

Detecting Site-Specific Biochemical Constraints Through Substitution Mapping

Julien Dutheil

Received: 29 January 2008 / Accepted: 9 June 2008 / Published online: 12 August 2008
© Springer Science+Business Media, LLC 2008

Abstract The neutral theory of molecular evolution states that most mutations are deleterious or neutral. It results that the evolutionary rate of a given position in an alignment is a function of the level of constraint acting on this position. Inferring evolutionary rates from a set of aligned sequences is hence a powerful method to detect functionally and/or structurally important positions in a protein. Some positions, however, may be constrained while having a high substitution rate, providing these substitutions do not affect the biochemical property under constraint. Here, I introduce a new evolutionary rate measure accounting for the evolution of specific biochemical properties (e.g., volume, polarity, and charge). I then present a new statistical method based on the comparison of two rate measures: a site is said to be constrained for property X if it shows an unexpectedly high conservation of X knowing its total evolutionary rate. Compared to single-rate methods, the two-rate method offers several advantages: it (i) allows assessment of the significance of the constraint, (ii) provides information on the type of constraint acting on each position, and (iii) detects positions that are not proposed by previous methods. I apply this method to a 200-sequence data set of triosephosphate isomerase and report significant cases of positions constrained for polarity, volume, or charge. The three-dimensional localization of these positions shows that they

are of potential interest to the molecular evolutionist and to the biochemist.

Keywords Constraint · Protein · Phylogeny · Substitution mapping · Markov model

Introduction

The level of constraint—the number of distinct amino acids acceptable at a particular position—varies along protein sequences. Positions located in an active site, for instance, are directly involved in the function of the protein, and allow for only a few types of residues. Other typically constrained positions include sites controlling the conformation of the protein, e.g., sites involved in the hydrophobic core of globular proteins. Conversely, positions less critical from a structural or functional point of view are less constrained, like some positions belonging to external loops, for instance (Goldman et al. 1998). From a molecular evolutionary point of view, most of the mutations occurring at a constrained position are deleterious, and hence removed from the population genetic pool, resulting in a low rate of amino acid substitution. Unconstrained positions experience a higher rate because most mutations do not affect the structure and/or function of the molecule and, therefore, have a chance to increase in frequency in the population and reach fixation (Kimura 1983).

Comparing homologous sequences from extant species, different constraint levels hence result in different variability patterns. Unconstrained positions show a high variability, whereas highly constrained positions are conserved across species. The reciprocal of this property has important practical implications: since the variability of a position is directly related to its level of constraint, one may measure

J. Dutheil
Institut des Sciences de l'Évolution (UM2-CNRS),
Université Montpellier 2, Place Eugène Bataillon, CC064,
34 095 Montpellier Cedex 5, France

J. Dutheil (✉)
BiRC—Bioinformatics Research Center, University of Aarhus,
Høegh-Guldbergs Gade 10, Building 1090, 8000 Aarhus C,
Denmark
e-mail: jdutheil@daimi.au.dk

this variability to infer structurally and/or functionally important positions from a set of homologous sequences (Lichtarge et al. 1996; Pupko et al. 2002; Mayrose et al. 2005). Using the raw site variability index is nonoptimal because different evolutionary processes can lead to indistinguishable patterns. For instance, a low site variability may result from either a low substitution rate or a high rate with several reverse mutations. The usual interpretation is that only slow-evolving sites are a signature of evolutionary constraint. Several methods have been developed to account for phylogenetic relationships between sequences when detecting functionally important positions (Mayrose et al. 2004), the most recent ones using state-of-the-art evolutionary models and Bayesian inference (Yang 1994; Lichtarge et al. 1996; Pupko et al. 2002). All these methods disregard fast-evolving sites as potentially constrained. Such positions, however, may contain relevant constraint information, potentially even more than slow-evolving positions. Consider, for instance, a position which is constrained for the volume of the residue, that is, a position which needs to be small for the protein to be well conformed. All conservative mutations of the kind “small residue” → “small residue” will be neutral, whereas nonconservative mutations of the kind “small” → “large” will be deleterious. Positions showing a slow “volume evolutionary rate” are hence also of potential interest to the biochemist. Besides slow positions, which have a slow volume evolution by definition, one may be interested in detecting positions with many changes but a conserved volume. Such positions are informative since they reflect a constraint pattern, but also carry information on the biochemical nature of the constraint, which is not the case for fully constrained sites.

Here I introduce a new method to detect positions in an alignment that show significant conservation of a given biochemical property, fully accounting for the underlying phylogeny and rate of evolution. This method is based on a new evolutionary measure that enables assessment of the evolutionary rate of a biochemical property using standard models of amino acid substitution. Using a triosephosphate isomerase (TIM) example data set, I show that this method provides new information on constrained positions and is a useful complement to existing methods based only on raw evolutionary rates.

Materials and Methods

A New Method for Detecting Constrained Sites

Outline

We aim at detecting sites (positions) at which a given biochemical property has been conserved during evolution,

even though the total amino acid substitution rate is not particularly low. We hence need two rate measures: one for all substitutions and one for substitutions affecting the biochemical property of interest. I introduce a new measure of evolutionary rate based on the mapping of substitution events onto a phylogeny (Dutheil et al. 2005; Dutheil and Galtier 2007). Probabilistic substitution mapping provides a way to estimate, for each site, the average number of substitutions that occurred on each branch of the tree, accounting for uncertainty about ancestral states (Nielsen 2002; Dutheil et al. 2005). Branch-specific estimates of numbers of substitutions are stored in a site-specific “substitution vector” (Fig. 1), which summarizes the substitution history of the site (Dutheil et al. 2005), and can be computed using standard phylogenetic models (Felsenstein 2004; Yang, 2006). The norm of the substitution vector (noted N in the following) is a measure of global evolutionary rate: the lower the norm, the lower the number of substitutions that occurred at this particular position. Substitution vectors are generalized by weighting amino acids changes according to their biochemical properties (Dutheil and Galtier 2007) (Fig. 1). One may, for instance, give a positive weight to substitutions modifying the volume of the residue and ignore substitutions that do not affect it (and similarly for polarity, charge, etc.). Using the norm of the weighted substitution vector (noted N^* in the following) hence provides a measure of the evolutionary rate for a given biochemical property.

By contrasting the different rates of evolution, one may gain information on the constraints acting across sites. Of particular interest are sites with a high value of N (high substitution rate) and a low weighted norm N^* (conserved biochemical property). The two norms are correlated and linearly linked. We want to exhibit significant outliers from this correlation. The statistic to be used is hence the distance of each point from the (model II) regression line, i.e., the coordinate on the second axis of a principal component

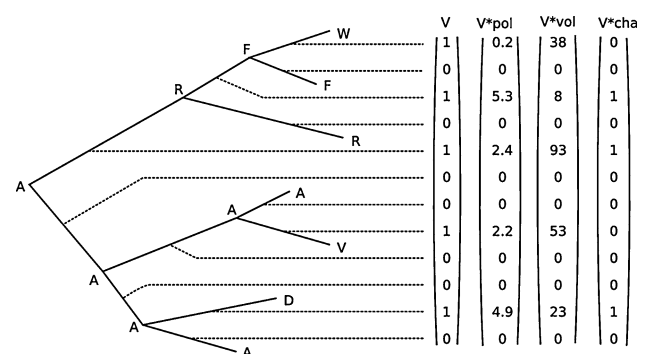


Fig. 1 Substitution vectors. V: simple substitution vector. V*: weighted substitution vectors, according to polarity (pol), volume (vol), or charge (cha). For the sake of simplicity, the uncertainty over ancestral states is not taken into account here. True substitution vectors, however, are averaged over all possible ancestral states, according to their likelihood

analysis (Sokal and Rohlf 1995). The significance is assessed thanks to simulations, and correction for multiple testing is performed.

Algorithm

The method takes as input a set of aligned sequences and proceeds in several steps. It is exemplified using a data set of TIM sequences. The corresponding results are discussed under Results.

Step 1: reconstruct phylogeny and estimate parameters. A phylogenetic tree and other evolutionary parameters are estimated from the data using a maximum likelihood approach (Felsenstein 2004; Yang 2006). A model of evolution is used, including a Markov model of substitutions accounting for different rates of change between amino acids (Felsenstein 1981) and a discretized gamma law modeling the differences in site-specific rates (Yang 1994). It is also possible to use as input a tree taken from the literature. This tree and estimated parameters (noted Θ) are used in the following steps.

Step 2: estimate substitution vectors. We note $V_i = (v_{i,1}, \dots, v_{i,k}, \dots, v_{i,m})$ the substitution vector for site i , m being the total number of branches in the phylogenetic tree. All substitution numbers for each site i and each branch k are then computed using the formula:

$$v_{i,k} = \sum_c \sum_x \sum_y \Pr(x, y, r_c | D_i, \hat{\Theta}) \times w_{x,y} \tag{1}$$

where x and y are the states at the bottom and top nodes of branch k , and r_c is the relative rate associated with class c . The first factor in eq. 1 is deduced from Bayes' theorem:

$$\begin{aligned} \Pr(x, y, r_c | D_i, \hat{\Theta}) &= \frac{\Pr(x, y, r_c, D_i | \hat{\Theta})}{\Pr(D_i | \hat{\Theta})} \\ &= \frac{\Pr(x, y, D_i | \hat{\Theta}, r_c) \times \Pr(r_c)}{\Pr(D_i | \hat{\Theta})}. \end{aligned} \tag{2}$$

$\Pr(x, y, r_c | D_i, \hat{\Theta})$ is the joint probability for the pair (x, y) . It is computed as a ratio of likelihoods (Yang et al. 1995): the denominator is the likelihood for site i , computed using Felsenstein's (1981) recursion. The numerator of eq. 2 is the likelihood of site i computed by fixing states x and y at the top and bottom nodes of branch k , and $\Pr(r_c)$ is the prior probability of rate class c . The second factor in eq. 1, $w_{x,y}$, is a measure of change over branch k , with initial state x and final state y . Several expressions of $w_{x,y}$ are possible. When mapping all substitutions, $w_{x,y}$ is the number of jumps of the Markov chain during time $t \times r_c$. This number depends only on the substitution model, and can be computed analytically (Dutheil et al. 2005). A useful approximation consists in ignoring multiple substitutions events by taking:

$$w_{x,y} \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases} \tag{3}$$

This procedure is referred to as the “simple” substitution mapping. $w_{x,y}$ can also be set to measure the difference between the two states x and y , according to a given biochemical property. $W = \{w_{x,y}\}$ is then a symmetrical weight matrix containing absolute differences in the property index of interest. Any weighting scheme of interest may be used, like the one in the AAindex database (Kawashima and Kanehisa 2000). In this study we used simple weights accounting for:

- difference of volume (as defined by Grantham [1974]; AAIndex1 entry GRAR740103);
- difference of polarity (as defined by Grantham [1974]; AAIndex1 entry GRAR740102); and
- difference of charge (K and R have +1, D and E have -1, 0 for the others; AAIndex1 entries FAUJ880111 and FAUJ880112);

Step 3: compute rates for each site i . The norm N_i of substitution vectors V^i is computed using the formula

$$N_i = \left(\sum_k v_{i,k}^2 \right)^{1/2} \tag{4}$$

In the following, the N_i notation refers to norms of simple substitution vectors, and N_i^* to weighted substitution vectors. N_i is linearly linked to the posterior rate estimate, as computed following Pupko et al. (2002) (Fig. 2, left). Using the approximation in eq. 3 appears to underestimate the highest rates, although the relationship remains monotonous and increasing (Fig. 2, right).

Step 4: estimate the null distribution. One hundred thousand sites are simulated according to the estimated phylogeny. The simulation process uses the same model as the one estimated from the real data. All sites are simulated independently, and without any constraint.

The corresponding substitution vectors are estimated, and their norms (N and N^*) computed (Fig. 3a). All resulting points are plotted in the (N, N^*) coordinate system and a principal component analysis is performed. Let (A, B) be the coordinate system defined by the resulting principal components. Coordinates of observed points (real data) are then evaluated in (A, B) , and their coordinates on the second axis b are recorded (Fig. 3b). If (A, B) is positively rotated from the first one, then negative values of b correspond to sites with an evolutionary rate of the biochemical property slower than expected knowing their total substitution rate. The p -value for a given site is then defined as

$$p = \Pr(b \leq b_{obs}) = (N_1 + 1)/(N_2 + 1) \tag{5}$$

where N_2 is the total number of simulated points and N_1 is the number of simulated sites showing a value of b lower

Fig. 2 Relationship between the norm of substitution vectors and the posterior rate of sites. *Left* Norm of the substitution vectors. *Right* Norm computed while ignoring multiple substitution events

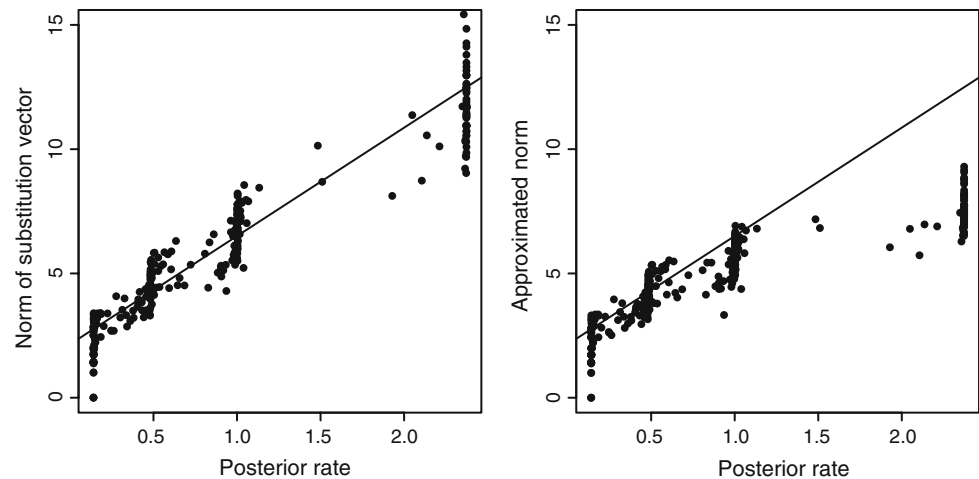
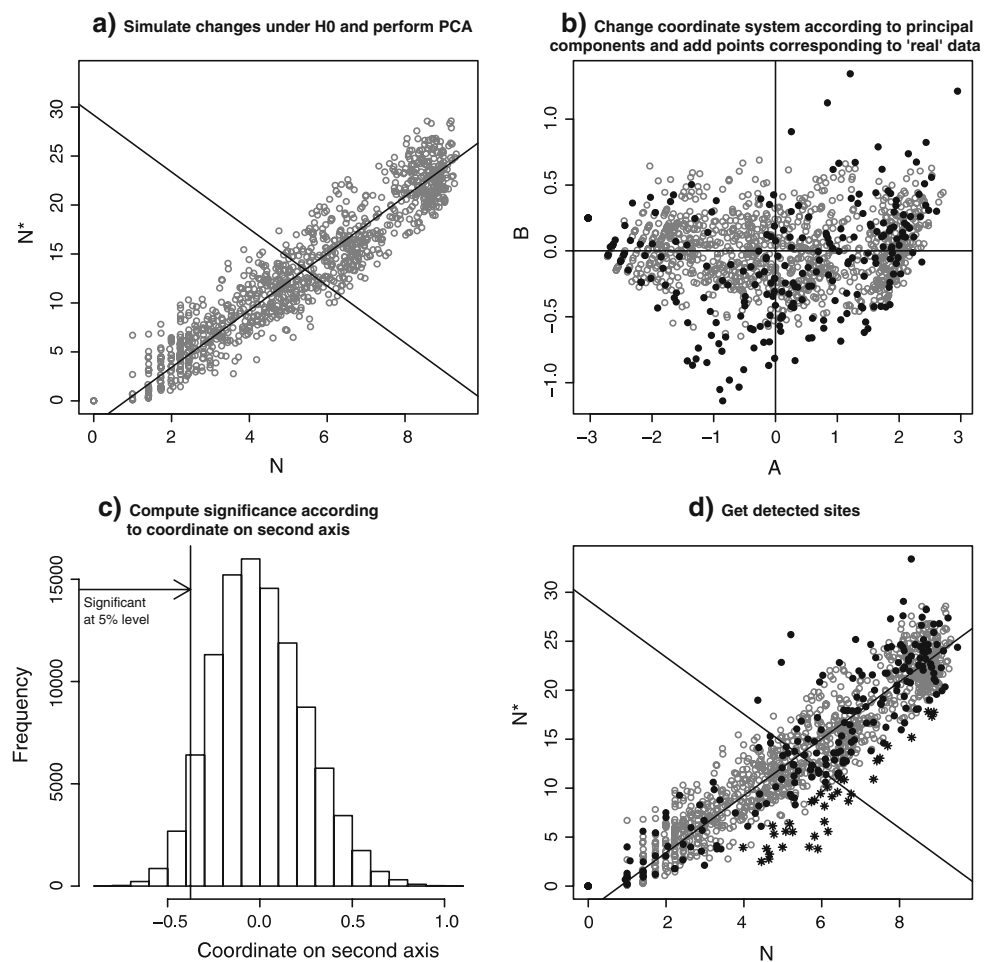


Fig. 3 Detecting significant constrained sites: principle of the method. **a, b, c, d:** four steps of the detection procedure (see text for details). *N*, simple norm; *N**, weighted norm; *A*, first component axis; *B*, second component axis. Significant sites (5% level after correction for multiple testing) are labeled with an asterisk in **(d)**. Data and simulations are from the TIM data set. The biochemical property used here is the amino acid polarity



than b_{obs} , the observed value of b for the position tested (Fig. 3c).

Step 5: correct for multiple testing. When trying to detect constrained sites, one must test every candidate position, leading to the multiple testing problem. The false discovery rate (FDR) method was used (Benjamini and Hochberg 1995) to correct for this issue (Fig. 3d). The

p -values of all unit tests are ordered by ascending order, $p_1 \leq p_2 \leq \dots \leq p_i \leq \dots \leq P_n$, n being the number of positions tested. For a global false discovery rate of α , we define k as the largest i verifying the relation

$$p_i \leq \alpha \times i/n \quad (6)$$

and consider all tests from 1 to k as significant.

Data Sets and Technical Details

PFAM entry PF00121 was used as the initial data set. It contained 979 sequences of TIM. Identical sequences were removed. The alignment was inspected by eye, and badly aligned sequences were removed, leaving 474 sequences of bacteria, archaea, and eukaryotes. To save calculation time, 200 sequences were then sampled and used for the analysis. A phylogenetic tree was built using the PhyML program (Guindon and Gascuel 2003), using the dcmut protein substitution model (Kosiol and Goldman 2005) and a gamma distribution with four rate classes (Yang 1994). PhyML-estimated parameters (tree topology and branch lengths + α parameter of the gamma distribution) were used in the constraint analysis. To estimate the distribution of the b statistics under the null hypothesis, 100,000 independent, unconstrained sites were simulated using the estimated phylogeny and parameters (parametric bootstrap approach).

The 1TIM structure was retrieved from the Protein Data Bank, together with the corresponding DSSP file from the CMBI server (<ftp://ftp.cmbi.ru.nl/pub/molbio/data/dssp>). Blanks in the “structure” section were replaced by ‘C’ (coil). The molscript (Kraulis 1991) and raster3d programs (Merritt and Bacon 1997) were used for making pictures of protein structures.

All statistical analyses were conducted with the R statistical software (R Development Core Team 2006), using the ade4 package (Chessel et al. 2004). (Weighted) substitution mapping and simulations were performed using the Bio ++ libraries (Dutheil et al. 2006). A program named ConTest (Constraint Testing) was written and is available at <http://home.gna.org/contest/>, together with the data set used in this article.

Results

The new detection method was applied to a data set of 200 sequences of TIM, to detect positions constrained for volume, polarity, or charge. TIM (EC 5.3.1.1) is an enzyme involved in glycolysis. It catalyzes the interconversion of the two triosephosphate isomers, dihydroxyacetone phosphate and D-glyceraldehyde 3-phosphate. The structure involves a height-stranded α/β barrel, containing the active site with prosite pattern

[AVG]-[YLV]-E-P-[LIVMEPKST]-[WYEAS]-[SAL]-[IV]-[GN]-[TEKDVS]-[GKNAD].
 163 164 165 166 167 168 169 170 171 172 173

Position numbers refer to the 1TIM PDB entry, which was used for three-dimensional localization. The loop near the active site is called “flexible” since it has no electron

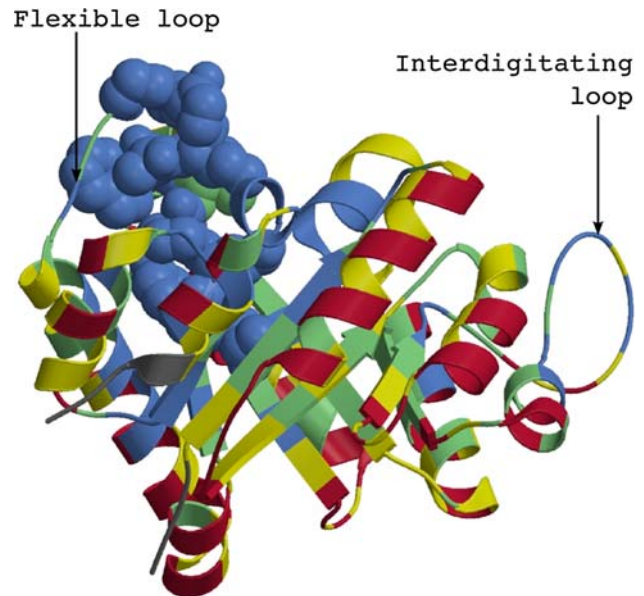


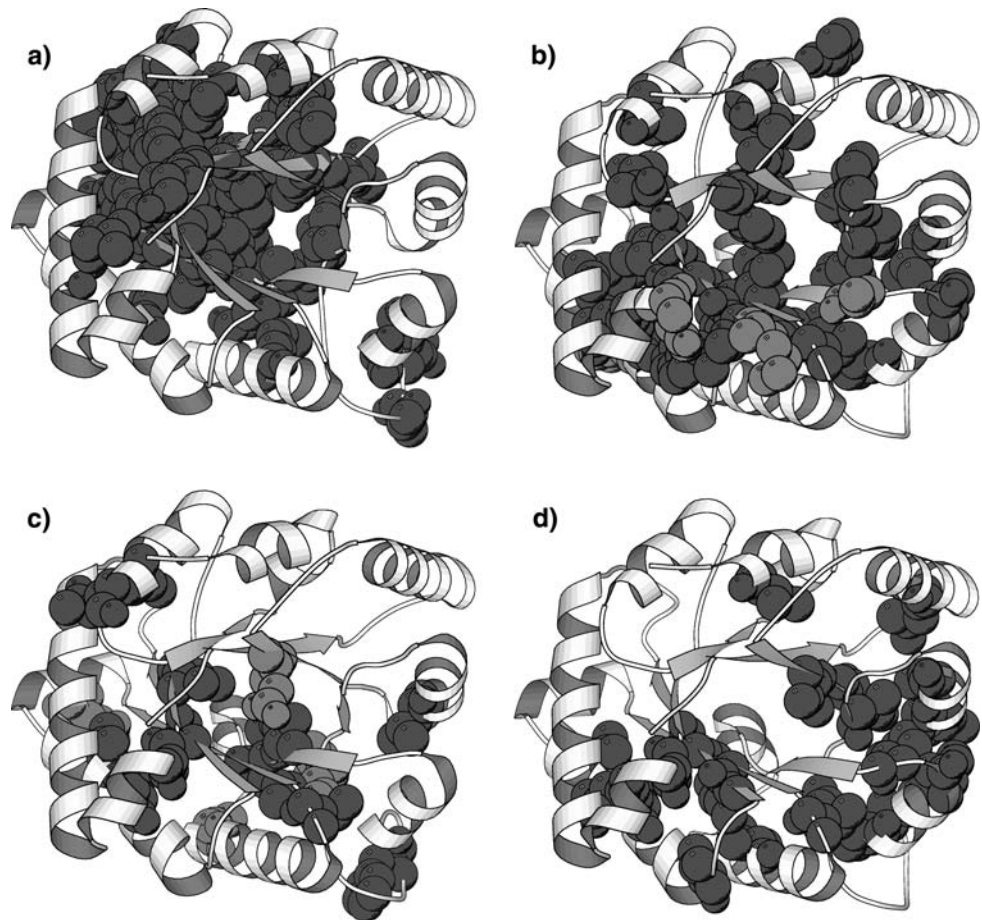
Fig. 4 Evolutionary rate heat map, in the following order, from slow to fast-evolving positions: blue, green, yellow, red. Full structure is shown for residues in the active site. PDB entry: 1TIM

density (Lolis et al. 1990). It undergoes a conformational change when fixing the substrate. Furthermore, the TIM enzyme is a homodimer, the two subunits interacting in the region of the interdigitating loop.

The evolutionary rate of each position was computed using the empirical Bayesian approach (Pupko et al. 2002) for comparison with the method presented here. Contrary to the newly introduced constraint detection method, the evolutionary rate measurement does not make it possible to perform any significance test. This method typically produces a three-dimensional heat map, with residues colored according to their rate (Fig. 4) (Glaser et al. 2003). Posterior rate appears to be positively correlated with solvent accessibility (Kendall’s $\tau = 0.37$, $p < 1e-6$), a general trend in globular proteins (Goldman et al. 1998). The user then has to set a threshold to exhibit conserved sites. For instance, a threshold of 0.11 yields 46 sites, from zero change (totally conserved sites) up to sites with 12 substitutions along the phylogeny (Fig. 5a). These sites contain all the positions in the active site, excepted VAL167, which is the most variable in the prosite pattern. Other slowly evolving sites appear to be mostly clustered around the active site region and the “flexible loop,” but also in the interdigitating loop (Figs. 4 and 5a).

A 5% false discovery rate was used for the constraint detection method. Detected positions are shown in Fig. 6. Thirty-five positions were detected as constrained for their

Fig. 5 Localization of constrained sites. (a) Positions with a rate lower than 0.11. (b) Polarity constrained sites: dark gray, nonpolar residues; light gray, polar residues. (c) Volume-constrained residues: dark gray, large residues; light gray, small residues. (d) Charge constrained residues: dark gray, neutral. PDB entry: 1TIM



polarity level, among which 31 are hydrophobic and 4 polar. Significantly constrained hydrophobic sites are buried, whereas constrained polar sites are located in external turns (see Fig. 5b). Detected hydrophobic sites are mainly located in the α/β -barrel, with a two-amino-acid periodicity (Fig. 7), corresponding mainly to residues with a lateral chain pointing inward to the wall of the barrel. Other detected nonpolar sites are located in amphiphilic helices, also pointing inward to the wall of the barrel, and facing residues of the β -sheet. It is noteworthy that position VAL167 in the active site is detected as constrained for its hydrophobicity.

Seventeen positions are detected as constrained for their volume, among which 12 conserve a large residue (Fig. 5c). Seven of them were also detected as constrained for their hydrophobicity. Position PHE74 is located in the flexible loop, and ILE92, VAL143, ILE207, and GLN223 are near the active site. Small constrained residues show no particular location. It is noteworthy, however, that position ASN65, constrained for a small volume, is in contact with position ILE92, constrained for a large volume.

Finally, 23 positions are significantly constrained for their charge—or, actually, for their absence of charge (Fig. 5d). It appears that charged residues are either slowly

evolving (and in most cases totally conserved) or not conserved at all. No $D \Leftrightarrow E$ or $K \Leftrightarrow R$ pattern was observed. Thirteen sites constrained to a neutral charge are also detected as constrained for their hydrophobicity, and one is also constrained for being large. VAL154 and VAL113 are both located in helices, and are in contact. Other positions are buried and clustered near the region of contact between the two subunits.

Discussion

A New Evolutionary Rate Measure

The use of the norm of the substitution vectors, while sharing all the benefits of the usual empirical Bayesian approach (full accounting of phylogeny, different substitution rates, and rate variation across sites), offers several new advantages. It is less sensitive to the prior distribution of rates, and precise estimates can be obtained with only a few rate classes. Compared to the usual rate, the newly introduced measure is not relative to the global gene rate, making the measure more similar in spirit to the parsimony score. The possibility of weighting substitution vectors

Fig. 6 Sites significant at the 5% level after correction for multiple testing

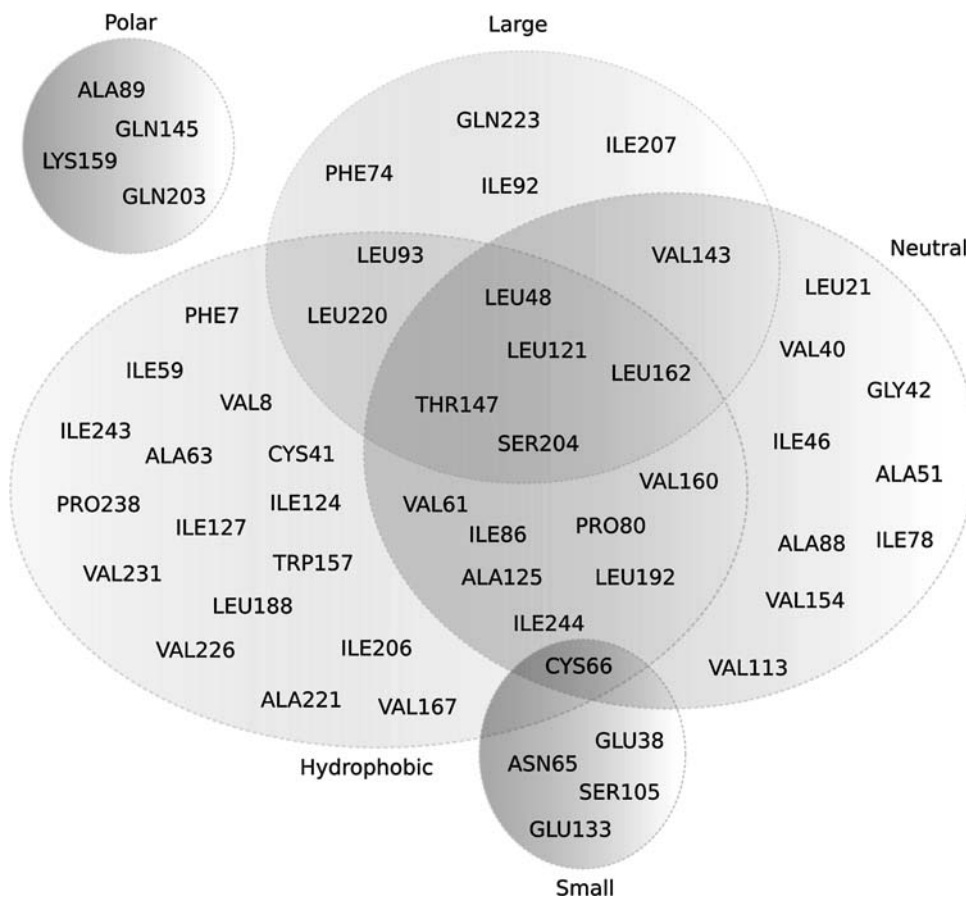
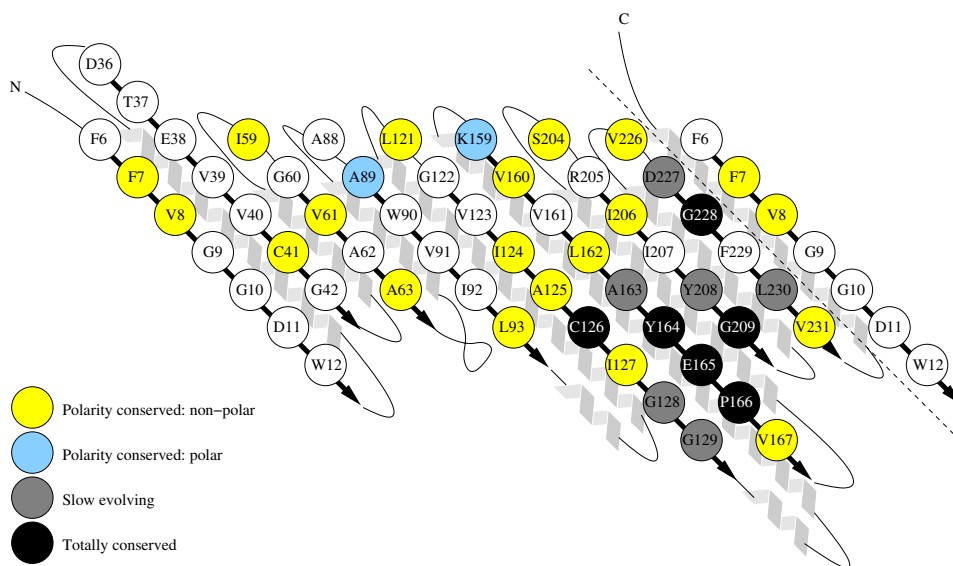


Fig. 7 Conserved and polarity-constrained sites in the β -barrel of triosephosphate isomerase. Black arrows represent β -strands



with potentially any biochemical property of interest provides the unique opportunity to assess the evolutionary rate of such properties.

The new measure is model-based and, hence, sensitive to violation of model assumptions. It has previously been shown that substitution mapping is quite robust regarding the substitution model used, and also regarding the

underlying phylogeny, as long as it is sensible. Another concern is the independence of sites. The substitution maps and their norms are constructed independently for each site, whereas some of them may have undergone correlated evolution, resulting in correlated norms. Whereas this phenomenon is expected to have little effect on the mapping procedure itself, this is of greater importance when

accounting for multiple testing. The correlation between single site tests is expected to be positive, and the FDR method used here was shown to correct for multiple testing in this case (Benjamini and Yekutieli 2001, cited in Verhoeven et al. 2005).

Constraint Detection

By comparing different rate measures, I have shown that we can detect significantly constrained positions that are not detected by single-rate approaches. The methodology introduced improves existing methods in at least two ways.

- It provides a significance measure of the level of constraint, and does not rely on a user-specified threshold. Significant positions appear to have an intermediate substitution rate, probably because they are partly, but not totally, constrained (see Fig. 3).
- It provides information about the type of constraint acting on the detected positions.

Contrary to single-rate methods, which generally detect positions directly involved in active sites and hence highly conserved, the two-rate method exhibits positions which are important for the protein structure, like residues involved in the hydrophobic core of the protein. Here I tested three amino acid properties—volume, polarity, and charge of the residues—and found the constraint for hydrophobicity to be prominent in this data set. This is consistent with previous findings that hydrophobicity is a major force of evolutionary constraints (Koshi and Goldstein 1997). More amino acid properties also deserve to be tested, as they have been reported to be important for protein evolution (Xia and Li 1998).

Positive Selection Detection

The method presented here could potentially be used to infer positively selected positions, by looking for sites having experienced more nonconservative substitutions than expected by chance under the neutral hypothesis, similar to the TreeSAAP method (Woolley et al. 2003). Such positions would correspond to sites with a positive coordinate on the second axis in the principal component analysis. In the TIM data set, only a few such positions remain significant after correction for multiple testing, which is expected since positive selection is typically less frequent than negative selection. Positions GLY16, ARG18, LEU131, GLU140, ILE207, and GLN223 are detected as changing their polarity, and GLU135 is detected for charge selection. Positions ILE207 and GLN223 are detected as significantly constrained for the large volume of the residue. The negative selection on small residues at these positions may hence explain the

apparent high rate of polarity change. Positions GLY16 and ARG18 are in the last turn of an α -helix, in the contact region between the two subunits, LEU131 and GLY140, and are exposed to the surface as well.

Some of the most commonly used methods for detecting positive selection using interpopulation sequences are based on codon sequences, and aim at comparing the synonymous versus nonsynonymous rates of substitution (Nielsen and Yang 1998). Recent work has successfully accounted for biochemical properties when computing these rates and has proved to be useful to infer positive selection (Woolley et al. 2003; Sainudiin et al. 2005; Wong et al. 2006). These methods are, however, restricted to evolutionary time scales for which the third position of codons is not saturated. By comparing rates of substitution at the protein level, one may hence gain extra information on the selection pattern, at a larger time scale. This approach deserves more investigation to rule out possible confounding factors, as, for instance, negative selection on an untested biochemical property.

Conclusion

By comparing different evolutionary rate measures, the new method introduced in this article succeeds in detecting constrained positions not detected by single-rate methods and, hence, appears to be a useful complement to existing tools. Detected positions are of potential interest to the molecular evolutionist and to the biochemist: when the structure is not known, this method can be used to make robust predictions on functionally and/or structurally important positions. In relation to structural data, this method provides information about the key positions of the structure. It may hence prove useful in helping to design mutagenesis studies, by providing information on the positions to mutate and on the type of mutation to perform.

I have presented some evidence for constraint for polarity, volume, and charge, but other properties also need to be tested. A large-scale approach would be interesting, to assess which properties generally tend to be constrained in proteins, in relation to the structure of molecules. Another promising perspective would be to detect positions that have undergone a constraint shift in their evolutionary history, for instance, after a duplication event.

Acknowledgments This work was supported by Centre National de la Recherche Scientifique and Action Concertée Incitative “Informatique, Mathématiques et Physique pour la Biologie.” The author would like to thank Nicolas Galtier, Tal Pupko, Itay Mayrose, Adi Stern, Adi Doron, Eyal Privman, Nimrod Rubinstein, Ofir Cohen, Osnat Penn, David Burnstein, and Guillaume Achaz for helpful suggestions on this work, Nicolas Galtier for help with the writing of the manuscript, and Karine Jacquet for help with the ade4 package.

This publication is contribution 2008-051 of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554—CNRS).

References

- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met* 57:89–300
- Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29(4):1165–1188
- Chessel D, Dufour A, Thioulouse J (2004) The ade4 package. I. One-table methods. *R News*:5–10
- Dutheil J, Galtier N (2007) Detecting groups of co-evolving positions in a molecule: a clustering approach. *BMC Evol Biol* 7:242–242
- Dutheil J, Pupko T, Jean-Marie A, Galtier N (2005) A model-based approach for detecting coevolving positions in a molecule. *Mol Biol Evol* 22:1919–1928
- Dutheil J, Gaillard S, Bazin E, Glémin S, Ranwez V, Galtier N, Belkhir K (2006) Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics* 7:188–188
- Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17:368–376
- Felsenstein J (2004) *Inferring phylogenies*. Sinauer Associates, Sunderland, MA
- Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N (2003) ConSurf: identification of functional regions in proteins by surfacemapping of phylogenetic information. *Bioinformatics* 19:163–164
- Goldman N, Thorne JL, Jones DT (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149:445–458
- Grantham R (1974) Amino acid difference formula to help explain protein evolution. *Science* 185:862–864
- Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704
- Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. *Nucleic Acids Res* 28:374–374
- Kimura M (1983) *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge
- Koshi JM, Goldstein RA (1997) Mutation matrices and physicochemical properties: correlations and implications. *Proteins* 27:336–344
- Kosiol C, Goldman N (2005) Different versions of the Dayhoff rate matrix. *Mol Biol Evol* 22:193–199
- Kraulis PJ (1991) Molscript—a program to produce both detailed and schematic plots of protein structures. *J Appl Crystal* 24:946–950
- Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 257:342–358
- Lolis E, Alber T, Davenport RC, Rose D, Hartman FC, Petsko GA (1990) Structure of yeast triosephosphate isomerase at 19-A resolution. *Biochemistry* 29:6609–6618
- Mayrose I, Graur D, Ben-Tal N, Pupko T (2004) Comparison of site specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol Biol Evol* 21:1781–1791
- Mayrose I, Mitchell A, Pupko T (2005) Site-specific evolutionary rate inference: taking phylogenetic uncertainty into account. *J Mol Evol* 60:345–353
- Merritt EA, Bacon DJ (1997) Raster3d: photorealistic molecular graphics. *Methods Enzymol* 277:505–524
- Nielsen R (2002) Mapping mutations on phylogenies. *Syst Biol* 51:729–739
- Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936
- Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18(Suppl 1):S71–S77
- R Development Core Team (2006) *R: a language and environment for statistical computing*
- Sainudiin R, Wong WS, Yogeewaran K, Nasrallah JB, Yang Z, Nielsen R (2005) Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *J Mol Evol* 60:315–326
- Sokal RR, Rohlf FJ (1995) *Biometry*, 3rd edn. W. H. Freeman, New York
- Verhoeven KJF, Simonsen K, McIntyre LM (2005) Implementing false discovery rate control: increasing your power. *Oikos* 108:643–647
- Wong WS, Sainudiin R, Nielsen R (2006) Identification of physicochemical selective pressure on protein encoding nucleotide sequences. *BMC Bioinformatics* 7:148–148
- Woolley S, Johnson J, Smith MJ, Crandall KA, McClellan DA (2003) TreeSAAP: selection on amino acid properties using phylogenetic trees. *Bioinformatics* 19:671–672
- Xia X, Li WH (1998) What amino acid properties affect protein evolution? *J Mol Evol* 47:557–564
- Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* 39:306–314
- Yang Z (2006) *Computational molecular evolution*. Oxford University Press, Oxford, UK
- Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141:1641–1650