

TULIP Software and Web Server: Automatic Classification of Protein Sequences Based on Pairwise Comparisons and Z-Value Statistics

Delphine Grando¹, Philippe Ortet², Fourie Joubert³, Eric Maréchal^{*1} and Olivier Bastien^{*1}

¹UMR 5168 CNRS-CEA-INRA-Université Joseph Fourier; Institut de Recherches en Technologies et Sciences pour le Vivant, CEA Grenoble, 17 rue des Martyrs, 38054, Grenoble Cedex 09, France

²UMR 6191 CNRS-CEA-Université Aix-Marseille II, Institut de Biologie Environnementale et Biotechnologies, CEA Cadarache, 13108 Saint Paul-lez-Durance, France

³Bioinformatics and Computational Biology Unit, Department of Biochemistry, University of Pretoria, 0002, Pretoria, South Africa

Abstract: A configuration space of homologous protein sequences (or CSHP) has been recently constructed based on pairwise comparisons, with probabilities deduced from *Z-value* statistics (Monte Carlo methods applied to pairwise comparisons) and following evolutionary assumptions. A *Z-value* cut-off is applied so as proteins are placed in the CSHP only when the similarity of pairs of sequences is significant following the Theorem of the Upper Limit of a score Probability (TULIP theorem). Based on the positions of similar protein sequences in the CSHP, a classification can be deduced, which can be visualized as trees, called TULIP trees. In previous case studies, TULIP trees were shown to be consistent with phylogenetic trees. To date, no tool has been made available to allow the computation of TULIP trees following this model. The availability of methods to cluster proteins based on pairwise comparisons and following evolutionary assumptions should be useful for evaluation and for the future improvements they might inspire. We developed a web server allowing the local or online computation of TULIP trees based on the CSHP probabilities. The input is a set of homologous protein sequences in multi-FASTA format. Pairwise comparisons are conducted using the Smith-Waterman method, with 100-1,000 sequence shuffling to estimate pairwise *Z-values*. Obtained *Z-value* matrix is used to infer a tree which is then written to a file. Output consists therefore of a *Z-value* matrix, a distance matrix, a TULIP treefile in NEWICK format, and a TULIP tree visualisation. The TULIP server provides an easy-to-use interface to the TULIP software, and allows a classification of protein sequences based on pairwise alignments and following evolutionary assumptions. TULIP trees are consistent with phylogenies in numerous cases, but they can be inconsistent for multi-domain proteins in which some domains have been conserved in all branches. Thus TULIP trees cannot be considered as conventional phylogenetic trees, following the MIAPA (Minimum Information About a Phylogenetic Analysis) recommendations. A major strength of the TULIP classification is its statistical validity when analysing samples including compositionally unbiased and biased sequences (i.e. with biased amino acid distributions), like sequences from *Plasmodium falciparum*. The TULIP web server is a service of the Malaria Portal of the University of Pretoria, South Africa, and is available at <http://malport.bi.up.ac.za/TULIP/>

INTRODUCTION

Evolutionary analysis of genes or proteins is based on sequence comparisons. Since Felsenstein introduced the PHYLogeny Inference Package (PHYLIB) in the 1980's [1], phylogeny is classically predicted based on multiple sequence alignments. In this paper, these methods are called 'multiple alignment-based' (MAB) methods, also known as 'multiple sequence alignment' (MSA) methods. In the mid-1990's, Doolittle [2] proposed a possible alternative to infer the molecular phylogeny of proteins based on pairwise sequence alignments. Here, these methods are called 'pairwise alignment-based' (PAB) methods.

MAB approaches are currently the standard for molecular phylogeny inference and are advised for publication of

phylogenetic trees following the MIAPA (Minimum Information About a Phylogenetic Analysis) checklist (<http://www.mibbi.org/index.php/projects/MIAPA>; [3]). A well known property of MAB methods is that the addition of sequences helps the reconstruction of the phylogeny of sequences that have strongly diverged [4]. This property is an advantage, when one is able to increase the number of sequences used for a phylogeny inference (improving the output by adding input sequences). The MAB methods rely on different hypotheses regarding the evolution of sequences and the validity of the mathematic approaches used to reconstruct phylogenies. This prevents methods to be theoretically compared: it is difficult to assess that one method is better than another, based on theoretical arguments, and usually different methods are pragmatically applied to a given set of protein sequences, and a consensus result is considered as a valid. The comparison of MAB methods and others that do not use multiple alignments shows that no method "recovers the correct phylogeny as accurately as does an approach based on maximum

*Address correspondence to these authors at the UMR 5168 CNRS-CEA-INRA-Université Joseph Fourier; Institut de Recherches en Technologies et Sciences pour le Vivant, CEA Grenoble, 17 rue des Martyrs, 38054, Grenoble Cedex 09, France; E-mail: eric.marechal@cea.fr; olivier.bastien@cea.fr

likelihood distance estimates of multiply aligned sequences” [5]. Although comparative analysis supports the use of MAB methods to reconstruct phylogenies, and helps defining an unambiguous standard that biologist can rely on for publications, the availability of alternative methods is still useful for evaluation, and to help future improvements they might inspire.

In PAB approaches, the addition of new sequences does not alter the pre-existing pairwise comparisons and the outputs are therefore intrinsically stable regarding addition or removal of data. This can be considered as a drawback since classifications cannot be ‘improved’ by addition of new data; however, if the mutual information shared by pairs of sequences is conserved, PAB classification is expected to reflect the complete information of the system, and simply not require improvement by adding more samples. This property is an advantage for the clustering of large databases of biological sequences, since the addition of new sequences does not necessarily requires the recalculation of previous alignments. This is why different clustering methods based on pairwise comparisons of proteins have been proposed, using either *E-value* (COG [6], TribeMCL [7], ProtoNet [8], ProtoMap [9], SIMAP [10], SYSTEMS [11]) or *Z-value* statistics (Decryphon [12], TeraProt [12], PhytoProt [13], CluStr [14]). Recent use of PAB classification for an automatic inference of phylogeny includes OrthoMCL [15], based on pairwise BLAST comparisons and the computation of evolutionary distance based on *E-value* statistics (for review, [12]).

Numerous excellent tools have been developed for MAB phylogeny reconstructions. Popular methods for phylogeny reconstructions include PHYLIP [1], PAUP [16], MEGA [17], PhyML [18], MAFFT [19], RAxML [20], MrBayes [21], GARLI [22] etc. Outputs are treefiles. Nodes are positioned when two ancestral sequences were predicted to have diverged. Software for multiple alignments, phylogeny re-

constructions and tree representations are provided independently or as parts of packages. They can be accessed online *via* repository sites that allow users to design workflows combining some of the most popular programs.

To explore the potential of PAB approaches to classify proteins following evolutionary assumptions [2], we designed a spatial representation of protein sequences (the Configuration Space of Homologous Proteins or CSHP), with probabilities deduced from *Z-value* statistics (Monte Carlo methods applied to pairwise comparisons) [23]. A sequence is placed in the CSHP based on pairwise alignments with other sequences. A *Z-value* cut-off is applied so as proteins are placed in the CSHP only when the similarity of pairs of sequences is significant following the Theorem of the Upper Limit of a score Probability (TULIP theorem) [24]. By default this cut-off value is 8. Based on the positions of similar protein sequences in the CSHP, a classification can be deduced, which can be visualized as trees, called TULIP trees [23]. In previous case studies, TULIP trees were shown to be consistent with phylogenetic trees [23]. The higher accuracy of *Z-value* over *E-value* statistics has been discussed and tested [12, 23-25]. In particular, *Z-value* statistics are valid when comparing sequences of very different amino acid compositions, an interesting feature to help the analysis of compositionally biased sequences. Calculations of *Z-values* are quite CPU intensive compared to *E-values*, and some limitations of the *Z-value* have been reported [25]. The probability deduced from *Z-value* statistics to build TULIP trees has been recently refined [26]. Trees calculated from this PAB model are called TULIP trees [23].

To date, no tool has been made available to allow the computation of TULIP trees following this model. The TULIP web server was therefore developed to allow an easy computation of *Z-values* and deduced classifications. It also provides statistics on amino acid distribution of the submitted sequences.

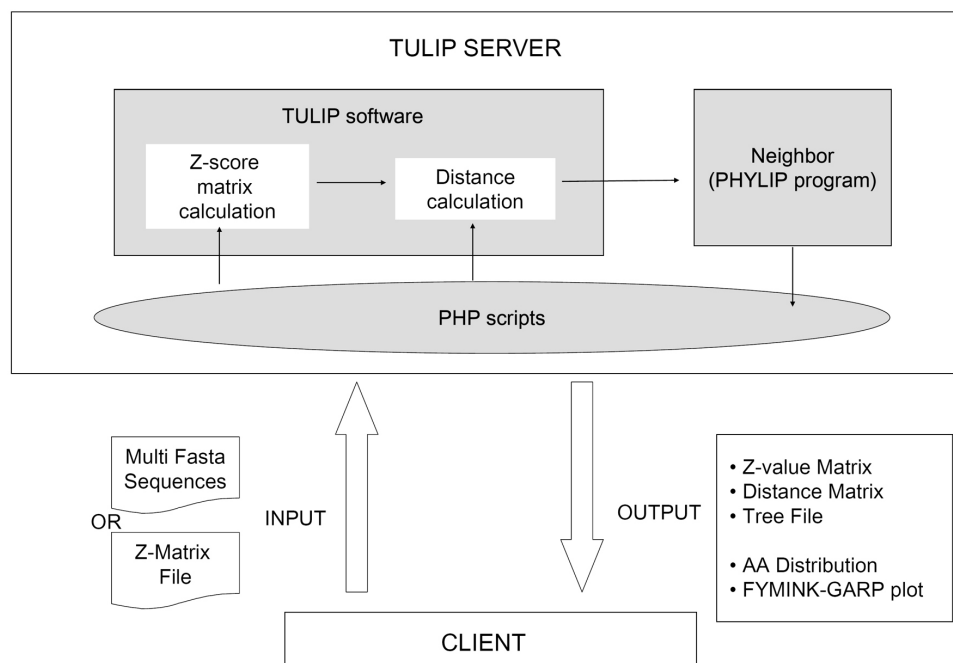


Fig. (1). Outline of the TULIP web server.

PRINCIPLES OF THE TULIP SOFTWARE

Given a set of homologous sequences, the TULIP method consists of computing the optimal pairwise alignment of each pair of sequences a and b , using the Smith and Waterman algorithm [27], measured by a score $s(a,b)$. Alignments of shuffled sequences from a and b (variables corresponding to the shuffled sequences are termed a^* and b^* respectively) allow the estimate of an empirical mean score ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) from the distribution of the random variable $\tilde{S}(a^*, b^*)$. A Z -value (also termed Z -score; [28]) is then defined as:

$$Z = \frac{s(a,b) - \hat{\mu}}{\hat{\sigma}}$$

The computation of Z depends on the estimation of μ and σ , and on the number of shuffling, ranging from 100 to 1,000. The asymptotic law of Z -values was shown to be in-

dependent of the length and amino acid distribution of sequences [24].


The TULIP theorem [24] further assesses that $1/Z^2$ is an upper limit to the probability of the alignment score and that Z -values can be used as a statistical test and a robust single-linkage clustering criterion for sequences' classification. Furthermore, the table of Z -values obtained from all pairwise comparisons allows the computation of a distance matrix and a reconstruction of a tree. Evolutionary distance $t(a,b)$ between two sequences a and b is defined as:

$$t(a,b) = -[\log(p_{id/a}(b^*)) + \log(p_{id/b}(a^*))]$$

with symmetric expressions for $p_{id/a}(b^*)$ and $p_{id/b}(a^*)$:

$$p_{id/a}(b^*) \approx \exp\left(-\frac{\pi}{\sqrt{6}}(z(a,a) - z(a,b))\right)$$

where γ is the Euler-Mascheroni constant ($\gamma \approx 0.5772$) [26]. The TULIP software computes this distance between two sequences a and b .

 **TULIP tree reconstruction from a set of sequences**

Title
If you wish, give a title to this query (up to 10 letters)

Sequences
Paste a set of protein sequences (up to 50 sequences) in FASTA format into the field below: ?

```
>osMGD2
MVISVATPRRSIRDVAVLGGVGLGAGGRQLYQPLRCAFYDGAAGGG
LTAALSSEDGAEGGVPLPCGRKTAATAAKNVILMSDTGGGHRASAEALRDAFRLEFGDAY
QVFFVDLWKEYGGWPLNDMERSYKFMIRHVRLWKVAFHGTSPRVVHGMVLAALAYFYA
NEVVAGIMRYPNDIIISVHPLMQHIPLWVWKWQSLHPKVPFVTVITDLNLTCHPTWFHH
GVTRCYCPSAEVAKRALLRGLLEPSQIRVYGLPIRPSFCRAVLDDKDELKELDMDDPLP
AVLLMGGGEGMGPVEETARALSDELIDRRRRRPPVQIVVICGRNQVLRSTLQSSRWNV
PVKIRGFQKQMEKWMGACDCIITKAGPGTIAEALIRGLPIILNDFIPGQEVGNVPYVV
DNGAGVFSKDPREAAARQVAVRWFTHHTNELRRYSNLAKLAQPEAVFDIVKDIHKLQQQ
PATVTRIPYSLTSSFSYSI
>osMGD1
MPAPTASSLAAAADPALPAFLSLPSPLLPASPLPAAAPSSNAFCVPRGPARAVAVSVS
AAASRLHRMWAEFVRLHGNQIAPLGFASLGLGVGGGGGSGEGAGGGGGGGGEVDGI
EAPKKVLILMSDTGGGHRASAEAIKAAF IQEFGDDYQVFTDLWTDHTPVPFNQLPRSYF
MTYYGTAPRVVHQPHEAATSTFIAREVAKGLMKYQPDVIISVHPLMQHVPLRILRSKGLLI
```

or Submit a file (up to 50 sequences) in FASTA format:

Substitution matrix for pairwise comparison

BLOSUM80 PAM30
 BLOSUM62 (default) PAM70
 BLOSUM45

Number of shuffling for Z-value computation

100 (default)
 1000 (higher quality, takes longer, for longer sequences and high Z-scores)

Valid email address (this field is required)

Input:
Up to 50 protein sequences in FASTA format:
• either copied/pasted
• or uploaded as a multifasta file.

Parameters (optional):
• substitution matrix for pairwise Smith Waterman comparison (default = BLOSUM 62)
• number of shuffling for Monte Carlo based estimates of the Z-value matrix (default = 100)

Valid email address (required)

Fig. (2). TULIP web server main input: TULIP tree reconstruction from a set of protein sequences.

Parameters for sequence alignments (substitution matrices) and Monte Carlo simulations (number of sequence shuffling) can be defined. Up to 50 sequences can be submitted. Larger sample sets can be analyzed upon request or using the free downloadable version of the software.

OUTLINE OF THE TULIP SERVER

The TULIP server is the web interface to the TULIP software, deducing trees from pairwise *Z-value* matrices. As shown in Fig. (1), the user can obtain the TULIP tree file by submitting a set of protein sequences (pasting sequences in FASTA format or uploading a FASTA file) or by directly uploading a *Z-value* matrix file. In the first case, the TULIP software computes the *Z-value* matrix, using the SIM program (Smith and Waterman algorithm, [27]) for the sequence comparison. The *Z-value* matrix is used to infer a tree which is then written to a file. (using the neighbor program of the PHYLIP package). Additional amino acid profile analyses are performed to help users visualize whether the initial set of sequences had divergent amino acid distributions, and therefore help indicating compositionally biased sequences that might coincide with evolutionary divergences in the returned tree. Result files are sent *via* e-mail, together with a link to a result page on the server, with a unique identifier. This page displays a graph representation of the computed TULIP tree, provides links to all results and links to alternative methods for tree representations.

IMPLEMENTATION

TULIP interface is implemented in the PHP language. The TULIP software is implemented in Perl and C. A first module runs the TULIP software by a set of PERL scripts. Sequence randomization, SIM pairwise alignments, *Z-value*

and distance calculations are made *via* C and PERL scripts. The core module for local *Z*-matrix computations is freely downloadable for Linux and Windows. Graphical displays of amino acid profile analyses provided by a second module are created using GD and JpGraph libraries. The server is 2x Quad-core CPU Intel system with 8GB RAM, running SuSE Linux 10.2.

Input Form

The input is a set of protein sequences in multi-FASTA format (as a pasted text or uploaded file) (Fig. 2). The input page also allows the definition of some parameters for the Monte Carlo simulation: for each pairwise comparison, the substitution matrix (default = BLOSUM 62) and the number of randomizations (100 or 1,000; default = 100) can be determined by users. Different substitution matrices can be selected, of the PAM and BLOSUM series, helping users to compare results with alternative methods utilizing the same matrices, and allowing the future implementation of novel matrices. A derived *Z-value* matrix is computed. The user may alternatively provide a pre-calculated *Z-value* matrix (Fig. 3). The *Z-value* matrix is then used to compute a distance matrix following [26].

Outputs

Main outputs include the *Z-value* table (Fig. 4A), the computed distance matrix (Fig. 4B) and the protein classifi-

The TULIP 1.1 server estimates molecular phylogenies based on pairwise protein sequence alignments and the corresponding Z-score probabilities, according to the TULIP theorem (Theorem of the Upper Limit of a score Probability). The TULIP theorem allows the estimate of the probability of an alignment even with protein sequences of very different lengths, complexity and amino acid distributions. Using a set of homologous sequences, pairwise alignments are computed using the Smith and Waterman method. A phylogenetic tree or TULIP tree, is reconstructed from a distance matrix derived from the pairwise Z-score matrix. TULIP based phylogenetic reconstruction is advised when using sets including compositionally biased sequences (i.e. with biased amino acid distributions). The TULIP 1.1 server also allows the reconstruction of phylogenies from a Z-score matrix in a BioFace@format.

SUBMISSION: From a set of sequences | From a Z-Score matrix

TULIP tree reconstruction from a Z-score matrix

Title
If you wish, give a title to this query (up to 10 letters)

Z-score matrix
Submit a complete Z-score matrix in ZMATRIX format from a pairwise all-by-all comparison (up to 100 sequences)

Valid email address (this field is required)

Other possible input:
 Pre-calculated Z-value matrix (up to 100 protein sequences).

Example (Z-score_matrix.txt):

```

ENO_ARATH ENO_ORYSA ENO2_MAIZE ENOA_HUMAN ENOA_MOUSE ENO_DROME
465.726775201868 430.413921152173 376.339734349411 270.904037098422 324.033195985057 265.997029033339
405.529856544394 442.267446990473 422.612510402257 299.356006988195 283.937234455606 351.273433723435
305.26716993702 418.991880130524 450.072410373909 325.17887892338 335.115949307428 259.807297069726
377.718118413861 307.965525993729 302.687894680689 473.179371896571 415.19436590644 357.142376568042
334.259751443431 285.720457208938 300.962998437341 465.105522106238 412.915176258847 392.054832991878
257.005612292031 302.352661090633 283.095934734874 324.635624136084 418.685764927084 446.100117972037

```

Fig. (3). TULIP web server alternative input: TULIP tree reconstruction from a *Z-value* matrix.

A file corresponding to all pre-calculated pairwise *Z-values* of up to 100 sequences can be submitted. Larger sample sets can be analyzed upon request or using the free downloadable version of the software.

cation based on this distance matrix. The TULIP tree is provided as a treefile in NEWICK format (Fig. 4C) and a simple graphical visualisation (Fig. 4D). Links to other servers to obtain different graphical representations of the TULIP tree are also provided.

An analysis of the length (Fig. 5A) and amino acid distribution (Fig. 5B, 5C) of the input sequences is additionally returned. Both a global amino acid profile for each of the submitted sequences and a "GARP vs FYMINK" statistical repartition are created, by a set of PHP scripts, using GD and JpGraph libraries. "GARP" stands for the amino acid markers of GC-rich codons, i.e. Glycine, Alanine, Aspartic acid

and Proline; "FYMINK" stands for the amino acid markers of AT-rich codons, i.e. Phenylalanine, Tyrosine, Methionine, Isoleucine, Asparagine and Lysine [29]. The "GARP vs FYMINK" plot allows therefore the detection of possible compositional biases, due to trends in the AT/GC ratio in the initial protein set [29]. Additional outputs provided by the TULIP server consist therefore of radar plot graphs showing the amino acid profiles of each protein (Fig. 5B) and a "GARP vs FYMINK" plot for the complete set of sequences (Fig. 5C). This information, which is usually not provided by other protein clustering servers, are valuable to point some features in the TULIP tree that might be related to important length alterations and/or strong nucleotidic compositional

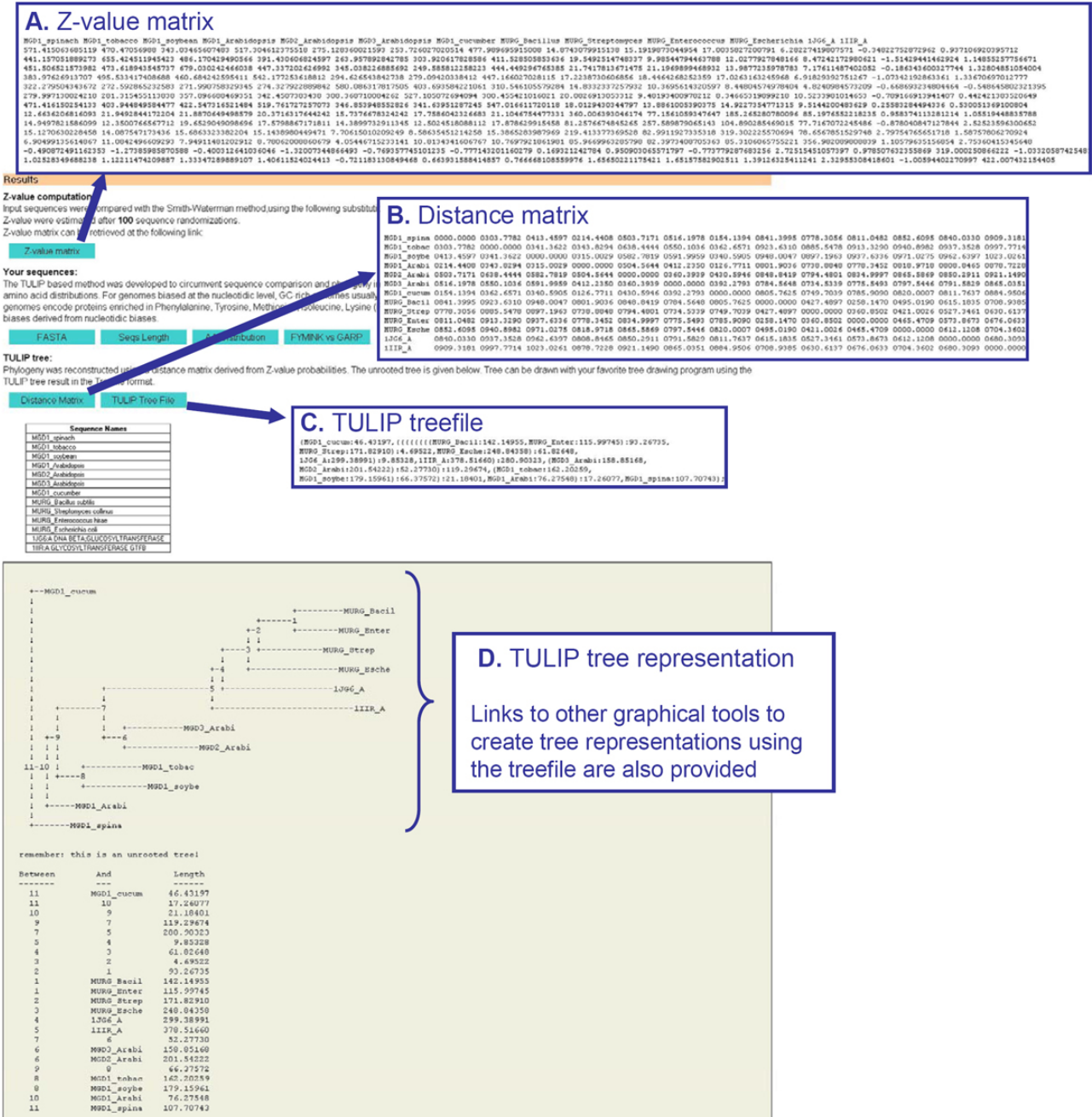


Fig. (4). TULIP web server main outputs: a classification of proteins based on pairwise sequence alignments. (A) Z-value matrix. (B) Distance matrix deduced from the Z-value matrix. (C) TULIP treefile in NEWICK format. (D) TULIP tree graphical representation. Links to other tools allowing alternative graphical representations of the TULIP treefile are provided.

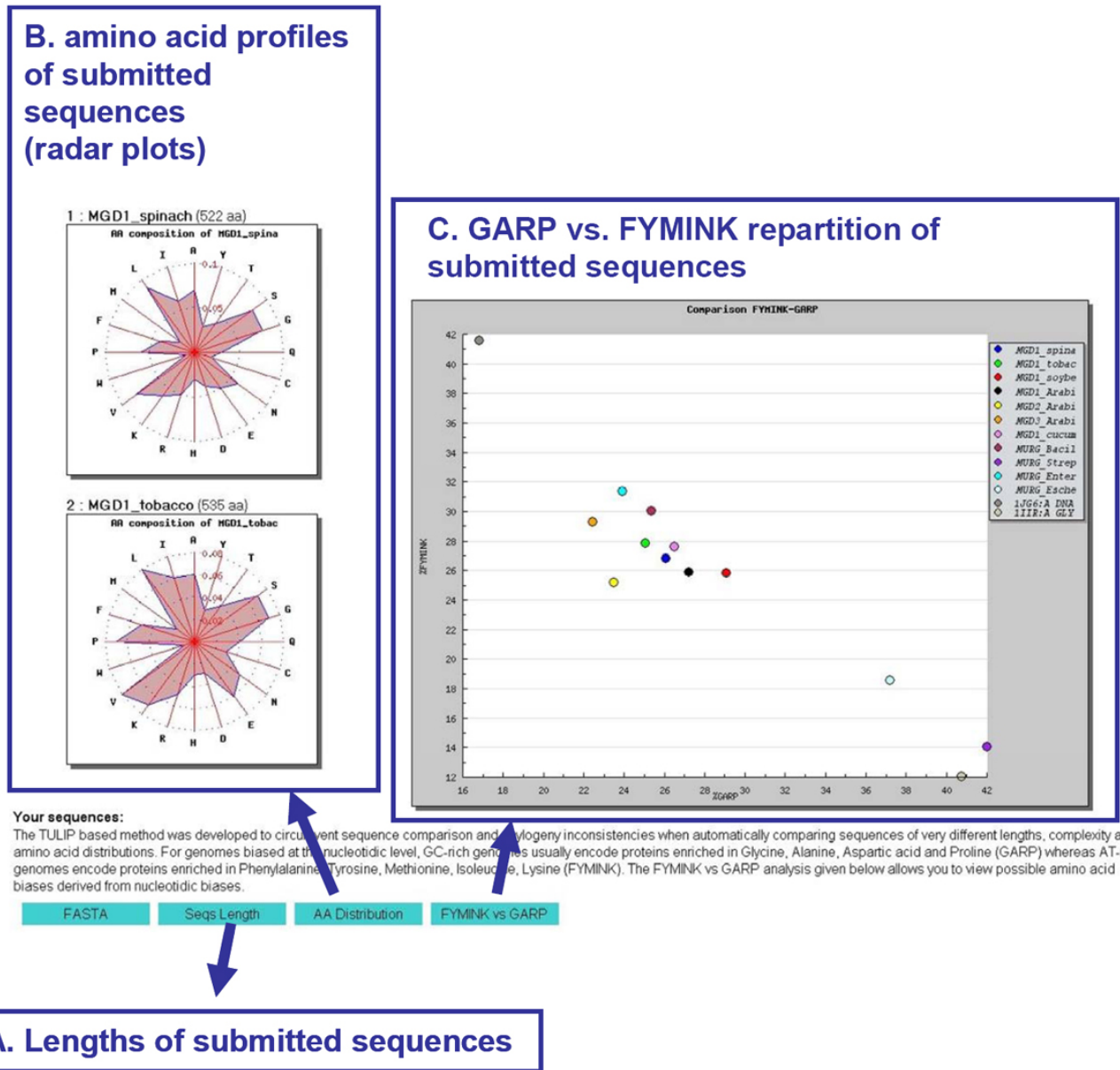


Fig. (5). TULIP web server additional outputs: analyses of possible heterogeneity of the length and amino acid composition of submitted sequences. (A) Length of submitted sequences. (B) Radar plot graphs of the amino acid distributions of all submitted sequences. (C) GARP vs FYMINIK plot. GARP stands for the amino acid markers of GC-rich codons, i.e. Glycine, Alanine, Aspartic acid and Proline; FYMINIK stands for the amino acid markers of AT-rich codons, i.e. Phenylalanine, Tyrosine, Methionine, Isoleucine, Asparagine and Lysine. The GARP vs FYMINIK plot allows therefore the detection of possible compositional biases, due to trends in the AT/GC ratio in the initial protein set.

trends (GC or AT enrichment), underlying divergences at the amino acid level.

ACCESS, TESTING AND PERFORMANCE

The TULIP server has been tested on Microsoft Internet Explorer, Netscape and Mozilla Firefox. The server is available at <http://malport.bi.up.ac.za/TULIP/> as one of the services of the Malaria Portal of the University of Pretoria. The number of sequences for submission is restricted to 50, but larger sample sets can be analyzed upon request. Output from 12 sequences (~500-1000 amino acid-length; 100 sequence shuffling), is returned in less than 10 min. Accuracy is gained by setting the number of shuffling to 1,000. If users

submit pre-calculated *Z-value* matrices, the number of analyzed sequences is restricted to 100. Output from a 50 x 50 *Z-value* matrix is returned in less than 5 seconds. Larger sample sets can be analyzed upon request or using the free downloadable version of the software (<http://malport.bi.up.ac.za:7070/downloads/tulip>). The TULIP software is available for Linux and for Windows.

CONCLUSIONS

The TULIP server is an easy-to-use web interface to the TULIP program and the first online PAB method for protein classification following evolutionary assumptions, based on the TULIP theorem and corollaries. The TULIP server was

initially developed to allow the comparative analyses of proteins including sequences of *Plasmodium falciparum*, the malaria causative agent, which are atypical due to their strong amino acid compositional bias, low complexity and being 20% longer than their homologues. The TULIP server therefore finds a specific use for samples including sequences of different lengths, complexity and amino acid distributions such as malaria proteins. TULIP trees are consistent with phylogenies in numerous cases reported earlier, but they can be inconsistent for multi-domain proteins in which some domains have been conserved in all branches. For example, in some cases, it is possible that after a comparison of three sequences *a*, *b* and *c*, the *ab*, *ac* and *bc* may not overlap, being a clear limit of the method. Thus TULIP trees cannot be considered as conventional phylogenetic trees, following the MIAPA (Minimum Information About a Phylogenetic Analysis) recommendations. The availability of methods to cluster proteins based on pairwise comparisons and following evolutionary assumptions should therefore be used with caution, be useful for evaluation and for future improvements they might inspire. A major strength of the TULIP classification is its statistical validity when analysing samples including compositionally unbiased and biased sequences (i.e. with biased amino acid distributions), like sequences from *Plasmodium falciparum*.

AUTHORS' CONTRIBUTIONS

DG, PO and OB contributed to the development of the software and web server and drafted the manuscript. FJ contributed to the development of the web server and helped to draft the manuscript. EM participated in the design of the web server, coordinated its development and helped to draft the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

Authors wish to thank Tjaart de Beer for technical support. This publication was made possible through financial supports from South Africa National Research Foundation, South Africa National Bioinformatics Network, the French National Research Agency (PlasmoExplore project), the Rhône-Alpes Regional Council, and the French and South African Ministries of Foreign Affairs (SAFE-ICT program). This publication was further supported by the New Partnership for Africa's Development, African Union.

REFERENCES

- [1] J. Felsenstein, "Evolutionary trees from DNA sequences: a maximum likelihood approach", *J. Mol. Evol.*, vol. 17, no. 6, pp. 368-376, 1981.
- [2] R. F. Doolittle, D. F. Feng, S. Tsang, G. Cho, and E. Little, "Determining divergence times of the major kingdoms of living organisms with a protein clock", *Science*, vol. 271, no. 5248, pp. 470-477, January 1996.
- [3] J. Leebens-Mack, T. Vision, E. Brenner, et al. "Taking the first steps towards a standard for reporting on phylogenies: Minimum Information About a Phylogenetic Analysis (MIAPA)", *Omic*, vol. 10, no. 2, pp. 231-237, June 2006.
- [4] J. D. Thompson, F. Plewniak, and O. Poch, "A comprehensive comparison of multiple sequence alignment programs", *Nucleic Acids Res.*, vol. 27, no. 13, pp. 2682-2690, July 1999.
- [5] M. Höhl, and M. A. Ragan, "Is multiple-sequence alignment required for accurate inference of phylogeny?", *Syst. Biol.*, vol. 56, no. 2, pp. 206-221, April 2007.
- [6] R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, et al. "The COG database: new developments in phylogenetic classification of proteins from complete genomes", *Nucleic Acids Res.*, vol. 29, no. 1, pp. 22-28, January 2001.
- [7] A. J. Enright, V. Kunin, and C. A. Ouzounis, "Protein families and TRIBES in genome sequence space", *Nucleic Acids Res.*, vol. 31, no. 15, pp. 4632-4638, August 2003.
- [8] O. Sasson, A. Vaaknin, H. Fleischer, et al. "ProtoNet: hierarchical classification of the protein space", *Nucleic Acids Res.*, vol. 31, no. 1, pp. 348-352, January 2003.
- [9] G. Yona, N. Linial, and M. Linial, "ProtoMap: automatic classification of protein sequences and hierarchy of protein families", *Nucleic Acids Res.*, vol. 28, no. 1, pp. 49-55, January 2000.
- [10] R. Arnold, T. Rattei, P. Tischler, M. P. Truong, V. Stumflen, and W. Mewes, "SIMAP--The similarity matrix of proteins", *Bioinformatics*, vol. 21(Suppl 2), pp. 42-46, September 2005.
- [11] A. Krause, J. Stoye, and M. Vingron, "Large scale hierarchical clustering of protein sequences", *BMC Bioinformatics*, vol. 6, p. 15, January 2005.
- [12] O. Bastien, P. Ortet, S. Roy, and E. Maréchal, "The configuration space of homologous proteins: a theoretical and practical framework to reduce the diversity of the protein sequence space after massive all-by-all sequence comparisons", *Future Generation Comput. Syst.*, vol. 23, no. 3, pp. 410-427, 2007.
- [13] A. Louis, E. Ollivier, J. C. Aude, and J. L. Risler, "Massive sequence comparisons as a help in annotating genomic sequences", *Genome Res.*, vol. 11, no. 7, pp. 1296-1303, July 2001.
- [14] R. Apweiler, M. Biswas, W. Fleischmann, et al. "Proteome analysis database: online application of InterPro and CluStr for the functional classification of proteins in whole genomes", *Nucleic Acids Res.*, vol. 29, no. 1, pp. 44-48, January 2001.
- [15] OrthoMCL, [Online]. Available: <http://orthomcl.cbil.upenn.edu/> [Accessed Feb. 27, 2009].
- [16] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis, "Phylogenetic inference", in *Molecular Systematics*, D. M. Hillis, C. Moritz, B. K. Mable, Eds. Sunderland: Sinauer, 1996, pp. 407-514.
- [17] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0" *Mol. Biol. Evol.*, vol. 24, no. 8, pp. 1596-1599, August 2007.
- [18] S. Guindon, and O. Gascuel, "PhyML – A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood", *Syst. Biol.*, vol. 52, no. 5, pp. 696-704, October 2003.
- [19] K. Katoh, and H. Toh, "Recent developments in the MAFFT multiple sequence alignment program", *Brief Bioinform.*, vol. 9, no. 4, pp. 286-298, July 2008.
- [20] A. Stamatakis, T. Ludwig, and H. Meier, "RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees", *Bioinformatics*, vol. 21, no. 4, pp. 456-463, February 2005.
- [21] F. Ronquist, and J. P. Huelsenbeck, "MrBayes 3: Bayesian phylogenetic inference under mixed models", *Bioinformatics*, vol. 19, no. 12, pp. 1572-1574, August 2003.
- [22] D. Zwickl, "Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion", PhD Thesis, University of Texas, Austin, TX, USA, 2006.
- [23] O. Bastien, P. Ortet, S. Roy, and E. Maréchal, "A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise Z-value probabilities", *BMC Bioinform.*, vol. 6, p. 49, March 2005.
- [24] O. Bastien, J. C. Aude, S. Roy, and E. Maréchal, "Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics", *Bioinformatics*, vol. 20, no. 4, pp. 534-537, March 2004.
- [25] T. Hulsen, J. de Vlieg, J. A. Leunissen, and P. M. Groenen, "Testing statistical significance scores of sequence comparison methods with structure similarity", *BMC Bioinformatics*, vol. 7, p. 444, October 2006.
- [26] O. Bastien, and E. Maréchal, "Evolution of biological sequences implies an extreme value distribution of type I for both global and local pairwise alignment scores", *BMC Bioinformatics*, vol. 9, p. 332, August 2008.
- [27] T. F. Smith, and M. S. Waterman, "Identification of common molecular subsequences", *J. Mol. Biol.*, vol. 147, no. 2, pp. 195-197, April 1981.

- [28] D. J. Lipman, and W. R. Pearson, "Rapid and sensitive protein similarity searches", *Science*, vol. 227, no. 4693, pp. 1435-1441, March 1985.
- [29] O. Bastien, S. Lespinats, S. Roy, *et al.* "Analysis of the compositional biases in Plasmodium falciparum genome and proteome using Arabidopsis thaliana as a reference", *Gene*, vol. 336, no. 2, pp. 163-173, July 2004.

Received: March 11, 2009

Revised: April 03, 2009

Accepted: April 15, 2009

© Grando *et al.*; Licensee Bentham Open.

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.