BMC Bioinformatics

BioMed Central

Research article

A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise Z-score probabilities

Olivier Bastien[1,2], Philippe Ortet[3], Sylvaine Roy[4] and Eric Maréchal*[1]

Address: [1]UMR 5019 CNRS-CEA-INRA-Université Joseph Fourier, Laboratoire de Physiologie Cellulaire Végétale; Département Réponse et Dynamique Cellulaire; CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France, [2]Gene-IT, 147 avenue Paul Doumer, F-92500 Rueil-Malmaison, France, [3]Département d'Ecophysiologie Végétale et de Microbiologie; CEA Cadarache, F-13108 Saint Paul-lez-Durance, France and [4]Laboratoire de Biologie, Informatique et Mathématiques; Département Réponse et Dynamique Cellulaire, CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France

Email: Olivier Bastien - obastien@cea.fr; Philippe Ortet - portet@cea.fr; Sylvaine Roy - sroy@cea.fr; Eric Maréchal* - emarechal@cea.fr

* Corresponding author

Published: 10 March 2005

BMC Bioinformatics 2005, **6**:49 doi:10.1186/1471-2105-6-49

Received: 25 November 2004
Accepted: 10 March 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/49

Abstract

Background: Popular methods to reconstruct molecular phylogenies are based on multiple sequence alignments, in which addition or removal of data may change the resulting tree topology. We have sought a representation of homologous proteins that would conserve the information of pair-wise sequence alignments, respect probabilistic properties of Z-scores (Monte Carlo methods applied to pair-wise comparisons) and be the basis for a novel method of consistent and stable phylogenetic reconstruction.

Results: We have built up a spatial representation of protein sequences using concepts from particle physics (configuration space) and respecting a frame of constraints deduced from pair-wise alignment score properties in information theory. The obtained configuration space of homologous proteins (CSHP) allows the representation of real and shuffled sequences, and thereupon an expression of the TULIP theorem for Z-score probabilities. Based on the CSHP, we propose a phylogeny reconstruction using Z-scores. Deduced trees, called TULIP trees, are consistent with multiple-alignment based trees. Furthermore, the TULIP tree reconstruction method provides a solution for some previously reported incongruent results, such as the apicomplexan enolase phylogeny.

Conclusion: The CSHP is a unified model that conserves mutual information between proteins in the way physical models conserve energy. Applications include the reconstruction of evolutionary consistent and robust trees, the topology of which is based on a spatial representation that is not reordered after addition or removal of sequences. The CSHP and its assigned phylogenetic topology, provide a powerful and easily updated representation for massive pair-wise genome comparisons based on Z-score computations.

Background

Past events that gave birth to biological entities can be tentatively reconstructed based on collections of descriptors traced in ancient or present-day creatures. Using genomic sequences, an estimate of the relative time separating branching events, previously supported by geological records, could be formalized using mathematical models. The use of proteins for evolutionary reconstructions was vastly explored as soon as the first amino acid sequences were made available [1-9]. The rich biological information contained in protein sequences stems from their being, on the one hand, translation of genes that reflect the history of genetic events to which the species has been subjected, and on the other hand, effectors of the functions constituting a living creature [10] Since protein sequences are encoded in a 20-amino acid alphabet, they are also considered to embody more *information-per-site* than DNA or RNA [11]; they also exhibit smaller compositional trends [12,13]. When compared, sequences that share substantial features are considered as possible homologues [14], based on the fundamental postulate that can be simply stated as "the closer in the evolution, the more alike and conversely, the more alike, *probably* the closer in the evolution".

As summarized by Otu and Sayood [15], the techniques of molecular phylogenetic analyses can be divided into two groups. In the first case, a matrix representing the distance between each pair of sequences is calculated and then transformed into a tree. In the second case, a tree is found that can best explain the observed sequences under evolutionary assumptions, after evaluation of the fitness of different topologies. Some of the approaches in the first category utilize distance measures [16-19] with different models of nucleotide substitution or amino acid replacement. The second category can further be divided into two groups based on the optimality criterion used in tree evaluation: parsimony [20,21] and maximum likelihood methods [22,23]. For a detailed comparison of these methods see [24].

In phylogeny inference based on distance methods, features separating related proteins are used to estimate an observed distance, also called the p-distance, the simplest measure of which is just the number of different sites between proteins. Divergence time (t), also called genetic distance or evolutionary time, is calculated from the p-distance, depending on assumptions derived from evolutionary models [11,24]. For example, the assumption that mutational events happen with equal probability at each site of any sequence leads to the molecular clock model [2]. Although widely used, it is well-known to be unrealistic and numerous corrections have been proposed to refine it [19,25,26]. By definition, the distance matrix is given as $T = (t_{ab})$ where a and b represent the homologous sequences from the analyzed dataset. Tree reconstruction algorithms are then applied to these matrices [11,24]. Eventually, phylogenetic trees corresponding to the classified sequences are statistically evaluated with bootstrap methods and, when available, calibrated using dated fossils [25,26].

Doolittle et al. [27] have proposed methods for converting amino acid alignment scores into measures of evolutionary time. Similarity between amino acids [28-30] provides a way to weight and score alignments [31]. In practice the optimal alignment of two sequences (a and b) is determined from the optimal score $s(a,b)$ [25,27], computed with a dynamic programming procedure [32,33]. In aligned sequences, conservation is measured at identical sites, whereas variation is scaled at substituted sites. To estimate the variation/conservation balance, the *p-distance* can be given as a function of f_{id}, the fraction of identical residues: *p-distance* = $1 - f_{id}$. To take into account that multiple mutations can happen at the same site, an expression of f_{id} was proposed by Doolittle et al. [27] using $s(a^*,b^*)$, the score obtained from randomized a and b sequences [34] and s_{id}, the average score of the sequences compared with themselves [19,25,27]:

$$f_{id}\left(a,b\right) = \frac{s(a,b) - s(a^*,b^*)}{s_{id} - s(a^*,b^*)} \qquad (1)$$

To connect pair-wise alignments and phylogeny, divergence time has been approximated:

$$t(a,b) = -\lambda \log[f_{id}(a,b)] \qquad (2)$$

introducing a Poisson correction [2] as a reasonable stochastic law relating amino acid changes and elapsed time. As mentioned earlier, adjustments and corrections of equation (2) were proposed to fit more realistically the complexity of evolution [11,25,35]. This attempt of unification helped reconstructing phylogeny of major lineages [27]. However, detailed phylogenic trees obtained from evolutionary close sequences are not satisfactory. In practice, phylogenies are reconstructed based on multiple alignments. Multiple alignment based (MAB) trees are recalculated when incremented with additional sequences; although MAB methods are usually considered accurate, numerous cases of inconsistencies (incongruence) between observed data and deduced MAB trees are recorded (see [15,36]).

Here, we re-examine the estimate of the *p-distance* between two homologous sequences, based on f_{id}, as a source for geometric positioning, divergence time calculations and evolutionary reconstruction. We based our model on mathematical properties that alignment scores should respect; i) information theory [37,38] applied to sequence similarity, ii) algorithmic theory applied to alignment optimization [28] and iii) alignment probability, particularly in conformity with the TULIP theorem [39]. We used these properties as a framework of constraints to build a geometric representation of a space of probably homologous proteins and define a theoretically explicit measure of protein proximity. This unified model conserves information in the way physical models conserve matter or energy. The obtained representation of protein sequences is unaltered by adding or removing sequences. Applications include therefore the reconstruction of evolutionary consistent and robust trees, the topology of which is based on a spatial representation that is not reordered after addition or removal of sequences.

Results and discussion

Pair-wise sequence alignment scores in information theory

Criteria to measure the variation/conservation balance between proteins should embody as much as possible the structural and functional potentiality within sequences of amino acids. In the absence of explicit physical criteria, amino acid similarity was solved empirically by measuring amino acid substitution frequencies in alignments of homologous sequences [30,40]. Given two amino acids i and j, the similarity function $s(i,j)$ was set as:

$$s(i,j) = \log \frac{\varpi_{ij}}{\pi_i \pi_j} \qquad (3)$$

where ϖ_{ij} is the observed frequency of substitution of i by j or j by i, and π_i and π_j are the frequencies of i and j in the two aligned sequences. The ϖ_{ij} frequency is the estimate of the probability of substitution of i by j in real alignments; whereas $\pi_i\pi_j$ is the estimate of the probability of substitution under the independency hypothesis. The similarity function gives a 20 × 20 similarity matrix usable to score protein sequence alignments, that can be interpreted in the information theory [37,38] according to the following proposition.

Proposition 1

Amino acid substitution matrix values are estimates of the mutual information between amino acids in the sense of Hartley [37,38]. Consequently, the optimal alignment score computed between two biological sequences is an estimate of the optimal mutual information between these sequences.

Proof

Given a probability law P that characterizes a random variable, the Hartley self-information h is defined as the amount of information one gains when an event i occurred, or equivalently the amount of uncertainty one loses after learning that i happened:

$$h(i) = -\log(P(i)) \qquad (4)$$

The less likely an event i, the more we learn about the system when i happens. The mutual information I between two events, is the reduction of the uncertainty of one event i due to the knowledge of the other j:

$$I_{j\to i} = h(i) - h(i/j) \qquad (5)$$

Mutual information is symmetrical, *i.e.* $I_{j\to i} = I_{i\to j}$, and in the following will be expressed by $I(i;j)$. The self and mutual information of two events i and j are related:

$$h(i \cap j) = h(i) + h(j) - I(i;j) \qquad (6)$$

If the occurrence of one of the two events makes the second impossible, then the mutual information is equal to - ∞. If the two events are fully independent, mutual information is null. The empirical measure of the similarity between two amino acids described in equation (3) can therefore be expressed in probabilistic terms:

$$s(i,j) = \log(\frac{\varpi_{ij}}{\pi_i \pi_j}) = \log(P_{\varpi}(i \cap j)) - \log(P_{\pi}(i)) - \log(P_{\pi}(j)) \qquad (7)$$

where P_{ϖ} is the joint probability to have i and j aligned in a given alignment and P_{π} the measure of probability that amino acids occur in a given sequence. From equations (4) and (6), equation (7) becomes:

$$s(i,j) = h(i) + h(j) - h(i \cap j) \qquad (8)$$

that is

$$s(i,j) = I(i;j) \qquad (9)$$

As a consequence, the similarity function (or score) is the mutual information between two amino acids. Additionally, the score between sequences (the sum of elementary scores between amino acids, [32,33,41,42]) is, according to the hypothesis of independence of amino acid positions, the estimated mutual information between the two given biological sequences.

Once two sequences are aligned, we pose the question whether the alignment score is sufficient to assess that the proteins are conceivably alike and thus evolutionarily related? The theorem of the upper limit of a sequence alignment score probability (TULIP theorem, [39]), sets the upper bound of an alignment score probability, under a hypothesis less restrictive than the Karlin-Altschul model [43]. Given two real sequences a and b ($a = a_1a_2...a_m$ and $b = b_1b_2...b_n$), where $s = s(a,b)$ the maximal score of a pair-wise alignment obtained with any alignment method, b^* the variable corresponding to the shuffled sequences from b, and given $P\{S(a,b^*)\geq s\}$ the probability that an alignment by chance between a and b^* has a higher score than s, then whatever the distribution of the random variable $S(a,b^*)$ the TULIP theorem states:

$$s \geq \mu + k\sigma \Rightarrow P\left\{ S(a,b^*) \geq s \right\} \leq \frac{1}{k\dagger} \qquad (10)$$

with $k > 1$, μ the mean of $\tilde{S}(a,b^*)$ and σ its standard deviation. The unique restriction on $S(a,b^*)$ is that it has a finite mean and a finite variance. A first corollary of the TULIP theorem is that the Z-score is a statistical test for the probability of a sequence alignment score. We additionally state the following new corollary.

TULIP corollary 2

Given the TULIP theorem conditions, let $z(a,b^*) = \dfrac{s(a,b) - \mu}{\sigma}$ be the Z-score [44]. Then, $z(a,b^*)$ is the greatest possible value for k ($k \in]1,+\infty[$), which holds relation (10) true. In consequence, with $k = z(a,b^*)$, then

$$P\left\{ S(a,b^*) \geq s \right\} \leq \frac{1}{z(a,b^*)\dagger} \qquad (11)$$

The best upper bound value for $P\{S(a,b^*) \geq s\}$ is termed $l_\upsilon(a,b^*) = \dfrac{1}{z(a,b^*)\dagger}$. From the TULIP theorem and corollaries, the comparison of a protein to a given reference a, weighed by an alignment score, is characterized by a bounded probability that the alignment is fortuitous.

Question of the proximity between protein sequences in the light of information theory

Since the optimized alignment score of two protein sequences allows an access to both the mutual information between proteins and an upper bound that the alignment is not fortuitous, one would expect that it is an accurate way to spatially organize proteins sets. A simple relation would be "the higher the mutual information, the nearest". There are three ways to assess the proximity between two objects a and b in a given space E [41]. The first is dissimilarity, a function $f(a,b)$: $E \times E \rightarrow \Re^+$ such that $f(a,b) = 0 \Leftrightarrow a = b$ and $f(a,b) = f(b,a)$; the second is the distance *per se*, that is a dissimilarity such that the triangle inequality is respected: $\forall\ a,b,c \in E, f(a,c) \leq f(a,b) + f(b,c)$; and the third is the similarity defined as a function $f(a,b)$: $E \times E \rightarrow \Re$ such that $f(a,a) = \max\limits_{b} f(a,b)$ and $f(a,b) = f(b,a)$. Representing objects in a space is convenient using the notion of distance. When the optimal alignment is global, *i.e.* requiring that it extends from the beginning to the end of each sequence [32], it is theoretically possible to define a distance *per se*, that is to spatially organize the compared sequences [41]. However, from a biological point of view, global alignment algorithms are not reliable to assess homology of protein domains. Local alignments are better suited, using scoring matrices to find the optimum local alignment and maximizing the sum of the scores of aligned residues [28,31]. In contrast with global alignments, local alignments do not allow any trivial definition of distances [41].

Although amino acid similarity is a function $f(i, j)$: $E \times E \rightarrow \Re$, owing to the local alignment optimization algorithms, the computed score is a function $f(a,b)$: $E \times E \rightarrow \Re^+$, requiring the existence of at least one positive score in the similarity matrices. Thus, when constructing an alignment with the Smith and Waterman [33] method, the constraint that $s(a,b)>0$ (*i.e.* $I(a;b) > 0$) is imposed. This condition is consistent with proposition 1: if two sequences are homologous, knowledge about the first has to bring information about the second, that is to say, the mutual information between the two sequences cannot decrease below zero: $I(a;b) > 0$ (*i.e.* $s(a,b) > 0$). As a consequence, in the following geometric construction we sought a refined expression for the proximity of proteins.

Geometric construction of a configuration space of homologous proteins (CSHP) conserving mutual information

In a set of homologous proteins, any sequence a can be selected as a reference, noted a_{ref}, in respect to which the others are compared. A geometric representation of objects relatively to a fixed frame is known as a configuration space (CS). In physics, a CS is a convenient way to represent systems of particles, defined by their positional vectors in some reference frame. Here, given n similar sequences, it is therefore possible to consider n references of the CSHP. In a given (CSHP, a_{ref}), each amino acid position aligned with a position in the a_{ref} sequence, corresponds to a comparison dimension (CS dimension). Proteins are simply positioned by a vector, the coordinates of which are given by the scores of aligned amino acids. Gaps are additional dimensions of the CS. When considering that local algorithms identify the space of biological interest, *i.e.* a CSHP, the gap penalty is a parameter that maximizes the shared informative dimensions. Thus, given the amino acids mutual information, alignment optimization methods define the relative positions of proteins.

At this point in our construction, a first important property of the CSHP can be deduced. Since mutual information with a_{ref} is sufficient for the full positioning, then positioning of proteins in a given (CSHP, a_{ref}) is unambiguous, unique, and is not altered when proteins are added or removed. In other words, a (CSHP, a_{ref}) is a univocal space.

Given two sequences a and b, if b occurs in (CSHP, a_{ref}), then a also occurs in (CSHP, b_{ref}). The pair-wise alignment of a and b having no order (symmetry of the mutual information), the positions of b in (CSHP, a_{ref}) is dependent of the position of a in (CSHP, b_{ref}). Thus, once a (CSHP, a_{ref}) has been built, $\forall b \in$ (CSHP, a_{ref}), part of the geometry of (CSHP, b_{ref}) is learnt. Thus, in a CSHP, information needed for the position of n sequences is totally contained in the geometry of the n (CSHP, a_{ref}). This geometric stability is not observed with multiple alignments, which can be deeply modified by addition or removal of sequences. In the CSHP, protein position is unaltered by additions or removals of other proteins. In practice, the construction of CSHP is therefore completely deduced from any all-by-all protein sequence comparison [45,46] and can be easily updated.

The q-dissimilarity, a proximity notion for a geometric representation of the CSHP

In the CSHP, the definition of a distance *per se* based on mutual information is reduced *ad absurdum* (For demonstration, see methods). To define a proximity function i) sharing properties of distance, *i.e.* increasing when objects are further apart, ii) deriving from similarity and iii) relying on mutual information, particularly the property "$f(a,a) \neq f(b,b)$ is possible", we introduce a fourth notion of proximity. Such proximity was called *q-dissimilarity* (for quasi-dissimilarity), a function $f(a,b)$: $E \times E \rightarrow \Re^+$ is defined such that

$$\forall a \in E, f(a,a) = \min_{b \in E} f(a,b) \qquad (12)$$

$$\forall\ a \in E, \forall\ b \in E, f(a,b) = f(b,a) \qquad (13)$$

Let s be a similarity, then $q = e^{-s}$ is a q-dissimilarity, named the 'canonical q-dissimilarity' associated to s. Accordingly, the TULIP theorem allows a statistical characterization of $q(a,b)$ the canonical q-dissimilarity between two sequences a and b.

TULIP corollary 3

From the TULIP corollary 2, relation (14) is simply deduced:

$$P\{Q(a,b^*) \leq q(a,b)\} \leq \frac{1}{z(a,b^*)\dagger} \qquad (14)$$

with $Q(a,b^*)$ being the random q-dissimilarity variable associated with $S(a,b^*)$. Given a (CSHP, a_{ref}), each sequence b aligned with a is characterized by a q-dissimilarity $q(a,b)$. In geometric terms, b can be represented as a point contained in a hyper-sphere B of radius $q(a,b)$.

The representation of a (CSHP, a_{ref}) shown in Figure 1 is therefore in conformity with all constraints listed earlier and can also serve as a Venn diagram for the setting of events realized following a continuous random variable $Q(a,b^*)$. When a is compared to itself, it is set on a hypersphere A of radius $q(a,a)$, which is not reduced to one point. In the context of information theory, it is therefore possible to express that the proximity respects the property "$q(a,a) \neq q(b,b)$ is possible". Considering Figure 1, $P\{Q(a,b^*) \leq q(a,b)\} \leq \frac{1}{z(a,b^*)\dagger}$ is the probability for a random sequence b* to be in the hyper-sphere B. In conclusion, the *q-dissimilarity* is therefore a proximity notion that allows a rigorous geometric description of the configuration space of homologous proteins, real or simulated, (CSHP, a_{ref}, q).

Unification of pair-wise alignments theory, information theory, p-distance and q-dissimilarity in the CSHP model

A geometric space is a *topological* space when endowed with characterized *paths* that link its elements. Here, paths can be defined as the underlying evolutionary history separating sequences [11]. Given u the common unknown ancestor, then the divergence time $t(a,b)$ is theoretically the summed elapsed times separating u to a and to b. Without any empirical knowledge of u, the simplest approximation for $t(a,b)$ was sought as a function of the fraction of identical residues f_{id}, thus of the p-distance. With the hypothesis of the molecular clock, this function can be given as equation (2), where the transmutation of a and b is a consequence of a Poisson process. By using relation (9) on the equivalence between score similarity and mutual information, then the fundamental postulate "the closer in the evolution, the more alike and conversely, the more alike, *probably* the closer in the evolution" can be reformulated:

Fundamental postulate

Given two homologous proteins a and b, the closer in the evolution, the greater the mutual information between a and b (*i.e.* the optimal computed score $s(a,b)$) and conversely, the greater the mutual information between a and b, *probably* the closer in the evolution.

Whereas the first part of the postulate is a consequence of the conservational pressure on mutual information, the second assertion founds the historical reconstruction underlying a set of biological sequences on statistical concepts. A corollary is that evolution of two homologous proteins is characterized by a loss of mutual information.

In the CSHP, this formulation of the fundamental postulate allows a novel mathematical formalization of the p-distance in probabilistic terms. Basically, the p-distance is the divergence observed between two sequences *knowing* that they share some features (the observed sequences a and b) and that they were identical before the speciation event (sequence u).

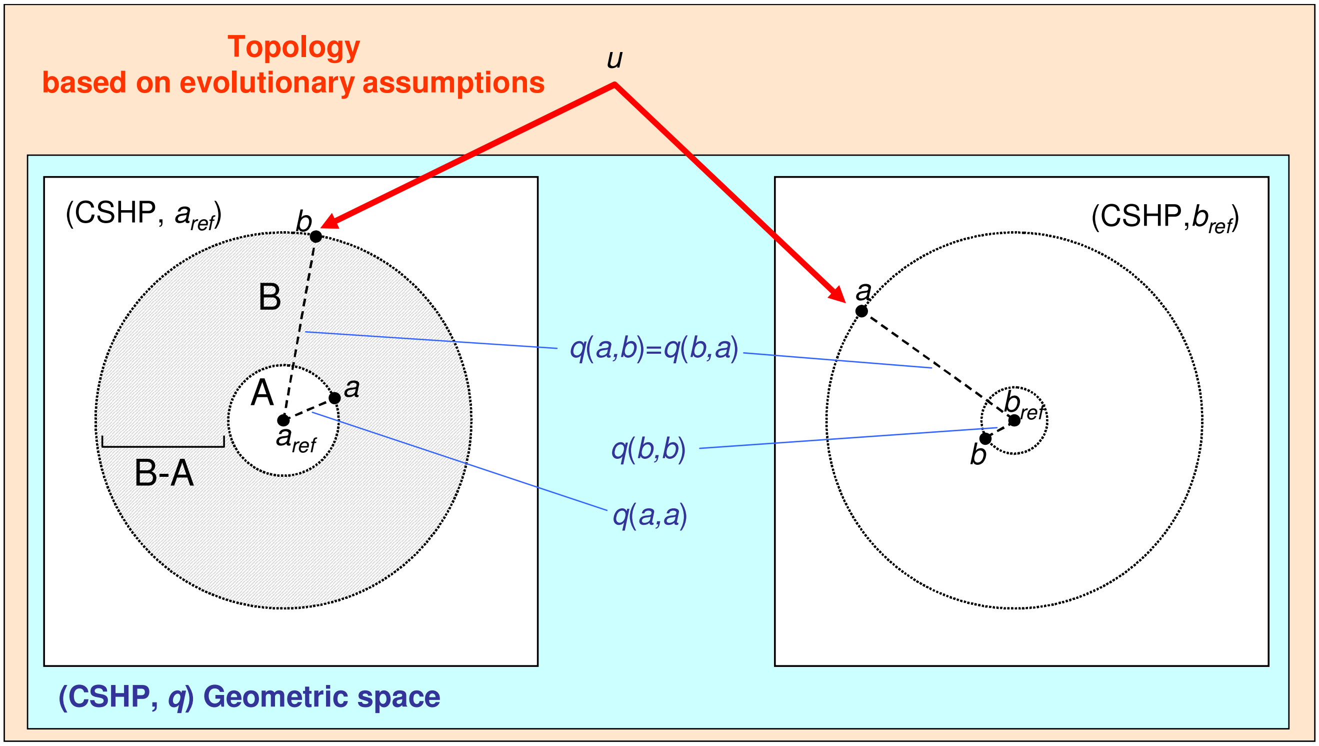


Figure 1
Geometric and probabilistic representation of the configuration space of homologous proteins (CSHP). For any sequence *a* taken as a reference (a_{ref}), one can build a configuration space (CSHP, a_{ref}) where all sequences that are homologous to *a* can be set. When two sequences *a* and *b* are aligned with a score *s*(*a*,*b*), then *b* is positioned in (CSHP, a_{ref}) and *a* in (CSHP, b_{ref}). The sequence alignment length determines the number of configuration dimensions; pair-wise amino acid scores determine the unique solution for its positioning. The q-dissimilarity ($q = e^{-s}$) defines a proximity between sequences allowing a geometric representation (CSHP, *q*). Remarkable properties are i) the conservation of mutual information, [$I(a;b) = I(b;a) \Rightarrow q(a,b) = q(b,a)$], between (CSHP, a_{ref}) and (CSHP, b_{ref}), ii) a probabilistic representation of homologies based on q-dissimilarities by Venn diagrams (A and B) and iii) the assignment of a topology relying on protein evolution assumptions. Evolutionary paths for *a* and *b* lineages, sharing an unknown ancestor *u*, have a probabilistic expression, bounded above (see text), supporting a phylogenetic topology (TULIP trees).

Looking back to equation (1), we can re-formulate f_{id} in probabilistic terms, considering the fraction of shared features (identical sites) *knowing* the observed data and the existence of a common ancestor. Given two proteins *a* and *b*, let us consider the random variable Q(a,b*), defined in TULIP corollary 3. In (CSHP, a_{ref}), shown in Figure 1, one can define the probability law $P\{Q(a,b^*)\leq\rho\}$ as the probability that the q-dissimilarity between b^* and a_{ref} is lower than ρ. The hyper-sphere of radius ρ contains therefore the b^* random sequences sharing informative features with *a* accordingly. The probability $p_{id/a}$ that b^* shares identity with *a*, *knowing* that the q-dissimilarity between b^* and *a* is lower than that between the real sequences *b* and *a*, is:

$$p_{id/a}(b^*) = P\{Q(a,b^*) \leq q(a,a) \,/\, Q(a,b^*) \leq q(a,b)\} \qquad (15)$$

which is a probabilistic expression of f_{id} in respect to the reference a_{ref}. According to the Venn diagram in Figure 1: $p_{id/a}(b^*) = P(A/B)$

Using the Bayes theorem, equation (15) can be expressed as:

$$p_{id/a}(b^*) = P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} \qquad (16)$$

In consequence:

$$p_{id/a}(b^*) = \frac{P\{Q(a,b^*) \leq q(a,a)\}}{P\{Q(a,b^*) \leq q(a,b)\}} \qquad (17)$$

which can be expressed as

$$p_{id/a}(b^*) = \frac{P\{S(a,b^*) \geq s(a,a)\}}{P\{S(a,b^*) \geq s(a,b)\}} \qquad (18)$$

Assuming that substitution rates are independent of lineages [35], then random sequence models a* and b* are equivalent, that is to say $Q(a,a^*) \approx Q(a,b^*)$ and

$$p_{id/a}(b^*) = \frac{P\{Q(a,b^*) \leq q(a,a)\}}{P\{Q(a,b^*) \leq q(a,b)\}} \approx \frac{P\{Q(a,a^*) \leq q(a,a)\}}{P\{Q(a,b^*) \leq q(a,b)\}} \qquad (19)$$

Thus $p_{id/a}$, and symmetrically $p_{id/b}$, provide a probabilistic expression of f_{id} *knowing* the data, *i.e.* the observed mutual information between a and b expressed as $Q(a,b)$.

Given two homologous sequences a and b, when their optimal score is $s(a,b) \geq \mu + \psi$ with ψ being a critical threshold value depending on the score distribution law (See Methods for the demonstration for the critical threshold), owing to the TULIP corollary 2, we can state that $p_{id/a}$ is bounded above:

$$p_{id/a}(b^*) \approx \frac{P\{S(a,a^*) \geq s(a,a)\}}{P\{S(a,b^*) \geq s(a,b)\}} \leq \frac{l_\upsilon(a,a^*)}{l_\upsilon(a,b^*)} = \frac{z(a,b^*)\dagger}{z(a,a^*)\dagger} \qquad (20)$$

This expression can also be developed as:

$$p_{id/a}(b^*) \leq \frac{\left(\dfrac{s(a,b)-\mu_1}{\sigma_1}\right)^2}{\left(\dfrac{s(a,a)-\mu_2}{\sigma_2}\right)^2} \qquad (21)$$

where μ_1, σ_1, μ_2 and σ_2 are the mean and the standard deviation of $S(a,b^*)$ and $S(a,a^*)$ respectively. The right term in relation (21) exhibits analogies with f_{id} given by equation (1), showing that the pragmatic approach by Feng and Doolittle [19] could be supported and generalized in a theoretical elaboration.

Using the Poisson correction, an expression of $t(a,b)$ is given as the linear combination of the two corrections of the p-distance deduced from $p_{id/a}$ and $p_{id/b}$:

$$t(a,b) = -[\log(p_{id/a}(b^*)) + \log(p_{id/b}(a^*))] \qquad (22)$$

with a^* and b^* the random variables corresponding to the shuffled sequences of a and b respectively. The sum of the logarithms corresponds to the product of the two probabilities, an expression of the hypothesis of independence of lineage. Interestingly, equation (22) provides an expression of the symmetric effect of time on the variations that independently affected a and b.

From relation (20), $t(a,b)$ appears as a function of Z-score ratios. For any set of homologous proteins, it is therefore possible to measure a table of pair-wise divergence times and build phylogenetic trees using distance methods.

Reconstruction of protein phylogeny: first example, case study of the glucose-6-phosphate isomerase phylogeny

We compared the trees we obtained, called TULIP trees, to phylogenetic trees built using classical methods, for instance the popular PHYLIP [47] or PUZZLE-based [48] methods, termed here MAB trees (for multiple alignment-based trees). Firstly, because MAB trees are constructed from multiple alignments, removals or additions of proteins modify the multiple alignments. Inclusion of sequences is considered as a way to improve the quality of multiple alignments and to increase the sensitivity of the comparison of distant sequences [49,50]. By contrast, the protein space used to build TULIP trees is not reordered when data sets are incremented or decremented (drawing of the TULIP tree may apparently change due to the tree graphic representation methods; nevertheless the absolute tree topology is not reordered). This remarkable property is due to both the geometrical construction by pairwise comparison and the convergence of the distance matrix elements estimated by equation (21). Indeed, the estimate of the right-hand term of equation (21) relies on a Monte Carlo method, after randomization of the biological sequences [39,44,51] and is therefore dependent on the sequence randomization model [52] and convergent in respect to the weak law of large numbers [53]. Convergence is proportional to $1/\sqrt{numb_{rand}}$, where $numb_{rand}$ is the number of randomizations. In the case studies presented here, we set num_{rand} = 2000 (see Methods). By contrast, stability of MAB trees is sought by bootstrapping approaches and consensus tree reconstruction. MAB trees appear as the result of a complex learning process including possible re-adjustment of the multiple alignments after eye inspection pragmatically applied to assist the reconstruction. Alternatively, Bayesian analyses have been recently proposed for phylogenetic inference [54], estimating posterior probability of each clades to assess most likely trees. Still, in a recent comparative study, Suzuki et al. [55] and Simmons et al. [56] provided evidence supporting the use of relatively conservative bootstrap and jacknife approaches rather than the more extreme overestimates provided by the Markov Chain Monte Carlo-based Bayesian methods. In the absence of any decisive methods to assess the validity of the trees obtained after such different approaches, no absolute comparison with the TULIP classification trees can be rigorously provided.

Whenever a TULIP classification was achieved on a dataset that led to a consensual MAB tree, both were always consistent. For example, Figure 2 shows the phylogenetic PHYLIP [47] and TULIP trees obtained for glucose-6-

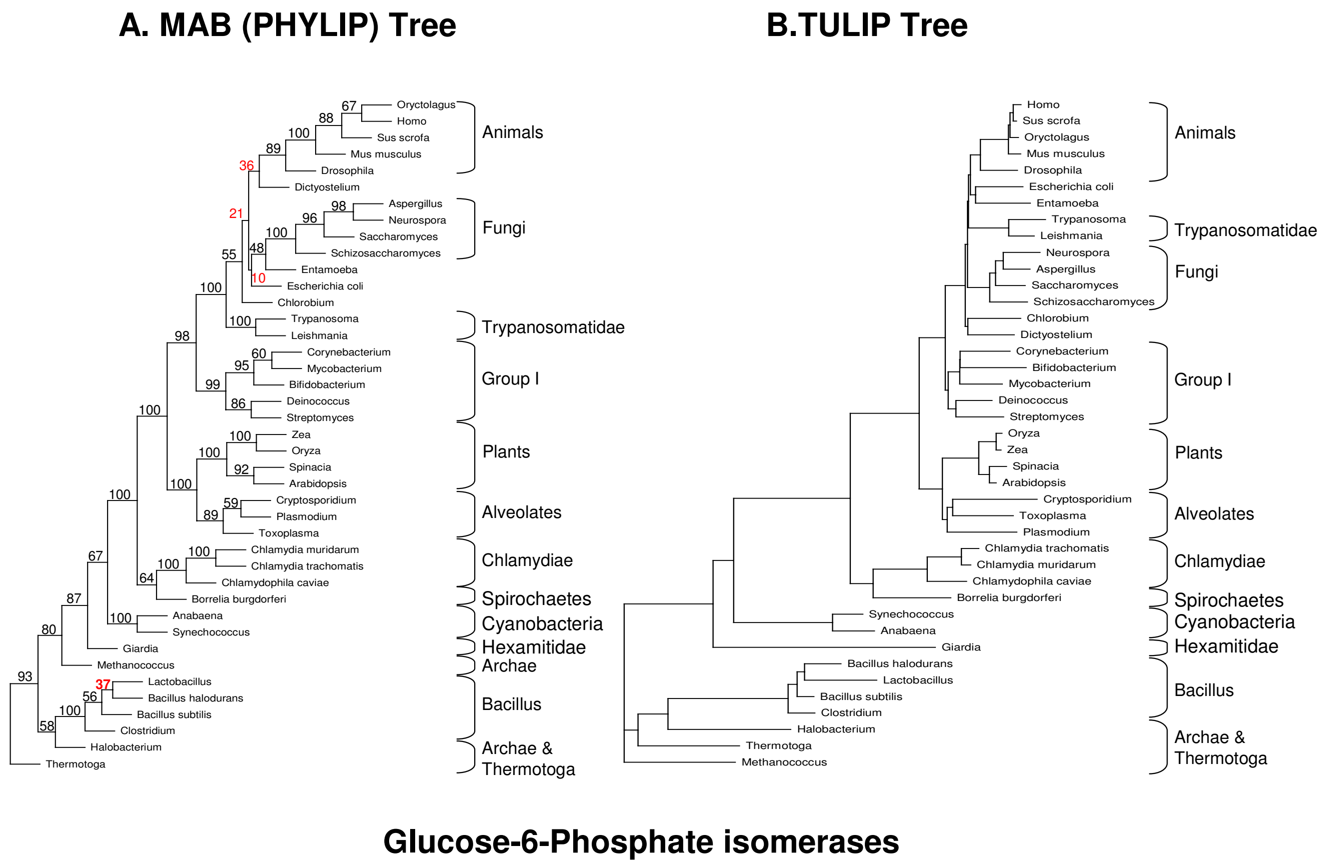
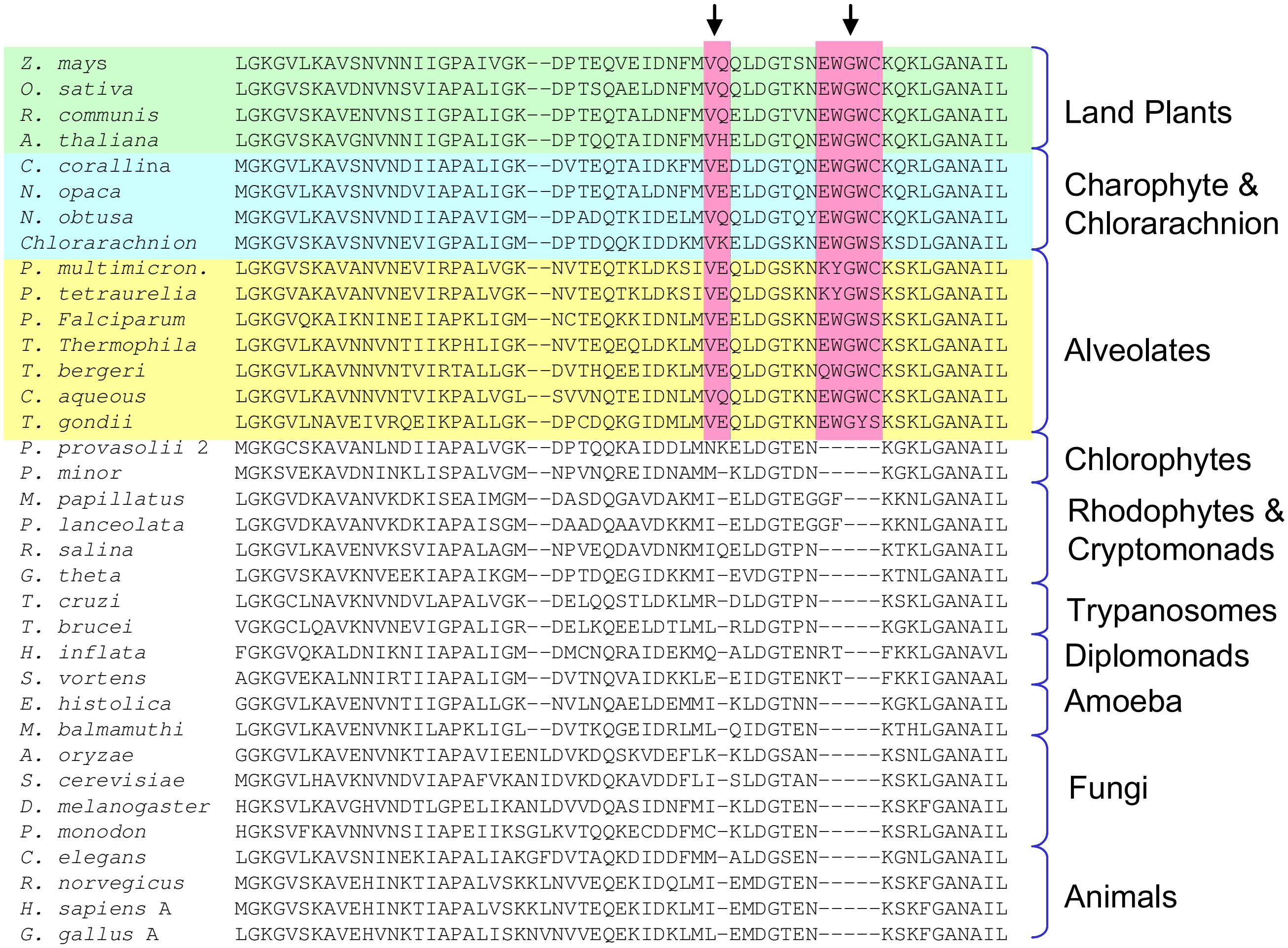


Figure 2
Glucose-6-phosphate isomerase phylogeny. **(A)** Multiple alignment based (MAB) tree. **(B)** TULIP tree. Both trees were constructed using the BLOSUM 62 similarity matrix. For MAB tree construction, bootstrap support was estimated using 1000 replicates. To build TULIP trees, Z-scores were estimated with 2000 sequence shuffling. Topology supported by high bootstrap results in the MAB tree (figures in black), are consistently recovered in the corresponding pair-wise alignement based TULIP tree.

phosphate isomerases (G6PI). Phylogeny of the G6PI enzyme has been studied by Huang et al. [57] in order to demonstrate the horizontal transfer of this enzyme in the apicomplexan phylum due to a past endosymbiosis [57]. Owing to the neighbor-joining analysis used by Huang et al. [57] (see methods) Figure 2A shows that apicomplexan G6PI is "plant-like". The TULIP tree shown in Figure 2B is consistent with this conclusion. Interestingly, differences between the two trees are found only when the bootstrap values on the MAB tree are not strong enough to unambiguously assess branching topology.

Reconstruction of protein phylogeny: second example, case study of the enolase phylogenic incongruence

TULIP classification tree further helps in solving apparent conflicting results obtained with MAB methods. In a comprehensive study from Keeling and Palmer [36] the PUZZLE-based reconstruction of the enolase phylogeny led to incongruent conclusions. Enolase proteins from a wide spectrum of organisms were examined to understand the evolutionary scenario that might explain that enolases from land plants and alveolates shared two short insertions. Alveolates comprise apicomplexan parasites, known to contain typical plant features as mentioned above, particularly a plastid relic. In this context, the shared insertion in apicomplexan and plant enolases (Figure 3) has been interpreted as a possible signature for



```
Z. mays          LGKGVLKAVSNVNNIIGPAIVGK--DPTEQVEIDNFMVQQLDGTSNEWGWCKQKLGANAIL
O. sativa        LGKGVSKAVDNVNSVIAPALIGK--DPTSQAELDNFMVQQLDGTKNEWGWCKQKLGANAIL
R. communis      LGKGVSKAVENVNSIIGPALIGK--DPTEQTALDNFMVQELDGTVNEWGWCKQKLGANAIL
A. thaliana      LGKGVSKAVGNVNNIIGPALIGK--DPTQQTAIDNFMVHELDGTQNEWGWCKQKLGANAIL
C. corallina     MGKGVLKAVSNVNDIIAPALIGK--DVTEQTAIDKFMVEDLDGTQNEWGWCKQRLGANAIL
N. opaca         MGKGVLKAVSNVNDVIAPALIGK--DPTEQTALDNFMVEELDGTQNEWGWCKQRLGANAIL
N. obtusa        MGKGVLKAVSNVNDIIAPAVIGM--DPADQTKIDELMVQQLDGTQYEWGWCKQKLGANAIL
Chlorarachnion   MGKGVSKAVSNVNEVIGPALIGM--DPTDQQKIDDKMVKELDGSKNEWGWSKSDLGANAIL
P. multimicron.  LGKGVSKAVANVNEVIRPALVGK--NVTEQTKLDKSIVEQLDGSKNKYGWCKSKLGANAIL
P. tetraurelia   LGKGVAKAVANVNEVIRPALVGK--NVTEQTKLDKSIVEQLDGSKNKYGWSKSKLGANAIL
P. Falciparum    LGKGVQKAIKNINEIIAPKLIGM--NCTEQKKIDNLMVEELDGSKNEWGWSKSKLGANAIL
T. Thermophila   LGKGVLKAVNNVNTIIKPHLIGK--NVTEQEQLDKLMVEQLDGTKNEWGWCKSKLGANAIL
T. bergeri       LGKGVLKAVNNVNTVIRTALLGK--DVTHQEEIDKLMVEQLDGTKNQWGWCKSKLGANAIL
C. aqueous       LGKGVLKAVNNVNTVIKPALVGL--SVVNQTEIDNLMVQQLDGTKNEWGWCKSKLGANAIL
T. gondii        LGKGVLNAVEIVRQEIKPALLGK--DPCDQKGIDMLMVEQLDGTKNEWGYSKSKLGANAIL
P. provasolii 2  MGKGCSKAVANLNDIIAPALVGK--DPTQQKAIDDLMNKELDGTEN-----KGKLGANAIL
P. minor         MGKSVEKAVDNINKLISPALVGM--NPVNQREIDNAMM-KLDGTDN-----KGKLGANAIL
M. papillatus    LGKGVDKAVANVKDKISEAIMGM--DASDQGAVDAKMI-ELDGTEGGF---KKNLGANAIL
P. lanceolata    LGKGVDKAVANVKDKIAPAISGM--DAADQAAVDKKMI-ELDGTEGGF---KKNLGANAIL
R. salina        LGKGVLKAVENVKSVIAPALAGM--NPVEQDAVDNKMIQELDGTPN-----KTKLGANAIL
G. theta         LGKGVSKAVKNVEEKIAPAIKGM--DPTDQEGIDKKMI-EVDGTPN-----KTNLGANAIL
T. cruzi         LGKGCLNAVKNVNDVLAPALVGK--DELQQSTLDKLMR-DLDGTPN-----KSKLGANAIL
T. brucei        VGKGCLQAVKNVNEVIGPALIGR--DELKQEELDTLML-RLDGTPN-----KGKLGANAIL
H. inflata       FGKGVQKALDNIKNIIAPALIGM--DMCNQRAIDEKMQ-ALDGTENRT---FKKLGANAVL
S. vortens       AGKGVEKALNNIRTIIAPALIGM--DVTNQVAIDKKLE-EIDGTENKT---FKKIGANAAL
E. histolica     GGKGVLKAVENVNTIIGPALLGK--NVLNQAELDEMMI-KLDGTNN-----KGKLGANAIL
M. balmamuthi    LGKGVLKAVENVNKILAPKLIGL--DVTKQGEIDRLML-QIDGTEN-----KTHLGANAIL
A. oryzae        GGKGVLKAVENVNKTIAPAVIEENLDVKDQSKVDEFLK-KLDGSAN-----KSNLGANAIL
S. cerevisiae    MGKGVLHAVKNVNDVIAPAFVKANIDVKDQKAVDDFLI-SLDGTAN-----KSKLGANAIL
D. melanogaster  HGKSVLKAVGHVNDTLGPELIKANLDVVDQASIDNFMI-KLDGTEN-----KSKFGANAIL
P. monodon       HGKSVFKAVNNVNSIIAPEIIKSGLKVTQQKECDDFMC-KLDGTEN-----KSRLGANAIL
C. elegans       LGKGVLKAVSNINEKIAPALIAKGFDVTAQKDIDDFMM-ALDGSEN-----KGNLGANAIL
R. norvegicus    MGKGVSKAVEHINKTIAPALVSKKLNVVEQEKIDQLMI-EMDGTEN-----KSKFGANAIL
H. sapiens A     MGKGVSKAVEHINKTIAPALVSKKLNVTEQEKIDKLMI-EMDGTEN-----KSKFGANAIL
G. gallus A      LGKGVSKAVEHVNKTIAPALISKNVNVVEQEKIDKLML-EMDGTEN-----KSKFGANAIL
```

Figure 3
Enolase phylogenic incongruence. When aligned, the enolase region corresponding to amino acids 73–118 of the *Oryza sativa* gene, exhibit two insertions (red boxes) that are only present in land plants, charophytes and alveolates. In alveolates, these insertions are consistent with a horizontal gene transfer. However, to date, evolutionary reconstructions based on enolase sequences did not allow any phylogenetic branch gathering for these clades [36].

some evolutionary relationship between apicomplexans and plants [58,59] and a likely sign of a lateral transfer. From the distribution of this insertion in enolases from several key eukaryotic groups, Keeling and Palmer [36] postulated that lateral transfer had been an important force in the evolution of eukaryotic enolases, being responsible for their origin in cryptomonads, *Chlorarachnion* and *Arabidopsis*. However, they could not conclude about alveolates, finding a conflict between the distribution of the insertion and the MAB phylogenetic position (Figure 4A). The authors had to admit that lateral gene transfers failed to explain apicomplexa enolases, and were compelled to suppose that the lack of congruence between insertion and phylogeny could be because of a parallel loss of insertions in lineages, or to more complex transfers of gene portions.

Based on our theoretical model, we constructed the corresponding TULIP tree. TULIP trees given with BLOSUM 62 or PAM 250 matrices, Fitch-Margoliash or neighbor-joining methods led indistinctly to a unique tree topology (Figure 4B). Separation of great phyla (Archaebacteria, Eubacteria, Diplomonads, Trypanasomes, Animals, Fungi and Amoeba) is recovered. A plant-like cluster is additionally reconstructed, in which a distinct separation occurred between {Rhodophytes ; Cryptomonads} and {Land

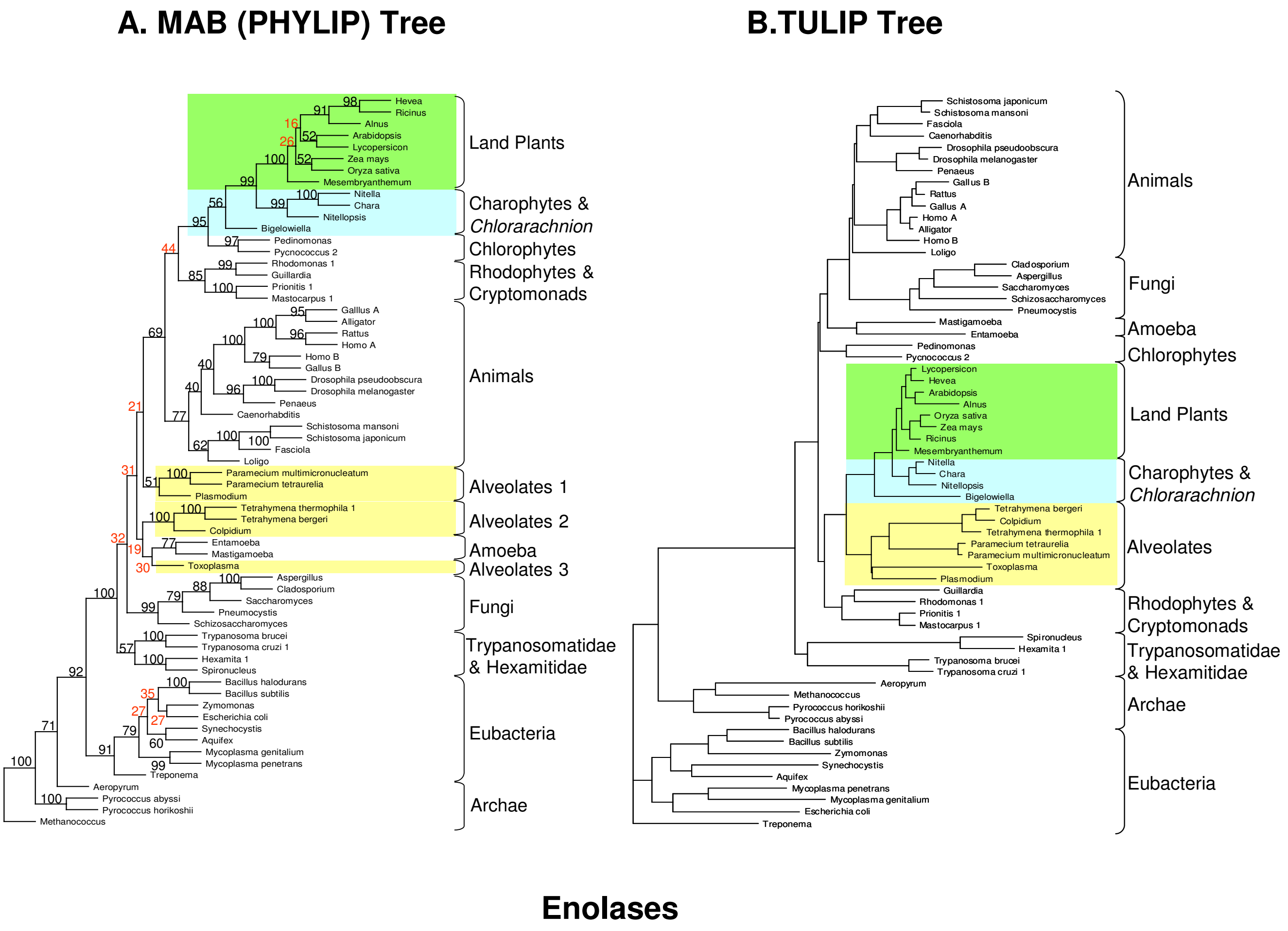


Figure 4
Solution of the enolase phylogenic incongruence. **(A)** Multiple alignment based (MAB) tree. **(B)** TULIP tree. Both trees were constructed using the BLOSUM 62 similarity matrix. For MAB tree construction, bootstrap support was estimated using 1000 replicates. To build TULIP trees, Z-scores were estimated with 2000 sequence shuffling. Clades that contain a unique insertional signature (Land plants, green box; Charophytes and Chlorarachnion, blue box; Alveolates; yellow box) are not gathered in the MAB tree, as previously reported [36]. By contrast, in the TULIP tree, the phylogeny of enolase proteins is reconciled with the insertional signature detection in Land plants, Charophytes, Chlorarachnion and Alveolates.

Plants ; Charophytes ; Chlorarachnion ; Alveolates} main clusters. It is remarkable that this latter cluster is that characterized by the enolase insertion.

This topology corresponds to the observed distribution of the enolase short insertions and provides therefore a solution to the apparent enolase phylogeny incongruence: the phylogenetic position of alveolates is not in conflict with the distribution of enolase insertion and the apicomplexa enolase is possibly a consequence of a lateral transfer, like in cryptomonads.

Large scale phylogeny based on a CSHP built from massive genomic pair-wise comparisons

A CSHP containing large sets of protein sequences can be built after any all-by-all massive comparison providing Z-score statistics. Because the space elaboration is explicit, then quality of the mutual information conservation depends on the choice of the scoring matrix, the geometric positioning depends on the local alignment method, the homology assessment depends on the alignment score and probabilistic cutoffs and the phylogenetic topology on the choice of the stochastic law correction. Eventually, genome-scale pair-wise comparisons [39,36] find in the present CSHP a robust, evolutionary consistent and easily updatable representation.

Methods

Glucose-6-Phosphate Isomerase sequences

The 41 Glucose-6-phosphate isomerase (EC 5.3.1.9) sequences studied in the paper are taken from several representative groups, as provided from the Swiss-prot database. Group I: Archae ([Swiss-prot:G6PI_HALN1], *Halobacterium* sp.; [Swiss-prot:G6PI_METJA], *Methanococcus jannaschii*). Group II: Bacteria Actinobacteria ([Swiss-prot:G6P1_STRCO], *Streptomyces coelicolor*; [Swiss-prot:G6PI_COREF, *Corynebacterium efficiens*; [Swiss-prot:G6PI_MYCTU], *Mycobacterium tuberculosis*). Group III: Bacteria Cyanobacteria ([Swiss-prot:G6PI_ANASP], *Anabaena sp.*; [Swiss-prot:G6PI_SYNEL], *Synechococcus elongates*). Group III: Bacteria Bacillus ([Swiss-prot:G6PI_LACFE], *Lactobacillus fermentum*; [Swiss-prot:G6PI_BACHD], *Bacillus halodurans*; [Swiss-prot:G6PI_BACSU], *Bacillus subtilis*; [Swiss-prot:G6PI_CLOPE], *Clostridium perfringens*). Group IV: Bacteria Proteobacteria ([Swiss-prot:G6PI_BIFLO], *Bifidobacterium longum*; [Swiss-prot:G6PI_ECOLI], *Escherichia coli*). Group V: Bacteria Chlamydiae ([Swiss-prot:G6PI_CHLTR], *Chlamydia trachomatis*; [Swiss-prot:G6PI_CHLCV], *Chlamydophila caviae*; [Swiss-prot:G6PI_CHLMU], *Chlamydia muridarum*). Group VI: Others Bacteria ([Swiss-prot:G6PI_CHLTE], *Chlorobium tepidum*; [Swiss-prot:G6PI_DEIRA], *Deinococcus radiodurans*; [Swiss-prot:G6PI_BORBU], *Borrelia burgdorferi*; [Swiss-prot:G6PI_THEMA], *Thermotoga maritime*). Group VII: Fungi ([Swiss-prot:G6PI_SCHPO], *Schizosaccharomyces pombe*; [Swiss-prot:G6PI_YEAST], *Saccharomyces cerevisiae*; [Swiss-prot:G6PI_NEUCR], *Neurospora crassa*; [Swiss-prot:G6PI_ASPOR], *Aspergillus oryzae*). Group VII: Eukaryota Viridiplantae ([Swiss-prot:G6PI_ARATH], *Arabidopsis thaliana*; [Swiss-prot:G6PI_MAIZE], *Zea mays*; [Swiss-prot:G6PI_SPIOL, *Spinacia oleracea*; [Swiss-prot:G6PA_ORYSA], *Oryza sativa*). Group VIII: Eukaryota Alveolata Apicomplexa ([Swiss-prot:G6PI_PLAFA], *Plasmodium falciparum*; [Swiss-prot:Q9XY88], *Toxoplasma Gondii*; [Swiss-prot:269_185], *Cryptosporidium parvum*). Group IX: Animals ([Swiss-prot:G6PI_DROME, *Drosophila melanogaster*; [Swiss-prot:G6PI_MOUSE], *Mus musculus*; [Swiss-prot:G6PI_HUMAN], *Homo sapiens*; [Swiss-prot:G6PI_PIG], *Sus scrofa*; [Swiss-prot:G6PI_RABIT], *Oryctolagus cuniculus*; [Swiss-prot:G6PI_TRYBB], *Trypanosoma brucei brucei*). Group X: Other Eukaryota ([Swiss-prot:AY581147], *Entamoeba histolytica*; [Swiss-prot:G6PI_LEIME], *Leishmania mexicana*; [Swiss-prot:AY581146], *Dictyostelium discoideum*; [Swiss-prot:Q968V7], *Giardia intestinalis*).

Enolase sequences

Enolase sequences used for the case-study presented in this paper were taken from eight major groups previously studied by [36]. Group I: Land Plant, Charophytes, Chlorophytes, Rhodophytes and Cryptomonads ([Swiss-prot:CAA39454], *Zea mays*; [Swiss-prot:Q42971], *Oryza sativa*; [Swiss-prot:Q43130], *Mesembryanthemum crystallinum*; [Swiss-prot:P42896], *Ricinus communis*; [Swiss-prot:Q43321], *Alnus glutinosa*; [Swiss-prot:Q9LEJ0], *Hevea brasiliensis* 1; [Swiss-prot:P25696], *Arabidopsis thaliana*; [Swiss-prot:P26300], *Lycopersicon esculentum*; [Swiss-prot:AF348914], *Chara corallina*; [Swiss-prot:AF348915], *Nitella opaca*; [Swiss-prot:AF348916], *Nitellopsis obtusa*; [Swiss-prot:AF348918], *Pycnococcus provasolii* 2; [Swiss-prot:AF348919], *Bigelowiella natans* – Chlorarachnion -; [Swiss-prot:AF348920], *Mastocarpus papillatus* 1; [Swiss-prot:AF348923], *Prionitis lanceolata* 1; [Swiss-prot:AF348931], *Rhodomonas salina* 1; [Swiss-prot:AF348933], *Guillardia theta*; [Swiss-prot:AF348935], *Pedinomonas minor*). Group II : Animals and Fungi ([Swiss-prot:P04764], *Rattus norvegicus*; [Swiss-prot:P51913], *Gallus gallus* A; [Swiss-prot:P07322], *Gallus gallus* B; [Swiss-prot:Q9PVK2], *Alligator mississippiensis*; [Swiss-prot:P06733], *Homo sapiens* A; [Swiss-prot:P13929, *Homo sapiens* B; [Swiss-prot:P15007], *Drosophila melanogaster*; [Swiss-prot:AF025805], *Drosophila pseudoobscura*; [Swiss-prot:O02654], *Loligo pealeii*; [Swiss-prot:AF100985], *Penaeus monodon*; [Swiss-prot:Q27527], *Caenorhabditis elegans*; [Swiss-prot:Q27877], *Schistosoma mansoni*; [Swiss-prot:P33676], *Schistosoma japonicum*; [Swiss-prot:Q27655], *Fasciola hepatica*; [Swiss-prot:P00924], *Saccharomyces cerevisiae* 1; [Swiss-prot:Q12560], *Aspergillus oryzae*; [Swiss-prot:P42040], *Cladosporium herbarum*; [Swiss-prot:P40370], *Schizosaccharomyces pombe* 1; [Swiss-prot:AF063247], *Pneumocystis carinii* f.). Group III: Amoebae ([Swiss-prot:P51555], *Entamoeba histolytica*; [Swiss-prot:Q9U615], *Mastigamoeba balamuthi*). Group IV: Alveolates ([Swiss-prot:AF348926], *Paramecium multimicronucleatum*; [Swiss-prot:AF348927], *Paramecium tetraurelia*; [Swiss-prot:AF348928], *Colpidium aqueous*; [Swiss-prot:AF348929], *Tetrahymena thermophila* I; [Swiss-prot:AF348930], *Tetrahymena bergeri*; [Swiss-prot:Q27727], *Plasmodium falciparum*; [Swiss-prot:AF051910], *Toxoplasma gondii*). Group V: *Trypanosomatidae* ([Swiss-prot:AF159530], *Trypanosoma cruzi* eno1 partial; [Swiss-prot:AF152348], *Trypanosoma brucei* complete). Group VI: Hexamitidae ([Swiss-prot:AF159519], *Hexamita inflata* eno1 partial; [Swiss-prot:AF159517], *Spironucleus vortens* partial). Group VII: Archaebacteria ([Swiss-prot:Q9UXZ0], *Pyrococcus abyssi*; [Swiss-prot:O59605], *Pyrococcus horikoshii*; [Swiss-prot:Q60173], *Methanococcus jannaschii*; [Swiss-prot:Q9Y927], *Aeropyrum pernix*). Group VII: Eubacteria ([Swiss-prot:O66778], *Aquifex aeolicus*; [Swiss-prot:P37869],*Bacillus subtilis*; [Swiss-prot:Q9K717], *Bacil-*

lus halodurans; [Swiss-prot:P77972], *Synechocystis* sp.; [Swiss-prot:P33675], *Zymomonas mobilis*; [Swiss-prot:P08324], *Escherichia coli*; [Swiss-prot:P47647], *Mycoplasma genitalium*; [Swiss-prot:Q8EW32, *Mycoplasma penetrans*; [Swiss-prot:P74934], Treponema pallidum).

Demonstration that distance of a protein to itself cannot be defined in the CSHP

In the simplest case, building a distance between amino acids (that would lead to distance between sequences) on the basis of computed similarity values would have to respect the condition:

$$\forall i \in E, \forall j \in E, d(i,j) = 0 \Rightarrow i = j \qquad \text{(a)}$$

for *i* and *j*, two given amino acids and *d* the distance function. Using this condition in the proposition, any organization of the CSHP with a geometric distance is reduced *ad absurdum*.

Proposition

Building a distance between amino acids derived from the composed function $d(i,j) = (\phi \bigcirc s)(i,j)$, where *s* is a similarity function and ϕ a bijection, is impossible without a loss of mutual information. Moreover, two proteins from distinct organisms can have the same configuration, being like "twins", and $d(i,j) = 0$ does not imply $i = j$.

Proof

Condition (a) implies that $\phi(s(i,i)) = \phi(s(j,j)) = 0$. This equality imposes that $s(i,i) = s(j,j)$ and, following equation (7) of main text, that $I(i;i) = I(j;j)$. Considering for example tryptophan (W) and glutamic acid (E), if W occurs in a sequence, the mutual information gained about the occurrence of W at the aligned position would be the same as that gained in the case of E about the occurrence of E at the aligned position in the homologous protein. This statement is easily rejected on the basis of biochemical concerns. On one hand, aspartic acid (D) shares common biochemical properties with E, particularly a carboxylic acid, and easily substitutes in homologous sequences. By contrast W, exhibiting a unique biochemical feature, is less substitutable without altering the function. Thus the mutual information $I(E;E)$ is necessarily lower than $I(W;W)$. This that can be checked in scoring matrices such as BLOSUM 62 [30] where $I(E;E) = 5$, $I(D;D) = 6$ and $I(W;W) = 11$. Condition $d(i,i) = 0$ leads to an obvious loss of information. The second assertion of the proposition is obvious.

Determination of the threshold value ψ, for topological reconstructions in the CHSP based on pair-wise alignment score probabilities

An important basis of the reconstruction of a probabilistic evolutionary topology in the CSHP is based on the demonstration that, given *S* the random variable corresponding to the alignment scores of pairs of shuffled sequences and μ the mean of *S*, given two homologous sequences *a* and *b*, when their optimal score is $s(a,b) \geq \mu + \psi$ (with ψ a critical threshold value depending on the score distribution law), owing to the TULIP corollary 2, we can state that $p_{id/a}$ is bounded above

$$p_{id/a}(b^*) \approx \frac{P\{S(a,a^*) \geq s(a,a)\}}{P\{S(a,b^*) \geq s(a,b)\}} \leq \frac{l_\upsilon(a,a^*)}{l_\upsilon(a,b^*)} = \frac{z(a,b^*)\dagger}{z(a,a^*)\dagger}$$

To the purpose of this demonstration, we considered the cumulative distribution function $F(s) = P(S \leq s)$, its derivative $f(s)$ known as the probability density function defined as $dF(s) = f(s)ds$, and the positive delta function $\delta(s) = (s - \mu)^2(1 - F(s))$. Since $\delta(s) = (s - \mu)^2(1 - F(s))$ is null for $s = \mu$ and $\lim_{s \to +\infty} \delta(s) = 0$, the Rolle's theorem implies that $\exists s_0 \in]\mu,+\infty[$ such as $\frac{\partial \delta}{\partial s}(s_0) = 0$ [60]; s_0 corresponds to a maximum of $\delta(s)$ and is therefore the solution of the equation

$$2(1 - F(s)) - (s - \mu) f(s) = 0 \qquad \text{(b)}$$

one can express as

$$s - \mu = \frac{2(1-F(s))}{f(s)} \qquad \text{(c)}$$

The $\frac{2(1-F(s))}{f(s)}$ term corresponds to a continuous function. [illegible]ly, $\phi(s) = \frac{f(s)}{(1-F(s))}$ is known as the hazard function [61], th[illegible]lity of *s*, per score unit (*i.e.* mutual information), conditional to the fact that the pair-wise alignment score is *at least* equal to *s*. The hazard [illegible] defined by $\phi(s) = \lim_{ds \to 0} \frac{P(s \leq S \leq s + ds | S \geq s)}{ds}$. A critical hypothesis is $2(1-F(s))$ 2 [illegible]sing and conversely that [illegible]tly decreasing. Considering $\psi = \lim_{s \to \mu} \frac{2}{\phi(s)} = \frac{2(1-F(\mu))}{f(\mu)}$, equation 2 has only one [illegible] is bounded above:

$$s_0 = \mu + \frac{2(1-F(s_0))}{f(s_0)} \leq \mu + \psi \qquad \text{(d)}$$

In consequence, $\delta(s)$ reaches its maximum for a s_0 ($s_0 \leq \mu + \psi$) and it is strictly decreasing on $]\mu + \psi,+\infty[$.

The estimation of s_0 is not trivial because it depends on the knowledge of the cumulative distribution function. Extensive studies provided experimental and theoretical supports for an extreme value distribution of alignment scores [31,43,44]. Using the extreme value distribution of

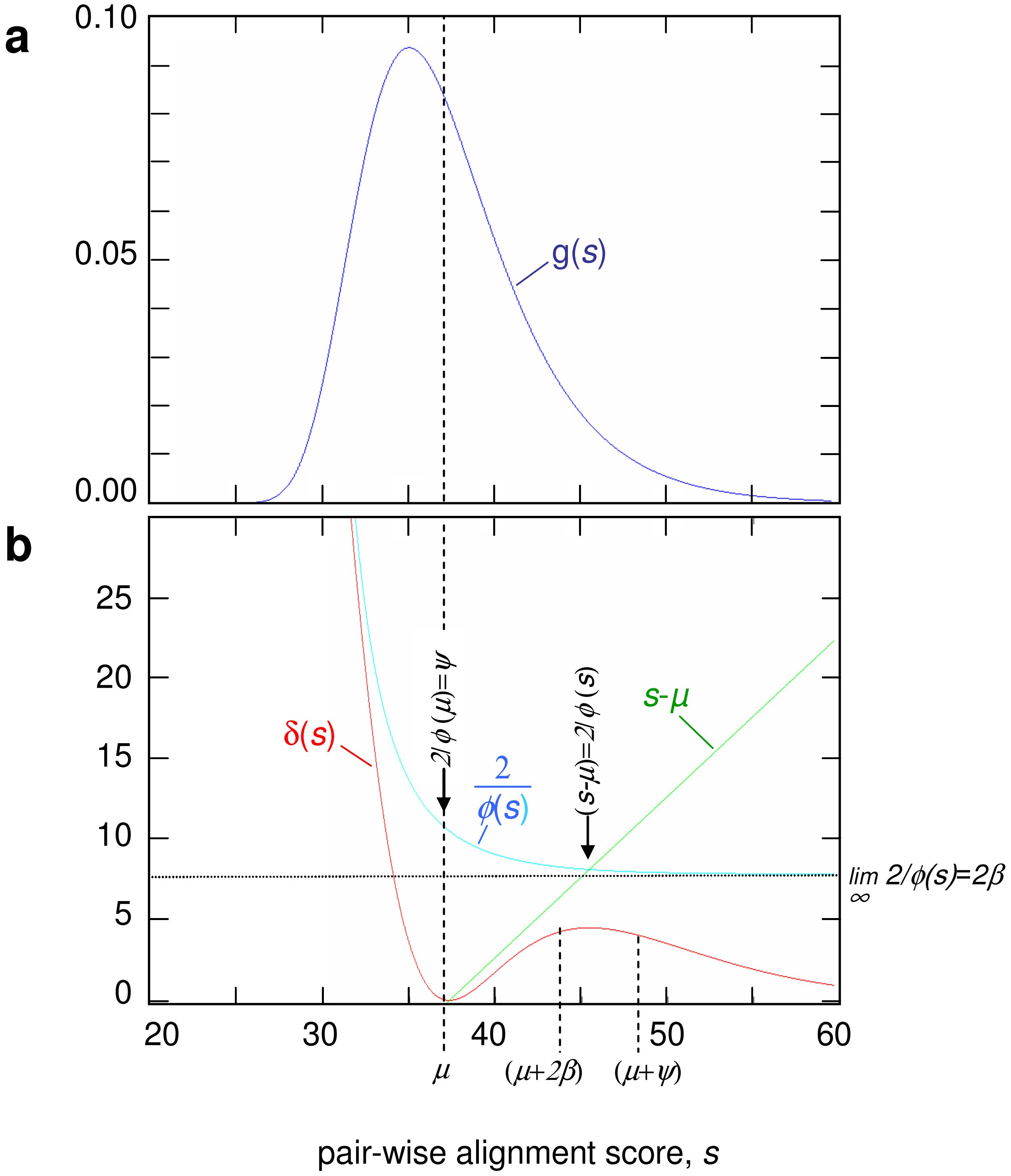


Figure 5
(A) Gumbel score distribution simulated for enolases used in the present paper **(B)** graphical determination of ψ

type I, *i.e.* the Gumbel distribution [62], the cumulative distribution is given by

$$G(s) = P(s \le S) = \exp\left\{-\exp(-\frac{s-\theta}{\beta})\right\} \qquad \text{(e)}$$

with θ and β ($\beta > 0$) the location and scale parameters. The probability density function $g(s)$ is defined by $dG(s) = g(s)ds$. We observe with $\varepsilon(s) = -\exp(-\frac{s-\theta}{\beta})$ that $\lim_{s\to+\infty} \varepsilon(s) = 0$. Using the Taylor's polynomial formula, *i.e.* $\exp x \underset{0}{\approx} 1 + x$:

$$\lim_{s\to+\infty} \frac{2}{\phi(s)} = \lim_{\varepsilon\to 0} \frac{2(1-\exp\varepsilon)}{-\frac{1}{\beta}\varepsilon\exp\varepsilon} = \lim_{\varepsilon\to 0} \frac{2(1-1-\varepsilon)}{-\frac{1}{\beta}\varepsilon(1+\varepsilon)} = \lim_{\varepsilon\to 0} \frac{2}{\frac{1}{\beta}+\frac{\varepsilon}{\beta}} = 2\beta \qquad \text{(f)}$$

In consequence, for a Gumbel score probability distribution:

$$\mu + 2\beta \le \mu + \frac{2(1-G(s_0))}{g(s_0)} \le \mu + \psi \qquad \text{(g)}$$

A graphical determination of ψ from a Gumbel distribution is illustrated in Figure 5.

If a pair-wise alignment score of two sequences *a* and *b* is relatively high, that is $s(a,b) \ge \mu + \psi$, then the trivial inequality $s(a,a) \ge s(a,b)$ implies

$$(s(a,b) - \mu)^2(1 - F(s(a,b))) \ge (s(a,a) - \mu)^2(1 - F(s(a,a))) \qquad \text{(h)}$$

that is to say

$$\frac{P\{S(a,a^*) \ge s(a,a)\}}{P\{S(a,b^*) \ge s(a,b)\}} \le \frac{(s(a,b)-\mu)\dagger}{(s(a,a)-\mu)\dagger} = \frac{z(a,b^*)\dagger}{z(a,a^*)\dagger} \qquad \text{(i)}$$

From inequality (i), we deduce that $p_{id/a}$ is bounded above.

Construction of PHYLIP multiple alignment based trees and pair-wise alignment based TULIP trees

To build PHYLIP trees, multiple sequence alignments were created with ClustalW [63]. PHYLIP trees where constructed using the protpars and neighbor modules from the PHYLIP package [47] and the BLOSUM 62 substitution matrix. Bootstrap support was estimated using 1000 replicates. To build TULIP trees, for each couple of sequences *a* and *b*, alignment was achieved with the Smith-Waterman method and the BLOSUM 62 scoring matrices, using the BIOFACET package from Gene-IT, France [64]. We computed estimated z-scores $z(a,b^*)$, $z(a,a^*)$, $z(a^*,b)$, $z(b^*,b^*)$, with 2000 sequence shuffling. For all computations, an estimation of the Gumbel parameters θ and β was made using the computed μ and σ of any $S(a,b^*)$ and the formula $\beta = \frac{\sigma}{\pi}\sqrt{6}$ and $\theta = \mu - \beta\Gamma'(1)$, where $\Gamma'(1) \approx 0.577216$ is the Euler constant. In all computations, both Gumbel parameters were very close (in the case of enolases, $mean(\theta) = 35.04$, $SD(\theta) = 0.12$, $mean(\beta) = 3.92$, $SD(\beta) = 0.08$). As a consequence, the assumption $Q(a,a^*) \approx Q(a,b^*)$ was verified for any pair of sequences. We used the parameters to estimate $\mu = \theta + \beta\Gamma'(1)$ (in the case of enolases, $\mu = 37.33$), and $\mu + \psi \approx \mu + 10.5178 \approx 47.85$. As any pairs of computed scores are higher than this critical threshold, we used relation [20]. Estimation of evolutionary time was achieved according to equations [20] and [22]. Trees were constructed using Fitch-Margoliash and Neighbor-Joining methods [47].

List of abbreviations

CSHP, configuration space of homologous proteins, TULIP, theorem of the upper limit of a score probability

Authors' contributions

OB conceived the main theoretical model, designed and developed the method to build phylogenetic trees and drafted the manuscript. SR and PO participated in the theoretical model refinement and in the design and development of computational methods to build TULIP trees. EM contributed to the conception of this study, participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank John Dunlop for copy editing and Joël Bleuse and Jacques Demongeot for critical reading of this manuscript. We are indebted to Jean-Paul Eynard, Jean-Michel Pabiot and Jean-Jacques Codani for supportive help. Part of this work was financed by Ministère de la Recherche et de la Technologie and by Agence Nationale de la Valorisation de la Recherche Rhône-Alpes.

References

1. Zuckerkandl E, Pauling L: **Molecules as documents of evolutionary history.** *J Theor Biol* 1965, **8:**357-366.
2. Zukerkandl E: **The evolution of hemoglobin.** *Sci Am* 1965, **212:**110-118.
3. Fitch WM, Margoliash E: **Construction of phylogenetic trees.** *Science* 1967, **155:**279-284.
4. Arnheim N, Taylor CE: **Non-Darwinian evolution: consequences for neutral allelic variation.** *Nature* 1969, **223:**900-903.
5. Dayhoff MO: **Computer analysis of protein evolution.** *Sci Am* 1969, **221:**86-95.
6. Arnheim N, Steller R: **Multiple genes for lysozyme in birds.** *Arch Biochem Biophys* 1970, **141:**656-661.
7. DeLange RJ, Smith EL: **Histones: structure and function.** *Annu Rev Biochem* 1971, **40:**279-314.
8. Zuckerkandl E: **Some aspects of protein evolution.** *Biochimie* 1972, **54:**1095-102.
9. Dayhoff MO, Barker WC, McLaughlin PJ: **Inferences from protein and nucleic acid sequences: early molecular evolution, divergence of kingdoms and rates of change.** *Orig Life* 1974, **5:**311-330.

10. Wu TT, Fitch WM, Margoliash E: **The information content of protein amino acid sequences.** *Annu Rev Biochem* 1974, **43:**539-566.
11. Brocchieri L: **Phylogenetic inferences from molecular sequences: review and critique.** *Theor Popul Biol* 2001, **59:**27-40.
12. Singer GA, Hickey DA: **Nucleotide bias causes a genomewide bias in the amino acid composition of proteins.** *Mol Biol Evol* 2000, **17:**1581-1588.
13. Bastien O, Lespinats S, Roy S, Metayer K, Fertil B, Codani JJ, Maréchal E: **Analysis of the compositional biases in Plasmodium falciparum genome and proteome using Arabidopsis thaliana as a reference.** *Gene* 2004, **336:**163-173.
14. Doolittle RF: **Similar amino acid sequences: chance or common ancestry?** *Science* 1981, **214:**149-159.
15. Otu HH, Sayood K: **A new sequence distance measure for phylogenetic tree construction.** *Bioinformatics* 2003, **19:**2122-2130.
16. Jukes TH, Cantor CR: *Mammalian Protein Metabolism* New York: Academic Press; 1969.
17. Kimura M: **A simple model for estimating evolutionary rates of base substitiutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980, **16:**111-120.
18. Lake JA: **Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances.** *Proc Natl Acad Sci USA* 1994, **91:**1455-1459.
19. Feng DF, Doolittle RF: **Converting amino acid alignment scores into measures of evolutionary time: a simulation study of various relationships.** *J Mol Evol* 1997, **44:**361-370.
20. Camin J, Sokal R: **A method for deducing branching sequences in phylogeny.** *Evolution* 1965, **19:**311-326.
21. Fitch WM: **Toward defining the course of evolution: minimum change for a specific tree topology.** *Syst Zool* 1971, **35:**406-416.
22. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981, **17:**368-376.
23. Felsenstein J, Churchill GA: **A hidden Markov model approach to variation among sites in rate of evolution.** *Mol Biol Evol* 1996, **13:**93-104.
24. Salemi M, Vandamme AM: *The Phylogenetic Handbook* Cambridge University Press; 2003.
25. Feng DF, Cho G, Doolittle RF: **Determining divergence times with a protein clock: update and reevaluation.** *Proc Natl Acad Sci USA* 1997, **94:**13028-13033.
26. Nei M, Xu P, Glazko G: **Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms.** *Proc Natl Acad Sci USA* 2001, **98:**2497-2502.
27. Doolittle RF, Feng DF, Tsang S, Cho G, Little E: **Determining divergence times of the major kingdoms of living organisms with a protein clock.** *Science* 1996, **271:**470-477.
28. Dayhoff MO, Barker WC, Hunt LT: **Establishing homologies in protein sequences.** *Methods Enzymol* 1983, **91:**524-545.
29. Risler JL, Delorme MO, Delacroix H, Henaut A: **Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix.** *J Mol Biol* 1988, **204:**1019-1029.
30. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89:**10915-10919.
31. Waterman MS: *Introduction to computational biology* CRC Press; 1995.
32. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol* 1970, **48:**443-453.
33. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147:**195-197.
34. Fitch WM: **Random sequences.** *J Mol Biol* 1983, **163:**171-176.
35. Grishin NV: **Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites.** *J Mol Evol* 1995, **41:**675-679.
36. Keeling PJ, Palmer JD: **Lateral transfer at the gene and subgenic levels in the evolution of eukaryotic enolase.** *Proc Natl Acad Sci USA* 2001, **98:**10745-10750.
37. Hartley RVL: **Transmission of Information.** *The Bell System Technical Journal* 1928, **3:**535-564.
38. Shannon CE: **A Mathematical Theory of Communication.** *The Bell System Technical Journal* 1948, **27:**379-423.
39. Bastien O, Aude JC, Roy S, Maréchal E: **Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics.** *Bioinformatics* 2004, **20:**534-537.
40. Dayhoff MO, Schwartz RM, Orcutt BC: **A Model of Evolutionary Change in Proteins.** *Atlas of Protein Sequence and Structure* 1978, **5:**345-352.
41. Setubal J, Meidanis J: *Introduction to Computational Molecular Biology* PWS Publishing Compagny; 1997.
42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
43. Karlin S, Altschul SF: **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.** *Proc Natl Acad Sci USA* 1990, **87:**2264-2268.
44. Comet JP, Aude JC, Glemet E, Risler JL, Henaut A, Slonimski PP, Codani JJ: **Significance of Z-value statistics of Smith-Waterman scores for protein alignments.** *Comput Chem* 1999, **23:**317-331.
45. Bacro JN, Comet JP: **Sequence alignment: an approximation law for the Z-value with applications to databank scanning.** *Comput Chem* 2001, **25:**401-410.
46. Louis A, Ollivier E, Aude JC, Risler JL: **Massive sequence comparisons as a help in annotating genomic sequences.** *Genome Res* 2001, **11:**1296-1303.
47. Felsenstein J: **PHYLIP – Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164-166.
48. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18:**502-504.
49. Thompson JD, Plewniak F, Poch O: **A comprehensive comparison of multiple sequence alignment programs.** *Nucleic Acids Res* 1999, **27:**2682-2690.
50. Simmons MP, Freudenstein JV: **The effects of increasing genetic distance on alignment of, and tree construction from, rDNA internal transcribed spacer sequences.** *Mol Phylogenet Evol* 2003, **26:**444-451.
51. Manly BFJ: *Randomization, Bootstrap and Monte Carlo Methods in Biology* CRC Press; 1997.
52. White S: **Global statistics of protein sequences: implications for the origin, evolution, and prediction of structure.** *Annu Rev Biophys Biomol Struct* 1994, **23:**407-439.
53. Capinski M, Kopp E: *Measure, Integral and Probability* New-York: Springer-Verlag; 1999.
54. Rannala B, Yang Z: **Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference.** *J Mol Evol* 1996, **43:**304-311.
55. Suzuki Y, Glazko GV, Nei M: **Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics.** *Proc Natl Acad Sci U S A* 2002, **99:**16138-16143.
56. Simmons MP, Pickett KM, Miya M: **How meaningful are Bayesian support values?** *Mol Biol Evol* 2004, **21:**188-199.
57. Huang J, Mullapudi N, Lancto CA, Scott M, Abrahamsen MS, Kissinger JC: **Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in Cryptosporidium parvum.** *Genome Biol* 2004, **5:**R88.
58. Read M, Hicks KE, Sims PF, Hyde JE: **Molecular characterisation of the enolase gene from the human malaria parasite Plasmodium falciparum. Evidence for ancestry within a photosynthetic lineage.** *Eur J Biochem* 1994, **220:**513-520.
59. Dzierszinski F, Popescu O, Toursel C, Slomianny C, Yahiaoui B, Tomavo S: **The protozoan parasite Toxoplasma gondii expresses two functional plant-like glycolytic enzymes. Implications for evolutionary origin of apicomplexans.** *J Biol Chem* 1999, **274:**24888-24895.
60. Lang S: *Undergraduate analysis* New-York: Springer-Verlag; 1997.
61. Valleron AJ: *Introduction à la Biostatistique* Paris: Masson; 1998.
62. Coles S: *An introduction to Statistical Modeling of Extreme Values* New-York: Springer-Verlag; 2001.
63. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
64. Codani JJ, Comet JP, Aude JC, Glémet E, Wozniak A, Risler JL, Hénaut A, Slonimski PP: **Automatic analysis of large-scale pairwize alignments of protein sequences.** *Methods in Microbiology* 1999, **28:**229-244.