive alignment of amino acid sequences and construction of phylogenetic trees from them, Methods Enzymol. 266 (1996) 368–382.

[20] G. Gonnet, C. Korostensky, Optimal scoring matrices for estimating distances between aligned sequences, manuscript, 1999.

[21] P. McCaldon, P. Argos, Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences, Proteins: Struct. Funct. Genet. 4 (1988) 99–122.

[22] M. Linial, N. Linial, N. Tishby, G. Yona, Global self-organization of all known protein sequences reveals inherent biological signatures, J. Mol. Biol. 268 (1997) 539–556.

[23] J. Felsenstein, Phylip: Phylogeny Software Package, University of Washington, WA, USA.

[24] C. Kurland, Something for everyone: Horizontal gene transfer in evolution, EMBO Rep. 11 (2) (2000) 92–95.

[25] C. Brochier, H. Phylippe, Phylogeny: A non-hyperthermophilic ancestor for bacteria, Nature 417 (6886) (2002) 244–247.

[26] R. Page, Component, Tree Comparison Software, The Natural History Museum, London.

[27] G. Andrews, The Theory of Partitions, Cambridge University Press, 1998.

[28] J. Thorley, Cladistic information, leaf stability and supertree construction, Phd thesis, Department of Biological Sciences, Faculty of Science, University of Bristol, UK, 2001.

[29] D. Swofford, Paup, Phylogeny Software Package, Sinauer Associates.

[30] M. Steel, A. McKenzie, D. Bryant, The expected size of a mast under two stochastic speciation models, in: Posters from JOBIM, 2002.

[31] A. Sekowska, A. Danchin, J.-R. Risler, Phylogeny of related functions: the case of polyamine biosynthetic enzymes, Microbiology 146 (2000) 1815–1828.

[32] D. Sankoff, J. Nadeau (Eds.), Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families, Computational Biology Series, vol. 1, Kluwer Academic Publishers, 2001.

[33] A. Gambin, R. Otto, Contextual multiple sequence alignment, Posters from RECOMB, J. Biomed. Biotechnol. (in press).