BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 2 2008, pages 225–233
doi:10.1093/bioinformatics/btm580

Systems biology

Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor

Jean-Loup Faulon[1,*], Milind Misra[1], Shawn Martin[2], Ken Sale[3] and Rajat Sapra[3]

[1]Sandia National Laboratories, Computational Biosciences Department, P.O. Box 5800, Albuquerque, NM 87185-1413, [2]Sandia National Laboratories, Computer Science & Informatics Department, P.O. Box 5800, Albuquerque, NM 87185-1316 and [3]Sandia National Laboratories, Biosystems Research, P.O. Box 969, Livermore, CA 94551-9291, USA

Received on June 21, 2007; revised on October 21, 2007; accepted on November 19, 2007

Advance Access publication November 23, 2007

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Identifying protein enzymatic or pharmacological activities are important areas of research in biology and chemistry. Biological and chemical databases are increasingly being populated with linkages between protein sequences and chemical structures. There is now sufficient information to apply machine-learning techniques to predict interactions between chemicals and proteins at a genome scale. Current machine-learning techniques use as input either protein sequences and structures or chemical information. We propose here a method to infer protein–chemical interactions using heterogeneous input consisting of both protein sequence and chemical information.

Results: Our method relies on expressing proteins and chemicals with a common cheminformatics representation. We demonstrate our approach by predicting whether proteins can catalyze reactions not present in training sets. We also predict whether a given drug can bind a target, in the absence of prior binding information for that drug and target. Such predictions cannot be made with current machine-learning techniques requiring binding information for individual reactions or individual targets.

Availability and Contact: For questions, paper reprints, please contact Jean-Loup Faulon at jfaulon@sandia.gov. Additional information on the signature molecular descriptor and codes can be downloaded at: http://www.cs.sandia.gov/~jfaulon/publication-signature.html

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

1 INTRODUCTION

Molecular recognition is the primary event involved in the interaction of proteins with other proteins and with small molecules such as metabolites and therapeutics. The ability to predict these interactions on a genome-wide scale is vital in determining the enzymes that catalyze a given metabolic reaction, the reactions a given enzyme catalyzes, the targets a particular drug binds and the drugs that will bind a given target. Predicting these interactions has direct application when completing genome annotations, finding enzymes for synthetic chemistry, and predicting drug specificity, promiscuity and polypharmacology.

Molecular recognition has been studied using a variety of approaches. At the molecular level, biophysical simulation techniques ranging from *ab inito* quantum calculations (Fukuzawa *et al*., 2005) to rapid virtual screening (Warren *et al*., 2006) have been used. These techniques can be very accurate but require knowledge of the three dimensional structure of the chemical–protein interface and can be computationally expensive, giving them limited applicability at the whole genome level. On the other hand, statistical machine-learning approaches can be applied on a much larger scale. The increasing amount of information linking chemical and biological data (Austin *et al*., 2004; Brooksbank *et al*., 2005; Kanehisa *et al*., 2006; Wishart *et al*., 2006) has provided the required input for these machine-learning tools. Machine-learning approaches are computationally efficient, and they do not necessarily require three-dimensional structural information.

To date, machine-learning methods have been applied to molecular recognition questions based on whether they are used with biological sequences or with chemical connectivity data. In bioinformatics, for instance, there are techniques making use of Enzyme Commission (EC) numbers to predict the metabolites a given sequence can catalyze (Borgwardt *et al*., 2005; Cai *et al*., 2003; Kunik *et al*., 2005). There are also sequence and structure-based methods for locating homologous ligand binding sites (Johnson and Church, 2000; Kalinina *et al*., 2004). In cheminformatics, methods have been developed to find compounds binding to a given target (Gasteiger and Engel, 2003), as well as to predict the EC number of a given metabolic reaction (Kotera *et al*., 2004). However, none of these techniques take advantage of the increasing linkage between biological sequences and molecular information. Further, there are no methods available that can predict when an unclassified (in the EC nomenclature) chemical reaction will be catalyzed by a given enzyme, or when a chemical will bind to a sequence in the absence of binding information for either the sequence or

*To whom correspondence should be addressed.

the chemical. To overcome these problems, we have developed an approach making use of a Support Vector Machine (Noble, 2006) (SVM) kernel-based method that operates on both protein and chemical information.

Our approach uses a graph-based representation of molecules known as signature (Faulon, 1994). This representation describes a molecular graph by decomposition into canonical subgraphs (Faulon *et al.*, 2004). Using this type of representation, the similarity between two molecules can be evaluated by comparing their subgraphs (Bender *et al.*, 2004). Similar comparisons have been previously used by exploiting graph kernel SVMs (Gartner *et al.*, 2003; Kashima *et al.*, 2003; Mahe *et al.*, 2006; Swamidass *et al.*, 2005) for the prediction of mutagenecity, toxicity and anti-cancer activities (Swamidass *et al.*, 2005).

In our case, we are interested in chemicals as well as proteins. Perhaps unsurprisingly, functional predictions using protein sequences are also based on pair-wise comparison. These comparisons are typically made using sequence homology, as computed by sequence alignment algorithms, or, in the case of SVMs, remote homology detection with string kernels (Leslie *et al.*, 2002). Protein sequence- based approaches and in particular sequence alignment are routinely used to assign EC numbers in newly sequenced genomes (White, 2006).

The similarities between the chemical-based and sequence-based approaches described above provided the motivation for combining the different methods into a general framework. To accomplish this goal, we used molecular signature as a common representation for both chemicals and protein sequences. Chemicals cannot be coded as linear sequences of well-defined units, but protein sequences are molecules and as such can be represented by molecular graphs. A common representation should thus come from cheminformatics. Our choice of molecular signature (among the myriad cheminformatics descriptors) is motivated by the fact that signatures are molecular fragments. When applied to proteins, these fragments correspond to short sequences, and short sequences correspond to protein domains, which are extensively used in bioinformatics applications (Mulder *et al.*, 2007).

We implemented our approach within the framework of SVMs by developing a kernel that uses the signature descriptor in the context of protein–chemical pairs. Our kernel is based on the signature product kernel developed for protein–protein interaction (Martin *et al.*, 2005) and encompasses ideas related to string kernels (Leslie *et al.*, 2002), graph kernels (Gartner *et al.*, 2003; Kashima *et al.*, 2003; Mahe *et al.*, 2006; Swamidass *et al.*, 2005) and pairwise kernels (Ben-Hur and Noble, 2005; Martin *et al.*, 2005). Details of our techniques are found in the Methods section. The Results section presents enzyme-metabolites binding predictions based on metabolic reaction information, sequence information and the product of both. Next, enzyme–metabolite and drug–target interactions are predicted for independent test sets. In these two latter cases, we investigate the possibilities of predicting enzyme–metabolite and drug–target interactions in the absence of prior binding information in the training sets for the metabolites, enzymes, drugs or targets of the test sets.

2 METHODS

2.1 Molecular signature

The signature of a molecule is a vector whose components correspond to atomic signatures (Faulon, 1994). Each component of a molecular signature counts the number of occurrences of a particular atomic signature in the molecule. 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. Formally, we let ${}^{h}\mathrm{S}=\mathrm{Z}^{N\ (h)}$, where N (h) is the number of possible atomic signatures of height h and the unit basis vectors $\mathrm{Z}^{N\ (h)}$ correspond uniquely to the possible atomic signatures. If $G=(V, E)$ is a molecular graph, with vertex (atom) set V, and edge (bond) set E, then the molecular signature of G is given by

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

where ${}^{h}\sigma_{G}$ (x) is the unit basis vector of ${}^{h}\mathrm{S}$ corresponding to the atomic signature in G rooted at x of height h.

Algorithms and codes to compute atomic and molecular signatures have been previously documented (Faulon *et al.*, 2004). Examples of atomic and molecular signatures are provided in Supplementary Figures S1 and S2. The signature heights used for chemicals typically range between 0 and 6. The signature height used for proteins ranges between 6 and 18 and corresponds roughly to amino acid strings with 1–7 residues. Computationally, the most expensive step in computing a molecular signature is the subgraph canonization step (Faulon *et al.*, 2004). While the computational complexity for canonizing subgraphs (and graphs in general) is unknown, a computational running times study for various types of chemical structures indicate the running time to be linearly proportional to the input size (Faulon *et al.*, 2004). This computational efficiency allows us to process large chemical structures such as proteins with more than 100 000 atoms. We find the computational complexity to only slightly increase with the signature height as on average it takes 6 s CPU time (on a Dual 2.3 GHz PowerPC Macintosh) to compute protein signatures for height 6 and 8 s for height 18. We were thus able to process datasets of 1000 proteins (which was the maximum training and test set size used in this study) in less than 3 h. While not applied in the present study, further saving in running time could potentially be gained by computing signature only for specific atoms along the protein backbone (only C-α atoms, for instance).

2.2 Reaction signature

Reaction signatures are computed for enzymatic reactions. We assume that all enzymatic reactions take the general form R: $s_1\ S_1+s_2\ S_2+\ldots+s_n\ S_n\rightarrow p_1\ P_1+p_2\ P_2+\ldots+p_m\ P_m$, where s_i and p_j are the stoichiometric coefficients of substrates S_i and products P_j. The height h signature of reaction R is then defined by

$$ {}^{h}\sigma(R)=\sum_{j}p_{j}{}^{h}\sigma(P_{j})-\sum_{i}s_{i}{}^{h}\sigma(S_{i}) \tag{2}$$

where ${}^{h}\sigma$ (P_j) and ${}^{h}\sigma$ (S_i) are the height h molecular signatures of substrate S_i and product P_j computed using Equation (1).

2.3 Support vector machines

Support Vector Machines (Noble, 2006) (SVMs) are classifiers that have been widely used in both bioinformatics and cheminformatics. Relevant to this study, SVMs have been used to predict the EC numbers of protein sequences (Borgwardt *et al.*, 2005; Cai *et al.*, 2003; Kunik *et al.*, 2005). In the chemistry literature, SVMs have been used to make property predictions based on molecular connectivity (Kashima *et al.*, 2003; Swamidass *et al.*, 2005).

A SVM can be described as follows. Suppose our data are given as pairs $\left\{(x_i,y_i)\right\} \subseteq \mathrm{R}^n\{\pm 1\}$. In other words, suppose our data consists of two classes with labels 1 and -1. Using this notation, a SVM assumes the form: $f(x) = \sum_i \alpha_i y_i k(x_i, x) + b$ where $f : \mathrm{R}^n \rightarrow \mathrm{R}$ is a decision function [**x** belongs to class 1 if $f(\mathbf{x})$ is greater than some threshold b, or to class –1 otherwise], $k : \mathrm{R}^n \times \mathrm{R}^n \rightarrow \mathrm{R}$ is a kernel function, otherwise known as a dot product in some vector space, and the constants b and α_i are obtained by solving a quadratic programming problem.

A kernel function should measure the similarity between two inputs as a dot product. For sequences, the most commonly used kernels are string kernels (Leslie *et al.*, 2002). A string kernel is a dot product between two vectors containing occurrence numbers of short subsequences within the main sequence. In addition to string kernels, kernels based on feature vectors are also used (Bock and Gough, 2001). These vectors compile physico-chemical parameters collected for each amino acid in a sequence. Various kernels for chemicals have been proposed including path and tree kernels (Swamidass *et al.*, 2005), marginalized kernels (Kashima *et al.*, 2003) and pharmacophore kernels (Mahe *et al.*, 2006). Path and tree kernels count the occurrence of paths and trees and in a molecular graph, the marginalized kernel is based on random walks in a graph, and the pharmacophore kernel takes into account the three dimensional structure of the chemicals. The above kernels have been used to classify chemicals for mutagenicity, toxicity and anti-cancer activities (Swamidass *et al.*, 2005). While kernels already exist for protein sequences and chemicals, our goal in this article is to develop a kernel that will process both simultaneously.

2.4 Signature kernels

Different kernels based on signatures are used for comparing chemicals, reactions and sequences. In all cases, kernels are scalars while signatures are vectors. In the case of chemicals, the kernel function computes the similarity between two chemicals represented by their molecular signatures. In this article, to measure similarity we use a simple dot product between the molecular signature vectors. In the case of chemical reactions, the kernel function computes the similarity between the signatures of the reactions. In the case of protein sequences, the sequences are first transformed into molecular graphs by replacing the amino acids with their chemical structures. The proteins are then processed as chemicals.

When using both chemicals and proteins, as in the case of a drug binding to a target, or a metabolic reaction catalyzed by an enzyme, the kernel is defined for two pairs of chemical–protein interactions. It is the product of the signature kernel between the two chemicals and the signature kernel between the two proteins. We define all of these kernels in Equations. (3–4) below.

For a given height h, the signature kernel between two chemical structures A and B is defined by:

$$^{h}k(A,B) = \frac{^{h}\sigma(A) \cdot {}^{h}\sigma(B)}{\left|^{h}\sigma(A)\right|\left|^{h}\sigma(B)\right|} \tag{3}$$

where $^{h}\sigma$ (A) and $^{h}\sigma$ (B) are computed using Equation (1), and $|\ ^{h}\sigma\ (A)|$ denotes the norm of $^{h}\sigma$ (A). This kernel also applies to proteins considered as atomic structures and reactions, where reaction signatures are used instead of molecular signatures. Examples of signature kernels are given in Supplementary Figure S2.

Let P be a protein and C be a chemical. The height l signature of P can be expressed as: $^{l}\sigma(P) = (p_1, p_2, \ldots, p_n)$ where $n = |\ ^{l}\Sigma\ |$. Similarly the height h signature of C can be written as: $^{h}\sigma(C) = (c_1, c_2, \ldots, c_m)$where $m = |\ ^{h}\mathrm{S}\ |$. Following the definition found in Martin *et al.* (2005) and Ben-Hur and Noble (2005) for protein–protein interaction kernel, we define the signature of the complex $P \otimes C$ as the tensor product of the signatures of P and C:

$$^{l,h}\sigma(P \otimes C) = (p_1c_1, \ldots, p_1c_m, p_2c_1, \ldots, p_2c_m, \ldots, p_nc_1, \ldots, p_nc_m)$$

If we let (P,C) and (Q,D) be two pairs of chemical–protein interactions, using simple algebra we have:

$$\begin{aligned}
&^{l,\,h}\sigma(P \otimes C) \cdot {}^{l,\,h}\sigma(Q \otimes D)\\
&= (p_1c_1,\ \ldots,\ p_1c_m,\ p_2c_1,\ \ldots,\ p_2c_m,\ \ldots, p_nc_1,\ \ldots, p_nc_m)\\
&\quad \cdot (q_1d_1,\ \ldots, q_1d_m, q_2d_1,\ \ldots, q_2d_m,\ \ldots, q_nd_1,\ \ldots, q_nd_m)\\
&= p_1q_1(c_1d_1 + \ldots + c_md_m) + p_2q_2(c_1d_1 + \ldots + c_md_m)\\
&\quad + \ldots + p_nq_n(c_1d_1 + \ldots + c_md_m)\\
&= (p_1q_1 + p_2q_2 + \ldots + p_nq_n)(c_1d_1 + c_2d_2 + \ldots + c_md_m)\\
&= (^{l}\sigma(P) \cdot {}^{l}\sigma(Q))\ (^{h}\sigma(C) \cdot {}^{h}\sigma(D))
\end{aligned}$$

Similarly it can easily be shown that:

$$\left|^{l,h}\sigma(P \otimes C)\right|\left|^{l,h}\sigma(Q \otimes D)\right| = \left|^{l}\sigma(P)\right|\ \left|^{l}\sigma(Q)\right|\ \left|^{h}\sigma(C)\right|\ \left|^{h}\sigma(D)\right|$$

Thus the signature product kernel between two pairs (P,C) and (Q,D) of protein–chemical interactions, is simply defined as:

$$^{l,h}k_p((P, C), (Q, D)) = {}^{l}k(P, Q)^{h}k(C, D) \tag{4}$$

where ^{l}k (P,Q) is the signature kernel of height l for the protein pair (P, Q) and ^{h}k (C,D) is the signature kernel of height h for the chemical pair (C,D). Both kernels are defined using Equation (3). In other words, the similarity of two protein–chemical pairs is simply the product of the similarity between the two proteins and the similarity between the two chemicals. An example of signature product kernel is given in Supplementary Figure S3.

2.5 Cross-validation

Several metabolite-enzyme and drug-target datasets were processed in this study, these are detailed in the Result sections. Once a dataset was generated, prediction accuracy was assessed using cross-validation. X-fold cross-validation is performed by dividing a dataset into X equal non-intersecting subsets. A given subset is treated as a test set, and the complement serves as a training set. A classifier is trained using the training set and predictions are made on the test set. This procedure is repeated for each of the original subsets to obtain predictions on the entire dataset. In this work, we used either leave-one-out (LOO) or 5-fold cross-validation. LOO cross-validation uses each element of the dataset separately as a test set and 5-fold cross-validation splits the dataset into five equal subsets.

Statistics were compiled for each dataset using the cross-validation predictions. Accuracy, sensitivity, specificity and precision were computed. Using *TP*, *FP*, *TN* and *FN* to denote true positive, false positives, true negatives and false negatives, accuracy is defined by (TP + TN)/(TP + FP + TN + FN), sensitivity is defined by TP/(TP + FN), specificity is defined by TN/(TN + FP) and precision is defined by TP/(TP + FP). In addition, we computed the Jaccard coefficient $J = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$ and the area under the receiver-operator-characteristic (ROC) curve. The area under the curve (auc) is obtained by integrating under the ROC curve. The ROC curve is obtained by varying the threshold b (see SVM section) separating positives from negatives and plotting the TP rate (sensitivity) versus the FP rate (1-specificity). For all of these statistics, a larger number indicates a better result.

3 RESULTS

Our approach rests on the assumption that we can replace protein homology calculations based on sequence with similar calculations based on the underlying atomistic representation. Our first result is therefore related to the investigation of this claim. Next, we compare our approach to existing methods using only chemical or only sequence information. Then, we combine both sequence and chemical information for the prediction of reaction–enzyme and drug–target interactions. Finally, we investigate the efficacy of our method for predictions in the absence of prior binding information for reaction, enzyme, drug or target.

3.1 Signature similarity measurements and sequence alignment scores

Sequence alignment algorithms such as Altschul *et al.* (1997) use substitution matrices to compute similarity scores. The procedure is straightforward and consists of summing up the scores found in substitution matrices (such as BLOSUM62) for the amino acids that are aligned. BLOSUM62 is a 20×20 matrix of amino acids substitution scores (Henikoff and Henikoff, 1992). A score between two amino acids is the logarithm of the ratio of the likelihoods of these two amino acids to be substituted. In BLOSUM62 the substitution likelihood between two given amino acids was originally computed from the frequencies observed for finding the two amino acids aligned in a large set of trusted alignments having at least 62% identity.

To investigate the relationship between sequence alignment scores and similarity obtained using the signature kernel [c.f. Equation (3)], the kernel was computed for every pair of amino acids found in BLOSUM62. The results shown in Figure 1 demonstrate a clear correlation, indicating that residues that are chemically similar generally have a high BLOSSUM62 value and conversely. For example, while the signature kernel value between Leucine and Isoleucine is 1.0 (c.f. Supplementary Fig. S2) and its BLOSUM62 value is 2, the BLOSUM62 value for Glycine and Isoleucine is −4 and the signature kernel value is 0.42 (c.f. Supplementary Fig. S2). Figure 1 confirms previous findings relating amino acids substitution scores and their physico–chemical properties (Atchley *et al.*, 2005). Indeed, since properties are structure dependent, if BLOSUM62 scores are related to amino acid properties, some relationship should be found between BLOSUM62 scores and amino acid atomistic structures. Our relationship is interesting because it occurs despite the fact that BLOSUM62 scores were obtained without regard to the amino acid chemical structures.

3.2 Chemicals classification

To benchmark our signature kernel with other state-of-the-art kernel we used the Predictive Toxicology Challenge (PTC) dataset (Helma *et al.*, 2001), which reports the carcinogenicity of 417 chemical compounds for male mice (MM), female mice (FM), male rats (MR) and female rats (FR). Because toxicology is generally hard to predict, this dataset has been used previously to evaluate various kernel methods (Swamidass *et al.*, 2005). Supplementary Table S1 gives a comparison of the signature kernel with a frequent pattern discovery approach (Kashima *et al.*, 2003), a marginalized kernel (Kramer and De Raedt, 2001), SMILES string kernels, pathcount-based graph kernels and a kernel based on three dimensional atomic distances (Swamidass *et al.*, 2005). Our kernel compares well with these state-of-the-art kernels for chemical heights ranging between 1 and 4.

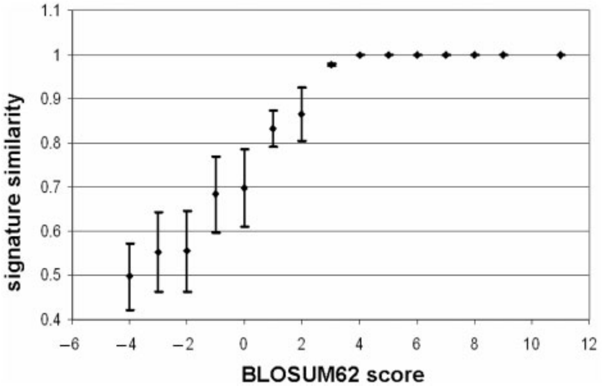

Fig. 1. BLOSUM62 score versus signature kernel similarity. This figure shows the average height 1 signature kernel similarity [computed using Equation (3)] for every pair of amino acids having a given BLOSUM62 score. Error bars indicate SD for each averaged signature similarity value. The elements in the diagonal of the BLOSUM62 matrix are identical (kernel value = 1) but have different scores. The score for diagonal elements is not related to substitution frequencies between two different amino acids, but to the abundance of the amino acids in the set used to build the matrix (Henikoff and Henikoff, 1992). The regression coefficient omitting the diagonal elements is $r^2 = 0.97$.

3.3 EC number classification using metabolic reactions

Enzymes are organized according to the Enzyme Commission (EC) system, a hierarchical classification that assigns unique four-field numbers to different enzymatic activities (Webb, 1992). The first field of an EC number indicates the general class of catalyzed reaction: 1 denotes oxidoreductases, 2 denotes transferases, 3 denotes hydrolases, 4 denotes lyases, 5 denotes isomerases and 6 denotes ligases. The second and third fields depend on different criteria related to the chemical features of the substrate and the product of the reaction. The fourth field is substrate and product specific. As an example, the tripeptide aminopeptidases have the number ‘EC 3.4.11.4’. Level 1 ‘EC 3’ enzymes are hydrolases. Level 2 ‘EC 3.4’ enzymes are hydrolases that act on peptide bonds. Level 3 ‘EC 3.4.11’ enzymes are hydrolases that cleave off the amino-terminal amino acid from a polypeptide, and level 4 ‘EC 3.4.11.4’ enzymes are those that cleave off the amino-terminal end from a tripeptide.

To predict the EC numbers of metabolic reactions, we created a training set by downloading the KEGG database. As of 21 November 2006, this database contained 10 951 compounds and 6556 reactions involving compounds with molecular structures stored in the database. Positive examples were compiled using all reactions in the KEGG database having a specified EC number. Wild cards were allowed in order to generate datasets at various EC levels. As an example, EC level 2 ‘1.1.*.*’ consisted of 409 reactions having an EC number starting with 1.1. Positive example sets having less than 50 elements were not processed. If the positive set included more than 500 examples, then excess examples were removed at random. Next, the datasets were completed by taking equal numbers of negative examples at random. Negative examples were composed of reactions not present in the positive class. The fourth EC level could not be processed because it usually contains only one reaction. The signature reaction kernel [Equation (2)] was applied in order to classify metabolic reactions. Cross-validation results are presented in Supplementary Figure S4a and Table S2. Maximum accuracies are 91% for level 1, 84% for level 2 and 88% for level 3. These accuracies, obtained for heights ranging between 1 and 6, are lower than those published in Kotera *et al.* (Kotera *et al.*, 2004). However, the signature kernel can process a larger number of reactions (c.f. Supplementary Table S2).

3.4 EC number classification using protein sequence

To predict EC numbers using protein sequence, we downloaded protein sequences from the KEGG database having one or more EC number assignments, resulting in 308 094 sequences. Training sets were constructed for all four EC levels as described in Section 3.3 using protein sequences instead of reactions. The signature kernel [Equation (1)] was computed using several heights for each protein sequence using an atomistic description of the protein. Cross-validation results are reported in Supplementary Figure S4b. We find the accuracies peak at height 8 for levels 1, 2 and 3 (values are 60%, 79% and 81%, respectively), and at height 6 for level 4 (value is 95%). These accuracies are generally lower than those reported using protein structure graph kernels (Borgwardt *et al.*, 2005), feature vectors [i.e. SVM-PROT (Cai *et al.*, 2003)] or motifs (Kunik *et al.*, 2005). Nevertheless, the accuracies are high enough that the signature kernel can be used with protein sequence alone to predict EC numbers (except for EC level 1). When both the sequence and the reaction are used, the accuracy improves as shown in the next section.

3.5 EC number classification using both protein sequences and metabolic reactions

Using the reactions and sequences with assigned EC numbers in the KEGG database, a set of 855 772 pairs (out of 3905 reactions and 255 304 enzymes) was compiled. Training sets were constructed for all four EC levels as described in Section 3.3 taking reaction-protein pairs instead of reaction only. Signatures of various heights were computed for both the reactions and the atomistic representations of the proteins. The signature product kernel [c.f. Equation (4)] was applied for each EC level. Cross-validation results are reported for classes 1, 1.1, 1.1.1 and 1.1.1.1 in Supplementary Figure S5. We find the accuracy for class 1 peaks at height 4 for chemicals and height 8 for proteins. For class 1.1, the largest accuracy is obtained at heights 1 and 8 for reactions and proteins; for class 1.1.1, the accuracy peaks at heights (3,10) and for class 1.1.1.1, maximum accuracy is reached at heights (3,6). Taking the above optimum heights, the signature product kernel is compared with the protein structure graph kernel (Borgwardt *et al.*, 2005) in Supplementary Table S3, the SVM-PROT feature vectors method (Cai *et al.*, 2003) in Supplementary Table S4, the motifs technique of Kunik *et al.* (2005) in Supplementary Tables S5. Additional results are reported for all classes 1.1.1.* in Supplementary Table S6. Table 1 (which summarized results reported in Supplementary Tables S3–S6) shows that the signature product performs comparably with the other competing techniques. Better sensitivity (accuracy on positives) and Jaccard coefficient are obtained with the signature product because the competing techniques were trained on unbalanced training sets comprising more negatives than positives. This imbalance appears to increase the number of false negative predictions thus reducing sensitivity and Jaccard coefficient.

Table 1. Statistics for the signature product kernel [Equation (4)] compared with other SVM techniques for predicting EC numbers at the different hierarchical levels

Level	Method	Acc.	Auc	Prec.	Sens.	Spec.	J
L1	Graph	89.9			40.0	99.9	
	kernel[1]	±1.3			±7.4	±0.1	
	Signature	88.0	91.8	87.1	89.6	86.4	79.4
	Product	±6.1	±4.6	±6.8	±6.1	±7.2	±9.6
L2	SVM-	95.2		97.4	77.4	88.6	70.6
	PROT[2]	±5.4		±3.8	±14.1	±7.6	±13.8
	Signature	94.2	96.0	93.6	95.2	93.3	89.6
	Product	±4.9	±4.0	±5.7	±4.6	±6.3	±8.1
L3	MEX-						89.3
	motifs[3]						±8.0
	Signature	97.9	98.9	97.9	97.9	97.9	96.0
	Product	±2.3	±1.2	±3.1	±2.1	±3.4	±4.1
L4	Signature	99.0	99.2	98.7	95.5	98.7	98.2
	Product	±2.3	±2.0	±2.7	±2.9	±2.7	±4.1

All values are averaged from Supplementary Tables S3-6, with SDs indicated by ±. Acc. stands for accuracy, Auc for area under the curve, Prec. for precision, Sens. for sensitivity, Spec. for specificity and J for Jaccard coefficient. All these statistical parameters are defined in Section 2.5.

[1] From Borgwardt *et al.* (2005).

[2] From Cai *et al.* (2003).

[3] From Kunik *et al.* (2005).

3.6 Predicting new enzyme–metabolite interactions

To test the ability of the signature product kernel to predict enzyme–metabolite interactions not present in training sets, all enzymes and reactions corresponding to EC numbers accepted in September 2006 by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) were removed from the KEGG database. Test sets composed of removed enzymes and reactions were created for each selected EC numbers. For each test set, a training set composed of 500 positive examples was constructed from the remaining KEGG database. An equal number of negative interactions were added to both training and test sets by choosing at random interactions between reactions and enzymes not present in the reduced KEGG database. SVMs were trained with the signature product kernel using

Table 2. Prediction statistics for reaction–enzyme interactions not classified by an EC number

EC class	Number of Pairs	Acc.	Auc	Prec.	Sens.	Spec.	J
EC 1.1.1.290 4-phosphoerythronate dehydogenase	59	88.7	87.8	82.6	98.3	79.1	81.4
EC 1.13.11.52 indoleamine 2,3-dioxygenase	13	76.9	71.4	76.9	76.9	76.9	62.5
EC 1.13.11.53 acireductone dioxygenase (Ni2+-requiring)	11	86.4	79.4	83.3	90.9	81.8	76.9
EC 1.2.1.71 succinylglutamate-semialdehyde dehydrogenase	55	87.5	88.6	82.3	95.5	79.5	79.3
EC 1.2.1.72 erythrose-4-phosphate dehydrogenase	46	88.0	90.0	80.8	100.0	76.1	80.8
EC 1.8.4.11 peptide-methionine (S)-S-oxide reductase	390	79.5	99.8	99.8	59.1	99.9	59.0
EC 2.6.1.81 succinylornithine transaminase	21	81.0	95.4	72.7	100.0	61.9	72.7
EC 3.1.3.77 acireductone synthase	160	89.4	96.1	82.5	100.0	78.8	82.5
EC 3.3.2.9 microsomal epoxide hydrolase	17	84.3	90.5	82.3	88.2	80.4	73.9
EC 3.5.1.96 succinylglutamate desuccinylase	49	88.6	84.7	81.6	100.0	77.1	81.6
EC 3.5.3.23 N-succinylarginine dihydrolase	51	90.2	91.9	83.7	100.0	80.4	83.7
EC 4.2.1.109 methylthioribulose 1-phosphate dehydratase	12	87.5	92.9	80.0	100.0	75.0	80.0
Average		85.7	89.0	82.4	92.4	78.9	76.2
		±4.3	±7.7	±6.3	±12.6	±8.4	±7.9

All reactions and enzymes corresponding to the EC numbers listed in the first column were removed from the KEGG database and stored in test sets (c.f. Table 1 for definition of statistical parameters).

height 3 for reactions, and heights 6, 8 and 10 for proteins. These heights were selected because they lead to high accuracies in cross-validation studies (c.f. Section 3.5). Results are given in Table 2 using height 3 for reactions and 10 for proteins, similar results were obtained for heights (3,6) and (3,8). The results show that it is possible to predict (with accuracies above 80%) whether or not a given enzyme will catalyze a given reaction, even when the EC class of the enzyme reaction pair is not present in the training set.

3.7 Predicting new drug–target interactions

A dataset linking drugs with protein targets was created from the KEGG database. This dataset contained 873 drug–target pairs taken from 121 targets and 551 drugs. A training set was constructed using all 873 positive pairs. An equal number of negatives were added by selecting random pairs that did not belong to the set of positives. Cross-validation results were obtained using 5-fold cross-validation with the signature product kernel, as reported in Supplementary Figure S6 for various drug and target signature heights. Best accuracies (>85%) are obtained when the drug height is in the range 2–3 and the target height in the range 10–14.

The same training set was used to predict if any of the drug target binding pairs stored in the DRUGBANK database could be predicted using the SVM trained on the KEGG database. As of 1 December 2006, DRUGBANK was composed of 1133 drugs and 509 targets, with 1849 drug–protein pairs. A test set was created from DRUGBANK by including drugs having a KEGG reference name. Drugs having no KEGG references were removed to avoid discrepancies in chemical structure representations, e.g. KEGG stores hydrogen suppressed structures in mol files, while DRUGBANK has full hydrogen structures. From the initial 1113 drugs, 124 had a KEGG reference. These 124 drugs were bound to 117 targets resulting in 298 pairs. Out of the 124 drugs and 117 targets, 24 drugs and 50 targets were present in the training set. Further classification of the DRUGBANK test pairs, depending on whether or not the target or the drug were in the KEGG training set, is given in the caption of Table 3. Negatives were constructed forming pairs at random between the 124 drugs and 117 targets of the DRUGBANK test set. Calculations were run using the signature product kernel [Equation (4)] with a height 2 for drugs and 10 for targets, these heights were selected because of their high cross-validation accuracies (c.f. Supplementary Fig. S6). The prediction efficiencies on this test set are presented in Figure 2 and Table 3.

Table 3. DRUGBANK test set predictions

Class	POS	NEG	TP	FP	TN	FN
I	32	0	32	0	0	0
II	11	40	9	10	30	2
III	145	144	118	64	80	27
IV	16	26	5	7	19	11
V	94	88	67	46	42	27
All	298	298	231	127	171	67

The table reports prediction results for an FP rate of 0.42. This rate maximizes the overall accuracy (67.5%). The results are divided into five possible cases, depending on whether or not the targets or the drugs from the DRUGBANK test set were present in the KEGG training set. Class I is composed of the 32 interactions common between KEGG and DRUGBANK. Class II contains cases where both the drug and the target are in the training set, albeit with different partners. Class III contains cases where the targets are in the training set, but the drugs are absent. Class IV contains cases where the drugs are in the training set, but not the targets. Finally, class V contains cases where neither the target nor the drug is in the training set.

Despite the small number of common interactions between KEGG and DRUGBANK (32), $231 - 32 = 199$ additional interactions not in KEGG predicted by the signature product kernel were found in DRUGBANK, including 67 interactions between drugs and targets not present in KEGG (c.f. Table 3).

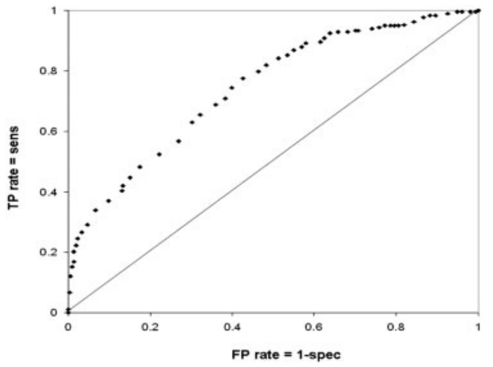

Fig. 2. ROC curve for predicted DRUGBANK drug–target interactions using a drug-target KEGG training set. The ROC curve in this figure was generated by varying the threshold for calling a test pair positive or negative. The area under the ROC curve is 0.74.

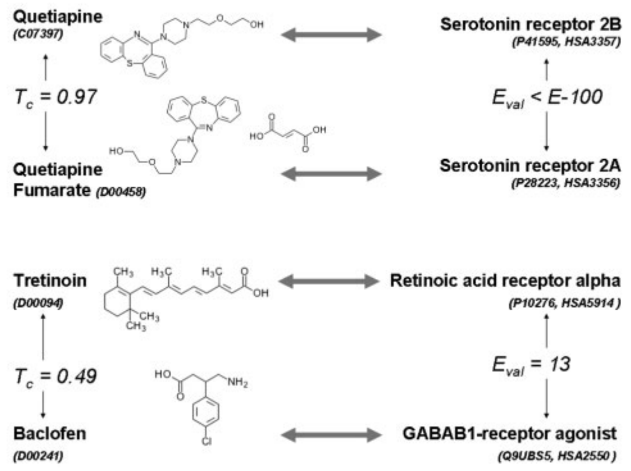

Fig. 3. Homologous drug–target pairs between DRUGBANK test set and KEGG training set. (Top) The closest homolog of the test pair (Quetiapine, Serotonine receptor 2B) is the training set pair (Quetiapine Fumarate, Serotonine receptor 2B). KEGG compound names and target names are indicated in parentheses. (Bottom) The closest homolog of the test pair (Tretinoin, retinoic acid receptor alpha) is the pair (Baclofen, GABAB1-receptor agonist). T_c stands for Tanomito coefficient and E_{val} is the expected value of BLASTP alignment score (computational details for T_c and E_{val} are given in the text).

Some of the predictions made by the signature product kernel even when the drugs and the targets are not present in the training set (class V in Table 3) are rather obvious. As shown in Figure 3, predicting that Quetiapine binds serotonin receptor 2B could have easily been guessed from the training set pair Quetiapine Fumarate binds serotonin receptor 2A. However, this is not the case when predicting Tretinoin binds to retinoic acid receptor alpha since the closest homolog in the training set (Baclofen binds GABAB1-receptor agonist) is rather dissimilar to the tested pair. To systematically evaluate all class V predictions made by the signature product kernel, we computed for each of the 67 pairs its closest homolog in the training set. Precisely, for each pair we searched the pair in the training leading to the highest Tanimoto coefficient (T_c) between the drugs, and the lowest BLAST E-value between the targets. T_c is a widely used measure of similarity between two chemicals (Gasteiger and Engel, 2003); T_c ranges between 0 and 1 and the closer to 1 the more similar are the chemicals. In this study, T_c was evaluated using a standard procedure (Swamidass and Baldi, 2007), which consists of computing 512 bits fingerprints counting paths up to 7 atoms. BLAST E-values were returned running BLASTP (Altschul *et al.*, 1997) using BLOSUM62 as the scoring matrix. Two examples of homologous pairs are given in Figure 3. Results for all 67 tested pairs are listed in Supplementary Table S7. Setting up rather weak thresholds of $T_c \geq 0.50$ and E-value $\leq$ 10, to maximize the number of potential homologs, we find that 17 of the 67 predictions could have been guessed from similarity and homology calculations. The fact that the signature product kernel can find interactions even when homologs cannot be found in the training set should not come as a surprise. First, the signature product kernel is used in a supervised learning context, where learning is performed on protein–chemical pairs, while homology and similarity calculations are unsupervised. Second, related kernels such as the string kernel (Leslie *et al.*, 2002) and the motif kernel (Kunik *et al.*, 2005) are known to perform well to classify remote homologues (c.f. Table S5 where motifs kernel is compare with sequence alignment).

4 DISCUSSION

We have proposed a unified method for predicting protein–chemical interactions based on the representation of a protein using its atomistic structure. We found by a comparison with sequence alignment scores that our method could be used to detect protein homology, even though (strictly speaking) we ignore sequence information in favor of molecular structure. Looking at the atomistic structures of the amino acids, it can be demonstrated that a height 2 atomic signature can uniquely characterize each amino acid, every dimer can be characterized using a height 8 atomic signature, every trimer with a height 9, every tetramer with a height 11 and so on. Thus, when two proteins have many signatures in common, they share many amino acid strings and thus should have some degree of homology. Sharing strings certainly increases alignment score, but there are some instances where strings made of different amino acids can lead to high homology scores. Sequence alignment algorithms incorporate this information in substitution matrices such as BLOSUM62. The surprising implication of our comparison with the sequence alignment score is that an *atomistic structure representation of proteins encompasses information stored in substitution matrices*. This is the key observation that underlies our method.

Because of the key observation just discussed, we find that the signature kernel and the product kernel can be used to predict whether a given enzyme sequence will catalyze a given reaction. This problem can be solved by other machine-learning methods depending on the database available to construct the training sets. There are five cases. In the first case, both the sequence and the reaction are in the database and they have the same EC number; then the problem is solved and no prediction is needed. In the second case, the sequence and the reaction are in the database albeit with different EC numbers. Note that this situation can arise because enzymes can process several reactions. In such an instance, a training set can be collected taking all the enzymes known to catalyze the given reaction and testing using machine learning the given sequence of interest. If the given reaction has several known enzymes then sequence-based bioinformatics techniques (Borgwardt *et al.*, 2005; Cai *et al.*, 2003; Kunik *et al.*, 2005) can perform the task. The third case arises when the reaction is in the database but not the sequence; as in second case, sequence-based bioinformatics techniques can be used, provided several enzymes are known to catalyze the reaction. The last two cases correspond to situations where the sequence is in the database set but not the reaction, or neither the sequence nor the reaction are in the database. In such instances, training sets cannot be constructed because no enzymes are known to catalyze the reaction, and classical bioinformatics techniques cannot be used. The signature product kernel can process all five cases, it has accuracies comparable to other techniques (c.f. Table 1) in the first three cases and reaches accuracies above 80% in cases where other techniques cannot be used (c.f. Table 2).

The traditional method for constructing a metabolic map of a newly sequenced organism is to assign EC numbers to its proteins. Many proteins remain un-annotated not only because their sequences have not been mapped to an already classified enzyme, but also because the reactions catalyzed by the proteins have not been characterized in the EC nomenclature. EC number assignment requires published evidence and full characterization of the enzymatic reaction. For this reason many reactions, although occurring in various pathways, do not have an assigned EC number. Again as shown in Table 2 our method can predict when an enzyme will catalyze a metabolic reaction, even in the absence of any EC nomenclature information.

A similarly important result, again following from our key observation, is that signature and signature product kernels can be used to predict drug–target interactions. The fact that the signature kernel can predict when a new drug will bind to a given target when other binders are known is not unexpected. Indeed, it has previously been demonstrated that signature and signature-like descriptors can be used to derive structure–activity relationships, and thus predict other drugs binding to a specific target (Churchwell *et al.*, 2004; Faulon *et al.*, 2003). The real strength of the signature product kernel is its ability to detect new interactions when the targets and the drugs are not present in the training set (c.f. Table 3 and Fig. 3). Such predictions cannot be made with classical cheminformatics techniques such as chemical similarity and structure–activity relationships.

As already mentioned, several cheminformatics and bioinformatics methods exist to predict protein-chemical binding. When used in a machine-learning context, traditional cheminformatics methods require training sets composed of chemicals binding to the same protein; bioinformatics methods require sequences catalyzing the same chemicals. The technique we have presented here does not outperform these traditional methods provided training sets can be constructed and the methods can be applied (c.f. Sections 3.1 and 3.2 for cheminformatics methods and Sections 3.4 and 3.5 for bioinformatics methods). The real strength of the proposed technique is its ability to handle both proteins and chemicals using a common representation (e.g. the signature). This common representation allows us to train machine learning directly on protein–chemical pairs, rather than on protein sequences and chemical structures alone.

While the technique we have proposed deals with situations where traditional bioinformatics and cheminformatics methods fail, it has its own limitation. First, the technique requires training, and its accuracy depends on the quality and completeness of the training set. Second, as we have shown, the technique makes use of the signature product kernel for measuring similarity between protein–chemical pairs. The similarity between two chemicals (or two proteins) is based on the number of signatures the chemicals (or proteins) have in common. Thus, a given protein–chemical pair can be predicted accurately only when proteins and chemicals can be found in the training set having signatures in common with the tested pair.

ACKNOWLEDGEMENTS

This work was funded by Sandia Laboratory Directed Research and Development. Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Atchley,W.R. *et al.* (2005) Solving the protein sequence metric problem. *Proc. Natl Acad. Sci. USA*, **102**, 6395–6400.

Austin,C.P. *et al.* (2004) NIH molecular libraries initiative. *Science*, **306**, 1138–1139.

Ben-Hur,A. and Noble,W.S. (2005) Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **21** (Suppl. 1), i38–i46.

Bender,A. *et al.* (2004) Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.*, **44**, 1708–1718.

Bock,J.R. and Gough,D.A. (2001) Predicting protein-protein interactions from primary structure. *Bioinformatics*, **17**, 455–460.

Borgwardt,K.M. *et al.* (2005) Protein function prediction via graph kernels. *Bioinformatics*, **21** (Suppl. 1), i47–i56.

Brooksbank,C. *et al.* (2005) The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.*, **33** (Database issue), D46–D53.

Cai,C.Z. *et al.* (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.*, **31**, 3692–3697.

Churchwell,C.J. *et al.* (2004) The signature molecular descriptor. 3. Inverse-quantitative structure-activity relationship of ICAM-1 inhibitory peptides. *J. Mol. Graph. Model.*, **22**, 263–273.

Faulon,J.L. (1994) Stochastic generator of chemical structure. 1. Application to the structure elucidation of large molecules. *J. Chem. Inf. Comput. Sci.*, **34**, 1204–1218.

Faulon,J.L. *et al.* (2003) The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. *J. Chem. Inf. Comput. Sci.*, **43**, 721–734.

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

Fukuzawa,K. *et al*. (2005) Ab initio quantum mechanical study of the binding energies of human estrogen receptor alpha with its ligands: an application of fragment molecular orbital method. *J. Comput. Chem.*, **26**, 1–10.

Gartner,T. *et al*. (2003) On graph kernels: hardness results and efficient alternatives. In *Proceedings of the Sixteenth Annual Conference on Computational Learning Theory and Seventh Kernel Workshop*. Springer Verlag, New York.

Gasteiger,J. and Engel,T. (2003) *Chemoinformatics*. Wiley-VCH, Weinheim.

Helma,C. *et al*. (2001) The predictive toxicology challenge 2000–2001. *Bioinformatics*, **17**, 107–108.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Johnson,J.M. and Church,G.M. (2000) Predicting ligand-binding function in families of bacterial receptors. *Proc. Natl Acad. Sci. USA*, **97**, 3965–3970.

Kalinina,O.V. *et al*. (2004) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res.*, **32** (Web Server issue), W424–W428.

Kanehisa,M. *et al*. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34** (Database issue), D354–D357.

Kashima,H. *et al*. (2003) Marginalized kernels between labeled graphs. In *Proceedings of the Twentieth International Conference on Machine Learning*. AAAI Press, Washington DC.

Kotera,M. *et al*. (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.*, **126**, 16487–16498.

Kramer,S. and De Raedt,L. (2001) Feature construction with version spaces for biochemical applications. In *Eighteenth International Conference on Machine Learning Table of Contents*. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Kunik,V. *et al*. (2005) Motif extraction and protein classification. *Proc. IEEE Comput. Syst. Bioinform. Conf.*, **4**, 80–85.

Leslie,C. *et al*. (2002) The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomput.*, 564–575.

Mahe,P. *et al*. (2006) The pharmacophore kernel for virtual screening with support vector machines. *J. Chem. Inf. Model.*, **46**, 2003–2014.

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

Mulder,N.J. *et al*. (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35** (Database issue), D224–D228.

Noble,W.S. (2006) What is a support vector machine? *Nat. Biotechnol.*, **24**, 1565–1567.

Swamidass,S.J. and Baldi,P. (2007) Mathematical correction for fingerprint similarity measures to improve chemical retrieval. *J. Chem. Inf. Model.*, **47**, 952–964.

Swamidass,S.J. *et al*. (2005) Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, **21** (Suppl. 1), i359–i368.

Warren,G.L. *et al*. (2006) A critical assessment of docking programs and scoring functions. *J. Med. Chem.*, **49**, 5921–5931.

Webb,E.C. (1992) *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Academic Press, San Diego, CA.

White,R.H. (2006) The difficult road from sequence to function. *J. Bacteriol.*, **188**, 3431–3432.

Wishart,D.S. *et al*. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34** (Database issue), D668–D672.