Relationships Of Bacterial Metallothioneins: Phylogenetic Tree Construction Method

Škutková H [1], Provazník I [1], Kizek R [2]
[1] Department of Biomedical Engineering, Brno University of Technology, [2] Department of Chemistry and Biochemistry, Mendel University in Brno
helena.skutkova@phd.feec.vutbr.cz

Phylogenetics is a central discipline in the modern life sciences aimed at describing sequences similarity through evolutionary relationships among organisms. Although it may seem that the graphical representation of phylogenetic connection between organism is a mastered problem, the results always depend on correct choice of method of computation, especially for diverse data as sequences of bacterial metallothioneins. We shown how our used methods and algorithms in a suitable form gives a good results. The coupling a simple Jukes Cantor model with approximation to Poisson distribution for amino acids and BioNJ algorithm witch is more effectively for alignment sequences, we obtained the resulting tree structure forms closed units, which corresponds to their affinity.

1 Introduction

Metallothioneins (MTs) are small cysteine-rich proteins that are involved in many diverse biological processes, e.g. MT transports metal ions in organism, or they bind and eliminate toxic metal ions. MT are often studied for their possible effectiveness on treatment tumour diseases. These multifunction proteins constitute a superfamily. They are found not only in eukaryotes, but also in prokaryotic microorganisms. The usually indicated parameters of MTs are length of sequences around 60-62 amino acids and just about 20 of them are cysteins. However, this specifications are relatively valid only for human MTs, but MTs from other organisms are very different. The goal of this study is evaluated the measure of sequences similarity of bacterial MTs by phylogenetic analysis.

Study of evolutionary relationships in whole MT superfamily is complicated. One of the possible approaches is to aim to a specific subfamily of bacterial MTs. Various methods lead to description of relationships usually depicted by a phylogenetic tree. Phylogenetic analysis [4] is an important subproblem of sequence analysis, in our case analysis of protein sequences.

2 Materials and methods

The protein sequences of bacterial MTs used in this study were retrieved from protein knowledgebase UniProtKB that is freely accessible via Internet. Selected bacterial MT sequences were aligned because the study group were homological (length of sequences varied) [4]. After global multiple alignment, biological distances between each pair of sequences were determined. Finally, the phylogenetic tree was constructed using the relationship among the distance data.

The measure of evolutionary distance

The used genetic distance was defined as a count of changes n that must be made in one sequence to turn into another sequence [2]. This value was converted into distance measurement called *p distance* [7] - average degree of changes per length N of aligned sequence given by $p = n/N$. *p distance* must be strictly proportional to time of evolution (t in

million year) for computation real evolution distance between two sequences. However, number of visible amino acids (AA) substitutions is not equal to true number of mutations because one of positions could be changed multiple times. This is the reason to use Poisson distribution to estimate a higher evolution distances (p>0.3), as shown in the Fig 1. *Poisson correction (PC)* [5] distance is given by

$$d = -\ln(1-p),$$

where d is total number of AA substitutions $\boldsymbol{r}$ during $\boldsymbol{t}$ years for the two sequences (d~2rt).

Results were normalized to use on amino acid sequences. A simple *Jukes-Cantor model* [4] was used, which assumed that each AA eventually has the same probability of change at each sequence position. That means the same frequency in protein sequence for all of 20 AAs. Thus, correction to the distance ${}_{JC}d$ between two sequences is given by

$${}_{JC}d = -B \cdot \ln\left[1-\frac{p}{B}\right],$$

where B = 19/20 for the assumption of equal AA representation. Distance measurement for each pair of sequences formed MxM distance matrix d_{ij}, where M is a number of sequences and i, j are taxa.

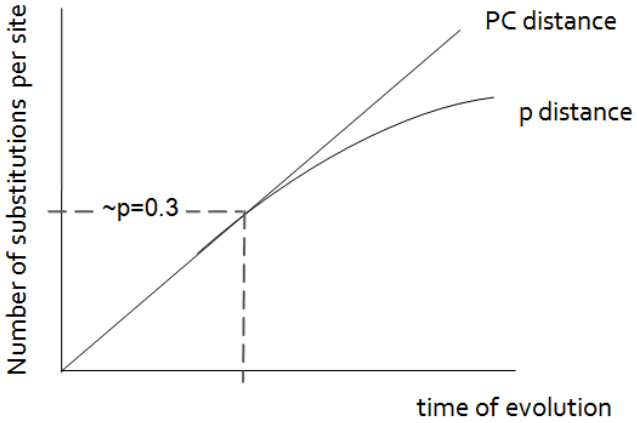

Fig 1. The relation among the p distance and the Poisson correction distance [5].

In the next step, *neighbour-joining* [6] *(BioNJ)* method for phylogenetic tree reconstruction is applied. This method is based on progressively adding the next most-alike sequence (or sequences) as additional branch to an existing tree using distances between the sequences.

BioNJ Algorithm

First, a pair of sequences which have minimum mutual distance in comparison to other sequences must be found [3]. Let Q_{xy} be the value of the criterion for the choice of the pair to be agglomerated, then minimum value of criterion Q_{xy} responding to neighbour taxa. For simplification, let x=1 and y=2, then criterion Q_{12} is given by

$$Q_{12} = (N-2) \cdot d_{12} - S_1 - S_2,$$

where $S_{1,2}$ are elements of the summing matrix defined by sums of each distance between taxa x (1 or 2) and all other taxa. The summing matrix is given by

$$S_x = \sum_{i=1}^{N} d_{xi}.$$

This pair creates a new node u, which represents new root for the cluster. Estimate of the branch lengths (1, u) is

$$d_{1u} = \frac{1}{2}(d_{12} + \frac{S_1 - S_2}{N-2}),$$

and d_{2u} is obtained similarly.

After estimating the length of a branch *u*-1 and *u*-2, BIONJ [3] reduces the distance matrix by deleting these taxes and by estimating the distances between the new node *u* and any other node *i*. The estimation is defined by

$$d_{ui} = \lambda d_{1i} + (1-\lambda)d_{2i} - \lambda d_{1u} - (1-\lambda)d_{2u},$$

where *d* are distances (branch-lengths) between appropriate nodes (or taxons) and adjusting value λ guarantees to find the correct tree with additive data. The value of λ is estimated with iteration method, which computes variances and covariances of evolutionary distance at each iteration and finds minimum of the sampling variance. The computation variances and covariances may be approximated with acceptable accuracy by variation matrix v_{ij}. At the start, the variance matrix is initialized as equal d_{ij} and reduction is applied at each step using

$$v_{ci} = \lambda v_{1i} + (1-\lambda)v_{2i} - \lambda(1-\lambda)v_{12},$$

where *c* is hypothetical centre of the cluster witch is farther from the root *u* at the start. The goal is to determine λ so that the distance between centre *c* and root *u* is decreased in each step. The parameter λ is given by

$$\lambda = \frac{1}{2} + \sum_{i=1}^{n} \frac{(v_{2i} - v_{1i})}{2(N-2)v_{12}}.$$

λ is adjusted at each step to value between 0 and 1.

This computing algorithm [3] is repeated until last three sequences (taxa) remain, because number of taxa decreases in each step *N*-1. For the remaining tree taxa, only branch length between taxa and the root *u* analogous to d_{1u} or d_{2u} is estimated. So this procedure gives the most correct phylogenetic tree.

3 Results

Figure 1 shows a phylogenetic tree of all selected bacterial MT sequences. 82 sequences from UniProtKB database compose MT superfamily of bacterial MTs (1081 data items accessible on 20 January 2010). Sequences which were duplicate (identical sequences) or evidently improper (sequences uploaded with an error) were filtered out from gene family. The selected sequences had various length. Therefore, global multiple sequence alignment using BLOSUM30 scoring matrix was applied. Conservation and variations in bacterial MTs sequences are shown in Fig 2. Only 25 representative sequences are shown because most of the other MTs are composed of a high number of AAs. It makes visualization difficult.

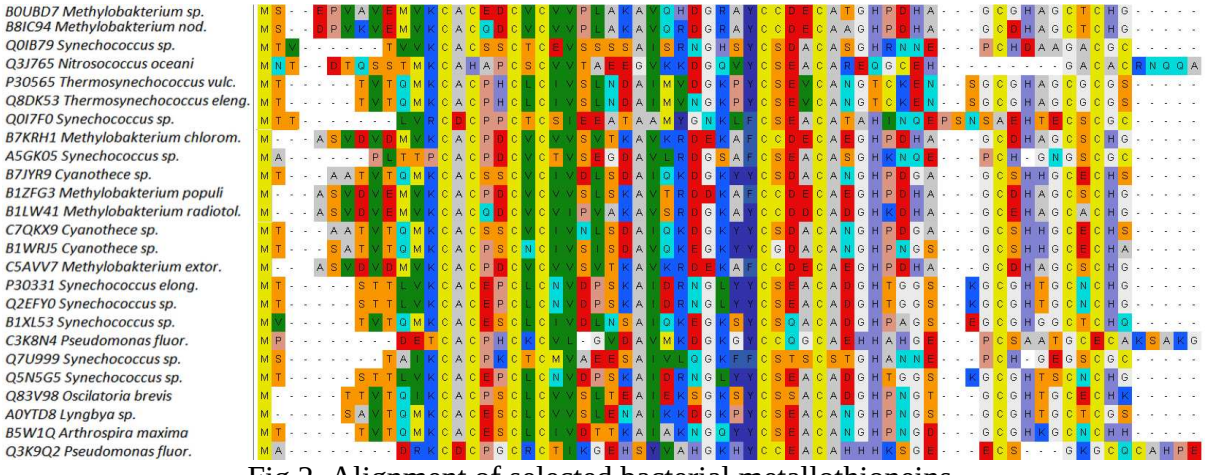

Fig 2. Alignment of selected bacterial metallothioneins.

Based on the results presented in Fig 3, the majority of bacterial MTs belongs to two large taxonomic categories: Cyanobacteria and Proteobacteria. Proteobacteria taxonomy is mainly represented in γ-proteobacteria, between them belong: pseudomonas, enterobacteria and helicobacteria. The largest subfamily in cyanobacteria is synechococus, which formed small isolated group in the tree, varying in their function. Delineation of particular groups in Fig 3 matches to taxonomic allocation in UniProt database.

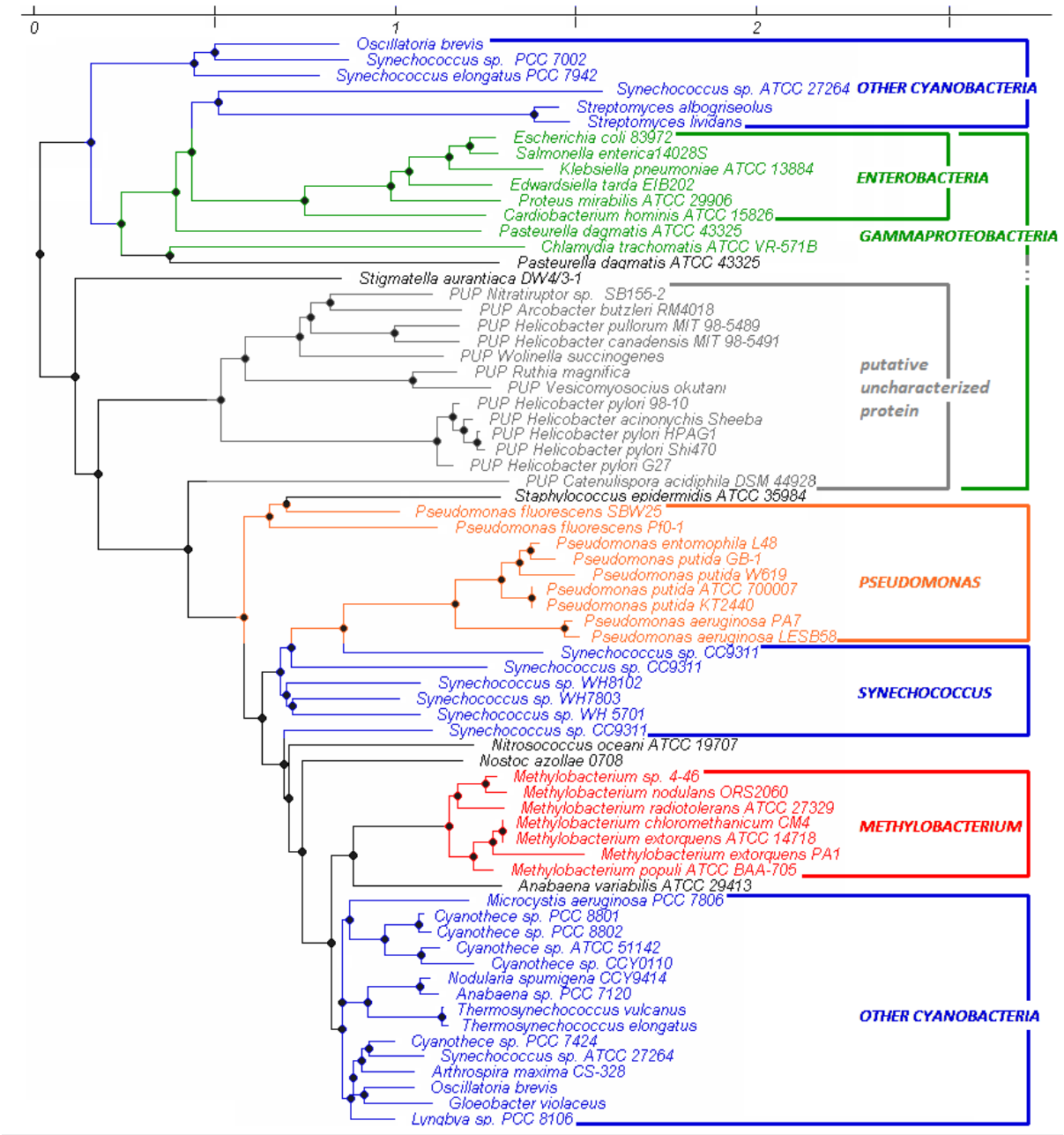

Fig 3. Structured phylogenetic tree of bacterial metallothioneins.

The phylogenetic tree in Fig 3 is descriptive for comparing evolutionary distance between organisms. However, the representation of tree topology is much better through unrooted dendrogram in the Fig 4. The colour marking in both of illustrating tree structures is corresponding for results comparison.

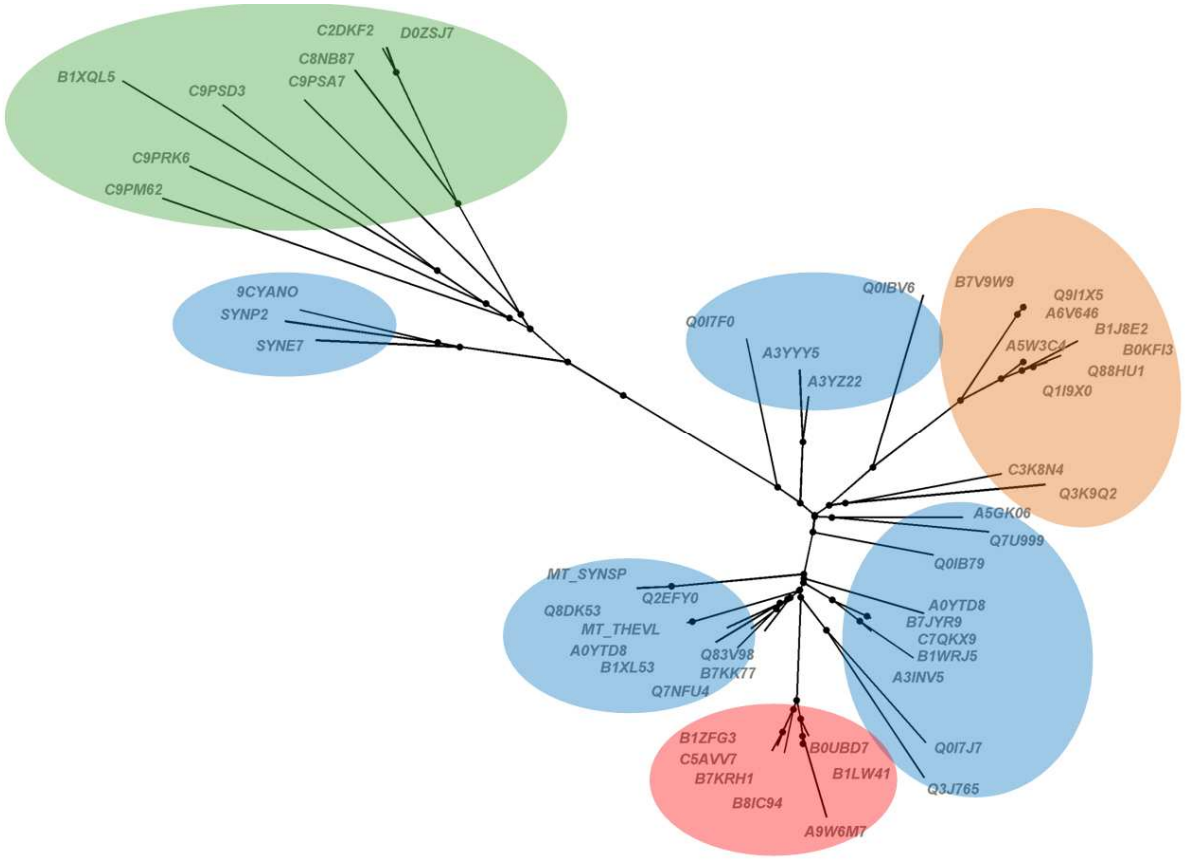

Fig 4. Unrooted dendrogram of bacterial metallothioneins.

4 Conclusions

The presented method for construction of phylogenetic tree combines several popular methods. We used Jukes-Cantor model which is proposed to compensate for some mutations that may have reverted or changed multiple times. Further, we used BioNJ method enhancing classic neighbour-joining algorithm by simple first-order model of variances and covariances of evolutionary distance estimates. This construction leads to more correct tree topology. Results were compared to similar tree of bacterial metallothioneins published by Blindauer [1]. Our results are slightly different because we used more recent version of the protein database.

The tree structure formed closed units which correspond in their affinity. Further improvement could be done by including probability of appearance of particular amino acids in particular sequences. More accurate results would require including a structural similarity of amino acids. However, the the developed method still provides a good overview on biological relationships of bacterial MTs.

Acknowledgement

This work was supported from: GAČR 102/07/1473, GAČR 102/09/H083, GA AV IAA401990701 and MSM0021630513.

References

[1] Blindauer CA. Metallothioneins and Related Chelators. Metal Ions in Life Sciences: Volume 5, 2009: 51-81

[2] Felsenstein J. Inferring phylogenies. Sinauer Associates Incorporated, Sunderland, Massachusetts 2004.

[3] Gascuel O. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. (1997) 14:685-695

[4] Mount, DW. Phylogenetic prediction. Pp. 281-325 in Bioinformatics: sequence and genome analysis – 2nd ed. Cold Spring Harbor, New York 2004.

[5] Nei M, Kumar S. Molecular Evolution and Phylogenetics. Oxford University Press, New York 2000.

[6] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4 (1987): 406–425.

[7] Wallace IM, Blackshields G, Higgins DG. Multiple sequence alignments. Curr Opin Struct Biol. 2005 Jun;15(3):261-6. Review.