Optimal Assignment Kernels For Attributed Molecular Graphs

Holger Fröhlich FROEHLIC@INFORMATIK.UNI-TUEBINGEN.DE
Jörg K. Wegner WEGNERJ@INFORMATIK.UNI-TUEBINGEN.DE
Florian Sieker FSIEKER@INFORMATIK.UNI-TUEBINGEN.DE
Andreas Zell ZELL@INFORMATIK.UNI-TUEBINGEN.DE
Centre For Bioinformatics Tübingen (ZBIT), Sand 1, 72076 Tübingen, Germany

Abstract

We propose a new kernel function for attributed molecular graphs, which is based on the idea of computing an optimal assignment from the atoms of one molecule to those of another one, including information on neighborhood, membership to a certain structural element and other characteristics for each atom. As a byproduct this leads to a new class of kernel functions. We demonstrate how the necessary computations can be carried out efficiently. Compared to marginalized graph kernels our method in some cases leads to a significant reduction of the prediction error. Further improvement can be gained, if expert knowledge is combined with our method. We also investigate a reduced graph representation of molecules by collapsing certain structural elements, like e.g. rings, into a single node of the molecular graph.

1. Introduction

In Chemoinformatics there has been a long history of work on the problem to infer chemical or biological properties of a molecule from the structure of the molecule, the so called *QSAR* approach (Kubinyi, 2003). The basic assumption is, that all molecular properties can be infered from the molecular structure. Classically, molecules are represented by a large amount of *descriptors* (= features in Machine Learning language) and then any data mining method, which works on vectorial data, can be applied. However, the problem here is to first find good descriptors and second to select the descriptors, which are best suited for the problem at hand. This can be quite difficult and computationally costly. More naturally, the topology of chemical compounds can be represented as labeled graphs, where edge labels correspond to bond properties like bond order, length of a bond, and node labels to atom properties, like partial charge, membership to a ring, and so on. This representation opens the opportunity to use graph mining methods (Washio & Motoda, 2003) to deal with molecular structures. Thereby a principal question is how different graph structures can be compared.

One way of doing so is the usage of a symmetric, positive semidefinite kernel – e.g. (Schölkopf & Smola, 2002). In (Kashima et al., 2003) the authors propose a kernel function between labeled graphs, which they call *marginalized graph kernel:* Its idea is to compute the expected match of all pairs of random walk label sequences up to infinite length. An efficient computation can be carried out in a time complexity proportional to the product of the size both graphs by solving a system of linear simultaneous equations. Kashima et al. show that also the *geometric* and the *exponential graph kernel* by (Gärtner et al., 2003) can be seen as special variants of the marginalized graph kernel. In contrast, the *pattern-discovery* (PD) kernel by De Raedt and Kramer (Raedt & Kramer, 2001) counts the set of all label sequences, which appear in more than p graphs with p being a so called *minimum support* parameter. Furthermore, it is possible to add extra conditions, for example selecting only the paths frequent in a certain class and scarce in another class. The PD method was especially designed for predicting toxicity of molecules, which from a chemical viewpoint mainly depends on the presence of certain functional groups in a molecule, and achieves about the same excellent performance there as the marginalized graph kernel (Kashima et al., 2003; Helma et al., 2001).

The goal of our work is to define a kernel for chemical compounds, which, like the marginalized graph kernel, is of general use for QSAR problems, but bet-

Appearing in *Proceedings of the 22^{nd} International Conference on Machine Learning*, Bonn, Germany, 2005. Copyright 2005 by the author(s)/owner(s).

ter reflects a chemists' point of view on the similarity of molecules. Rather than comparing label sequences, the main intuition of our approach is that the similarity of two molecules mainly depends on the matching of certain structural elements like rings, functional groups and so on (fig. 1). If we assume the membership of an atom to a structural element to be encoded in its labels, this leads to the idea of computing an optimal assignment from atoms in one structure to those in another one, including for each atom information on the neighborhood and other characteristic information, like e.g. charge, mass and so on. As a byproduct this leads to a new class of kernel functions, which to our knowledge has not been introduced so far. The optimal assignment allows an easy interpretation of the kernel from the chemistry side. We can extend the approach by considering *reduced graph representations* of molecules, i.e. we collapse certain structural elements, like a ring, a donor or an acceptor, into a single node and remove all remaining ones. In the literature this procedure is also called *pharmacophore mapping* (Martin, 1998). Furthermore, we investigate the effect of combining certain descriptor information provided by expert knowledge with our method.

This paper is organized as follows: In the next section we begin by defining so called "optimal assignment kernels" as a general class of kernel functions and prove their symmetry and positive semidefiniteness. Given this result we can introduce our optimal assignment kernel for chemical compounds in section 3 and show how it can be computed efficiently. In section 4 we investigate possible extensions of the optimal assignment kernel, namely by means of the reduced graph representation and by incorporating expert provided descriptor information. In section 5 we give experimental results of our method in comparison to marginalized graph kernels on several QSAR datasets and show that in some cases we can significantly outperform marginalized graph kernels. Finally, we conclude in section 6 and point out directions of future research.

2. Optimal Assignment Kernels

Let $\mathcal{X}$ be some domain of structured objects (e.g. graphs). Let us denote the parts of some object x (e.g. the nodes of a graph) by $x_1, ..., x_{|x|}$, i.e. x consists of $|x|$ parts, while another object y consists of $|y|$ parts. Let $\mathcal{X}'$ denote the domain of all parts, i.e. $x_i \in \mathcal{X}'$ for $1 \leq i \leq |x|$. Further let π be some permutation of either an $|x|-$subset of natural numbers $\{1, ..., |y|\}$ or an $|y|-$subset of $\{1, ..., |x|\}$ (this will be clear from context).

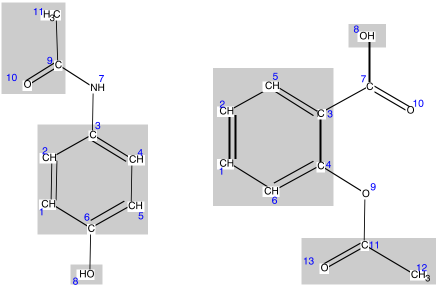

Figure 1. Matching regions of two molecular structures.

Definition 2.1. (Optimal Assignment Kernels) Let $k_1 : \mathcal{X}' \times \mathcal{X}' \to \mathbb{R}$ be some non-negative, symmetric and positive semidefinite kernel. Then $k_A : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with

$$k_A(x,y) := \begin{cases} \max_\pi \sum_{i=1}^{|x|} k_1(x_i, y_{\pi(i)}) & \text{if } |y| \geq |x| \\ \max_\pi \sum_{j=1}^{|y|} k_1(x_{\pi(j)}, y_j) & \text{otherwise} \end{cases}$$

is called an *optimal assignment kernel.*

This definition captures the idea of a maximal weighted bipartite matching (optimal assignment) of the parts of two objects.Each part of the smaller of both structures is assigned to exactly one part of the other structure such that the overall similarity score between both structures is maximized.

Lemma 2.2. *For all x: $k_A(x,x) = \sum_i k_1(x_i, x_i)$.*

Proof. For any π it is

$$\begin{aligned} & k_1(x_1, x_{\pi(1)}) + \ldots + k_1(x_{|x|}, x_{\pi(|x|)}) && (1) \\ \leq\ & \frac{1}{2}\big(k_1(x_1, x_1) + k_1(x_{\pi(1)}, x_{\pi(1)}) + \ldots && (2) \\ & + k_1(x_{|x|}, x_{|x|}) + k_1(x_{\pi(|x|)}, x_{\pi(|x|)})\big) \\ =\ & \sum_i k_1(x_i, x_i) && (3) \end{aligned}$$

because $2k_1(x_i, x_{\pi(i)}) \leq k_1(x_i, x_i) + k_1(x_{\pi(i)}, x_{\pi(i)})$ for all i. This is a direct consequence of the positive semidefiniteness of k_1. If we now take the maximum over all π, then $(1) = k_A(x,x) = (2) = (3)$ □

Theorem 2.3. *k_A is a symmetric and pos. def. kernel.*

Proof. Clearly, k_A is symmetric, because of the definition.

W.l.o.g. let $|y| \geq |x|$. Because of the lemma, we have $k_A(x,x) = \sum_i k_1(x_i, x_i), k_A(y,y) = \sum_j k_1(y_j, y_j)$. Further it holds for all $\alpha, \beta \in \mathbb{R}$ and i, j

$$2\alpha\beta k_1(x_i, y_j) \leq \alpha^2 k_1(x_i, x_i) + \beta^2 k_1(y_j, y_j) \qquad (4)$$

because k_1 is a positive semidefinite kernel. It is

$$\alpha^2 k_A(x,x) - 2\alpha\beta k_A(x,y) + \beta^2 k_A(y,y) = \qquad (5)$$
$$\alpha^2 \sum_i k_1(x_i, x_i) - 2\alpha\beta \max_\pi \sum_i k_1(x_i, y_{\pi(i)})$$
$$+\beta^2 \sum_j k_1(y_j, y_j)$$

By definition of k_A the second sum of (5) has $\min(|x|,|y|) = |x|$ addends. Let y'_i be the part of y to which x_i is assigned. Using (4) we have (5) $\geq \sum_{i=1}^{|x|}(\alpha^2 k_1(x_i,x_i) - 2\alpha\beta k_1(x_i,y'_i) + \beta^2 k_1(y'_i,y'_i)) \geq 0$. This proofs the positive semidefiniteness of each 2×2 kernel matrix. From this we can generalize the result to $n \times n$ matrices by induction using the assumption that k_1 is non-negative: Suppose we already know that each $n \times n$ kernel matrix $\mathbf{K} = (k_A(x^i, x^j))_{ij}$ for a set of objects $x^1, ..., x^n$ is positive semidefinite. Now assume we extend the matrix to size $n+1 \times n+1$ by adding an object x^{n+1}. It is

$$\sum_{i,j=1}^{n+1} \mathbf{v}_i \mathbf{v}_j \mathbf{K}_{ij} = \sum_{i,j=1}^{n} \mathbf{v}_i \mathbf{v}_j \mathbf{K}_{ij} \qquad (6)$$
$$+2\sum_{j=1}^{n} \mathbf{v}_{n+1} \mathbf{v}_j \mathbf{K}_{n+1,j} + \mathbf{v}_{n+1}^2 \mathbf{K}_{n+1,n+1}$$

By induction assumption we know the first part of (6) to be non-negative. Furthermore, by definition k_1 and thus also k_A is non-negative. Hence, we have $\mathbf{v}_{n+1}^2 \mathbf{K}_{n+1,n+1} \geq 0$. Therefore, in order to make (6) < 0 we have to suppose $2\sum_{j=1}^{n} \mathbf{v}_{n+1} \mathbf{v}_j \mathbf{K}_{n+1,j} < 0$. Using (4) this leads to $2\sum_{j=1}^{n} \mathbf{v}_{n+1} \mathbf{v}_j \mathbf{K}_{n+1,j} \leq \sum_{j=1}^{n} \mathbf{v}_{n+1}^2 \mathbf{K}_{n+1,n+1} + \mathbf{v}_j^2 \mathbf{K}_{jj} < 0$, which is a contradiction to the non-negativity of k_A. Hence, it is (6) ≥ 0, which proofs the theorem. □

3. Optimal Assignment Kernels for Chemical Compounds

Let us assume now we have two molecules m and m', which have atoms $a_1, ..., a_{|m|}$ and $a'_1, ..., a'_{|m'|}$. Let us further assume we have a kernel k_{nei}, which compares a pair of atoms $(a_h, a'_{h'})$ from both molecules, including information on their neighborhoods, membership to certain structural elements and other characteristics. Then, given our result from the last section, a valid kernel between m, m' is the optimal assignment kernel

$$k_A(m,m') = \qquad (7)$$
$$\begin{cases} \max_\pi \sum_{h=1}^{|m|} k_{nei}(a_h, a'_{\pi(h)}) & \text{if } |m'| \geq |m| \\ \max_\pi \sum_{h'=1}^{|m'|} k_{nei}(a_{\pi(h')}, a'_{h'}) & \text{otherwise} \end{cases}$$

That means we assign each atom of the smaller of both molecules to exactly one atom of the bigger molecule such that the overall similarity score is maximized. This can be computed efficiently in $\mathcal{O}(\max(|m|,|m'|)^3)$ (Mehlhorn & Näher, 1999). Although this seems to be a drawback compared to the quadratic time complexity of marginalized graph kernels, we have to point out, that marginalized graph kernels have to be iteratively computed until convergence, and thus in practice, depending on the size of the molecules, there might be no real difference in computation time.

In order to prevent larger molecules automatically to achieve a higher kernel value than smaller ones, we should further normalize our kernel (Schölkopf & Smola, 2002), i.e.

$$k_A(m,m') \leftarrow \frac{k_A(m,m')}{\sqrt{k_A(m,m) k_A(m',m')}} \qquad (8)$$

We now have to define the kernel k_{nei}. For this purpose let us suppose we have kernels k_{atom} and k_{bond} which compare the atom and bond features, respectively. These feature vectors should include various information, for instance, whether an atom belongs to a ring, if it is in a donor or acceptor, what partial charge it has and so on (see also experimental section). A natural choice for k_{atom} and k_{bond} would be the RBF-kernel, which computes the similarity of the feature vectors associated to a pair of atoms or bonds. Thereby we should normalize these feature vectors, e.g. to unit length. Let us denote by $a \rightarrow n_i(a)$ the bond connecting atom a with its ith neighbor $n_i(a)$. Let us further denote by $|N(a)|$ the number of neighbors of atom a. We now define a kernel R_0, which compares all direct neighbors of atoms (a, a') as the optimal assignment kernel between all neighbors of a and a' and the bonds leading to them, i.e.

$$R_0(a,a') = \qquad (9)$$
$$\frac{1}{|N(a')|} \max_\pi \sum_{i=1}^{|N(a)|} \Big(k_{atom}(n_i(a), n_{\pi(i)}(a'))$$
$$\cdot k_{bond}(a \rightarrow n_i(a), a' \rightarrow n_{\pi(i)}(a')) \Big)$$

were we assumed $|N(a')| \geq |N(a)|$ for the sake of simplicity of notation. As an example consider the C-atom 3 in the left and the C-atom 5 in the right structure of figure 2: If our only atom and bond features were the element type and bond order, respectively, and k_{atom} and k_{bond} would simply count a match by 1 and a mismatch by 0, our kernel $R_0(a_3, a'_5)$ would tell us that 2 of 3 possible neighbors of atom 3 in the left structure match with the neighbors of atom

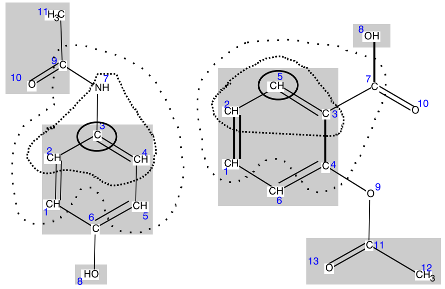

Figure 2. Direct and indirect neighbors of atom 3 in the left and atom 5 in the right molecule.

5 in the right structure. It is worth mentioning that the computation of R_0 can be done in $\mathcal{O}(1)$ as for chemical compounds $|N(a)| and |N(a')|$ can be upper bounded by a small constant (e.g. 4). Of course it would be beneficial not to consider the match of direct neighbors only, but also that of indirect neighbors and atoms having a larger topological distance. For this purpose we can evaluate R_0 not at (a, a') only, but also at all pairs of neighbors, indirect neighbors and so on, up to some topological distance L. In our example that would mean we also evaluate $R_0(a_2, a'_2), R_0(a_4, a'_2), R_0(a_7, a'_2), ...$ and so on. If we consider the mean of all these values and add them to $k_{atom}(a, a') + R_0(a, a')$, this leads to the following definition of the kernel k_{nei}:

$$k_{nei}(a, a') = \qquad (10)$$

$$k_{atom}(a, a') + R_0(a, a') + \sum_{\ell=1}^{L} \gamma(\ell) R_\ell(a, a')$$

Here R_ℓ denotes the mean of all R_0 evaluated at neighbors of topological distance ℓ, and $\gamma(\ell)$ is a decay parameter, which reduces the influence of neighbors that are further away and depends on the topological distance ℓ to (a, a'). It makes sense to set $\gamma(\ell) = p(\ell)p'(\ell)$, where $p(\ell), p'(\ell)$ are the probabilities for molecules m, m' that neighbors with topological distance ℓ are considered.

A key observation is, that R_ℓ can be computed efficiently from $R_{\ell-1}$ via the recursive relationship

$$R_\ell(a, a') = \frac{1}{|N(a)||N(a')|} \sum_{i,j} R_{\ell-1}(n_i(a), n_j(a')) \qquad (11)$$

I.e. we can compute k_{nei} by iteratively revisiting all direct neighbors of (a, a') only. In case that L is set to a constant we thus have a $\mathcal{O}(1)$ time complexity for the calculation of k_{nei}. In case that $L \to \infty$, we can prove the following theorem:

Theorem 3.1. *Let $\gamma(\ell) = (\hat{p}_1\hat{p}_2)^\ell$ with $\hat{p}_1, \hat{p}_2 \in (0, 1)$. If there exists a $C \in \mathbb{R}^+$, such that $k_{atom}(a, a') \le C$ for all a, a' and $k_{bond}(a \to n_i(a), a' \to n_j(a')) \le C$ for all $a \to n_i(a), a' \to n_j(a')$, then (10) converges for $L \to \infty$.*

Proof. It is $R_0(a, a') \le \frac{\min(|N(a)|,|N(a')|)}{\max(|N(a')|,|N(a')|)} C^2 \le C^2$ and thus $R_1(a, a') \le \frac{1}{|N(a)||N(a')|} \sum_{i=1}^{|N(a)|} \sum_{j=1}^{|N(a')|} C^2 = C^2$. Hence also $R_\ell(a, a') \le C^2$ for $\ell = 2, ..., L$. Therefore we have $(10) \le C + C^2 + (\hat{p}_1\hat{p}_2)^1 C^2 + ... + (\hat{p}_1\hat{p}_2)^L C^2 = C + C^2 + \sum_{\ell=1}^{L} (\hat{p}_1\hat{p}_2)^\ell$ which converges for $L \to \infty$. □

Note, that the boundedness of k_{atom} and k_{bond} is especially fulfilled, if we take the RBF-kernel for both.

4. Possible Extensions

4.1. Reduced Graph Representation

The main intuition of our method lies in the matching of structural elements from both molecules. In the previous section we achieved this by using structural, neighborhood and other characteristic information for each single atom and bond, and computing the optimal assignment kernel between atoms of both molecules then. A quite natural extension of this idea is to collapse structural elements, like rings, donors, acceptors and others, into one single node of the molecular graph in a precomputing step and then apply our method to this reduced graph representation. Atoms not belonging to a-priori defined types of structural elements can even be removed (Martin, 1998). This allows us to concentrate on important structural features, where the definition of what an important structural feature actually is may be given by expert knowledge, depending on the QSAR problem at hand. The high relevance of such a pharmacophore mapping for QSAR models is also reported e.g. in (Chen et al., 1999; Oprea et al., 2002). However, two principal problems have to be solved to implement this idea: Firstly, if certain atoms are removed from our graph, then we may obtain nodes, which are disconnected from the rest of the graph. They have to be reconnected by new edges again such that these new edges preserve the neighborhood information, i.e. if before we had $a \to b$ and $b \to c$ and node b is removed, we should obtain $a \to c$. These new edges should contain information on the topological and geometrical distance of their end nodes (fig. 3). Secondly, we have to define how the feature vectors for each single atom and bond included in a structural element can be transfered to the whole structural element. This can, for instance, be

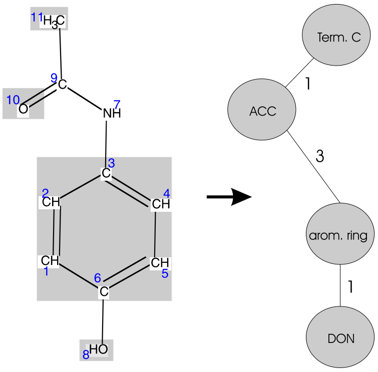

Figure 3. Example of a conversion of a molecule into its reduced graph representation with edge labels containing the topological distances.

solved by recursively applying our method from the last section, if two nodes representing structural elements have to be compared. From the computational side the reduced graph representation can be advantageous for larger molecules, because the effort for computing the optimal assignment is reduced.

4.2. Incorporating Expert Knowledge

For some QSAR problems it is known by experts that certain molecular properties are crucial. E.g. for human intestinal absorption the polar surface area of a molecule plays an important role (van de Waterbeemd & Gifford, 2003). In some sense these features describe global properties of a molecule, whereas our kernel relies on the graph structure and hence on local properties of a molecule. A natural question is, if it would be beneficial to combine both types of information. A straight forward way of doing so, is to consider the sum of a RBF-kernel for the descriptor information we have from our expert knowledge and the optimal assignment kernel.

5. Experimental Results

5.1. Datasets

We used the PTC dataset (Helma et al., 2001), which is the result of the following pharmaceutical experiments: Each of 417 chemical compounds is given to four types of test animals – Male Mouse (MM), Female Mouse (FM), Male Rat (MR) and Female Rat (FR). According to their carcinogenicity, each compound is assigned to one of the categories EE, IS, E, CE, SE, P, NE, N, where CE, SE and P indicate “relatively active” and NE and N “relatively inactive”, and EE, IS, E “can not be decided”. Following the approach in (Kashima et al., 2003), we simplified the problem by putting CE, SE and P into class “positive” and NE and N in class “negative”. The rest of the compounds was not considered. Hence, all in all we had four two-class problems. After removing the hydrogens (the hydrogen information can be encoded in the features “heavy valence”, “implicit valence” and “atom type” for each remaining atom - see table 1), the maximum size of a molecule in all four problems was 64 atoms, and the average size was 14 (FM/MM/MR) and 15 (FR) atoms, respectively.

The HIA (Human Intestinal Absorption) dataset consists of 164 structures from different sources in literature, which has been used in an earlier publication (Wegner et al., 2003). The molecules are divided into 2 classes “high oral bio-availability” and “low oral bio-availability”. The maximal molecule size was 57 and the average size 25 atoms after removing hydrogens. By expert knowledge 1 chemical descriptor (polar surface area) is provided, which is known to be relevant for the problem (van de Waterbeemd & Gifford, 2003).

The Yoshida dataset (Yoshida & Topliss, 2000) has 265 molecules divided into 2 classes “high bio-availability” and “low bio-availability”. After removing hydrogens, the maximum molecule size was 36 and the average size 20 atoms. For this problem 6 chemical descriptors are provided by expert knowledge describing the presence or absence of typical functional groups most likely to be involved in metabolic reactions (van de Waterbeemd & Gifford, 2003).

The BBB dataset (Feher et al., 2000) consists of 109 structures having a maximum molecule size of 33 and an average size of 16 atoms after removing hydrogens. The target is to predict the logBB value, which describes up to which degree a drug can cross the blood-brain-barrier. Two chemical descriptors (polar surface area and octane/water partition coefficient) are given by expert knowledge (van de Waterbeemd & Gifford, 2003).

5.2. Results

We compared the optimal assignment (OA) kernel from section 2 (7), (8) to the marginalized graph (MG) kernel using the same atom and bond features. All features were computed using the open source software JOELib[1] (table 1). The feature vectors for atoms and bonds were scaled to unit length. The kernels k_{atom} and k_{bond}, which compare atom and bond features, were the same for the OA and the MG kernel. In both cases we used a RBF kernel with $\sigma = 2^{-0.5}$. We ex-

[1] http://sourceforge.net/projects/joelib

Table 1. Atom and bond features chosen in our experiments

features	nominal	real valued
atom	element type, in donor, in acceptor, in donor or acceptor (Böhm & Klebe, 2002), in terminal carbon, in aromatic system (Bonchev & Rouvray, 1990), negative/positive, in ring (Figueras, 1996), in conjugated environment, free electrons, implicit valence, heavy valence, hybridization, is chiral, is axial	electro-topological state, conjugated topological distance (Todeschini & Consonni, 2000), Gasteiger/Marsili partial charge (Gasteiger & Marsili, 1978), mass
bond	order, in aromatic system, in ring, is rotor, in carbonyl/amide/primary amide/ester group	length

plicitly set k_{atom} to 0, if the element type of two atoms was different (e.g. if a C-atom is compared to a N-atom). This corresponds to the multiplication with a δ-kernel. The same was done for bonds, if one bond was in an aromatic system and the other not, or if both bonds had a different bond order.

For the OA kernel the probabilities $p(\ell), p'(\ell)$ to reach neighbors with topological distance ℓ was set to $p(\ell) = p'(\ell) = 1 - \frac{1}{L}\ell$ with $L = 3$. This allows us to consider the neighborhood of a whole 6-ring for an atom. We used a $C-$SVM on the classification problems and a $\epsilon-$SVR on the regression problem. The prediction strength was evaluated by means of 10-fold cross-validation, and on each training fold a model selection for the necessary parameters was performed by evaluating each candidate parameter set by an extra level of 5-fold cross-validation. On the classification problems the cross-validation procedures were stratified. For the MG kernel the model selection included testing the termination probabilities $p_t = 0.1, 0.3, 0.5, 0.7$. The parameter C was chosen from the interval $[2^{-2}, 2^{14}]$. Thereby on the classification problems we trained the SVM with asymmetric soft margin penalties $C_+ = w_+ \cdot C$ and $C_- = w_- \cdot C$, where $w_+ = 1$ and $w_- = \#negatives/\#positives$ in the actual training set. On the regression dataset (BBB) the parameter ϵ was chosen from the interval $[2^{-8}, 2^{-1}]$. The logBB values were normalized to mean 0 and standard deviation 1 on each training fold, and the calculated scaling parameters were then applied to normalize the logBB values in the actual testing set.

Table 2 shows the results we obtained. Our OA kernel on almost all data achieved better results than the MG kernel. The difference on the MR and BBB data was statistically significant (p-value $= 0.01$, p-value $= 0.06$) at significance level 10%. Thereby significance was tested by means of a two-tailed paired t-test. In tendency our OA kernel seems to better reflect the chemical and biological relevant aspects, which determine the similarity of chemical compounds. Comparing computation times, we could not find any significant difference between the MG kernel and the OA kernel on our data. Using our JAVA implementation one kernel function evaluation on our Pentium IV 3GHz desktop PC on average took 6 ± 9 (4 ± 10) ms on the BBB for the OA (MG) kernel, 10 ± 9 (7 ± 8) ms on the HIA, 7 ± 4 (5 ± 4) ms on the Yoshida, and 2 ± 3 (3 ± 4) ms on the PTC dataset.

We also investigated the effect of the reduced graph representation (OARG kernel - last column of table 2). Thereby in the reduced graph representation only direct neighbor subgraphs were considered to compute k_{nei} (i.e. $L' = 1$), whereas for the comparison of nodes representing structural elements we used $L = 3$ as before. Considered structural features were: ring, donor, donor or acceptor, acceptor and terminal carbon. Molecules, which did not contain any of these features and hence lead to an empty graph were removed: 2 compounds in the FM, FR, MR and BBB data, and 3 in the MM data. As represented in table 2 using the reduced graph representation we achieved similar error rates than with the original OA kernel. Compared to the MG kernel the improvement on the Yoshida dataset was statistically significant (p-value $=$ 0.05). This demonstrates that the reduced graph representation, although using less structural information than the original OA kernel, covers well the relevant biological aspects of the molecules in our data.

Next we investigated the effect of incorporating expert knowledge. We used the same experimental framework as described above. We first tested the prediction error when using the molecular descriptors provided by the expert only. The model selection thereby included the tuning of the width σ' of the RBF kernel in the range $\hat{\sigma}'/4, ..., 4\hat{\sigma}'$, where $\hat{\sigma}'$ was set such that $\exp(-D/(2\hat{\sigma}'^2)) = 0.1$ ($D =$ dimensionality of the data). The individual descriptors were normalized using the same procedure as described for the logBB values. In a second step we combined both kernels as described in 4.2. Thereby we just used a fixed $\sigma' = \hat{\sigma}'$.

Table 2. 10-fold CV error ± std. error. For classification problems the class loss (%) is reported, for the BBB data the mean squared error $\times 10^{-2}$. Significant wins of the OA/OARG kernels compared to the MG kernel are denoted by "*", losses by "-".

Data	MG	**OA**	**OARG**
FM	37.82 ± 3.01	38.69 ± 1.7	$\mathbf{35.99 \pm 3.31}$
MM	33.38 ± 2.19	$\mathbf{32.75 \pm 1.52}$	$\mathbf{32.16 \pm 2.25}$
FR	$\mathbf{33.06 \pm 1.73}$	$\mathbf{33.06 \pm 1.56}$	34.12 ± 1.47
MR	42.45 ± 1.15	$\mathbf{36.66 \pm 2.41}^*$	$\mathbf{40.99 \pm 2.91}$
HIA	17.72 ± 2.17	$\mathbf{15.33 \pm 1.44}$	$\mathbf{14.63 \pm 1.35}$
Yoshida	37.79 ± 2.33	$\mathbf{34.32 \pm 2.53}$	$\mathbf{32.21 \pm 2.82}^*$
BBB	54.66 ± 3.76	$\mathbf{39.91 \pm 6.55}^*$	$\mathbf{42.03 \pm 7.03}$

Table 3. Effect of incorporating expert knowledge. Significant improvements compared to the original OA/OARG kernels are marked by "*", deteriorations by "-".

Data	expert	**OA+exp.**	**OARG+exp.**
HIA	14.12 ± 2.13	$\mathbf{11.58 \pm 1.88}^*$	12.21 ± 1.79
Yoshida	32.45 ± 1.57	$\mathbf{30.2 \pm 1.52}$	31.3 ± 1.72
BBB	39.28 ± 4.41	40.42 ± 4.67	$\mathbf{35.55 \pm 5.14}$

Table 3 shows the results on this experiment. As one can see the prediction error by using the molecular descriptors, which are provided by expert knowledge, is almost identical to that achieved by our OA/OARG kernels in table 2. However, if both approaches were combined, on the HIA data with the OA kernel we obtained a significantly lower error rate (p-value $=$ 0.08) than with using the original OA kernel. Furthermore, in tendency the results on the other datasets are improved as well. This demonstrates that indeed it can be beneficial to incorporate further knowledge on "global" molecular properties additionally to the "local" properties encoded in the graph structure.

6. Conclusion

We introduced a new kernel for chemical compounds, which is based on the idea of computing an optimal assignment of atoms from one molecule to those of another one, including information on neighborhood, membership to a certain structural element and other characteristics. This leads to a new class of kernel functions, which we call *optimal assignment kernels*. The optimal assignment can be computed by means of the *Hungarian method*. We showed how the inclusion of neighborhood information for each atom can be done efficiently via a recursive update equation, even if not only direct neighbors are considered. Comparisons to the marginalized graph kernel by Kashima et al. in some cases lead to significantly lower error rates on our QSAR problems. We investigated two major extensions of our approach: the usage of a reduced graph representation, in which certain structural elements are collapsed into a single node of the molecular graph, and the incorporation of molecular descriptors provided by expert knowledge. We showed that the latter can lead to a further significant reduction of the error rate in comparison to the usual optimal assignment kernel, whereas the major benefit of the reduced graph representation lies in the fact that expert knowledge on important structural features can be included and that larger molecules can be handled more efficiently. All in all we see the main advantage of our approach that it better reflects a chemists' intuition on the similarity of molecules than marginalized graph kernels. The optimal assignment could also give the opportunity to deduce so called *pharmacophores* in future research. Especially for this purpose the reduced graph representation is advantageous. Other directions of future research include a more systematic investigation of the incorporation of expert knowledge, e.g. by means of kernel CCA – e.g. (Bach & Jordan, 2002), semidefinite programming (Lanckriet et al., 2004) or others.

Appendix: Preventing the "Tottering"

If we evaluate R_0 (Eq: 9) at all neighbors of a certain topological distance ℓ, we also revisit atoms and bonds that we have considered at topological distance $\ell - 1$. To prevent this "tottering", we can make our decay factor γ dependent not just on the topological distance, but also on the path of visited atoms and bonds. Thereby we have to explicitly forbid paths of the form $a \rightarrow n_i(a) \rightarrow a$. This can be achieved by setting

$$\gamma'(\ell, a, a', n_i(a), n_j(a')) = \begin{cases} 0 & \exists k : n_k(n_i(a)) = a \vee \exists t : n_t(n_j(a')) = a' \\ \gamma(\ell) & \text{otherwise} \end{cases} \tag{12}$$

This requires the following changes in our computation for k_{nei}:

$$k_{nei}(a, a') = k_{atom}(a, a') + \frac{1}{|N(a)||N(a')|} \sum_{i,j} r_0(a, a', n_i(a), n_j(a')) + \sum_{\ell=1}^{L} \left(\frac{1}{|N(a)||N(a')|} \sum_{i,j} \gamma'(\ell, a, a', n_i(a), n_j(a')) \right. \tag{13}$$

$$\cdot r_\ell(a,a',n_i(a),n_j(a'))\Big)$$

$$r_\ell(a,a',n_i(a),n_j(a')) = \frac{1}{|N(n_i(a))||N(n_j(a'))|}\cdot \tag{14}$$
$$\sum_{k,t} r_{\ell-1}(n_i(a),n_j(a'),n_k(n_i(a)),n_l(n_j(a')))$$

$$r_0(a,a',n_i(a),n_j(a')) = \tag{15}$$
$$k_{atom}(n_i(a),n_j(a'))k_{bond}(a \rightarrow n_i(a), a' \rightarrow n_j(a'))$$

That means we can compute k_{nei} by iteratively revisiting the direct neighbors and indirect neighbors of (a,a'). In contrast to (10) in (13) we do not use an optimal assignment kernel to for the direct neighbors of (a,a'), but compute the average match here. Up to now we did not recognize any significant effect on the error rate using the calculation described here instead of that described in section 3.

References

Bach, F., & Jordan, M. (2002). Kernel independent component analysis. *J. Machine Learning Research*, *3*, 1 – 48.

Böhm, M., & Klebe, G. (2002). Development of New Hydrogen–Bond Descriptors and Their Application to Comparative Molecular Field Analyses. *J. Med. Chem.*, *45*, 1585–1597.

Bonchev, D., & Rouvray, D. H. (Eds.). (1990). *Chemical Graph Theory: Introduction and Fundamentals*, vol. 1 of *Mathematical Chemistry Series*. London, UK: Gordon and Breach Science Publishers.

Chen, X., Rusinko, A., Tropsha, A., & Young, S. S. (1999). Automated Pharmacophore Identification for Large Chemical Data Sets. *J. Chem. Inf. Comput. Sci.*, *39*, 887–896.

Feher, M., Sourial, E., & Schmidt, J. (2000). A simple model for the prediction of blood-brain partitioning. *Int. J. Pharmaceut.*, *201*, 239 – 247.

Figueras, J. (1996). Ring Perception Using Breadth–First Search. *J. Chem. Inf. Comput. Sci.*, *36*, 986–991.

Gasteiger, J., & Marsili, M. (1978). A New Model for Calculating Atomic Charges in Molecules. *Tetrahedron Lett.*, *34*, 3181–3184.

Gärtner, T., Flach, P., & Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. *Proc. 16th Ann. Conf. Comp. Learning Theory and 7th Ann. Workshop on Kernel Machines.*

Helma, C., King, R., & Kramer, S. (2001). The predictive toxicology challenge 2000-2001. *Bioinformatics*, *17*, 107 – 108.

Kashima, H., Tsuda, K., & Inokuchi, A. (2003). Marginalized kernels between labeled graphs. *Proc. 20th Int. Conf. on Machine Learning.*

Kubinyi, H. (2003). Drug research: myths, hype and reality. *Nature Reviews: Drug Discovery*, *2*, 665–668.

Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L. E., & Jordan, M. (2004). Learning the kernel matrix with semidefinite programming. *J. Machine Learning Research*, *5*, 27 – 72.

Martin, Y. C. (1998). Pharmacophore mapping. *Des. Bioact. Mol.*, 121–148.

Mehlhorn, K., & Näher, S. (1999). *The LEDA Platform of Combinatorial and Geometric Computing.* Cambridge University Press.

Oprea, T. I., Zamora, I., & Ungell, A.-L. (2002). Pharmacokinetically based mapping device for chemical space navigation. *J. Comb. Chem.*, *4*, 258–266.

Raedt, L. D., & Kramer, S. (2001). Feature construction with version spaces for biochemical application. *Proc. 18th Int. Conf. on Machine Learning* (pp. 258 – 265).

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels.* Cambridge, MA: MIT Press.

Todeschini, R., & Consonni, V. (Eds.). (2000). *Handbook of Molecular Descriptors.* Weinheim: Wiley–VCH.

van de Waterbeemd, H., & Gifford, E. (2003). ADMET *In Silico* Modelling: Towards Prediction Paradise? *Nature Reviews: Drug Discovery*, *2*, 192–204.

Washio, T., & Motoda, H. (2003). State of the art of graph-based data mining. *SIGKDD Explorations Special Issue on Multi-Relational Data Mining*, *5.*

Wegner, J., Fröhlich, H., & Zell, A. (2003). Feature selection for Descriptor based Classification Models: Part II - Human Intestinal Absorption (HIA). *J. Chem. Inf. Comput. Sci.*, *44*, 931 – 939.

Yoshida, F., & Topliss, J. (2000). QSAR model for drug human oral bioavailability. *J. Med. Chem.*, *43*, 2575 – 2585.