

---

# Mapping gene family data onto evolutionary trees

**Boris Mirkin**

School of Computer Science and Information Systems, Birkbeck University of London  
Malet Street, London, WC1E 7HX, UK

---

*RÉSUMÉ. Des modèles pour placer les données des familles de gènes individuelles sur un arbre d'espèces évolutives sont présentés. Une famille de gènes est, typiquement, observée comme un ensemble de protéines homologues, qui appartient aux espèces différentes, qui veut dire que des données dans la forme de (a) une matrice de distance entre protéines, ou (b) un arbre de "gènes", dérivé de cette matrice, ou (c) le profil phylogénétique, un vecteur booléen qui indique la présence ou absence de la famille dans les feuilles du arbre. Des algorithmes seront présentés pour représenter tous ces types de données sur un arbre d'espèces évolutives donné pour reconstruire l'histoire évolutive du gène, avec quelques résultats expérimentaux*

*MOTS-CLÉS : Arbre évolutif, arbre de gènes, reconstruire l'histoire des gènes, parcimonie maximale.*

---

## 1 Introduction

The rooted tree with leaves labelled by taxa is a natural form of the biological taxonomy related, from the Darwin's times, to evolution. With the development of molecular biology and genomics, evolutionary trees, or phylogenies, became a major instrument for aggregated presentation and visualisation of interrelation among species. Methods for building evolutionary trees based on the premise: the more similar proteins the more recently they diverged, became an indispensable tool of a molecular biologist or bioinformatician (see, for instance [NEI 00], [FEL 01]).

To further advance into the understanding of natural phenomena and fuller exploit the aggregating capabilities of evolutionary trees, methods for mapping other data onto the trees should be advanced as well. This, however, is a twilight area at which only occasional efforts have been made so far.

In this paper some models for biologically meaningful mapping of data of individual gene families onto an evolutionary species tree are considered. A gene family, typically, is observed as a set of homologous proteins belonging to different species. The data of it can be presented either with a between-protein distance matrix, or a "gene" tree derived from the matrix, or just the phyletic profile, that is, a Boolean vector indicating presence/absence of the family in species under consideration. Theoretical and computational models will be presented for mapping of these data types to a pre-specified evolutionary tree so that the evolutionary history of individual genes can be reconstructed. Results found jointly with E. Koonin (NCBI NIH USA) and collaborators will be presented [MIR03], [MIR04].

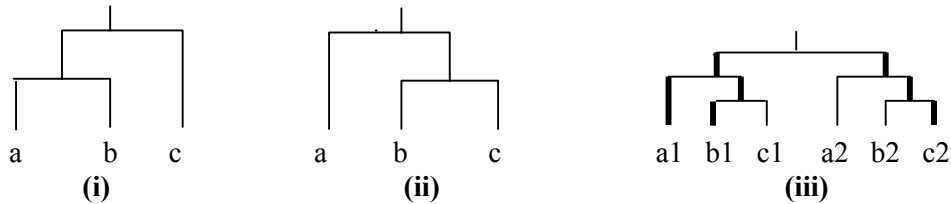
In our view, the approach may be more generally applicable to situations at which a general classification tree is present along with some partial data on the same entities, so that the partial data

could be meaningfully mapped to the tree. Perhaps such is the situation in which a number of texts is to be interpreted by mapping them onto a semantic structure tree (ontology).

## 2 Mapping individual gene trees

### 2.1 Duplication and reconciled tree

Duplication of genes is considered a major mechanisms of evolution: after a gene is duplicated on a chromosome, its copies may acquire different mutations so that eventually the copies may come to bear different functions [NEI 00]. Duplications of genetic material with follow-up losses may be used to explain the empirical fact that an evolutionary tree built on similarities between proteins from a gene family may differ from a species tree built using many sources of information.



**Figure 1. A gene tree (i) versus species tree (ii): the reconciled tree (iii) explains the difference by the duplication of the species tree with losses of lineages leading to c1, a2 and b2.**

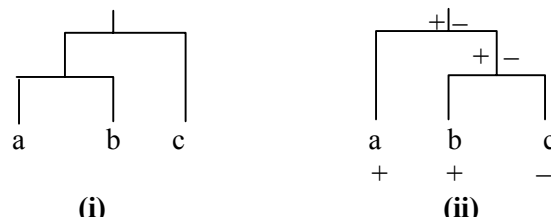
For example, gene tree (i) on Figure 1 may reflect similarities between haemoglobin proteins in horse (a), chimpanzee (b) and human (c) if those in horse and chimp belong to the haemoglobin alpha-lineage and that in human to the beta-lineage. Figure 1(iii) shows the "reconciled" correct evolutionary tree that would emerge if all alpha- and beta-haemoglobin proteins in the species have been collected. According to this tree, a duplication of the gene of haemoglobin occurred before the species' common ancestor appeared so that the "wrong" shape of gene tree (i) is an artefact caused by the data available at drawing it, which is expressed by solid lines in Figure 1(iii).

Here, as well as further on, all trees considered are binary rooted trees with leaves labelled by corresponding species.

The strategy of copying those subtrees of the species tree at which the gene tree differs from the species tree has been promoted in cladistics for building the so-called reconciled species tree [PAG 94, 97]. It allows to draw intuitively appealing pictures and, also, easily accommodates the presence of the so-called paralogs, protein products of duplicate copies of a gene. However, the reconciled tree is difficult to handle both theoretically and computationally in the situations at which the number of inconsistencies grows large as inevitably happens when treating a number of different gene trees.

### 2.2 Annotating model of duplication

Another formalisation of the concept of duplication was proposed in [MIR 95] according to the rules delineated in Figure 2.



**Figure 2. Species in the left and right subtrees of gene tree (i) annotated by symbols + and – in the species tree (ii), with the follow-up raising the symbols along the tree.**

A duplication is manifested in the difference between the contents of a divergence in the gene tree and species tree. If this is the case, the species tree can be annotated by differently labelling, on the species tree, the leaves of each of the two diverged subtrees of the gene tree. Then these labels, + and – on Figure 2, are extended bottom-up by annotating each internal node with the labels of its leaf contents

to represent the history of the corresponding duplication at the species tree. A loss event then is declared if a node has only one of the labels whereas its parent has both. This concept is less visually intuitive than that of the reconciled tree, but it allows to map different gene trees onto the same species tree without changing it and, moreover, can be analysed mathematically. In particular, let us refer to a pair of nodes  $(g,s)$ ,  $g$  being of the gene tree and  $s$  of the species tree, as a crossing if the leaf contents of  $s$  overlaps that of each of the children of  $g$ . Such is cluster  $s$  comprising leaves  $b$  and  $c$  at tree (ii) with regard to  $g$  being the root of tree (i) on Figure 2. If  $g$ -contents is part of  $s$ -contents, then crossing  $(g,s)$  corresponds to a duplication event, otherwise, pair  $(g,s)$  will be referred to as incompatible. It has been proven that the total number of losses at mapping a gene tree to the species tree is equal to the number of incompatible pairs plus twice the number of duplications. This implies that the total number of losses is at least three times greater than the number of duplications [MIR 95], [EUL 98].

### 2.3 Lca mapping as a computationally effective approach

Every node  $g$  of a gene tree  $G$  can be uniquely mapped to that node  $s(g)$  of a species tree  $S$  whose leaf contents is the minimal among those containing the leaf contents of  $g$ . Thus,  $s(g)$  is the last common ancestor (lca) in  $S$  to all species descending from  $g$  in  $G$ . The lca mapping is not computationally intensive: it can be done in a linear time over the size of the leaf set.

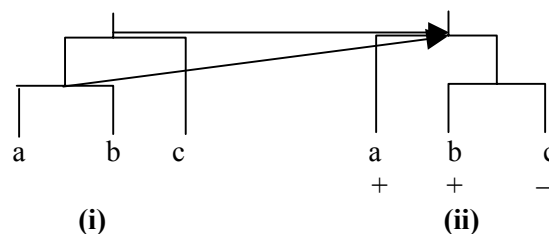


Figure 3. Lca mapping of internal nodes of gene tree (i) onto species tree (ii).

The lca mapping  $M: G \rightarrow S$  maps a pair node-parent  $(g, pg)$  of gene tree  $G$  to a pair  $(M(g), M(pg))$  of nodes in  $S$  in such a way that either: (a)  $M(pg)$  is the parent of  $M(g)$  in  $S$  (consistency), or (b)  $M(pg)=M(g)$  (contraction); a contraction is one-sided if  $M(g') \neq M(pg)$  where  $g'$  is sibling of  $g$  in  $G$ , or (c) there are intermediate nodes between  $M(pg)$  and  $M(g)$  in  $S$  (retraction). The inconsistencies of the lca mapping, contraction (b) and retraction (c), appear to be closely related to the duplications and losses above: the contractions correspond to duplications and intermediate nodes in the retractions correspond to those duplications for which the collateral children are losses [EUL 98]. The total number of losses in annotating  $S$  by  $G$ -duplication based events is the number of all the intermediate nodes plus the number of one-sided contractions under the lca mapping  $M$  [EUL 98]. The latter, lca mapping based measure, was used in a slightly different form by [GUI 96]; actually, the development of the annotating duplication model in [MIR 95] was motivated by the need to give a substantiation to the heuristic measure used in [GUI 96]. The fact that the lca mapping based measure is equal to the number of losses under the model of annotating duplication is proven in [ZHA 97], [EUL 98].

### 2.4 Consensus trees

Given a number of individual gene trees, the total number of losses and duplications can be used as the scoring function to derive a consensus tree with local search heuristics. This strategy was pursued, starting from the pioneering work [GUD 79], in [GUI 96], [PAG 97], [ARV 03] and others. Arguably, none of the attempts has been quite successful. This probably can be attributed to the combinatorial complexity of the problem, potentially biased representation of gene families, and missing of other important mechanisms of the evolution such as the horizontal transfer from the model. An attempt of incorporating the horizontal transfer to the setting is described in [ADD 03].

### 3 Parsimoniously mapping an individual gene set to the tree

#### 3.1 The problem

Given an evolutionary tree over species, the phyletic profile of a gene family is specified by the subset of tree leaves labelled by the extant species at which the family is present. An evolutionary scenario leading to the observed phyletic profile may involve the evolutionary events of *emergence*, *inheritance*, *horizontal transfer* and *loss*. No duplication concept has been used so far in this context, since the phyletic profile does not take into account the number of homologous proteins within a species. We refer to both emergence and horizontal transfer of a gene as a *gain*. The total number of loss and gain events in a scenario shows the extent of incompatibility between the evolutionary histories of the given gene according to this particular scenario and that implied by the topology of the species tree. Among all possible scenarios, we select those that are most parsimonious, i.e. require the minimum number of events to explain the observed phyletic profile. The two types of events, loss and gain, are likely to require different weighting in order to construct realistic evolutionary scenarios, which can be achieved by introducing the gain penalty  $gp$  that can differ from 1. Then a parsimonious scenario should minimize the total score; thus reaching the *minimum inconsistency* of the gene family.

#### 3.2 Algorithm PARS and its properties

This approach has been pursued in [SNE 02], [MIR 03], and [KUN 03] with different computational schemes, of which only [MIR 03] gives a genuine optimisation algorithm. This algorithm builds a parsimonious scenario for each parent node using parsimonious scenarios for its children. At each node of the tree, sets of loss and gain events are maintained under both the assumption that the gene has been inherited at the node and the assumption that it has not been inherited. A loss can occur only under the former assumption and a gain under the latter.

Consider the parent-children triple as shown in Figure 4, with each node assigned sets of loss and gain events under the above two inheritance assumptions. Let us denote the total number of events under the inheritance and non-inheritance assumptions by  $e_i$  and  $e_n$ , respectively, where gains are weighted by the gain penalty  $gp$ . An evolutionary scenario at a given node is defined by a pair of sets  $(G, L)$ , representing the gains and losses in the subtree rooted at the node. We use  $(G_i, L_i)$  and  $(G_n, L_n)$  to denote scenarios under the inheritance and non-inheritance assumptions, respectively. In a parsimonious scenario, the parental inconsistency score can be derived from those of its children as  $e_i = \min(e_{n1} + e_{n2} + 1, e_{i1} + e_{i2})$  or  $e_n = \min(e_{i1} + e_{i2} + g, e_{n1} + e_{n2})$ , under the inheritance or non-inheritance assumption, respectively. These lead to a recursive algorithm PARS for building parsimonious scenarios described in detail in [MIR 03]. At a leaf node the four sets  $G_i, L_i, G_n$  and  $L_n$  are empty, except that  $G_n = \{a\}$  if gene  $a$  is present in the given leaf or  $L_i = \{a\}$  if  $a$  is not present. The algorithm then computes parsimonious scenarios for parental nodes according to the topology of the tree using the rules given above, proceeding from the leaves to the root.

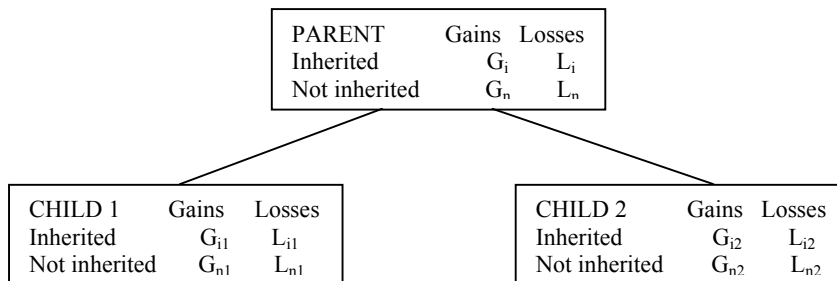


Figure 4. Patterns of events in a parent-children triple according to a parsimonious scenario.

The recursive structure of the algorithm PARS enables one to derive all parsimonious scenarios compatible with the phyletic profile of a gene. In some cases their number can be rather high. To have a unique outcome, a secondary criterion of minimising the number of gains was utilised in [MIR 03]. Some other properties of the criterion [MIR 03]:

- (1) The total number of gain and loss events in a parsimonious scenario is a monotone function of the difference between gain penalty  $gp$  and 1, getting its minimum at  $gp=1$ .
- (2) No gene family can emerge in a node being ascendant to the last common ancestor to the set of organisms to which the family belong.
- (3) No gene can be lost at the children of a node at which it emerged.

The method can incorporate a maximum likelihood approach, and take into account co-functioning of genes.

### 3.3 Reconstructing LUCA

This method has been used at an evolutionary tree involving 26 extant species of bacteria, archaea and yeast for reconstructing the evolutionary histories of about 2700 gene families represented by the so-called clusters of orthologous groups (COGs) in [MIR 03] at different gain penalty values ranging from 0.1 to 10. The reconstructed contents of the root represents the last ultimate common ancestor (LUCA); it was used as an external criterion for selecting the gain penalty value. At  $gp=1$  the set of 572 genes comprising LUCA was recognised as best approximating an organism capable of survival and reproduction; the relatively small inconsistencies were caused by mutual inconsistencies between the method and data, e.g., the mitochondrial origin of some gene families. This led the authors to conclude that the events of horizontal transfer in the process of evolution have been as frequent as loss events, which contradicted to previously expressed opinions, as well as to conclusions of [SNE 02] and [KUN 03]; note however that these authors did not use any external criterion for selecting the gain penalty in [SNE 02] or equivalent constants in [KUN 03].

### 3.4 Relation to reconstruction of ancestral characters: principles of maximum likelihood and parsimony

Let us treat a gene family as a character that may be present or not in any node of the tree, so that the character's state change from 1 to 0 corresponds to a loss and from 0 to 1 to a gain. Then the problem of building of an evolutionary scenario becomes equivalent to the problem of reconstructing the character's ancestral states. Two popular approaches to this latter problem are the principle of Maximum Parsimony (MP) and the principle of Maximum Likelihood (ML) [NEI 00].

The principle of Maximum Parsimony is exactly that used above. A method of doubly running through the tree, differing from that in [MIR 03], has been proposed in [FIT 71] and [HAR 73]. The principle is implemented as part of the maximum parsimony principle for building evolutionary trees [FEL 01], [MAD 92]. It is simple and intuitive. Its shortcoming is that it frequently leads to situations with several equally parsimonious scenarios drawing very different histories of the gene evolution.

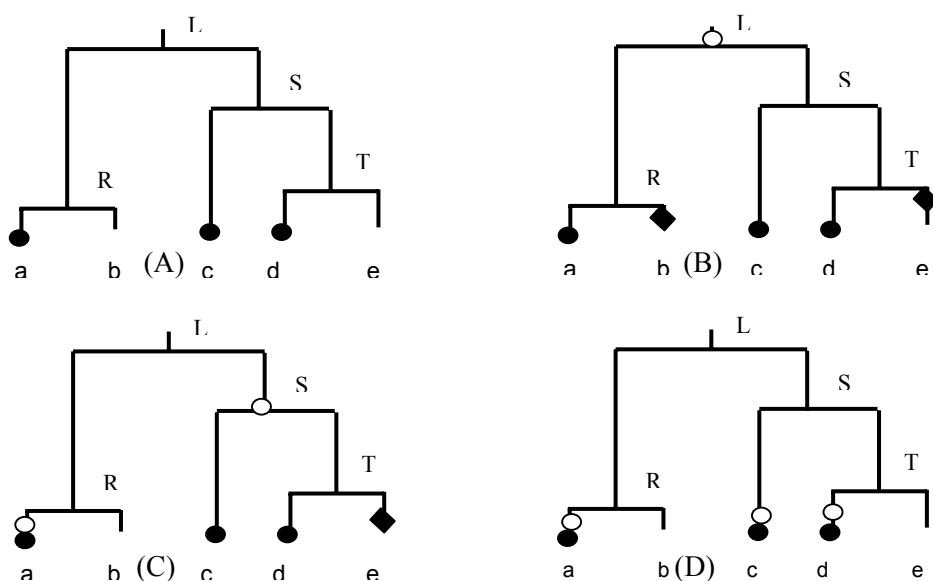


Figure 5. An evolutionary tree with a gene family's presence at a, c, and d denoted by a black circle (A). Equally parsimonious scenarios for the family's history are at (B), (C) and (D) with a white circle denoting a gain event and a black diamond denoting a loss.

Consider, for example, the case of a gene family that is present at three of the five extant species of Figure 5 (A). There are three equally parsimonious scenarios of its history so that an external criterion to choose one scenario from the set should be specified. A number of general selecting criteria were considered in [MAD 92], [MIR 03].

Another set of criteria emerges in the setting of the Maximum Likelihood (ML) approach. This approach involves two types of quantitative information, the evolutionary tree branch lengths and a probabilistic model of transformation of a distribution of presence/absence of the gene family along tree branches. Typically, a continuous-time Markov model is utilized involving constant rates  $q_{ij}$  of change of character states, that is, gene's presence or absence, from  $i$  to  $j$  where both  $i$  and  $j$  can be 1, for presence, or 0, for absence [YAN 95, 96].

In a typical situation such a model would lead to a unique most likely scenario for the gene history, which, at the first glance, supports the idea of superiority of the ML over MP. However, there are other issues that hinder the use of the Maximum Likelihood. First, the estimates of the change rates are solutions to a complex optimisation problem that can be solved only locally so that the solution much depends on the initial setting, thus biased towards a specific system of penalty weighting. Second, there is a clash between the maximum likelihood used for deriving the probabilistic model from data and the maximum likelihood used for deriving the gene histories from the model. The probabilistic model implies a probabilistic distribution over a number of possible evolutionary scenarios, of which only one, the most likely, is selected each time. This procedure, when applied to a multitude of gene families, obviously would lead to a biased empirical distribution of gene gains and losses.

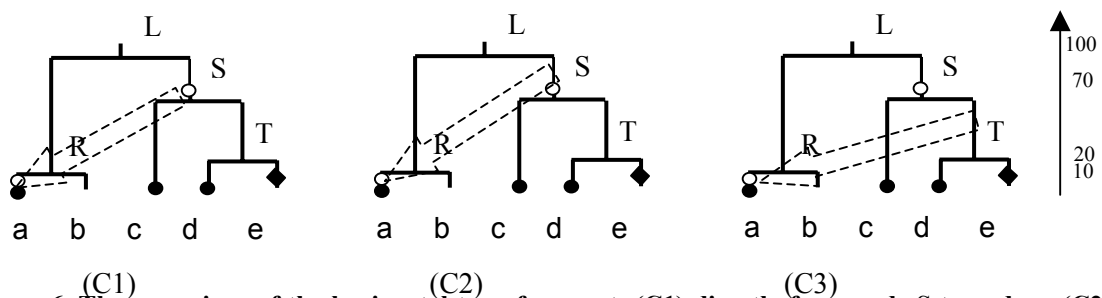
## 4 Within-family distances and principle of maximum correlation

### 4.1 Principle of maximum correlation

Deriving evolutionary histories should involve not the phyletic profile only but similarities between proteins constituting a gene family, the approach being currently developed jointly with E. Koonin, Y. Wolf and T. Fenner [MIR 04]. We consider that any gene family  $C$  embracing organisms of subset  $S$  of the set  $I$  of organisms under consideration is accompanied with matrix  $B_C$  of dissimilarity coefficients  $b_{jk}$  for all  $j, k \in S$ . Any evolutionary scenario  $h$  of gene  $C$  history should lead to a scenario-based distance matrix  $D_C(h)$  so that the quality of the scenario could be scored according to the correlation between  $D_C(h)$  and  $B_C$ . The higher the correlation, the better the history. This can be formulated as the principle of Maximum Correlation (MC). Obviously, the MC principle should supplement the MP and ML rather than substitute them.

### 4.2 Directed scenario and scenario-generated tree distance

To implement the MC principle, an evolutionary tree will be considered timed, that is, each node in it assigned with an estimate of time at which the corresponding ancestor species diverged in the process of evolution to give rise its "children" species. Given a set of gains, that is, nodes in the evolutionary tree, ordered over the times assigned to them, this can be further extended in what we call a directed scenario, as follows. The elder gain node is postulated to be the node at which the gene under consideration has emerged.



**Figure 6. Three versions of the horizontal transfer event: (C1) directly from node S to node a, (C2) from the middle of S's life span to the middle of a's life span, and (C3) from the middle of T's life span to the middle of a's life span.**

Each of the other gains,  $g$ , is assigned with its source,  $s$ , which must be a node belonging to a subtree generated by an older gain such that  $s$ 's life span either being earlier than that of  $g$  or overlapping with it. Figure 6 illustrates three possible directed scenarios for the case of a gene family comprising three extant species corresponding to the "static" scenario C from Figure 5. Each of these scenarios can be differently reflected in the matrix of distances between organisms in the family.

Indeed, if the horizontal transfer itself takes no time, the distance between  $a$  and  $c$ , which is 200 in Figure 5, becomes equal to just 70, the distance between  $S$  and  $c$ , according to C1, or, with mid-life-spans added,  $90=70+15+5$  according to C2, or  $100=70+25+5$  according to C3, with 25 being the distance from  $S$  to the middle point of its edge to  $T$ . Overall the within gene distance matrices will be

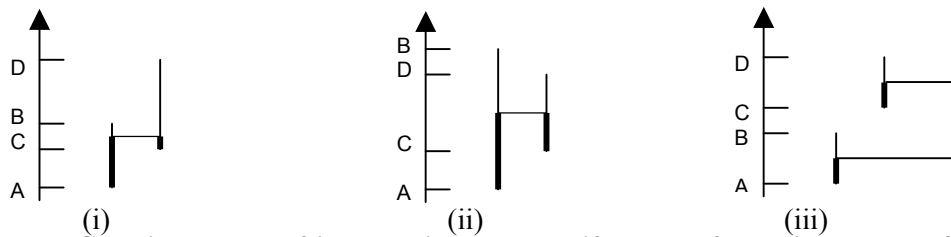
$$D = \begin{array}{c|c} c & d \\ \hline 200 & 200 \\ \hline | & 140 \\ \hline & c \end{array} \quad D1 = \begin{array}{c|c} c & d \\ \hline 70 & 70 \\ \hline | & 140 \\ \hline & c \end{array} \quad D2 = \begin{array}{c|c} c & d \\ \hline 90 & 90 \\ \hline | & 140 \\ \hline & c \end{array} \quad D3 = \begin{array}{c|c} c & d \\ \hline 100 & 50 \\ \hline | & 140 \\ \hline & c \end{array}$$

according to the original timed tree and scenarios C1, C2, C3, respectively.

This distance matrix can be proven to be uniquely defined by a directed scenario.

### 4.3 Synchronization of a horizontal transfer

The directed scenarios at Figure 6 are not entirely appropriate because the transfers between sources and targets in them are placed at different moments of time. To synchronize the target of a horizontal transfer and its source, one should take into account the pattern of interrelation between their life spans. Three generic patterns are shown on Figure 7: (i) and (ii), target's span overlaps that of the source, (iii) the source had died out before the target emerged. In case (iii) the transfer is to have been through a "relation lineage" which is absent from the tree as highlighted on the right-hand side of pattern (iii).



**Figure 7. Generic patterns of interrelation between life spans of a horizontal transfer target, AB, (vertical line on the left) and its source, CD (vertical line on the right), to generate the synchronising distance between them, AB at (i) or AD at (ii) and (iii).**

For the sake of simplicity, the midranges of the life span intervals are taken as the transfer time points.

With the synchronization, the within gene family distance matrices corresponding to scenarios C2 and C3 of Figure 6 will become  $D2'$  and  $D3'$  as follows:

$$D2' = \begin{array}{c|c} c & d \\ \hline 170 & 170 \\ \hline | & 140 \\ \hline & c \end{array} \quad D3' = \begin{array}{c|c} c & d \\ \hline 100 & 90 \\ \hline | & 140 \\ \hline & c \end{array}$$

These distances are relevant only at distances between extant organisms descending from the target and the source. Otherwise, the distances should take into account the transfer points only.

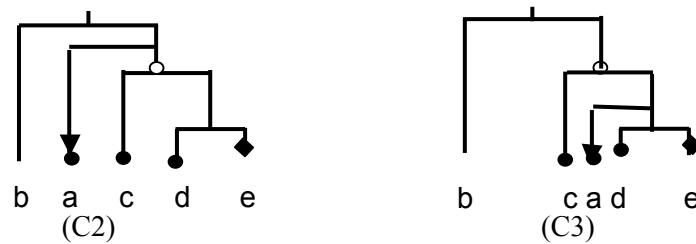
### 4.4 Virtual change of tree topology effected by a directed scenario

In many aspects, a directed scenario is equivalent to changing the evolutionary tree topology by joining each subtree rooted at a target to the edge above its source (as presented on Figure 8).

### 4.5 Experimental results

A newly developed algorithm MaCor implements the line of thought described above and involving procedures for (1) deriving directed-scenario-based tree distances between organisms within a family, (2) measuring correlation between protein-based within-family distance matrices and those scenario-based ones and for (3) greedy-wise building of the directed scenario maximally correlated with the protein based distance matrix. Our algorithms have been applied to the updated evolutionary tree built

by E. Koonin and Y. Wolf over 66 organisms and 4873 gene families (COGs) produced in NCBI NIH USA.



**Figure 8. Changed tree topologies corresponding to directed scenarios C2 and C3 on Figure 6.**

In particular, algorithm PARS produced a LUCA with 489 gene families at  $gp=1$ . The number decreased from the 572 found at the tree of 26 organisms (see section 3.3) because of more complex topology of the evolutionary tree at 66 organisms. This result was compared to that found with the algorithm MaCor applied to all parsimonious scenarios derived at  $gp$  varying through 1, 2 and 3. This way we have been able to select the gain penalty weight according to MC principle. Of 4873 gene families, the best correlation of the directed scenarios and protein-based distance matrices was achieved for 3975 of them at  $gp=1$ , 602 at  $gp=2$ , 295 at  $gp=3$  and 35 at greater gain penalties. Altogether, only 444 families appear to be in LUCA according to MaCor, of which only 325 are common to those found with algorithm PARS (note that the latter always selects the minimum number of gains thus forcing a family to belong to LUCA anytime when it is possible.) These results show that further improvements of MaCor should be made in all of its three major procedures above.

## 5 Conclusion

The frameworks presented can be seen as different approaches to mapping within gene family similarity data onto an evolutionary tree. One approach first produces a gene tree, then maps it to the species tree, whereas the other approach first takes the gene phyletic profile, maps it to the species tree and then fine tunes it into a directed scenario based on the similarity data. The first approach is natural for modelling duplications, the second for horizontal transfers, though there are ways for extending either to both types of events.

The difference between the approaches increases when looking at their scoring mechanisms: at the mapping of gene trees, differences at deeper parts of the species tree may formally translate into an excessive number of duplication and loss events. Consider, for example, a group of species descending, in the species tree, from a node  $p$  whose another child is just a leaf  $s$ . If the group and  $s$  have the same last common ancestor in the gene tree, but  $s$  now is not an outsider but sibling of another leaf in the group, then this difference will be translated into as many duplications as there are nodes between  $s$  and  $p$  in the gene tree, because each pair child-parent on the path will be contracted under the lca mapping. Intuitively, one may think that the difference between the species and gene trees should be attributed to just one event occurred at the level  $p$ , not many, which points to an inadequacy of the analysed concept of duplication. In terms of phyletic patterns such a difference would be attributed to a horizontal transfer event. It is quite clear, at least in principle, how to separate duplications from horizontal transfers in the phyletic profile approach. This is not so in terms of tree mapping.

On the data analysis level, the phyletic pattern concepts seems more flexible because it leads to easier mathematical problems and is better suitable to the biological intuition.

## 6 References

- [ADD 03] ADDARIO-BERRY, L., HALLETT, M.T., LAGERGREN, J., " Towards identifying lateral gene transfer events ", *Pacific Symposium on Biocomputing*, 2003, p. 279-290.
- [ARV 03] ARVESTAD, L., BERGLUND, A.-C., LAGERGREN, J., SENNOBLAD, B., " Bayesian gene/species tree reconciliation and orthology analysis using MCMC ", *Bioinformatics*, vol. 19, 2003, p. i7-i15.



- [EUL 98] EULENSTEIN, O., MIRKIN, B., VINGRON, M., " Duplication-based measures of difference between gene and species trees ", *Journal of Comp.Biology*, vol. 5, num. 1, 1998, p. 135-148.
- [FEL 01] FELSENSTEIN, J. *PHYLIP 3.6: Phylogeny inference package*, <http://evolution.genetics.washington.edu/phylip/>.
- [FIT 71] FITCH, W.M., " Towards defining the course of evolution: Minimum change for a specific tree topology ", *Systematic Zoology*, v. 20, 1971, p. 406-416.
- [GOO 79] GOODMAN, M., CZELUSNIAK, J., MOORE, G.W., ROMERO-HERRERA, A.E., MATSUDA, G., " Fitting the gene lineage into its species lineage. A parsimony strategy illustrated by cladograms constructed from globin sequences ", *Systematic Zoology*, vol. 28, 1979, p. 132-163.
- [GUI 96] GUIGÓ, R., MUCHNIK, I., SMITH, T.F., " Reconstruction of ancient molecular phylogeny", *Molecular Phylogenetics and Evolution*, vol. 6. 1996, p. 189-213.
- [HAR 73] HARTIGAN, J.A.; " Minimum evolution fits to a given tree ", *Biometrics*, vol. 19, 1973, p. 53-65.
- [KUN 03] KUNIN V., OUZONIS, C.A., " GeneTRACE – reconstruction of gene content of ancestral species ", *Bioinformatics*, vol. 19, 2003, p. 1412-1416.
- [MAD 92] MADDISON, W.P., MADDISON, D.R.; *MacClade 3.0*. Sunderland, MA: Sinauer Associates, 1992.
- [MIR 03] MIRKIN, B., FENNER, T., GALPERIN, M., KOONIN, E., " Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes ", *BMC Evolutionary Biology* 2003, 3:2 ([www.biomedcentral.com/1471-2148/3/2/](http://www.biomedcentral.com/1471-2148/3/2/)).
- [MIR 95] MIRKIN, B., MUCHNIK, I., SMITH, T. "A biologically consistent model for comparing molecular phylogenies ", *Journal of Comp.Biology*, vol. 2, num. 4, 1995, p. 493-507.
- [MIR 04] MIRKIN, B., WOLF, Y., FENNER, T., KOONIN, E., " Modelling horizontal transfer events with directed evolutionary scenarios and the principle of maximum correlation ", 2004, *in progress*.
- [NEI 00] NEI, M., KUMAR, S, *Molecular Evolution and Phylogenetics*, Oxford Univ. Press, 2000.
- [PAG 94] PAGE, R.D.M., " Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas ", *Systematic Biology*, vol. 43, 1994, p. 58-77.
- [PAG 97] PAGE, R.D.M., CHARLESTON, M.A., " From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem ", *Molecular Phylogenetics and Evol.*, vol. 7, 1997, p. 231-240.
- [SNE 02] SNEL, B., BORK, P., HUIJINEN, M.A., " Genomes in flux: The evolution of archaeal and proteobacterial gene content ", *Genome Research*, vol. 2, 2002, p.17-25.
- [YAN 96] YANG, Z., " Phylogenetic analysis using parsimony and likelihood methods ", *Journal of Molecular Evolution*, vol. 42, 1996, p. 294-307.
- [YAN 95] YANG, Z., KUMAR, Z., NEI, M., " A new method of inference of ancestral nucleotide and amino acid sequences ", *Genetics*, vol. 141, 1995, p. 1641-1650.
- [ZHA 97] ZHANG, L., " On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies ", *Journal of Comp. Biology*, vol. 4, num. 1, 1997, p. 177-187.