Ligand Prediction for Orphan Targets Using Support Vector Machines and Various Target-Ligand Kernels Is Dominated by Nearest Neighbor Effects

Anne Mai Wassermann, Hanna Geppert, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstrasse 2, D-53113 Bonn, Germany

Received July 21, 2009

Support vector machine (SVM) calculations combining protein and small molecule information have been applied to identify ligands for simulated orphan targets (i.e., targets for which no ligands were available). The combination of protein and ligand information was facilitated through the design of target-ligand kernel functions that account for pairwise ligand and target similarity. The design and biological information content of such kernel functions was expected to play a major role for target-directed ligand prediction. Therefore, a variety of target-ligand kernels were implemented to capture different types of target information including sequence, secondary structure, tertiary structure, biophysical properties, ontologies, or structural taxonomy. These kernels were tested in ligand predictions for simulated orphan targets in two target protein systems characterized by the presence of different intertarget relationships. Surprisingly, although there were target- and set-specific differences in prediction rates for alternative target-ligand kernels, the performance of these kernels was overall similar and also similar to SVM linear combinations. Test calculations designed to better understand possible reasons for these observations revealed that ligand information provided by nearest neighbors of orphan targets significantly influenced SVM performance, much more so than the inclusion of protein information. As long as ligands of closely related neighbors of orphan targets were available for SVM learning, orphan target ligands could be well predicted, regardless of the type and sophistication of the kernel function that was used. These findings suggest simplified strategies for SVM-based ligand prediction for orphan targets.

1. INTRODUCTION

The search for active compounds is one of the major focal points of chemoinformatics research and applications for which machine learning approaches are increasingly utilized.[1] The term Support Vector Machine (SVM) refers to an advanced machine learning methodology[2,3] that was originally developed for binary object classification. For SVM learning, objects with known class labels are projected into a feature space, and the learning process generally attempts to identify a hyperplane in this space that best separates objects belonging to two classes. The resulting linear function is then applied to predict class labels of other objects. The use of kernel functions[4,5] is a key feature of SVM learning because they make it possible to also solve nonlinear classification problems. In chemoinformatics, SVM has become increasingly popular for the classification of active versus inactive compounds and, in addition, the prediction of ligands for given protein targets.[6−9] Ligand prediction via SVM includes investigations that aim at predicting active small molecules for target proteins for which no ligand information is available. This search for ligands of so-called orphan targets is highly relevant for computer-aided drug discovery and chemical biology. Regardless of the methods that are applied for this purpose, finding ligands for orphan targets generally requires to relate biological target and chemical ligand information to each other. Using SVM, this task can be elegantly accomplished by designing separate kernel functions for pairs of proteins and pairs of ligands.[8,9] State-of-the-art protein kernels for ligand prediction include, for example, a sequence homology-based classification kernel.[9] Furthermore, among various ligand descriptors that can be used, 2D molecular fingerprints have been found to be efficient small molecule representations for SVM.[10] Fingerprint similarity can be captured, for example, by encoding the Tanimoto coefficient (Tc)[11] as a kernel function. However, many other types of kernel functions combining biological target and chemical ligand information can be envisioned, and one might expect that the information content of such kernels is a critical factor for SVM-based ligand prediction.

In this study, we have investigated the role of kernel functions for the prediction of ligands for orphan targets with a particular focus on target kernels capturing protein information at rather different levels, beyond sequence similarity. We have carried out calculations using three different SVM strategies, including standard SVM, the recently introduced SVM linear combination,[9] and target-ligand kernel SVM.[8,9] Three ligand kernels were tested that compare small molecules in different ways, utilizing 2D fingerprints as molecular representations. For the linear combination and target-ligand kernel SVM strategies that learn from multiple targets, eleven alternative and conceptually different target kernel functions were designed. The three SVM strategies were applied to search for

* Corresponding author phone: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

10.1021/ci9002624 CCC: $40.75

inhibitors of individual proteases in two different target sets that were regarded as orphan targets and hence not included during SVM learning. The results obtained in systematic search calculations using alternative kernel functions were rather unexpected, and their further analysis revealed a strong dependence of successful ligand prediction on the nearest neighbor reference target of a simulated orphan target, irrespective of the kernel functions and SVM search strategies that were used. Thus, the information provided by the reference target most closely related to an orphan target largely determined the success rate of ligand predictions.

2. SUPPORT VECTOR MACHINE THEORY

2.1. Simple SVM. SVMs are machine learning algorithms for binary object classification. A linear decision function is built based on a training data set to associate class labels of objects with feature vectors. SVM learning for the purpose of virtual compound screening makes use of training examples $(\mathbf{x}_i,y_i)$ with $\mathbf{x}_i$ being the feature vector (fingerprint representation) and $y_i \in\{-1,+1\}$ the class label (positive or negative; active or inactive) of a training compound. By solving a convex quadratic optimization problem, SVM derives the normal vector $\mathbf{w}$ and the scalar b to define a hyperplane $H =\{\mathbf{x}|\langle\mathbf{w},\mathbf{x}\rangle+b = 0\}$ that best separates positive from negative training examples. The classification of an unknown test molecule $\mathbf{x}$ is based on the decision function $f(\mathbf{x}) = \mathrm{sgn}(\langle\mathbf{x},\mathbf{w}\rangle+b)$, i.e. compounds with $f(\mathbf{x}) = +1$ are assigned to the positive class and those with $f(\mathbf{x}) = -1$ to the negative class. To permit classification functions that do not linearly depend on the training data, scalar products $\langle\cdot,\cdot\rangle$ occurring in the objective function of the SVM optimization problem can be replaced by a kernel function $K(\cdot,\cdot)$.[4,5]

Although SVM was originally developed for binary classification, the approach can also be adapted for multiclass predictions[12] or database ranking. In order to transform the classification approach into a ranking function, test molecules are sorted in descending order of $g(\mathbf{x}) = K(\mathbf{x},\mathbf{w})$. This is equivalent to ranking the test compounds by their signed distance from the hyperplane, i.e. from the most distant compound located on the side of the positive training class to the most distant compound on the side of the negative training class. For simple SVM ranking or what we call "homology-based" SVM, ligands of the most closely related target among the reference targets were used as the positive training class and a randomly chosen subset of the screening database as the negative training class to learn the ranking function $g(\mathbf{x})$.

2.2. Target-Ligand Kernel. In recent chemogenomics-oriented studies,[7−9] not only compounds but also protein-compound pairs were used to train SVM in order to enable learning and classification for multiple targets. For this purpose, a target-ligand kernel (TLK) was defined to compare two different target-ligand pairs and calculate the scalar products for target-ligand pairs during SVM optimization and ranking. Given the target-ligand pairs $(t, \mathbf{x}_i)$ and $(t', \mathbf{x}_j)$, the target-ligand kernel is generally defined as the product of two separate kernels for the target pair and the ligand pair

$$K((t, \mathbf{x}_i)(t', \mathbf{x}_j)) = K_{target}(t, t') \times K_{ligand}(\mathbf{x}_i, \mathbf{x}_j)$$

The design principle of target-ligand kernels is illustrated in Figure 1. Independent kernels for protein and ligand representations are combined to account for pairwise target and ligand similarities.

For SVM training, each reference target was combined with its true ligands in the training set to generate positive training examples and randomly selected screening database molecules were combined with each reference target to build negative training examples. For ranking, test compounds $\mathbf{x}$ were combined with the orphanized target t_{orphan} for which no known ligands were available during training and the pairs $(t_{orphan}, \mathbf{x})$ were ranked according to the signed distance from the hyperplane H, as described in section 2.1.

2.3. Linear Combination. The SVM linear combination (LC) technique was recently introduced.[9] For each reference target t_i, an individual weight vector $\mathbf{w}_i$ was derived by learning an SVM classification function with the known ligands of t_i as positive training examples and a randomly chosen subset of the screening database as negative examples. To then obtain a ranking function for the orphan target t_{orphan}, the weight vector $\mathbf{w}_{orphan}$ was generated by linearly combining the individual weight vectors $\mathbf{w}_i$ of the reference targets.

Values of the target kernels described in section 3 were used as linear factors so that the linear combination was directly comparable to the SVM TLK strategy

$$\mathbf{w}_{orphan} = \sum K_{target}(t_{orphan}, t_i)\mathbf{w}_i$$

For database ranking, all test compounds were sorted according to the value of

$$g(\mathbf{x}) = K_{ligand}(\mathbf{x}, \mathbf{w}_{orphan})$$

3. KERNEL DESIGN

Three different ligand kernels were applied for the search strategies described above, using fingerprints as ligand representation.

(**1**) *Gaussian kernel* (radial basis function kernel)[13]

$$K_{\mathrm{Gaussian}}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma|\mathbf{x}_i - \mathbf{x}_j|^2)$$

(**2**) *Tanimoto kernel*[14]

$$K_{\mathrm{Tanimoto}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\langle\mathbf{x}_i, \mathbf{x}_j\rangle}{\langle\mathbf{x}_i, \mathbf{x}_i\rangle + \langle\mathbf{x}_j, \mathbf{x}_j\rangle - \langle\mathbf{x}_i, \mathbf{x}_j\rangle}$$

(**3**) *Linear kernel* that corresponds to the standard scalar product.

In addition to the ligand kernel component, SVM TLK and LC require the calculation of a target component. The 11 target kernels considered in this study differ significantly in their design, information content, and complexity.

(**a**) *Uniform kernel* between two targets (t,t')

$$K_{\mathrm{uniform}}(t,t') = 1$$

In this case, differences between targets are not considered. For the TLK search strategy, using K_{uniform} corresponds to pooling the training molecules for all proteins and deriving the SVM on the pooled compounds.

(**b**) *Needle kernel* is the protein sequence identity *SI* for a protein pair (t,t') computed using the Needleman-Wunsch

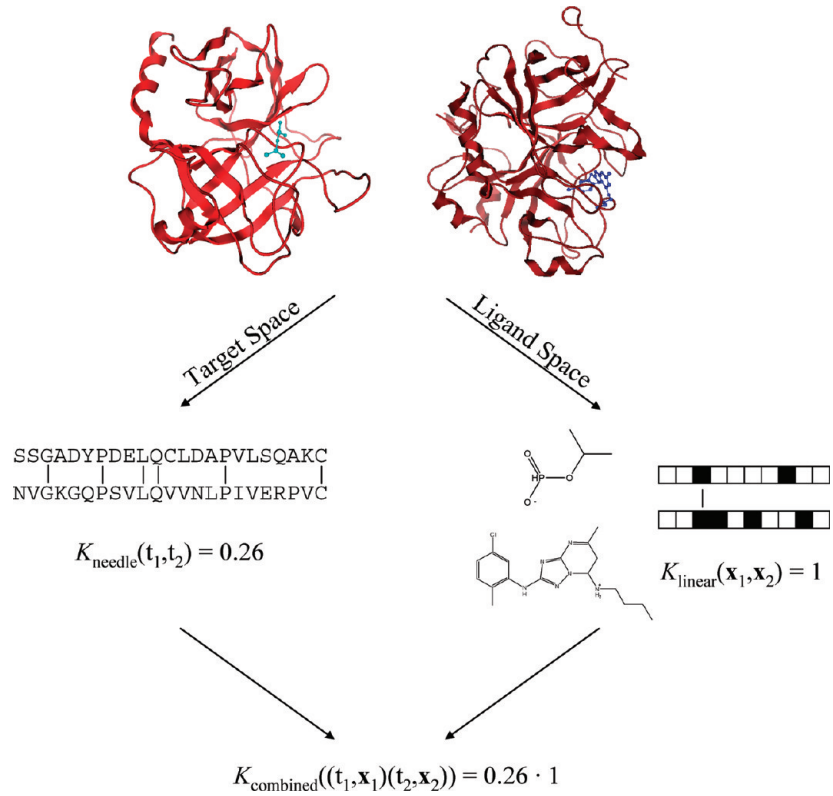

Figure 1. Target-ligand kernels. The comparison of two target-ligand pairs via a target-ligand kernel function is divided into two independent tasks. In this case, the similarity of protein targets is quantified by sequence comparison, while ligand similarity is assessed through comparison of fingerprint representations. The product of the two similarity scores is taken to recombine target and ligand information.

Table 1. Target and Ligand Data Set 1

target	abbreviation	MEROPS identifier	PDB entry	number of ligands	nearest neighbor target
angiotensin-converting enzyme 2	ace2	M02.006	1r42	28	mmp2
calpain 2	cal2	C02.002	1kfu	49	catL
caspase 3	cas3	C14.003	1cp3	264	catL
cathepsin D	catD	A01.009	1lyb	70	ren
cathepsin L	catL	C01.032	1mhw	78	cal2
glutamate carboxypeptidase 2	mgcp	M28.010	2oot	14	mmp8
methionine aminopeptidase 2	metap2	M24.002	1b6a	254	mgcp
matrix metalloprotease 2	mmp2	M10.003	1qib	83	mmp8
matrix metalloprotease 8	mmp8	M10.002	1bzs	16	mmp2
renin	ren	A01.007	2ren	164	catD
thrombin	thr	S01.217	1ppb	281	try
trypsin	try	S01.127	1trn	58	thr

For each target protein, its MEROPS identifier, a corresponding Protein Data Bank (PDB) entry,[39] the number of ligands, and the nearest neighbor target are reported. The MEROPS identifier is composed of a family-based component (e.g. thrombin and trypsin both belong to the family S01) and an individual target-based component.

algorithm for pairwise global sequence alignment implemented in EMBOSS[15]

$$K_{needle}(t,t') = SI(t,t')$$

(**c**) *Water kernel*. Each protein pair (t,t') is also subjected to pairwise local sequence alignment using the Smith-Waterman algorithm implemented in EMBOSS, and the alignment scores $S_{SW}(t,t')$ are expressed in logarithmic form

$$K_{water}(t,t') = \ln S_{SW}(t,t')$$

(**d**) *PROFEAT kernel*. The PROFEAT server[16] computes 1447 protein descriptors from protein sequence including descriptors developed by Dubchak et al.[17] that account for the composition, transition, and distribution of structural and physicochemical properties such as hydrophobicity, polarity, charge, and solvent accessibility. Each descriptor is separately normalized to the value range [0,1], and each target t is represented by a vector $\mathbf{\Phi}_P(t)$ of 1447 normalized descriptor values. The PROFEAT kernel is then defined as

$$K_{PROFEAT}(t,t') = \langle \mathbf{\Phi}_P(t), \mathbf{\Phi}_P(t') \rangle$$

(**e**) *Spectrum kernel* is a string kernel introduced by Leslie et al.[18] It compares sequence strings representing k-mers. Here conventional 3-mers were computed for target sequences. Each protein t is represented by a 20^3 dimensional vector $\mathbf{\Phi}_S(t)$ (for 20 amino acids) where each dimension corresponds to a possible string of three amino acids and reports the count of the number of occurrences of this fragment in the sequence of t. To account for

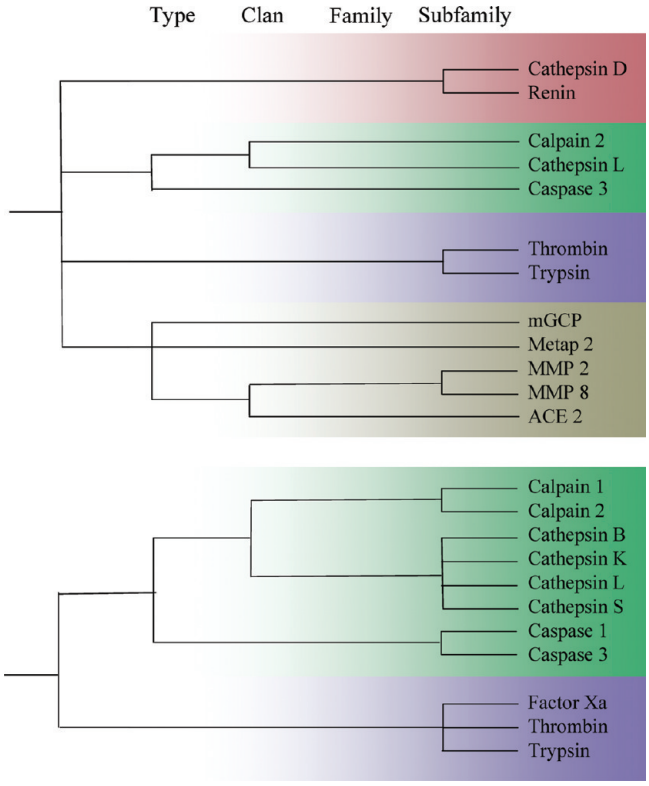

Figure 2. Target relationships. The relationships between the proteases in the target sets 1 (top) and 2 (bottom) are illustrated. The MEROPS classification scheme[25] (i.e., type, clan, family, and subfamily) is applied. From the "subfamily" to the "type" level, target similarity is fading away.

different lengths of protein sequences, the kernel is normalized as follows

$$K_{\text{spectrum}}(t,t') = \frac{\langle \mathbf{\Phi}_S(t), \mathbf{\Phi}_S(t') \rangle}{\sqrt{\langle \mathbf{\Phi}_S(t), \mathbf{\Phi}_S(t) \rangle \langle \mathbf{\Phi}_S(t'), \mathbf{\Phi}_S(t') \rangle}}$$

(**f**) *SSEA kernel*. For each target, the secondary structure is predicted by PSIPRED[19] resulting in a string of residues each represented by one of three letters for the states helix, strand, or coil. Strings for a target pair (t,t') are then globally aligned using the dynamic programming algorithm implemented in the SSEA Web server,[20] which yields a score $S_{SS}(t,t')$ in the range [0,100]. This score is directly used as target kernel

$$K_{\text{SSEA}}(t,t') = S_{SS}(t,t')$$

(**g**) The *GO kernel*. Gene Ontology (GO)[21] terms of the *Molecular Function* category are extracted for all protein targets from the UniProt Knowledgebase.[22,23] The GO kernel for a target pair (t,t') counts the number of identical GO terms in the GO term sets of t and t'.[24]

(**h**) *Cleavage kernel*. Peptidases act on specific substrates and their catalytic activity is often restricted to specific sequence recognition sites. For all targets, available cleavage sites of their substrates are extracted from the MEROPS[25] and CutDB[26] databases that collect cleavage sites in natural and synthetic substrates. Cleavage site patterns are reduced to two residues on either side of the scissile bond, and for each target, a position-specific frequency matrix is generated. The columns of the matrix are then concatenated to form a 4×20 dimensional feature vector $\mathbf{\Phi}_C(t)$, and the cleavage kernel is calculated as follows

$$K_{\text{cleavage}}(t,t') = \frac{\langle \mathbf{\Phi}_C(t), \mathbf{\Phi}_C(t') \rangle}{\sqrt{\langle \mathbf{\Phi}_C(t), \mathbf{\Phi}_C(t) \rangle \langle \mathbf{\Phi}_C(t'), \mathbf{\Phi}_C(t') \rangle}}$$

(**i**) *SCOP kernel*. The SCOP database[27] is hierarchically structured into protein folds, superfamilies, families, and domains and can be represented as a directed acyclic graph (DAG). Each target t is represented by an n-dimensional feature vector $\mathbf{\Phi}_{Sc}(t)$ where n is the number of nodes in the graph and each feature is assigned a value of 1 if the corresponding node occurs in t's SCOP hierarchy and 0 otherwise. The SCOP kernel is then defined as follows

$$K_{\text{SCOP}}(t,t') = 2^{\langle \mathbf{\Phi}_{Sc}(t), \mathbf{\Phi}_{Sc}(t') \rangle}$$

(**j**) *Topmatch kernel*. All protein targets are represented by a 3D substructure comprising all amino acids within an 8 Å radius of the target's catalytic residues. Residues falling within this radius are computed with MOE[28] and then subjected to structure comparison using TopMatch-web.[29,30] For a target pair (t,t'), TopMatch-web computes a relative similarity score S_T within the range [0,100] that is directly used as the target kernel

$$K_{\text{Topmatch}}(t,t') = S_T(t,t')$$

(**k**) *MEROPS kernel*. The MEROPS database[25] is hierarchically structured into catalytic types, so-called protein clans, families, and subfamilies and can also be visualized as a

Table 2. Target and Ligand Data Set 2

target	abbreviation	MEROPS identifier	PDB entry	number of ligands	nearest neighbor target
calpain 1	cal1	C02.001	1tlo	46	cal2
calpain 2	cal2	C02.002	1kfu	49	cal1
caspase 1	cas1	C14.001	1ice	21	cas3
caspase 3	cas3	C14.003	1gfw	264	cas1
cathepsin B	catB	C01.060	1gmy	17	catS
cathepsin K	catK	C01.036	1yk7	223	catS
cathepsin L	catL	C01.032	1mhw	78	catK
cathepsin S	catS	C01.034	1ms6	221	catK
factor Xa	faXa	S01.216	1mq5	783	thr
thrombin	thr	S01.217	1ppb	281	faXa
trypsin	try	S01.127	1trn	58	faXa

For each target protein, its MEROPS identifier, a corresponding Protein Data Bank (PDB) entry,[39] the number of ligands, and the nearest neighbor target are reported.

Table 3. Methods and Calculations

	search strategy		
	simple SVM	target-ligand kernel	linear combination
screening database	subset of ZINC7, 100,000 compounds		
reference targets	1 (nearest neighbor target)	all except the orphan target (i.e., 11 targets for data set 1 and 10 for set 2)	
inactive training class	1000 ZINC7 compounds		
active training class	5 inhibitors for the reference target	5 inhibitors per reference targets (i.e., a total of 55 inhibitors for data set 1 and 50 for set 2)	
ligand kernel	Gaussian, linear, Tanimoto	linear	
target kernel	-	11 different kernels (uniform, needle, water, PROFEAT, spectrum, SSEA, GO, cleavage, SCOP, Topmatch, MEROPS)	
fingerprints	MACCS, TGD		
trials	10 with different randomly selected compound reference sets		

DAG. Hence, $\mathbf{\Phi}_M(t)$ can be defined analogously to $\mathbf{\Phi}_{Sc}(t)$, and the MEROPS kernel is given by

$$K_{\text{MEROPS}}(t,t') = 2^{\langle \mathbf{\Phi}_M(t), \mathbf{\Phi}_M(t') \rangle}$$

Some of the above-mentioned functions are not generally valid kernel functions because they are not necessarily positive semidefinite for all protein sets (i.e., for a given set of proteins, an all-versus-all matrix of scores might have some negative eigenvalues). It has been shown that a symmetric matrix can be converted into a positive semidefinite matrix by subtracting its smallest negative eigenvalue from the diagonal.[31] However, for the analysis presented herein, this conversion was not required because all score matrices were positive semidefinite for our data sets.

4. TARGET AND LIGAND SYSTEMS

Two sets of reference targets were assembled that represented different degrees of intertarget relationships. The first target set included 12 proteases belonging to nine different families (Table 1) and showing four different catalytic mechanisms: cathepsin D and renin are aspartate proteases; thrombin and trypsin serine proteases; cathepsin L, calpain 2, and caspase 3 cysteine proteases; and matrix metalloproteases 2 and 8, methionine aminopeptidase 2, glutamate carboxypeptidase 2, and angiotensin-converting enzyme 2 are metalloproteases. Proteases possessing the same catalytic machinery can either be closely or distantly related in sequence, as illustrated in Figure 2, which organizes targets into clans, families, and subfamilies following the classification scheme of the MEROPS peptidase database. Based on the MEROPS hierarchy, the nearest neighbor target for each protease in our test set was determined. If several nearest neighbor candidates were suggested for a given target based on the MEROPS hierarchy, the protease with highest sequence identity to the target was chosen.

For all 12 targets, ligand sets were assembled from the MDL Drug Data Report (MDDR),[32] the BindingDB database,[33,34] and original literature sources. In total, 1359 different protease inhibitors were collected. As reported in Table 1, each ligand set contained between 14 and 281 compounds that had a potency of at least 1 μM (K_i or IC_{50}) against the target. Ligand sets were mutually exclusive in their composition, i.e. a compound reported to inhibit multiple protease targets was only assigned to the target it was most potent against.

The second target set included 11 proteases and was described previously.[9] These targets included the cysteine proteases calpain 1 and 2, caspase 1 and 3, and cathepsins B, L, K, and S, and the serine proteases factor Xa, thrombin, and trypsin, as summarized in Table 2. The relationship between these targets is also illustrated in Figure 2. Ligand sets for these targets were assembled as described above. For proteases shared among both target sets (i.e. calpain 2, caspase 3, cathepsin L, thrombin, and trypsin), the same ligand sets were used. As illustrated in Figure 2, the intertarget and nearest neighbor relationships differed between target sets 1 and 2. Whereas each target in set 2 had

Table 4. Search Results for Ligand Prediction Using Simple SVM (Set 1)[b]

	kernel					
	linear		Tanimoto		Gaussian	
	100[a]	1000[a]	100[a]	1000[a]	100[a]	1000[a]
	MACCS					
ace2	0.0	0.9	0.0	0.4	0.0	0.9
cal2	19.1	60.2	21.4	64.8	22.1	63.4
cas3	2.6	9.5	1.9	9.7	2.4	10.0
catD	15.4	51.5	17.9	56.5	16.0	54.5
catL	10.4	34.7	9.9	35.6	10.0	35.1
mgcp	0.0	0.0	0.0	1.1	0.0	0.0
metap2	0.0	0.5	0.0	0.0	0.0	0.2
mmp2	49.6	77.1	55.3	79.2	54.7	78.7
mmp8	49.1	59.1	49.1	60.0	50.9	59.1
ren	30.3	57.9	27.3	56.5	32.0	58.2
thr	27.7	70.3	26.5	69.8	28.0	71.3
try	36.9	61.4	37.3	61.5	38.5	61.2
average	20.1	40.3	20.5	41.3	21.2	41.0
	TGD					
ace2	0.9	14.8	0.4	14.4	0.4	17.4
cal2	4.8	30.7	3.6	28.4	5.0	30.2
cas3	0.6	8.5	0.1	3.9	0.2	5.4
catD	40.5	72.2	33.2	71.2	38.6	77.1
catL	8.0	23.6	7.3	27.0	6.2	25.2
mgcp	0.0	0.0	0.0	3.3	0.0	0.0
metap2	0.0	0.0	0.0	0.5	0.0	0.0
mmp2	75.6	94.0	78.0	94.7	77.6	95.3
mmp8	57.3	67.3	57.3	62.7	57.3	68.2
ren	44.3	59.1	41.6	60.0	41.3	59.9
thr	28.9	80.5	29.0	80.4	29.7	81.1
try	46.9	74.4	45.2	81.5	46.2	78.7
average	25.6	43.8	24.6	44.0	25.2	44.9

[a] Set size. [b] Recovery rates (in %) are reported for all targets in data set 1 averaged over 10 independent trials per target. The results reported for the Gaussian kernel were obtained with the parameter γ set to 0.01.

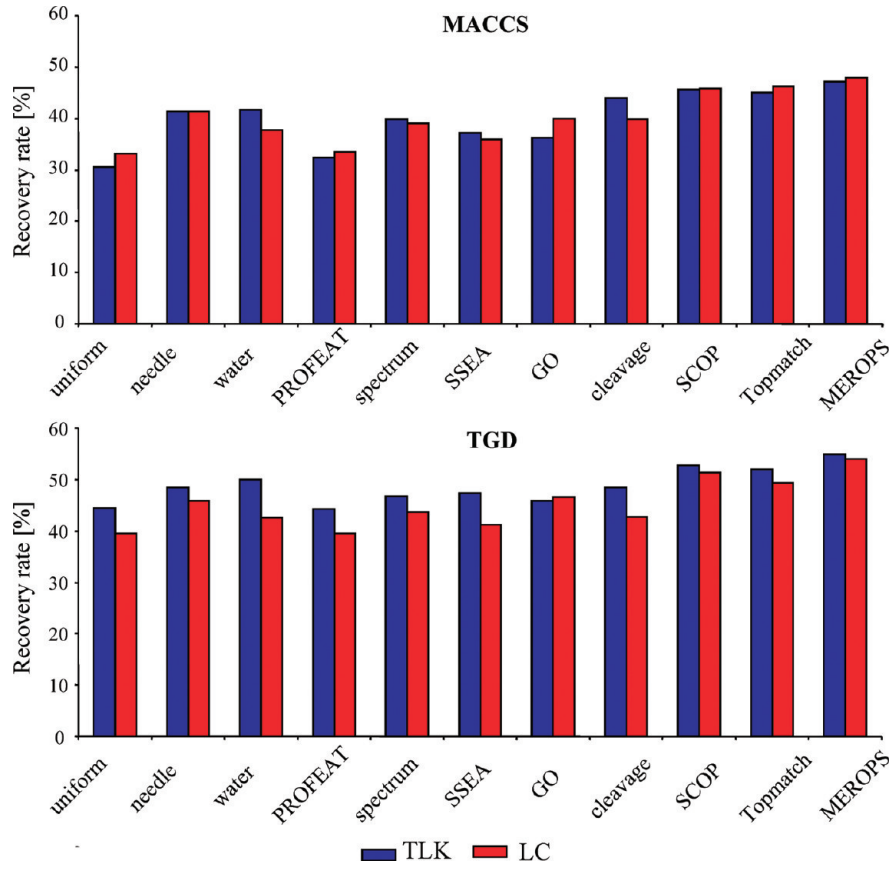

Figure 3. Ligand prediction for target set 1. For the MACCS and TGD fingerprints, recovery rates for SVM TLK and LC strategies are shown for 11 alternative target kernels and selection sets of 1000 database compounds. Recovery rates are averaged over all 12 targets and 10 independent search trials per target.

a nearest neighbor that belonged to the same subfamily, several targets in set 1 had nearest neighbors sharing the same catalytic mechanism but lacking further evidence of evolutionary relationships. The different intertarget relationships found in data sets 1 and 2 were explored in SVM modeling and ligand-target prediction.

5. SEARCH CALCULATIONS

The performance of alternative kernel functions and SVM ranking strategies was evaluated in systematic search calculations on our two protease systems using MACCS structural keys[35] and the TGD fingerprint[28] as ligand descriptors. TGD represents a two-point pharmacophore-type fingerprint that is calculated from the 2D connectivity table of a molecule. As a background database for SVM analysis, 100,000 compounds were randomly chosen from ZINC7.[36]

As negative training examples, 1000 database compounds were randomly selected in each case. For simple SVM calculations on each target, only five inhibitors of the nearest neighbor target were used as positive training molecules. For SVM LC and TLK, five inhibitors for each target were used. The inhibitor set of the orphanized target was not used during SVM learning but added to the background database as potential database hits during testing.

Kernel functions, SVM strategies, and fingerprint descriptors were systematically combined. For each investigated combination of a kernel function, SVM strategy, and fingerprint, 10 different randomly selected compound training and test sets were analyzed and the search results were averaged. Table 3 summarizes the different methods and calculation settings.

As a measure of performance, recovery rates (RR: number of correctly identified orphan target inhibitors divided by their total number) were calculated for database selection sets of increasing size and averaged over the 10 independent trials per target.

All calculations were carried out using SVM^{light},[37] a freely available SVM implementation.[38] All calculation parameters were SVM^{light} default settings to ensure reproducibility of the calculations. Perl scripts were applied to calculate SVM linear combinations.

6. GLOBAL KERNEL PERFORMANCE

We first investigated the relative performance of ligand kernels in simple SVM calculations searching for active compounds on the basis of randomly selected reference sets. Table 4 reports the compound recovery rates for activity classes of target set 1, the MACCS and TGD fingerprints, and database selection sets of 100 and 1000 compounds. In these calculations, all three ligand kernels produced comparable recovery rates. In some instances, the search calculations failed for any kernel, and, in others, high recovery rates were consistently observed. Because there was no apparent preference for a ligand kernel in our test calculations, we selected the linear kernel, which has the lowest computational complexity of the three, for further calculations and combined this kernel with the 11 different target kernels.

Table 5. Search Results for Ligand Prediction Using SVM LC (Set 1)[a]

	uniform	needle	water	PROFEAT	spectrum	SSEA	GO	cleavage	SCOP	Topmatch	MEROPS
						MACCS					
ace2	20.9	18.7	20.0	20.4	18.7	20.0	17.8	10.0	4.8	17.8	10.9
cal2	65.2	63.9	65.2	63.6	65.9	64.6	64.3	66.8	74.6	63.4	79.1
cas3	35.4	29.9	35.4	35.1	34.9	34.9	35.3	36.3	34.0	35.5	35.2
catD	76.0	78.0	86.3	77.1	92.5	80.0	89.5	90.2	73.9	85.2	74.2
catL	32.6	31.4	32.9	32.6	31.6	32.9	36.6	38.6	41.8	37.1	41.1
mgcp	0.0	0.0	0.0	0.0	1.1	0.0	2.2	0.0	0.0	1.1	1.1
metap2	9.6	7.6	9.5	9.3	8.8	9.8	9.9	8.5	10.1	9.0	9.8
mmp2	17.7	41.7	28.6	18.5	39.5	22.6	34.7	33.5	61.7	68.6	74.4
mmp8	19.1	50.0	30.9	18.2	32.7	24.6	36.4	33.6	55.5	55.5	57.3
ren	35.3	55.2	46.4	37.0	54.2	45.2	49.9	55.0	57.8	55.7	57.6
thr	49.5	66.6	57.2	51.2	50.7	55.5	62.3	58.8	74.0	67.5	73.7
try	37.1	54.4	41.4	38.1	37.3	42.1	40.2	46.2	61.0	57.5	61.4
average	33.2	41.4	37.8	33.4	39.0	36.0	39.9	39.8	45.8	46.2	48.0
						TGD					
ace2	30.0	32.6	31.7	30.9	30.9	31.7	46.5	25.7	33.0	27.4	37.4
cal2	33.2	24.3	32.5	29.3	27.3	31.4	42.1	38.6	36.1	32.3	44.1
cas3	45.8	45.1	46.2	45.8	46.2	45.4	47.8	49.5	45.1	48.8	49.0
catD	78.5	83.2	83.9	78.3	86.9	82.9	84.2	87.2	83.9	83.5	83.5
catL	14.8	13.7	14.8	14.8	15.1	15.1	17.0	18.6	19.0	15.2	18.5
mgcp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
metap2	33.4	29.4	33.8	33.6	34.0	33.5	32.8	18.3	35.9	27.3	39.6
mmp2	24.5	52.7	34.4	26.0	43.2	26.8	39.5	30.4	78.3	71.4	89.9
mmp8	33.6	66.4	43.6	33.6	56.4	36.4	49.1	46.4	70.9	73.6	72.7
ren	59.3	59.6	60.0	59.4	59.9	60.2	60.0	59.3	59.4	59.8	59.4
thr	60.3	74.7	64.8	61.1	61.8	66.0	76.5	76.3	80.9	79.2	80.7
try	60.8	67.7	65.0	61.2	62.7	65.0	63.3	63.7	72.9	73.1	73.7
average	39.5	45.8	42.6	39.5	43.7	41.2	46.6	42.8	51.3	49.3	54.0

[a] Recovery rates (in %) for all targets in data set 1 are reported for selection sets of 1.000 compounds averaged over 10 independent trials per target.

Table 6. Search Results for Ligand Prediction Using SVM TLK (Set 1)[a]

	uniform	needle	water	PROFEAT	spectrum	SSEA	GO	cleavage	SCOP	Topmatch	MEROPS
						MACCS					
ace2	20.9	20.4	24.4	23.5	20.4	23.0	0.4	3.9	4.4	17.8	11.3
cal2	46.4	55.5	59.3	40.9	54.3	43.6	61.8	79.6	74.1	62.5	78.0
cas3	27.9	28.2	32.8	30.4	34.9	31.4	25.9	34.9	36.0	35.9	36.1
catD	58.5	70.3	75.2	66.8	90.2	71.5	81.1	87.5	68.5	67.4	68.6
catL	28.8	30.8	33.8	29.7	30.8	27.3	31.0	39.7	41.6	35.3	42.3
mgcp	2.2	0.0	1.1	2.2	1.1	0.0	3.3	1.1	0.0	1.1	0.0
metap2	8.6	7.1	8.1	8.4	8.5	8.4	4.6	5.5	10.0	7.9	9.0
mmp2	22.7	49.2	46.7	23.6	48.3	34.4	40.6	50.4	65.3	70.1	73.1
mmp8	34.6	55.5	58.2	36.4	44.6	47.3	47.3	56.4	56.4	59.1	57.3
ren	21.2	54.3	42.1	27.6	54.0	39.3	58.6	49.1	57.5	55.7	57.4
thr	55.8	68.7	67.0	58.1	52.8	66.2	36.4	65.9	72.9	69.6	72.8
try	39.0	57.1	51.7	41.5	37.3	53.3	43.9	52.9	61.4	59.2	60.6
average	30.5	41.4	41.7	32.4	39.8	37.1	36.2	43.9	45.7	45.1	47.2
						TGD					
ace2	39.6	37.0	38.3	38.7	34.8	44.8	11.3	19.6	32.6	27.8	43.9
cal2	30.9	21.4	32.7	25.0	24.8	24.1	36.1	38.2	40.9	36.8	44.8
cas3	45.4	45.1	46.3	45.9	45.4	45.2	42.6	48.8	43.9	48.8	48.6
catD	80.0	78.6	85.7	79.2	83.2	83.9	83.2	82.9	80.6	80.8	77.9
catL	12.9	12.1	13.8	13.2	14.3	12.9	16.6	19.7	18.6	14.3	19.2
mgcp	1.1	0.0	0.0	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0
metap2	42.3	35.4	41.2	41.8	37.4	42.0	34.5	20.9	38.6	33.5	42.1
mmp2	34.0	71.9	56.9	35.0	56.0	43.9	53.2	60.4	87.7	88.6	91.8
mmp8	46.4	71.8	70.0	48.2	70.0	51.8	68.2	75.5	70.9	69.1	70.9
ren	57.8	59.5	60.6	58.6	60.1	60.8	60.2	59.7	59.8	59.7	59.6
thr	65.8	75.4	72.5	66.8	62.4	75.2	67.5	80.3	80.8	80.7	80.8
try	76.7	74.0	81.7	78.9	72.9	83.7	74.8	76.0	78.7	84.4	79.2
average	44.4	48.5	50.0	44.3	46.8	47.4	45.8	48.5	52.8	52.0	54.9

[a] Recovery rates (in %) for all targets in data set 1 are reported for selection sets of 1000 compounds averaged over 10 independent trials per target.

Next we tested the 11 target kernels in SVM TLK and LC ligand prediction calculations. Average results for target set 1 and selection sets of 1000 database compounds are shown in Figure 3, and Tables 5 and 6 report recovery rates on a per target basis. Supplementary Figure S1 and Tables S1 and S2 report corresponding results for database selection sets of 100 compounds. The SVM TLK and LC search strategies were found to produce similar average compound

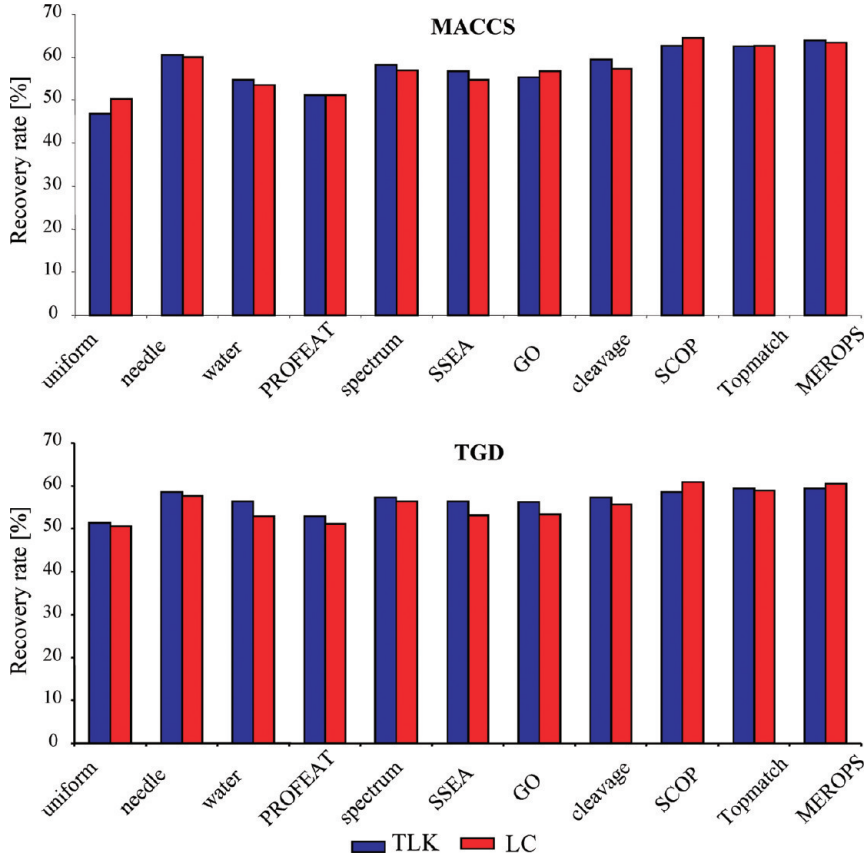

Figure 4. Ligand prediction for target set 2. Search results for target set 2 are shown corresponding to Figure 3.

Table 7. Search Results for Ligand Prediction Using SVM LC (Set 2)[a]

	uniform	needle	water	PROFEAT	spectrum	SSEA	GO	cleavage	SCOP	Topmatch	MEROPS
					MACCS						
cal1	76.8	67.6	77.3	77.3	67.8	77.1	76.1	76.8	73.2	75.6	69.0
cal2	90.5	92.5	95.2	92.3	87.3	94.1	98.6	98.9	99.6	98.0	95.7
cas1	35.0	30.0	36.3	35.0	36.9	36.3	39.4	40.0	50.0	40.0	50.0
cas3	36.7	46.3	40.3	37.0	43.9	40.6	32.1	45.0	47.0	45.8	47.0
catB	79.2	83.3	80.0	79.2	81.7	80.8	86.7	76.7	82.5	85.8	77.5
catK	45.2	55.5	46.3	46.3	54.6	48.5	49.1	48.1	53.6	53.8	52.1
catL	57.0	74.1	59.7	58.6	71.6	62.1	63.2	67.8	73.7	74.0	73.3
catS	44.4	62.1	43.2	46.1	60.6	50.3	50.9	48.0	57.4	55.8	60.7
faXa	10.7	27.8	15.3	11.3	24.3	15.9	16.1	22.8	38.0	33.8	38.0
thr	37.8	63.9	49.5	39.5	51.3	49.8	67.6	54.2	72.4	68.5	72.4
try	39.8	57.0	45.1	40.8	46.0	45.9	44.5	51.1	61.3	57.2	61.3
average	50.3	60.0	53.5	51.2	56.9	54.7	56.8	57.2	64.4	62.6	63.4
					TGD						
cal1	44.6	45.4	44.9	44.9	44.6	45.4	46.8	46.1	48.3	47.1	46.3
cal2	81.1	88.6	84.1	83.0	88.2	83.4	83.0	84.1	89.1	87.1	89.3
cas1	56.3	70.0	58.1	56.3	63.8	58.1	53.8	72.5	86.9	76.9	86.9
cas3	50.6	52.9	51.7	50.8	52.7	51.8	50.6	53.3	54.1	53.6	54.1
catB	65.8	66.7	65.8	65.8	67.5	67.5	70.8	64.2	65.0	66.7	65.0
catK	49.0	48.4	50.9	49.6	47.4	50.1	46.3	49.6	49.2	50.4	45.8
catL	22.6	27.0	24.0	22.5	29.2	23.6	23.4	25.5	28.9	28.5	27.7
catS	16.6	26.9	16.9	17.2	28.2	18.2	19.7	16.5	25.1	23.2	28.0
faXa	35.9	52.7	43.3	37.2	51.6	44.0	46.9	51.0	56.0	55.8	56.0
thr	76.5	80.4	80.1	77.1	80.3	80.0	81.4	80.1	82.0	80.8	82.0
try	58.3	75.3	63.0	59.4	66.2	63.4	63.6	69.3	84.5	76.8	84.5
average	50.7	57.7	53.0	51.2	56.3	53.2	53.3	55.6	60.8	58.8	60.5

[a] Recovery rates (in %) for all targets in data set 2 are reported for selection sets of 1000 compounds averaged over 10 independent trials per target.

recovery rates of approximately 30%−55% for both fingerprints and all target kernels for a database selection set size of 1000 compounds. Moreover, there was relatively little variation in kernel performance, much less so than anticipated given the significant differences in target kernel complexity and encoded protein information. Equivalent trends were observed for database selection sets of 100 compounds. The secondary structure-based SSEA kernel and the PROFEAT

Table 8. Search Results for Ligand Prediction Using SVM TLK (Set 2)[a]

	uniform	needle	water	PROFEAT	spectrum	SSEA	GO	cleavage	SCOP	Topmatch	MEROPS
						MACCS					
cal1	74.6	70.5	77.3	77.1	72.2	77.1	76.3	76.3	73.7	74.9	72.4
cal2	71.6	84.8	84.8	81.1	81.6	84.8	88.9	92.1	89.1	90.5	88.2
cas1	29.4	32.5	34.4	31.3	39.4	35.6	38.1	51.9	46.9	44.4	54.4
cas3	32.3	46.0	42.0	36.0	45.4	42.2	32.2	46.4	47.5	47.0	47.3
catB	78.3	80.0	80.8	82.5	83.3	83.3	77.5	77.5	75.8	84.2	79.2
catK	40.5	56.5	46.0	45.2	55.7	48.6	45.2	46.7	52.8	53.5	52.0
catL	53.3	70.7	58.2	57.3	70.1	61.2	59.0	64.7	68.4	71.0	71.0
catS	37.6	62.3	37.2	42.9	61.4	47.9	44.6	44.4	58.3	55.2	60.9
faXa	16.1	33.8	28.6	19.3	29.6	29.9	31.0	34.6	39.1	37.1	39.1
thr	42.4	67.8	60.3	45.6	54.2	60.2	64.6	62.1	72.4	70.1	72.2
try	40.2	60.9	52.1	43.6	47.4	53.6	50.2	58.1	66.0	58.9	66.0
average	46.9	60.5	54.7	51.1	58.2	56.8	55.2	59.5	62.7	62.4	63.9
						TGD					
cal1	42.9	44.9	46.6	45.1	45.4	47.1	47.8	48.5	49.0	49.0	46.6
cal2	61.1	83.9	76.8	68.4	84.8	74.3	74.1	75.0	80.2	78.9	84.3
cas1	70.6	75.6	79.4	72.5	76.3	80.0	75.0	83.8	80.6	88.1	85.0
cas3	51.5	53.0	52.3	51.6	53.0	52.8	50.7	53.9	54.2	54.1	54.1
catB	60.0	64.2	59.2	60.8	60.8	61.7	62.5	55.8	56.7	60.0	58.3
catK	40.5	40.4	42.2	40.8	43.2	41.3	40.9	43.1	40.9	42.6	41.7
catL	20.7	28.8	23.6	20.7	30.4	22.6	23.4	28.2	28.2	29.7	27.8
catS	14.9	30.3	16.7	17.2	29.6	20.6	23.8	18.0	30.8	28.9	31.4
faXa	44.1	53.8	50.9	45.6	53.4	51.1	53.3	52.9	53.2	54.6	54.0
thr	77.8	81.5	81.1	78.8	80.1	81.2	80.8	80.9	82.1	81.3	82.0
try	81.1	87.6	90.2	80.0	74.0	87.4	84.9	88.5	88.7	87.2	88.1
average	51.4	58.5	56.3	52.9	57.3	56.4	56.1	57.2	58.6	59.5	59.4

[a] Recovery rates (in %) for all targets in data set 2 are reported for selection sets of 1.000 compounds averaged over 10 independent trials per target.

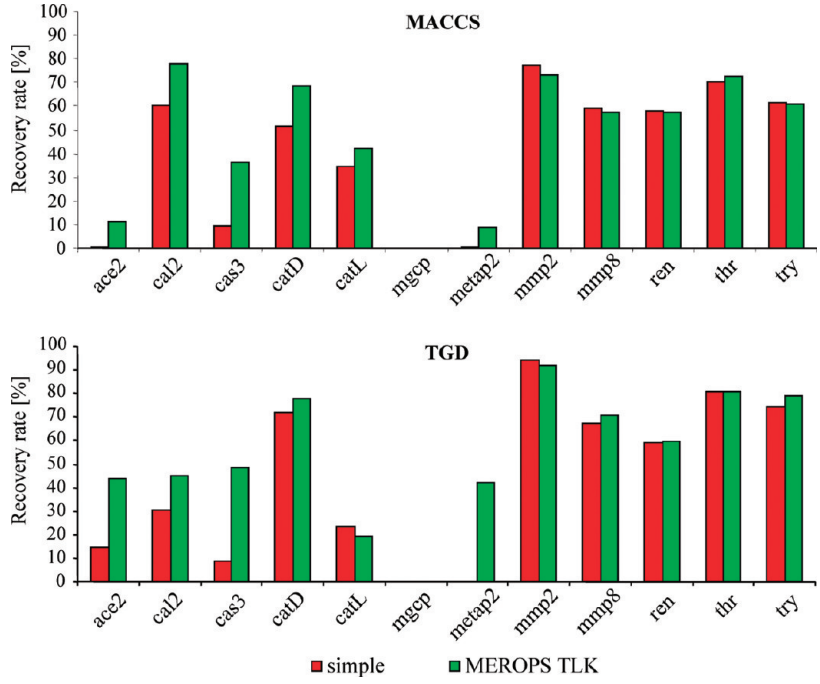

Figure 5. Target-dependent search performance (set 1). For all targets of set 1, recovery rates are shown for simple SVM ranking and for the SVM TLK search strategy in combination with the MEROPS kernel. Recovery rates are compared for selection sets of 1000 compounds averaged over 10 independent trials per target.

kernel, which is based on biophysical descriptors calculated from protein sequence, did not produce higher recovery rates than the uniform kernel that does not take protein similarity into account and hence served as a reference for target kernels. Differences in kernel performance were rather subtle, but the overall highest recovery rates were achieved with the MEROPS kernel that encodes a hierarchical protein organization scheme.

Equivalent observations were made for target set 2. Figure 4 (and Supplementary Figure S2) reports average results of the search calculations on set 2 (corresponding to those reported in Figure 3 and Supplementary Figure S1), and Tables 7 and 8 report recovery rates on a per target basis (Supplementary Tables S3 and S4 report corresponding results for database selection sets of 100 compounds). In this case, the recovery rates were generally higher than for set

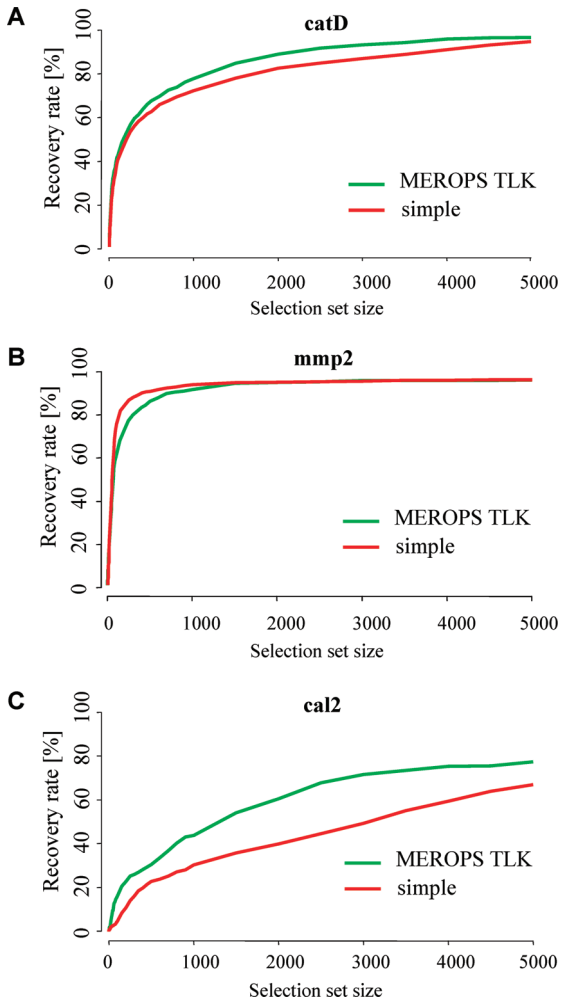

Figure 6. Cumulative recall curves. Representative recall curves for simple SVM and SVM TLK (MEROPS) calculations are shown for three targets, (**A**) catD, (**B**) mmp2, and (**C**) cal2, using TGD as the ligand descriptor. Recovery rates are averaged over 10 independent trials per target.

1, ranging on average from approximately 47%−64% for selection sets of 1000 compounds, but differences between SVM search strategies and alternative kernels were even smaller than those observed for target set 1. The uniform kernel produced an average recovery rate of close to 50%, and several kernels taking protein similarity at different levels into account performed only slightly better. Here, the hierarchical SCOP and MEROPS kernels and the Topmatch kernel that is based on active site structural similarity performed equally well but only slightly better than the sequence similarity-based needle kernel. Thus, taken together, the results of systematic SVM calculations on our two target sets revealed surprisingly little differences in search performance for target kernels of different design.

7. TARGET-DEPENDENT KERNEL PERFORMANCE

As described in the Methods section, the overall best-performing MEROPS kernel differs from other target kernels in that it assigns high weights to closely related targets, due to its exponential formalism. In order to explore the contributions of the most closely related targets to ligand recovery, we analyzed the search performance for all individual set 1 targets in SVM TLK calculations using the MEROPS kernel and, in addition, simple SVM control calculations. In the latter case, the SVM was trained on the ligands of the target most closely related to the orphanized target. The results of these SVM TLK and simple SVM calculations are shown in Figure 5 (and Supplementary Figure S3). Significant differences in target-dependent search performance were observed (consistent with the results reported in Tables 4 and 6). The search performance was found to be highly dependent on the degree of relatedness between the orphan target and its nearest neighbor. For those targets having a closely related nearest neighbor at the subfamily level (i.e., catD, mmp2, mmp8, ren, thr, try; see Figure 2), highest recovery rates were observed.

For these targets, simple SVM calculations using the ligands of the nearest neighbor as positive training examples matched the performance of SVM TLK calculations using the MEROPS kernel. By contrast, simple SVM search calculations produced only low recovery rates, or failed, for targets that had no closely related neighbor (i.e., all cysteine proteases in set 1, mgcp, ace2, and metap2; see Figure 2). The cumulative recall curves shown in Figure 6A,B illustrate the close correspondence between simple SVM and SVM TLK calculations when a closely related nearest neighbor target (see Table 1) was available. However, Figure 6C shows that taking additional target and ligand information into account when no closely related neighbor was available further improved the search performance, a trend that was especially observed for larger database selection sets.

Different from target set 1, each target in set 2 had a nearest neighbor at the subfamily level (Figure 2). Accordingly, one would expect better target-dependent search performance for targets in set 2 than in set 1. The SVM TLK (MEROPS) search calculations shown in Figure 7 (and Supplementary Figure S4) confirm this expectation. The majority of targets in set 2 produced recovery rates of at least 40% (with the MACCS and/or TGD fingerprints for ligand representation). In this case, simple SVM control calculations were also carried out after pooling the ligands of all members of the orphan target's subfamily for training. As illustrated in Figure 7, the recovery rates observed in these SVM control calculations were almost indistinguishable from those of SVM TLK calculations. Furthermore, in Supplementary Table S5, recovery rates for simple SVM calculations on set 2 targets are reported for selection sets of 100 and 1000 compounds when either only ligands of the nearest neighbor target were used for training or, alternatively, ligands of all subfamily members were pooled. The results demonstrate that recovery rates for targets having several closely related subfamily members further improved when ligands from all related targets were taken into account compared to ligands of only the most closely related target.

8. NEAREST NEIGHBOR EFFECTS

The findings discussed above reflect a strong influence of ligand information of nearest neighbor targets on ligand prediction for orphanized targets. In order to evaluate the magnitude of nearest neighbor effects, SVM TLK calculations using the MEROPS kernel were also carried out after removal of the ligands of the nearest neighbor target from SVM learning. The search results for set 1 targets are

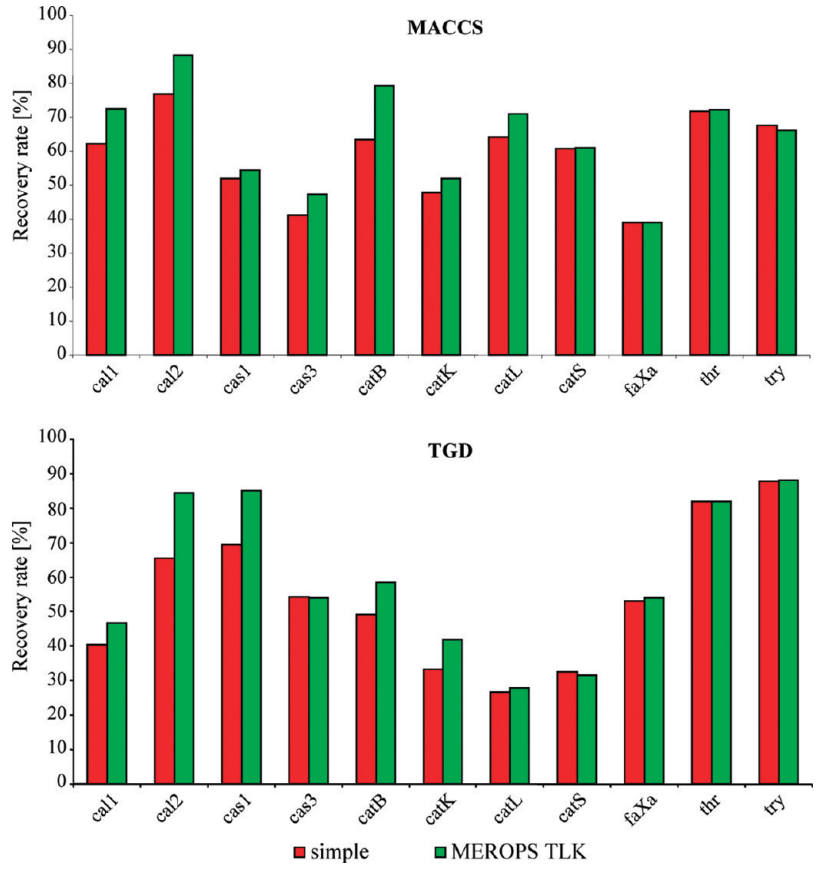

Figure 7. Target-dependent search performance (set 2). For all targets of set 2, recovery rates are shown for simple SVM ranking and for the SVM TLK search strategy in combination with the MEROPS kernel. For simple SVM ranking, ligands of all members of the orphanized target's subfamily were pooled and used as the positive training class. Recovery rates are shown for a selection set of 1000 compounds averaged over 10 independent trials per target.

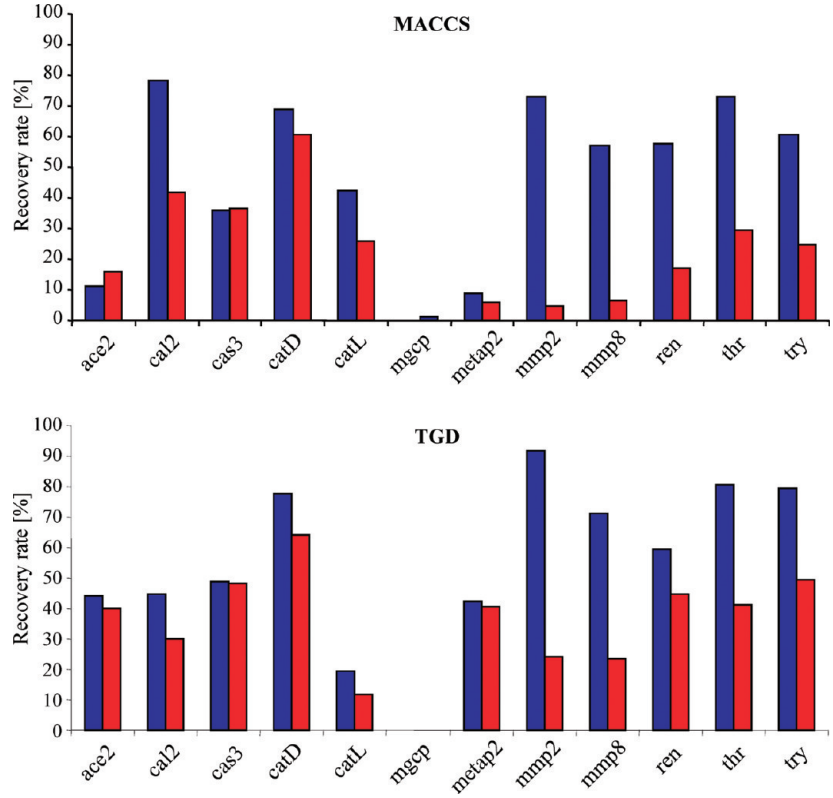

Figure 8. Dependence of search performance on ligands of nearest neighbor targets. For all set 1 targets, search results are reported for the SVM TLK (MEROPS) search strategy. Blue bars show recovery rates obtained by learning with all target set ligands, whereas red bars show recovery rates obtained when the ligands of the nearest neighbor are excluded from the training set. Recovery rates are shown for a selection set of 1000 compounds averaged over 10 independent trials per target.

shown in Figure 8 and Supplementary Figure S5, and Supplementary Table S6 reports the comparison of SVM TLK and LC calculations. As can be seen in Figure 8, removal of ligands led to a sharp decline in recovery rates when a nearest neighbor target was available at the subfamily level (effects observed in SVM TLK and LC calculations were similar). By contrast, removal of ligands for targets where no closely related neighbor was available had only little influence on the search performance. Thus, these findings further corroborated the crucial role of nearest neighbor ligand information for orphan target ligand prediction using SVM techniques.

9. CONCLUSIONS

In this study, we have investigated different strategies for SVM-based ligand prediction for simulated orphan targets with special emphasis on the evaluation of alternative target kernel functions that capture protein information at different levels. The approaches investigated here aimed at de novo ligand predictions. The way target information was taken into account presented a major variable in these calculations, and, accordingly, target kernels of different complexity and information content were designed and evaluated. Surprisingly, these alternative kernel functions influenced the calculations much less than one might anticipate. Rather, nearest neighbor effects were found to be the major determinant of ligand prediction performance. In particular, when ligand information from one or more closely related targets was available, simple SVM calculations utilizing this information met or exceeded the search performance of SVM TLK and LC calculations. For SVM-based ligand prediction on orphan targets, these findings have significant implications. Rather than focusing on information provided by reference systems capturing protein hierarchies, searching for targets with known ligands that are closely related to orphan targets (e.g., at the subfamily level) should be a primary objective. For this purpose, simple detection of sequence similarity might often be sufficient. In the presence of strong nearest neighbor relationships, SVM-based strategies for ligand prediction can be simplified. In these cases, simple SVM calculations using nearest neighbor ligands for learning, or corresponding SVM linear combinations, are expected to produce promising results. By contrast, if no closely related targets can be identified, SVM learning using target kernels capturing protein hierarchy information is likely to be a preferred approach. Thus, SVM strategies for ligand prediction can be adjusted based on an initial exploration of target relationships.

ACKNOWLEDGMENT

We wish to thank Dagmar Stumpfe and Jens Humrich for help with target and ligand sets and SVM^{light}, respectively.

Supporting Information Available: Supplementary Figures S1−S5 and Tables S1−S6 report the results of test calculations on individual targets using different SVM strategies. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225–233.
(2) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.
(3) Boser, B. E.; Guyon, I. M.; Vapnik, V. A training algorithm for optimal margin classifiers, In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, Pittsburgh, Pennsylvania, 1992; ACM, New York, 1992; pp 144−152.
(4) Müller, K.-R.; Rätsch, G.; Mika, S.; Tsuda, K.; Schölkopf, B. An introduction to kernel-based learning algorithms. *IEEE Neural Networks* **2001**, *12*, 181–201.
(5) Schölkopf, B.; Smola, A. *Learning with Kernels*; MIT Press: Cambridge, MA, 2002.
(6) Bock, J. R.; Gough, D. A. Virtual screens for ligands of orphan G protein-coupled receptors. *J. Chem. Inf. Model.* **2005**, *45*, 1402–1414.
(7) Erhan, D.; L'Heureux, P.-J.; Yue, S. Y.; Bengio, Y. Collaborative filtering on a family of biological targets. *J. Chem. Inf. Model.* **2006**, *46*, 626–635.
(8) Jacob, L.; Vert, J.-P. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **2008**, *24*, 2149–2156.
(9) Geppert, H.; Humrich, J.; Stumpfe, D.; Gärtner, T.; Bajorath, J. Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 767–779.
(10) Geppert, H.; Horváth, T.; Gärtner, T.; Wrobel, S.; Bajorath, J. Support-vector-machine-based ranking significantly improves the effectiveness of similarity searching using 2D fingerprints and multiple reference compounds. *J. Chem. Inf. Model.* **2008**, *48*, 742–746.
(11) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(12) Wassermann, A. M.; Geppert, H.; Bajorath, J. Searching for target-selective compounds using different combinations of multiclass support vector machine ranking methods, kernel functions, and fingerprint descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 582–592.
(13) Powell, M. J. D.; Mason, J. C.; Cox, M. G. Radial Basis Functions for Multivariable Interpolation: A review. In *Algorithm for Approximation*; Clarendon Press: 1987; pp 143−167.
(14) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093–1110.
(15) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277.
(16) Li, Z. R.; Lin, H. H.; Han, L. Y.; Jiang, L.; Chen, X.; Chen, Y. Z. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* **2006**, *34*, W32–W37.
(17) Dubchak, I.; Muchnik, I.; Holbrook, S. R.; Kim, S. H. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 8700–8704.
(18) Leslie, C.; Eskin, E.; Noble, W. S. The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomput.* **2002**, 564–575.
(19) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195–202.
(20) Fontana, P.; Bindewald, E.; Toppo, S.; Velasco, R.; Valle, G.; Tosatto, S. C. The SSEA server for protein secondary structure alignment. *Bioinformatics* **2005**, *21*, 393–395.
(21) Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat. Genet*. **2000**, *25*, 25−29.
(22) Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115–D119.
(23) Camon, E.; Magrane, M.; Barrell, D.; Binns, D.; Fleischmann, W.; Kersey, P.; Mulder, N.; Oinn, T.; Maslen, J.; Cox, A.; Apweiler, R. The Gene Ontology annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.* **2003**, *13*, 662–672.
(24) Lei, Z.; Dai, Y. Assessing protein similarity with Gene Ontology and its use in subcellular localization prediction. *Bioinformatics* **2006**, *7*, 491.
(25) Rawlings, N. D.; Morton, F. R.; Kok, C. Y.; Kong, J.; Barrett, A. J. MEROPS: the peptidase database. *Nucleic Acids Res.* **2008**, *36*, D320–D325.
(26) Igarashi, Y.; Eroshkin, A.; Gramatikova, S.; Gramatikoff, K.; Zhang, Y.; Smith, J. W.; Osterman, A. L.; Godzik, A. CutDB: a proteolytic event database. *Nucleic Acids Res.* **2007**, *35*, D546–549.

(27) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536−540.
(28) *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2007.
(29) Sippl, M. J.; Wiederstein, M. A note on difficult structure alignment problems. *Bioinformatics* **2008**, *24*, 426–427.
(30) Sippl, M. J. On distance and similarity in fold space. *Bioinformatics* **2008**, *24*, 872–873.
(31) Saigo, H.; Vert, J. P.; Ueda, N.; Akutsu, T. Protein homology detection using string alignment kernels. *Bioinformatics* **2004**, *20*, 1682−1689.
(32) *MDL Drug Data Report (MDDR)*; Symyx Software: San Ramon, CA, 2005.
(33) Liu, T.; Lin, Y.; Wen, X.; Jorrisen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.
(34) Chen, X.; Liu, M.; Gilson, M. K. Binding DB: a web-accessible molecular recognition database. *Comb. Chem. High Throughput Screening* **2001**, *4*, 719–725.
(35) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.
(36) Irwin, J. J.; Shoichet, B. K. ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(37) SVM^{light}. URL for the publicly available SVM tool. http://svmlight.joachims.org/ (accessed June 2009).
(38) Joachims, T. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*; Schölkopf, B., Burges, C., Smola, A., Eds.; MIT-Press: Cambridge, MA, 1999.
(39) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

CI9002624