BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 16 2010, pages 2012–2019
doi:10.1093/bioinformatics/btq317

Systems biology

Advance Access publication June 15, 2010

Molecular signatures-based prediction of enzyme promiscuity

Pablo Carbonell* and Jean-Loup Faulon

*i*SSB, Institute of Systems and Synthetic Biology, University of Evry, Genopole Campus 1, Genavenir 6, 5 rue Henri Desbruères, 91030 EVRY Cedex, France

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Enzyme promiscuity, a property with practical applications in biotechnology and synthetic biology, has been related to the evolvability of enzymes. At the molecular level, several structural mechanisms have been linked to enzyme promiscuity in enzyme families. However, it is at present unclear to what extent these observations can be generalized. Here, we introduce for the first time a method for predicting catalytic and substrate promiscuity using a graph-based representation known as molecular signature.

Results: Our method, which has an accuracy of 85% for the non-redundant KEGG database, is also a powerful analytical tool for characterizing structural determinants of protein promiscuity. Namely, we found that signatures with higher contribution to the prediction of promiscuity are uniformly distributed in the protein structure of promiscuous enzymes. In contrast, those signatures that act as promiscuity determinants are significantly depleted around non-promiscuous catalytic sites. In addition, we present the study of the enolase and aminotransferase superfamilies as illustrative examples of characterization of promiscuous enzymes within a superfamily and achievement of enzyme promiscuity by protein reverse engineering. Recognizing the role of enzyme promiscuity in the process of natural evolution of enzymatic function can provide useful hints in the design of directed evolution experiments. We have developed a method with potential applications in the guided discovery and enhancement of latent catalytic capabilities surviving in modern enzymes.

Availability: http://www.issb.genopole.fr/~faulon

Contact: pcarbonell@issb.genopole.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 17, 2010; revised on May 19, 2010; accepted on June 9, 2010

1 INTRODUCTION

Beyond the classical definition of enzyme as specific to the substrate and reaction that catalyzes, enzymes can show the capability of promiscuity, i.e. to catalyze more than one reaction (catalytic promiscuity) or show broad substrate specificity (substrate promiscuity) (Bornscheuer and Kazlauskas, 2004; Hult and Berglund, 2007). Enzyme promiscuity has drawn considerable attention in recent years due to its evolutionary and practical implications (Copley, 2003; O'Brien and Herschlag, 1999). In fact, enzyme promiscuity has been related to the evolvability of proteins, as a mechanism that can be traced in the divergence of enzyme families and superfamilies (Khersonsky *et al.*, 2006). Namely, the conservation of structural and catalytic features observed in enzyme superfamilies such as tautomerases (Poelarends *et al.*, 2008), enolases (Babbitt *et al.*, 1996) and many others, suggests that enzyme specialization arose via divergent evolution (gene duplication and subsequent mutations) driven by environmental selective pressure from promiscuous precursor enzymes. Inspired on this mechanism of natural evolution, enzyme promiscuity has been sought in the laboratory as a starting place for directed evolution of enzymes capable of new functions (Yoshikuni *et al.*, 2006), such as in biocatalysis (Bornscheuer and Kazlauskas, 2004), degradation of novel synthetic chemicals (Copley, 2003) or in other innovative ways in the context of synthetic biology (Andrianantoandro *et al.*, 2006). These potential applications of enzyme promiscuity in biotechnology have motivated the search of protein engineering methods adding or enhancing latent catalytic activities in existing enzymes, such as site-directed mutagenesis and directed evolution (Kazlauskas, 2005).

At the molecular level, several structural mechanisms have been related to enzyme promiscuity. For instance, promiscuous binding has been observed in a large hydrophobic cleft or cavity or in a region with some other unusual structural feature (Copley, 2003). Similarly, promiscuity has been linked in some cases to intrinsically disordered regions in the protein (Kim *et al.*, 2008; Tyagi *et al.*, 2009), while in other studies it has been related to conserved regions in the active site (Anandarajah *et al.*, 2000). A promiscuous active site might be able to accommodate different substrates, a feature that has been linked to protein flexibility, in particular to the mobility of active site loops, which appears to play a key role in mediating promiscuity (Ma *et al.*, 2005; Yasutake *et al.*, 2004). Furthermore, experiments have revealed that residues influencing enzyme promiscuity go beyond active residues contacting the substrate (O'Brien and Herschlag, 1999; Romero and Arnold, 2009). In fact, plasticity (Yoshikuni *et al.*, 2006) or neutral (Babbitt *et al.*, 1996) residues (residues which are not essential for core catalytic activity) lying along the contour of the active site may hasten divergence, as it has been observed in several α/β barrel superfamilies such as estareses and lipases (Hegyi and Gerstein, 1999), enolases (Petsko, 2000), ribulose-phosphate binding (Chan *et al.*, 2008), as well as in directed evolution of isomerases (Jurgens *et al.*, 2000; Kuper *et al.*, 2005) or aminotransferases (Rothman and Kirsch, 2003). Nevertheless, it is at present unclear to what extent these observations can be generalized to other enzyme families, since additional determinants such as post-translational modifications, the presence of multiple domains, and oligomeric states may be also affecting enzyme promiscuity, such as

*To whom correspondence should be addressed.

in HisF from Thermotoga maritima (Taglieber *et al.*, 2007), or the multidomain arom pentafunctional enzyme in yeast (Nobeli *et al.*, 2009).

The aim of this article is to introduce a new method for predicting enzyme promiscuity, which will allow us to investigate for the first time on genome-wide scales how residues participate as determinants of promiscuity and their interrelation with catalytic sites. Gaining insights into this question has important implications for the design in biocatalysis and synthetic biology applications. Predicting enzyme promiscuity, however, has been recognized as a problem of daunting complexity (Nobeli *et al.*, 2009). Even though some progress has been done in this field (Carbonell *et al.*, 2009; Gomez *et al.*, 2003; Macchiarulo *et al.*, 2004), there is still a need for better and more accurate methods of prediction. Our approach uses a graph-based representation known as molecular signature (Faulon *et al.*, 2008). On this method, similarity between two molecules is evaluated by comparison of their respective canonical subgraphs (Faulon *et al.*, 2004). Here, we propose to use the signature representation to build a tool for prediction of enzyme promiscuity by means of graph kernel support vector machines (SVM; Gartner *et al.*, 2003), which are machine-learning methods widely used as classifiers and in regression (Vapnik, 1995). Our dataset consists of the list of enzymes in the entire KEGG database (Kanehisa *et al.*, 2008). An enzyme is labeled as having promiscuous catalytic activity if it can process more than one different reaction. Similarity between reactions was evaluated by using reaction molecular signatures. After training and cross-validating a linear kernel SVM for the prediction of enzyme promiscuity, the set of molecular signatures contributing most to the determination of catalytic promiscuity was subsequently used to analyze the influence of structural properties on enzyme promiscuity. Potential applications in enzyme engineering and synthetic biology were illustrated by looking at the signatures distribution in the family of enolases and aminotransferases.

2 METHODS

Molecular signature: the molecular signature is a vector whose components correspond to atomic signatures. Initially developed for chemicals (Faulon *et al.*, 2004), the signature molecular descriptor was later extended to protein sequences (Faulon *et al.*, 2008; Martin *et al.*, 2005). Each component of a molecular signature counts the number of occurrences of a particular atomic signature in the molecule (see Supplementary Fig. S1). An atomic signature is a canonical representation of the subgraph surrounding a particular atom. This subgraph includes all atoms and bonds up to a predefined distance from the given atom. This distance is called the signature height h. If $G=(V,E)$ is a molecular graph, where vertices V correspond to atoms, and edges E to bonds, then the molecular signature of G is given by

$$^h\sigma(G)=\sum_{x_i\in V} {}^h\sigma(x_i) \tag{1}$$

where $^h\sigma(x_i)$ is the atomic signature of G rooted at atom x_i of height h.

Protein information can be encoded by using the signature molecular descriptor as well. For proteins, signatures are similar to the k-mer spectrum developed by Leslie *et al.* (2003). A residue signature consists of a residue and its sequence neighbors within a window specified by its height h. For a given sequence S, its signature of height h is defined as

$$^h\sigma(S)=\sum_{s_i\in S} {}^h\sigma(s_i) \tag{2}$$

where $^h\sigma(s_i)$ is the k-mer signature of height h centered at residue s_i of S.

Reaction signature: we assume that enzymatic reactions take the general form $R: s_1S_1+\ldots+s_nS_n \rightarrow p_1P_1+\ldots+p_mP_m$, where s_i and p_j are the stoichiometric coefficients of substrates S_i and products P_j. The signature of reaction R of height h is defined by

$$^h\sigma(R)=\sum_j p_j {}^h\sigma(P_j)-\sum_i p_i {}^h\sigma(S_i) \tag{3}$$

Reaction signature-based characterization of enzyme catalytic promiscuity: Enzyme catalytic promiscuity is defined as the ability of an enzyme to catalyze different type of reactions. Here, we use the reaction signature of height $h=1$ to characterize enzyme catalytic promiscuity (see Supplementary Fig. S2 for an illustrative example). An enzyme is classified as promiscuous if it can process dissimilar reactions, i.e. if the enzyme has been annotated with at least a pair of reactions R_A and R_B with different signatures.

Reaction signature-based characterization of enzyme substrate promiscuity: Substrate promiscuity of enzymes is defined as the ability of an enzyme to catalyze different substrates. It differs from catalytic promiscuity in the sense that it is possible for catalytic reactions to process more than one different substrate. We evaluated the substrate promiscuity of an enzyme by counting the number of unique reaction signatures of their annotated reactions extended over the height range of $h=0$ to 3 (see Supplementary Figs S3 and S4 for illustrative examples).

SVM: to train a two-level classifier SVM, data is presented as pairs $\{(\mathbf{x}_i, y_i)\} \in \mathbb{R}^n \times \{\pm 1\}$, where $\mathbf{x}_i$ is the vector of input features, and y_i is its observed class (labeled as -1 and $+1$). The SVM computes the following decision function for the input vector $\mathbf{x}$:

$$f(\mathbf{x})=\sum_i \lambda_i y_i k(\mathbf{x},\mathbf{x}_i)+b \tag{4}$$

where $k: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is a kernel function, and λ_i and b are typically obtained by solving a quadratic programming problem (Vapnik, 1995).

In our case, given an input sequence S_i, the input vector $\mathbf{x}_i$ corresponds to its signature descriptor $^h\sigma(S_i)$ of height h according to Equation (2) and the output y_i to one of the two classes $\{-1,+1\}$, which stand for non-promiscuous and promiscuous enzymes, respectively. By definition, the kernel function $k(\mathbf{x},\mathbf{x}_i)$ performs a dot product in some vector space. Following our previous work (Martin *et al.*, 2005), here we used the signature kernel function, which is given by the Euclidean dot product of signatures as vectors in the multidimensional space formed by all possible k-mers of height h

$$k(\mathbf{x},\mathbf{x}_i)={}^h\sigma(S)\cdot{}^h\sigma(S_i)=\sum_j {}^h\sigma_j(S)\,{}^h\sigma_j(S_i) \tag{5}$$

Since the SVM in Equation (5) is using a linear kernel, it is possible to express the decision function as a linear combination on the signature space

$$\begin{aligned} f({}^h\sigma_j(S)) &= \sum_i \lambda_i y_i \sum_j {}^h\sigma_j(S)\,{}^h\sigma_j(S_i)+b \\ &= \sum_j {}^h\sigma_j(S)\sum_i \lambda_i y_i\, {}^h\sigma_j(S_i)+b=\sum_j {}^h\sigma_j(S)\alpha_j+b \end{aligned} \tag{6}$$

Equation (5) expresses the dot product in the sequence data space, and it is a compact form for computing the SVM. Furthermore, Equation (6) based on the long and sparse signature vectors space is also used in our present study in order to analyze the individual contributions or α-values of each k-mer to the output of the decision function.

Other kernel functions based on k-mers that have been proposed are the mismatch string kernel (Leslie *et al.*, 2003) and the local string alignment kernel (Saigo *et al.*, 2004). We have included in our study both implementations from authors' sites in order to compare their performance with our proposed kernel, which is given by Equation (5). To compute the SVMs, we used a modification of the SVM_light software with a custom kernel based on signatures, as in Martin *et al.* (2005).

Sequence homology score: we implemented an enzyme promiscuity homology-based score by querying a blast database built from the non-redundant list of promiscuous and non-promiscuous enzymes in the

dataset. Predictions were given based on majority consensus of the input sequence homologs up to the specified sequence identity cut-off.

Disorder- and hydrophobicity-based predictors: disorder and hydrophobicity were computed for balanced sets of 1000 randomly selected sequences from the positive and negative sets, respectively. Disorder for each residue was measured based on the three parameters given by the disembl predictor (Linding *et al.*, 2003): loop/coils, hot loops and missing coordinates. Receiver-operator-characteristic (ROC) curves and accuracy were computed by varying the threshold of the ratio of disordered residues in the protein for each of the three parameters, and then averaged. Similarly, the ratio of non-hydrophobic residues was used to compute the ROC curve and accuracy for hydrophobicity.

Signature-based promiscuity predictor: predictions are given as the average of the promiscuity prediction score given by the output of the signature-based SVM and the homology score. *z*-score and *P*-value for the prediction are given based on the distribution of the predicted values in the training set.

Cross-validation and performance measures: a 10-fold random cross-validation of the performance of the SVM was implemented by using a balanced training and validation set containing 1000 example sequences, which was repeated 500 times over the dataset. Statistics were compiled for each dataset using the predictions from the cross-validation. Performance was measured as follows: let TP, FP, TN and FN to denote true positive, false positives, true negatives and false negatives, respectively, then accuracy is defined by (TP + TN)/(TP + FP + TN + FN), sensitivity is defined by TP/(TP + FN), and specificity is defined by TN/(TN + FP). The area under the ROC curve (AUC) is obtained by integrating under the ROC curve. The ROC curve is obtained by varying the discrimination threshold separating positives from negatives and plotting the TP rate (sensitivity) versus the FP rate (1-specificity).

Dataset: we downloaded from the KEGG database (release 50) the list of protein sequences with annotated catalytic activity. The dataset was selected by taking those sequences in the initial list with fully defined stoichiometrically balanced reaction, totaling 498 025 sequences (73% of the complete list in KEGG). Reaction similarity at catalytic and substrate level (see previous definitions) was computed between pairs of reactions annotated for each enzyme. Based on reaction similarity, the sequence list was split into positive set (promiscuous) and negative set (non-promiscuous). Redundancy was removed within each set by using the cd-hit program (Li and Godzik, 2006) at a threshold of 50% of sequence similarity.

Golden set: a *golden set* of 64 experimentally verified enzymes with promiscuous catalytic activity was built by selecting sequences from the BRENDA database (Chang *et al.*, 2009) , where detailed information about catalytic activities and the source of the information were available (see Supplementary Table S1).

3D structures assignment: we performed a blast search for non-redundant homolog PDB structures of enzyme sequences in the reference dataset, within a similarity threshold of 80%. Prior to the structural assignment, consistency between the annotated EC numbers in the KEGG database for the sequence and its homolog was checked. Following this strategy, we were able to identify model structures for 1035 different promiscuous enzymes, and 3491 non-promiscuous enzymes (see Supplementary Table S2).

Structural homology score: we downloaded the CATH (Orengo *et al.*, 1997) and SCOP (Murzin *et al.*, 1995) structural classifications for the dataset of enzyme structures. Structural homology predictions were given based on majority consensus of promiscuous and non-promiscuous homologs of the input structure.

Catalytic sites mapping: based on the Catalytic Site Atlas (Porter *et al.*, 2004), experimentally determined functional residues were mapped into enzyme model 3D structures. Information about catalytic site residues was identified for 1004 structures representing promiscuous enzymes and for 3406 non-promiscuous enzymes.

Secondary structure: secondary structure was assigned to the dataset of structures using the DSSP server (Kabsch and Sander, 1983) labeling each residue as helix, beta strand or loop (which includes turns or unassigned residues).

Table 1. Distribution of promiscuity annotations in the dataset before and after filtering sequence, reaction or substrate redundancy

Total sequences	498 025			
Single reaction	279 982	Multiple reactions	218 043	44%
Chemically unique reactions	356 342	Different reactions	141 683	28%
Non-redundant (0.5)	60 043	Non-redundant (0.5)	24 149	29%
Single substrate	303 709	Different substrates	194 316	39%
Non-redundant (0.5)	52 057	Non-redundant (0.5)	32 229	38%

First and second rows correspond to total sequences with reaction information in KEGG; third and fourth rows correspond to the classification based on catalytic promiscuity and after filtering redundancy; fifth and sixth rows correspond to the classification based on substrate promiscuity.

3 RESULTS AND DISCUSSION

Characterization of enzyme promiscuity in the reference dataset: catalytic promiscuity is defined as the ability of an enzyme to catalyze more than one reaction. To study catalytic promiscuity, we looked at the list of reactions annotated for enzyme sequences in KEGG (Kanehisa *et al.*, 2008). Even when reactions were labeled with different names, there might be cases where they correspond to the same chemical transformations. To prevent this redundancy, we compared pairwise the reactions catalyzed by each enzyme through the use of the molecular signature-based reaction similarity measure (see Section 2). An enzyme was labeled as promiscuous if the enzyme can catalyze at least two chemically different reactions. We found in the dataset that 28% of enzymes are catalytically promiscuous (Table 1). This percentage is substantially lower than the number of enzymes annotated with more than one reaction in KEGG (44%), as the study of promiscuous catalytic activity of the enzyme at the reaction level is able to avoid redundancies in reaction annotations. In turn, the number of promiscuous enzymes is significantly higher than the number of enzymes annotated with more than one EC number (2%), illustrating the fact that EC number annotations are in many cases a broader definition of catalytic activity. Furthermore, we observed that 96% of enzymes annotated with more than one EC number were actually catalyzing more than one different reaction. In contrast, 42% of enzymes with only one EC number were annotated with more than one different reaction.

Similarly, we used molecular signatures to compare substrates catalyzed by each enzyme (see Section 2). We found that the number of enzymes with substrate promiscuity is ∼10% higher than the number of enzymes with catalytic promiscuity, since the former might include cases where a unique reaction is used by the enzyme to process different substrates, whereas those cases were removed from the latter.

In summary, although different EC annotations generally encode for different reactions in KEGG, a single EC number can contain several different reactions according to our reaction similarity measures. For instance, EC 4.1.2.19 L-rhamnulose 1-phosphate aldolases (see Supplementary Fig. S2) are inducible not only by L-rhamnose but also by its pentose analog L-lyxose

Table 2. Average performance given by accuracy, (area under the ROC curve) of the catalytic promiscuity and substrate promiscuity predictors for different taxonomic groups, compared with other predictors and the random case

	Catalytic promiscuity										Substrate
Group	Signatures	Mismatch	Alignment	Random	Homology	CATH	SCOP	Length	Disorder	Hydrophobicity	Signatures
Dataset	0.85 (0.89)	0.84 (0.91)	0.84 (0.94)	0.51 (0.51)	0.73 (0.70)	0.73 (0.75)	0.77 (0.79)	0.51 (0.47)	0.54 (0.53)	0.52 (0.52)	0.84 (0.87)
Prokaryotes	0.87 (0.90)	0.88 (0.94)	0.87 (0.94)	0.53 (0.52)	0.71 (0.75)	0.74 (0.72)	0.75 (0.76)	0.54 (0.50)	0.55 (0.55)	0.52 (0.52)	0.87 (0.90)
Eukaryotes	0.88 (0.91)	0.89 (0.94)	0.89 (0.97)	0.52 (0.51)	0.80 (0.77)	0.77 (0.78)	0.77 (0.80)	0.51 (0.47)	0.50 (0.43)	0.51 (0.51)	0.89 (0.93)

The performance of catalytic promiscuity prediction is evaluated for several types of predictors: signatures, mismatch, alignment and random kernel-based SVMs, sequence and structural homology through CATH and SCOP, sequence length, disorder and hydrophobicity, respectively. Performance of the signatures-based SVM predictor of substrate promiscuity is also given.

(Badia *et al.*, 1991). Another example is EC 2.3.1.179, a β-ketoacyl-carrier-protein synthase II (see Supplementary Fig. S3) with a unique catalytic reaction signature that can process different substrates depending on the acyl group, acetyl or butyryl (Wu *et al.*, 2009). Our predictions, thus, rely on how well sequences are annotated with EC numbers and how accurately reactions have been mapped into these EC numbers. Without doubt, we can expect misannotations with both, the EC numbers and the reactions. However, even after removing all close homologs, the accuracies that we reached in our predictions (85%) are well above a random classifier or a homology-based classifier. These cross-validation accuracies indicate that even if the training set might not be perfect, promiscuity can still be learnt using k-mer signatures.

Effect of protein features on catalytic promiscuity: first, we consider the effect of sequence length on enzyme promiscuity, as is the case in multi-domain proteins where functional sites might be located in different domains of the protein. Catalytic promiscuity is defined as the ability of an enzyme to process different reactions through the same active site and, therefore, sequence length is not expected to be a determinant effect in the prediction of promiscuity. As a matter of fact, we tested sequence length as a predictive parameter of promiscuity in our dataset, obtaining an AUC (see definition in Section 2) of 0.47, which is a non-discriminant value for predicting purposes. Similarly, we tested the effect of hydrophobicity and protein disorder on enzyme promiscuity. In both cases, accuracy and AUC were again around 0.53 (Table 2).

Prediction of catalytic promiscuity using molecular signatures: our approach to enzyme promiscuity prediction is based on the use of protein molecular signatures as vector representations of protein sequences by its amino acid strings or k-mers, which can be used to develop a signature-based SVM (see details in Section 2). This approach has proven in the past to be information rich and efficient for predicting enzymatic function (Faulon *et al.*, 2008) such as our present study on enzyme promiscuity prediction. The performance of the proposed SVM as an enzyme catalytic promiscuity predictor was evaluated by means of a 10-fold cross validation of the dataset (Table 2), obtaining an overall average accuracy 85% (AUC = 0.89), with 87% (AUC = 0.90) for prokaryotes and 88% (AUC = 0.91) for eukaryotes. Similar performance values were obtained for the mismatch and local alignment kernels, as it is shown in Table 2. Indeed, after running the three kernels on 500 datasets, the SD of the accuracy and AUC for each kernel was found to be 0.02 and 0.03, respectively. This performance of the kernel-based predictor is significantly higher than the performance obtained for analogous sets of sequences randomly labeled as promiscuous/non-promiscuous, as shown in Table 2.

Prediction of substrate promiscuity using molecular signatures: similar to the previous section, we performed a cross-validation of the proposed SVM-based predictor for the prediction of substrate promiscuity. Accuracy values ranged from 88% (AUC = 0.91) for eukaryotes to 89% (AUC = 0.93) in prokaryotes (Table 2).

Comparison with a homology-based predictor: we compared the performance of our signature-based predictor with a sequence homology-based predictor of catalytic promiscuity, which we built from the non-redundant dataset by using blast (Altschul *et al.*, 1990) (details can be found in Section 2). A cross-validation test similar to the one performed on our predictor and described previously was applied to this homology predictor. Results are shown in Table 2. We performed the test for different cut-off levels in the sequence homology similarity, obtaining an increasing accuracy as we lowered the cut-off, i.e. increased the number of homologs included in the prediction, with maximum accuracy of 73% (AUC = 0.70) for a sequence similarity of 20% or below. A more accurate homology predictor might be obtained based on information which goes beyond sequence similarity, as is the case in the structural homology predictors CATH and SCOP. We used these databases in order to build a structural homology predictor, as it is described in Section 2. Results for the dataset (Table 2) show an improvement, specially for the SCOP-based predictor, which had an accuracy of 77% (AUC = 0.79).

Therefore, predictions based on sequence homology had lower performance compared with our predictor. Furthermore, from a technical point of view the main advantage of our signatures-based predictor over a homology-based predictor is that signatures might provide to protein designers a way to identify those residues that are contributing most as determinants of catalytic promiscuity through the use of the top k-mers, as it will be presented in the next sections.

Predictions for the promiscuity golden set: the performance obtained for the homology-based predictor might be due to the presence of homology annotations in KEGG. As a matter of fact, sequences in the non-redundant dataset might still have been annotated in KEGG by homology, because removing redundancy does not change the annotations of the remaining sequences. To circumvent this problem, we performed an additional validation by using the promiscuity golden set, a set of 64 enzymes with experimentally associated multiple reactions in BRENDA. After removing homologs of the golden set sequences from the dataset, we

repeatedly trained our promiscuity predictor, obtaining an average accuracy of 75% in the predictions for this subset.

Top signatures contributing to the prediction of promiscuity: we investigated the distribution of signatures in the positive and negative sets in order to identify those signatures or k-mers that are contributing most to the prediction of enzyme promiscuity. Identifying k-mers that most contribute to promiscuity will be subsequently used in this study to characterize promiscuity at sequence and structural level. The procedure is as follows: during the training of the SVM, the algorithm assigns a weight (α-value) to each k-mer in the training set (see Section 2). Thereafter these values are used to compute the prediction of a given sequence based on its signature or vector of occurrences of each particular k-mer. Thus, k-mers with top absolute α-values have been determined by the SVM algorithm either to contribute most to enzyme promiscuity or to correspond to non-promiscuous enzyme signatures with the lowest α-values. Selecting top k-mers based exclusively on the highest absolute α-values of the SVM might not be in some cases the best selection method, specially if there exists strong correlation between the input features. To test the efficiency of a selection based on k-mers with top α-values, we compared our selection with a more robust method known as recursive feature selection (Guyon *et al.*, 2002). We observed, however, that for a selection of 1000 top k-mers, there is a substantial overlap between both methods (>75%), indicating that a selection based on top α-values is representative for promiscuity determinants (see details in Supplementary Fig. S5). The discriminant power of these top k-mers was assessed by performing a 10-fold cross-validation test (see Supplementary Table S3), where it is shown that a set of 1000 top k-mers was significant enough in this test to give performance values close to the values achieved with the SVM at signature heights in the range $h=3$ to 5. A set of 1000 k-mers spans a small subspace of the total k-mer space, which has a total dimension of 20^h for height h. Therefore, for the list of enzyme sequences in the KEGG database, we observed that a relatively small set of k-mers with high α-values can act as determinants of enzyme promiscuity.

The amino acid content of the 1000 top k-mers differs from the general distribution in the dataset. In general, aromatic (Tyr, Trp, Phe) and other bulky amino acids such as Met, Hist are less present in promiscuous signatures, whereas small amino acids such as Gly and Ala are more frequent (see Supplementary Table S4). A detailed analysis of several physicochemical properties of top k-mers can be found in Supplementary Figs S6 and S7. In general, top k-mers are lighter and with smaller total solvent accessibility area, which might be associated with its lower catalytic or substrate specificity.

Signatures-based structural analysis of promiscuous enzymes: no prior structural or catalytic site information was used in the training of the signature-based SVM. Thus, it might be of interest to analyze how top predictive signatures locations in the protein structure relate to catalytic sites of promiscuous enzymes. We built for this purpose a database of 3D structures derived from the KEGG database, as described in the Section 2. Catalytic residues were mapped in the structures by using the Catalytic Site Atlas (CSA), which contains residues that have been determined to play a central role in the catalytic mechanism using defined criteria (Porter *et al.*, 2004). By computing contact maps of catalytic residues at different distance values, and the corresponding list of k-mers for each contacting residue, the average α-value of these k-mers gave an estimate of the contribution of residues around catalytic sites to the promiscuity of the enzyme. In Figure 1, we plotted the relationship between the distance radius and average α-values for the set of promiscuous and non-promiscuous enzymes. Signatures of k-mers are clearly split into positive and negative average values for promiscuous and non-promiscuous enzymes, respectively, with independence of the distance to the catalytic site. Furthermore, we observed a distinctive trend on the average α-value around catalytic sites in non-promiscuous enzymes that is significant compared with the average values observed around random residues. Namely, non-promiscuous catalytic sites or highly specific enzymes have lower α-values in k-mers as residues become closer to catalytic residues in the distance range of ∼5–20 Å. This is not the case in promiscuous enzymes, which appear to be more regularly distributed and essentially independent of the distance to the catalytic site. In this sense, the distribution of k-mers around catalytic sites in promiscuous enzymes is similar to the distribution around any other residue in the promiscuous enzyme and, on average, significantly higher than values around non-promiscuous catalytic sites. Therefore, we argue that from the molecular signature point of view, our results suggest that there exist some distinctive k-mers present in promiscuous enzymes, which are usually found distributed uniformly along the protein structure. In contrast, what confers specificity to non-promiscuous enzymes is precisely the existence of some functional regions in the protein containing specific signatures usually not found in promiscuous enzymes.

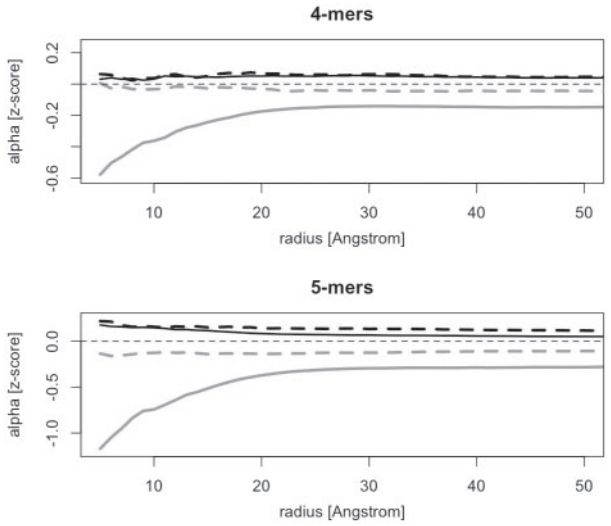

Fig. 1. Comparison between average α-value in 4-mers and 5-mers of contact residues around catalytic sites (solid lines) and around random residues (dotted lines) for different radii in promiscuous (black) and non-promiscuous enzymes (gray).

Secondary structure and enzyme promiscuity: despite the fact that evolutionary and structural properties of proteins have been pointed out by several authors as determinants of enzyme promiscuity; in general, there is no direct linkage between promiscuity and most of these properties measured for the entire sequences. Bartlett *et al.* (2002) already addressed the role of secondary structure in catalytic sites by looking at the residues in the CSA. We supplement here the study of these authors by looking at the secondary structure for the particular case of promiscuous catalytic sites. Results are shown in Table 3. The observed difference between the distribution

Table 3. Secondary structure distribution in promiscuous and non-promiscuous enzymes compared with the average distribution in the enzyme dataset

	Beta (P-value)	Helix (P-value)	Loop (P-value)
All residues	15.69%	40.64%	43.67%
Catalytic sites	23.79% (1.22×10^{-4})	32.15% (2.85×10^{-3})	44.05% (9.38×10^{-1})
Non-promiscuous	20.85% (4.96×10^{-2})	33.65% (4.614×10^{-2})	45.50% (6.42×10^{-1})
Promiscuous	30.00% (1.49×10^{-4})	29.00% (2.34×10^{-2})	41.00% (6.62×10^{-1})

The two first lines were obtained reproducing the work of Bartlett *et al.* (2002). The two last lines are new results obtained in this study. P-values are from a χ^2 test for proportionality with respect to random residues distribution (first row).

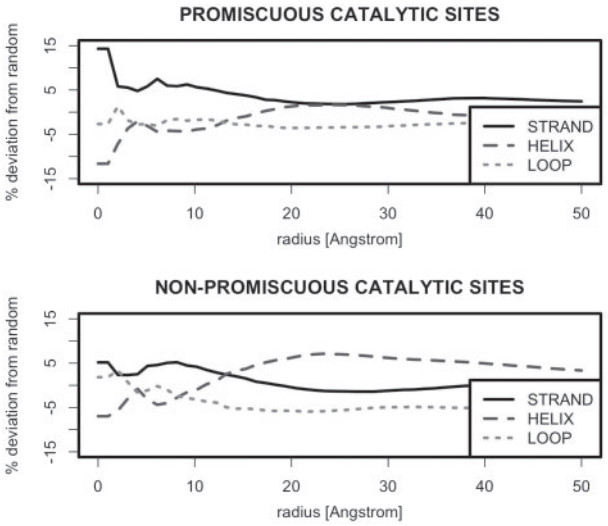

Fig. 2. Average deviation from random of the distribution of secondary structures around catalytic sites for different radii in promiscuous and non-promiscuous enzymes.

of secondary structures for all residues and catalytic sites is in good agreement with the results in Bartlett *et al.* (2002). That is, helices are in general underrepresented in catalytic residues. Interestingly, beta strands, which are known to be more conserved than other secondary structures (Sitbon and Pietrokovski, 2007), contain 30% of the catalytic residues, a percentage that is 10% higher than in non-promiscuous enzymes, and significantly 15% higher than in random residues. The percentage of helices in catalytic sites for both types of enzymes is around 30%, a 10% lower than in random residues, while the percentage of loops remains around 40% in all cases. Figure 2 plots the deviation from random of the distribution of secondary structures around catalytic sites for residues within different distances to catalytic sites. On average, enrichment in beta strands is clearly kept in promiscuous enzymes within the distance range 0–20 Å.

Application: catalytic promiscuous members of the enolase superfamily: enzymes mandalate racemase (EC 5.1.2.2) and muconate lactonizing (EC 5.5.1.1) are divergent members of the enolase superfamily that catalyze overall different reactions to produce their specific products (Hult and Berglund, 2007). In some cases, it has been also observed catalytic promiscuity as a *N*-acylamino acid racemase (annotated in KEGG as EC 4.2.1.113) (Glasner *et al.*, 2006). We selected from the KEGG database those non-redundant sequences in the enolase superfamily annotated with at least one of these catalytic reactions. Of the 346 sequences in the list, 33 out of 36 sequences were correctly classified as promiscuous enzymes, and 277 out of 310 in the case of non-promiscuous enzymes, which gives an accuracy of 90% and a sensitivity of 92%. This example illustrates the application of the method in order to identify potential members of an enzyme family or superfamily showing catalytic promiscuity.

Application: reverse engineering of a promiscuous aminotransferase: *Escherichia coli* aspartate aminotransferase (AATase, EC 2.6.1.1.) and tyrosine aminotransferase (TATase, EC 2.6.1.5) are two non-promiscuous enzymes of the same superfamily, which share a 43% sequence identity. Directed evolution of AATase to TATase activity (Rothman and Kirsch, 2003) showed that the new function can be acquired by mutagenesis without losing its original AATase function, suggesting that both enzymes arose by duplication from an ancestral promiscuous protein. We use here the promiscuity predictor to evaluate which signatures in the evolved mutant contributed most to the enhancement of its promiscuity. The SVM, trained with Prokaryote data and signature height $h=4$, predicted an increase in substrate promiscuity for the evolved mutant 8-2B (Rothman and Kirsch, 2003). We computed z-scores for α-values in wild type AATase and the selected variant after all rounds of directed evolution. Signatures centered in residues P13, I291 and A383 gave the highest contribution to the prediction of promiscuity ($z>1$). Obviously, changes in signatures α-values took place around point mutations, in this case P13T, A293V and A381V (Fig. 3). These three signatures are located in loops. It has been postulated that greater flexibility of the loops in TATase than in AATase is an important specificity determinant (Ishijima *et al.*, 2000). Signature in Pro13 is located on the mobile loop that precedes the substrate binding pocket. This signature also includes Asp15, which forms a salt bridge with Arg292, in the second signature, a contact residue for the inhibitor maleate (Rothman and Kirsch, 2003). Finally, signature in Ala383, which is located far away from the substrate is in the vincinity of Arg386, whose side chain directly interacts with the α-carboxylate group of the substrates (Yano *et al.*, 1998). This study illustrates the potential application of our predictor as a tool in protein engineering to analyze functional regions on the enzyme contributing to the mechanism of promiscuity.

4 IMPLEMENTATION

The Promis server: the Promis server is a web-based implementation of our method that provides predictions of catalytic and substrate promiscuity on enzyme sequences by an SVM trained on molecular signatures. The user inputs enzyme sequences in FASTA format and the server returns the predicted score (z-score and P-value) for both catalytic and substrate promiscuity. Users can select the taxonomic group and the molecular signature height used for training. In the predictor output, residues are coloured based on the α-value of the k-mers, i.e. the contribution of each residue to the prediction. These values can be downloaded as a text file and, for instance, incorporated into subsequent protein design pipelines.

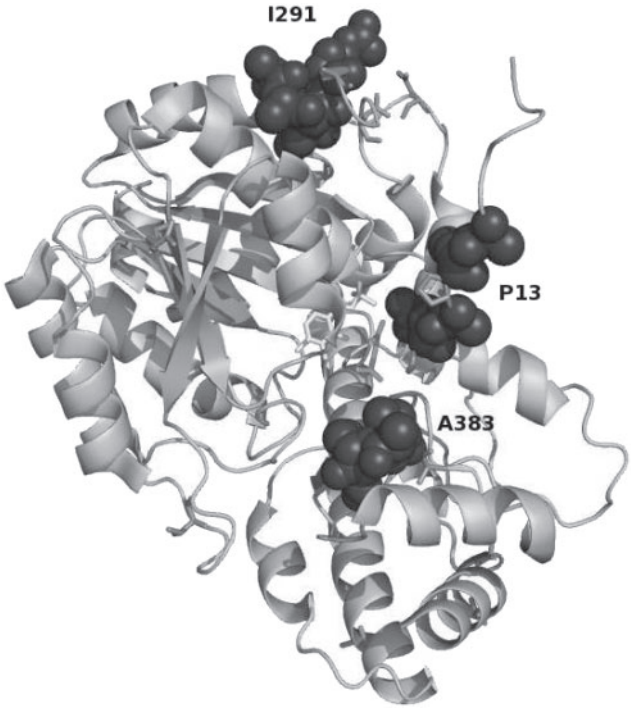

Fig. 3. Signatures (k-mers) with highest α-value change in directed evolved AATase after acquiring enhanced catalytic promiscuity. Residues with highest change ($z > 1$) and its k-mers are represented in spheres. Total substitutions in mutant: P13T, N69S, G72D, R129G, T167A, A293V, N297S, N339S, A381V, N396D, A398V. Mutations are depicted as sticks. (PDB ID: 1qis). This illustration was created using the program PyMOL (Delano, 2002).

5 CONCLUSIONS

Predicting enzyme promiscuity is a challenging task, which has remained elusive. Here, we introduced for the first time a method for predicting catalytic and substrate promiscuity using a graph-based representation known as molecular signature with potential applications in biocatalysis and protein engineering. Our analysis of the entire non-redundant list of enzymes sequences in KEGG revealed that, in general, there is no clear pattern between enzyme promiscuity and protein features such as sequence length, hydrophobicity, disorder or type of catalytic activity. Our approach to this problem was to implement a molecular signature-based SVM for the prediction of catalytic and substrate promiscuity. By training and cross-validating a kernel based-predictor using the entire non-redundant list of enzymes sequences in the KEGG database, we obtained an accuracy value of 85%. Furthermore, the use of linear kernels in the SVM allowed us to decompose its output into the individual contributions of signatures to the prediction of catalytic promiscuity. Looking at the significance of the set of signatures, which are enriched in the positive or promiscuous list of enzymes, we conclude that there exists a subset of signatures that act as determinants of enzyme promiscuity. Molecular signatures, therefore, can be used as an analytical tool to explore the role of residues contributing to enzyme promiscuity.

We studied some of the signature-based determinants of enzyme promiscuity. In general, signatures facilitating promiscuity contain lighter amino acids with smaller total solvent accessibility area, which might be associated with its lower catalytic or substrate specificity. Furthermore, we found that beta sheets, which are known to be more conserved than other secondary structures, play a role facilitating promiscuity. At genome-wide scales, inspection of the spatial neighbouring region around promiscuous and non-promiscuous catalytic sites in the CSA revealed a region of $\sim$20 Å surrounding catalytic sites where signatures are significantly contributing to the specificity in non-promiscuous enzymes, in contrast with promiscuous enzymes, which contain signatures distributed uniformly around catalytic sites.

The enolase and aminotransferase superfamilies were chosen as case studies to illustrate the ability of our method to predict catalytic and susbstrate promiscuity within members of the same family. In the case of aminotransferases, it is known that catalytic promiscuity can be achieved by directed evolution of residues in the area surrounding the catalytic site by means of increasing loop flexibility. How this process could be analyzed by reverse engineering using a signature kernel-based SVM prediction, illustrates some of the potential applications of our method, since it can provide useful hints in the design of directed evolution experiments and guide us to discover and enhance latent catalytic capabilities surviving in modern enzymes.

ACKNOWLEDGEMENTS

The authors want to acknowledge the assistance of Chloé Sarnowski in building the promiscuity golden set from BRENDA.

Funding: Genopole® through an ATIGE grant; ANR Chair de Excellence (to J.L.F.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Anandarajah,K. *et al*. (2000) Recruitment of a double bond isomerase to serve as a reductive dehalogenase during biodegradation of pentachlorophenol. *Biochemistry*, **39**, 5303–5311.

Andrianantoandro,E. *et al*. (2006) Synthetic biology: new engineering rules for an emerging discipline. *Mol. Syst. Biol.*, **2**, msb4100073–E1–msb4100073–E14.

Babbitt,P.C. *et al*. (1996) The enolase superfamily: A general strategy for enzyme-catalyzed abstraction of the protons of carboxylic acids. *Biochemistry*, **35**, 16489–16501.

Badia,J. *et al*. (1991) L-lyxose metabolism employs the L-rhamnose pathway in mutant cells of Escherichia coli adapted to grow on L-lyxose. *J. Bacteriol.*, **173**, 5144–5150.

Bartlett,G.J. *et al*. (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**, 105–121.

Bornscheuer,U.T. and Kazlauskas,R.J. (2004) Catalytic promiscuity in biocatalysis: using old enzymes to form new bonds and follow new pathways. *Angew. Chem. Int. Ed*, **43**, 6032–6040.

Carbonell,P. *et al*. (2009) Energetic determinants of protein binding specificity: insights into protein interaction networks. *Proteomics*, **9**, 1744–1753.

Chang,A. *et al*. (2009) Brenda, amenda and frenda the enzyme information system: new content and tools in 2009. *Nucleic Acids Res.*, **37**, gkn820.

Chan,K.K. *et al*. (2008) Structural basis for substrate specificity in phosphate binding (beta/alpha)8-barrels: D-allulose 6-phosphate 3-epimerase from Escherichia coli k-12. *Biochemistry*, **47**, 9608–9617.

Copley,S. (2003) Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol.*, **7**, 265–272.

Delano,W.L. (2002) *The PyMOL User's Manual*. DeLano Scientific, San Carlos.

Faulon,J.-L.L. *et al*. (2004) The signature molecular descriptor. 4. canonizing molecules using extended valence sequences. *J. Chem. Inf. Comput. Sci.*, **44**, 427–436.

Faulon,J.-L. *et al*. (2008) Genome scale enzyme metabolite and drug target interaction predictions using the signature molecular descriptor. *Bioinformatics*, **24**, 225–233.

Gartner,T. *et al*. (2003) On graph kernels: Hardness results and efficient alternatives. *Lect. Notes Comput. Sci.*, **2777**, 129–143.

Glasner,M.E. *et al*. (2006) Evolution of structure and function in the o-succinylbenzoate synthase/n-acylamino acid racemase family of the enolase superfamily. *J. Mol. Biol.*, **360**, 228–250.

Gomez,A. *et al*. (2003) Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins? *Bioinformatics*, **19**, 895–896.

Guyon,I. *et al*. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Hegyi,H. and Gerstein,M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.*, **288**, 147–164.

Hult,K. and Berglund,P. (2007) Enzyme promiscuity: mechanism and applications. *Trends Biotechnol.*, **25**(5), 231–238.

Ishijima,J. *et al*. (2000) Free energy requirement for domain movement of an enzyme. *J. Biol. Chem.*, **275**, 18939–18945.

Jurgens,C. *et al*. (2000) Directed evolution of a (alpha-beta)8-barrel enzyme to catalyze related reactions in two different metabolic pathways. *Proc. Natl Acad. USA*, **97**, 9925–9930.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kanehisa,M. *et al*. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Kazlauskas,R. (2005) Enhancing catalytic promiscuity for biocatalysis. *Curr. Opin. Chem. Biol.*, **9**, 195–201.

Khersonsky,O. *et al*. (2006) Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr. Opin. Chem. Biol.*, **10**, 498–508.

Kim,P.M. *et al*. (2008) The role of disorder in interaction networks: a structural analysis. *Mol. Syst. Biol.*, **4**, 179.

Kuper,J. *et al*. (2005) Two-fold repeated (betaalpha)4 half-barrels may provide a molecular tool for dual substrate specificity. *EMBO Rep.*, **6**, 134–139.

Leslie,C. *et al*. (2003) Mismatch string kernels for SVM protein classification. In Becker,S. *et al*. (eds), *Advances in Neural Information Processing Systems 15*. Cambridge. Proceedings of the Neural Information Processing (NIPS) 2002, MIT Press, pp. 1441–1448.

Linding,R. *et al*. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Macchiarulo,A. *et al*. (2004) Ligand selectivity and competition between enzymes in silico. *Nat. Biotechnol.*, **22**, 1039–1045.

Martin,S. *et al*. (2005) Predicting protein-protein interactions using signature products. *Bioinformatics*, **21**, 218–226.

Ma,W. *et al*. (2005) Specificity of trypsin and chymotrypsin: loop-motion-controlled dynamic correlation as a determinant. *Biophys. J.*, **89**, 1183–1193.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nobeli,I. *et al*. (2009) Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.*, **27**, 157–167.

O'Brien,P.J. and Herschlag,D. (1999) Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.*, **6**, 91–105.

Orengo,C.A. *et al*. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Petsko,G.A. (2000) Enzyme evolution: design by necessity. *Nature*, **403**, 606–607.

Poelarends,G. *et al*. (2008) The chemical versatility of the fold: catalytic promiscuity and divergent evolution in the tautomerase superfamily. *Cell. Mol. Life Sci.*, **65**, 3606–3618.

Porter,C.T. *et al*. (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32** (Suppl. 1), D129–D133.

Romero,P.A. and Arnold,F.H. (2009) Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, **10**, 866–876.

Rothman,S.C. and Kirsch,J.F. (2003) How does an enzyme evolved in vitro compare to naturally occurring homologs possessing the targeted function? tyrosine aminotransferase from aspartate aminotransferase. *J. Mol. Biol.*, **327**, 593–608.

Saigo,H. *et al*. (2004) Protein homology detection using string alignment kernels. *Bioinformatics*, **20**, 1682–1689.

Sitbon,E. and Pietrokovski,S. (2007) Occurrence of protein structure elements in conserved sequence regions. *BMC Struc. Biol.*, **7**, 3.

Taglieber,A. *et al*. (2007) Alternate-site enzyme promiscuity. *Angewandte Chemie*, **46**, 8597–8600.

Tyagi,M. *et al*. (2009) Exploring functional roles of multibinding protein interfaces. *Protein Sci.*, **18**, 1674–1683.

Vapnik,V.N. (1995) *The nature of statistical learning theory*. Springer, New York, NY.

Wu,B.-N.N. *et al*. (2009) Structural modification of acyl carrier protein by butyryl group. *Protein Sci.*, **18**, 240–246.

Yano,T. *et al*. (1998) Directed evolution of an aspartate aminotransferase with new substrate specificities. *Proc. Natl Acad. USA*, **95**, 5511–5515.

Yasutake,Y. *et al*. (2004) Crystal structure of the pyrococcus horikoshii isopropylmalate isomerase small subunit provides insight into the dual substrate specificity of the enzyme. *J. Mol. Biol.*, **344**, 325–333.

Yoshikuni,Y. *et al*. (2006) Designed divergent evolution of enzyme function. *Nature*, **440**, 1078–1082.