Assignment Kernels For Chemical Compounds

Holger Fröhlich, Jörg K. Wegner, Andreas Zell

Center For Bioinformatics Tübingen (ZBIT)
Sand 1, 72076 Tübingen, Germany
E-mail: {froehlic,wegnerj,zell}@informatik.uni-tuebingen.de

***Abstract*— During the last years Kernel Methods like the Support Vector Machine (SVM) have gained a growing interest in Machine Learning. One of the strengths of this approach is the ability to deal easily with arbitrarily structured data by means of the kernel function. In this paper we propose a kernel for chemical compounds which is based on the idea of computing optimal assignments between atoms of two different molecules including information about their neighborhood. As a byproduct this leads to a new class of kernel functions. We demonstrate how the necessary computations can be carried out efficiently. We compare our method against the marginalized graph kernels by Kashima et al. and show its good performance on classifying toxicological and human intestinal absorption data.**

I. 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 [9]. The basic assumption is, that in nature often there exists a relationship between structure and certain molecular properties. 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, etc, and node labels to atom properties, like atom type, partial charge, membership to a ring, and so on. This representation opens the opportunity to use graph mining methods [15] 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 definite kernel – e.g. [12], [13]. In [8] 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 [5] can be seen as special variants of the marginalized graph kernel. In contrast, the *pattern-discovery* (PD) kernel by De Raedt and Kramer [11] 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 [8], [7].

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 better 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.

This paper is organized as follows: We begin by defining so called *assignment kernels* as a general class of kernel functions and prove their positive definiteness. Given this result we can define our kernels for chemical compounds in section 3 and show how they can be computed efficiently. In section 4 we give experimental results on classifying toxicological and human intestinal absorption data. Finally, we conclude in section 5.

II. 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

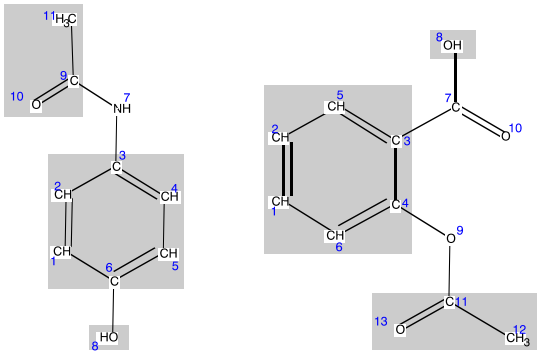

Fig. 1. Matching regions of two molecular structures.

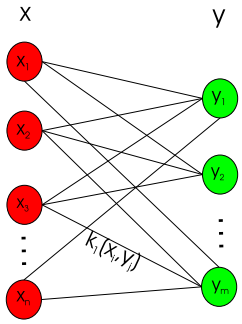

Fig. 2. Possible assignments between parts of two structured objects x, y with $|x| > |y|$. The goal is to find the maximum weighted bipartite matching (optimal assignment) from the parts of y to the parts of x.

parts, i.e. $x_i \in \mathcal{X}'$ for $1 \leq i \leq |x|$. Further let π be some permutation of an $|x|-$subset of natural numbers $\{1, ..., |y|\}$ or $|y|-$subset of $\{1, ..., |x|\}$, respectively (this will be clear from context).

Definition 2.1: (Assignment kernels) Let $k_1 : \mathcal{X}' \times \mathcal{X}' \to \mathbb{R}$ be some non-negative, symmetric, positive definite 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 *assignment kernel*.

This definition captures the idea of a maximal weighted bipartite matching (optimal assignment) of the parts of two objects (fig. 2). 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 definiteness 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 positive definite 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 definite kernel. It is

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

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 definiteness 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 definite. Now assume we extend the matrix to size $n + 1 \times n + 1$ by adding an object x^{n+1}. It is

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

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 $(6) \geq 0$, which proofs the theorem. ■

III. Assignment Kernels For Chemical Compounds

A. Construction of the kernel

We are now ready to construct an assignment kernel for chemical compounds. For each atom a in a molecule we have a set of certain real valued attributes $\phi_{real}(a)$ (like e.g. atom mass) and nominal attributes $\phi_{nom}(a)$ (like e.g. atom type). We define an atom kernel as

$$k_{atom}(a,a') = \\ k_{real}(\phi_{real}(a),\phi_{real}(a')) \cdot k_{nom}(\phi_{nom}(a),\phi_{nom}(a')) \tag{7}$$

Likewise, for each pair of bonds b, b' we have a bond kernel

$$k_{bond}(b,b') = \\ k_{real}(\phi_{real}(b),\phi_{real}(b')) \cdot k_{nom}(\phi_{nom}(b),\phi_{nom}(b')) \tag{8}$$

A natural choice for k_{real} is the RBF kernel of width σ while for the nominal kernel we take the normalized δ-kernel $k_{nom}(u,u') = \frac{1}{|u|}\sum_i \delta(u_i = u'_i)$. To have an accurate estimate of the similarity of a and a' we should also include information about their neighbors. Introducing the notation $< a >$ as the number of bonds of atom a, we thus define what we call the *base kernel* between two atoms (a,a') including their neighborhoods $N(a) = \{n_1(a),...,n_{<a>}(a)\}$, $N(a') = \{n_1(a'),...,n_{<a'>}(a')\}$, and all bonds $n_h(a) \to a, h = 1,...,|N(a)|$ and $n_{h'}(a') \to a', h' = 1,...,|N(a')|$:

$$k_{base}(a,a') = k_{atom}(a,a') \\ + \frac{1}{|N(a)||N(a')|}\sum_{h,h'}\Big(k_{atom}(n_h(a),n_{h'}(a')) \\ k_{bond}(n_h(a) \to a, n_{h'}(a') \to a')\Big) \tag{9}$$

That means the similarity between two atoms consists of two parts: first the similarity between the attributes of the atoms and second the similarity of the neighborhood structure. Thereby the similarity of each pair of neighbor atoms $(n_h(a), n_{h'}(a'))$ is weighted by the similarity of the bonds leading to them. The normalization factor before the sum is in order to ensure that atoms with a higher number of neighbors do not automatically achieve a higher similarity. Hence we divide by the number of addends in the sum. The definition of (9) is just a classical convolution kernel as introduced by D. Haussler [6].

As an example consider the C-atom 3 in the left and the C-atom 5 in the right structure of figure 1: Direct neighbors of atom 3 in the left structure are atoms 2, 4 and 7 (see fig. 3). Direct neighbors of atom 5 in the right structure are atoms 2 and 3. If we only concentrate on element and bond type and simply count a match by 1 and a mismatch by 0, clearly atoms 2 in the left and 2 in the right molecule match perfectly as well as atoms 4 and 3. They have the same element type and the same bond type leading to atoms 3 and 5, respectively. Atom 3 in the left molecule also has another neighbor, 7,

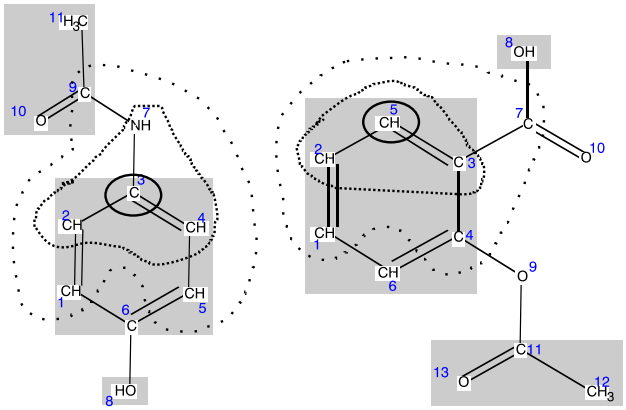

Fig. 3. Direct and indirect neighbors of atom 3 in the left and atom 5 in the right molecule

which does not match any neighbor of atom 5 in the right structure. Note that e.g. atom 2 in the left does not match atom 3 in the right molecule, because they have different bonds leading to atoms 3 and 5, respectively. The final kernel value for the C-atoms 3 and 5 would be computed as $k_{base}(a_3,a'_5) = 1 + \frac{1}{3\cdot 2}(1+1+0+0+0+0) = 1\frac{1}{3}$.

We may also want to consider not just direct neighbors, but also neighbors which are more far away up to some maximal topological distance $L+1$. For this purpose let us denote $n_{\mathbf{h}_l}(a) = n_{h_l}(\ldots n_{h_1}(a)\ldots)$. Let us further denote by

$$R_0(a,a') = \\ \frac{1}{|N(a)||N(a')|}\sum_{h,h'}\Big(k_{atom}(n_h(a),n_{h'}(a')) \\ k_{bond}(n_h(a) \to a, n_{h'}(a') \to a')\Big) \tag{10}$$

the second term in 9. Then we define the *extended base kernel* as

$$k'_{base}(a,a') = k_{base}(a,a') \\ +\gamma(1)\frac{1}{|N(a)||N(a')|}\sum_{h_1,h'_1} R_0(n_{h_1}(a),n_{h'_1}(a')) \\ +\gamma(2)\frac{1}{|N(a)||N(a')|}\Big(\sum_{h_1,h'_1}\frac{1}{|N(n_{h_1}(a))||N(n_{h'_1}(a'))|} \\ R_0(n_{\mathbf{h}_2}(a),n_{\mathbf{h'}_2}(a'))\Big) + \ldots \\ +\gamma(L)\frac{1}{|N(a)||N(a')|}\Big(\sum_{h_1,h'_1}\frac{1}{|N(n_{h_1}(a))||N(n_{h'_1}(a'))|}\Big(\\ \ldots\frac{1}{|N(n_{\mathbf{h}_{L-1}}(a))||N(n_{\mathbf{h'}_{L-1}}(a'))|} \\ \sum_{h_L,h'_L} R_0(n_{\mathbf{h}_L}(a),n_{\mathbf{h'}_L}(a'))\Big)\ldots\Big) \tag{11}$$

The first addend in (11) takes into account the direct neighbors of (a, a'), the next addend computes the average of the match of all neighbors which have topological distance 2 by evaluating R_0 for all direct neighbors of (a, a'). The third addend does the same for all neighbors with topological distance 3. Finally, the last addend considers all neighbors which have topological distance $L+1$ by evaluating R_0 for all neighbors at topological distance L. The factor $\gamma(l)$ is a decay parameter in order to reduce the influence of neighbors which are further away and depends on the topological distance $l+1$ to (a, a'). Like for the original base kernel we use the normalization factors to ensure that atoms with a higher number of neighbors do not automatically achieve a higher similarity.

As an example let us assume $L = 1$ in the previous example. We evaluate R_0 at all direct neighbors 2, 4 and 7 in the left, and 2 and 3 in the right structure, i.e. we compute $R_0(a_2, a'_2) = 0.5, R_0(a_2, a'_3) = \frac{1}{3}, R_0(a_4, a'_2) = 0.5, R_0(a_4, a'_3) = \frac{1}{3}$, $R_0(a_7, a'_2) = \frac{1}{6}$ and $R_0(a_7, a'_3) = \frac{1}{3}$. The average over the values of $R_0(a_2, a'_2), R_0(a_2, a'_3), ...$, weighted by the decay factor $\gamma(1)$ is added to $k_{base}(a_3, a'_5)$ in (11), i.e. $k'_{base}(a_3, a'_5) = k_{base}(a_3, a'_5) + \gamma(1)\frac{1}{6}\sum_{i=1}^{2}\sum_{j=1}^{3} R_0(a_i, a'_j) \approx 1\frac{1}{3} + \gamma(1) \cdot 0.36$. In this little example we just concentrated on element and bond type. As one can imagine the inclusion of more features can improve the results as well as higher values for L (see experimental section).

An interesting case is when $L \to \infty$. In this case we can prove the following theorem:

Theorem 3.1: Let be $\gamma(l) = (p_1 p'_2)^l$ and $p_1 p'_2 \in (0, 1)$. If there exists a $C \in \mathbb{R}^+$, such that $k_{atom}(a, a') \leq C$ for all a, a' and $k_{bond}(n(a) \to a, n(a') \to a') \leq C$ for all $n(a) \to a, n(a') \to a'$, then (11) converges.

Proof: $(11) \leq k_{base}(a, a') + C^2(p_1 p_2)^1 + ... + C^2(p_1 p_2)^L = k_{base}(a, a') + C^2 \sum_{l=1}^{L}(p_1 p_2)^l$ which converges for $L \to \infty$. ∎

The constants p_1, p_2 can be interpreted as continuation probabilities for random walks on molecules m and m'. Hence, the normalizing factors before the sums in (11) can be viewed as the probability to reach the corresponding pair of atoms. The boundedness of k_{atom} and k_{bond} can be ensured easily by taking the RBF kernel for k_{real} in both cases.

Now we want to construct the whole kernel between two molecules m, m' with atoms $a_1, ..., a_{|m|}$ and $a'_1, ..., a'_{|m'|}$. For the sake of simplicity of notation in the following let us assume $|m'| \geq |m|$. We define the *optimal assignment kernel* between two molecules as

$$k_{asn}(m, m') = \max_{\pi} \sum_{h} k'_{base}(a_h, a'_{\pi(h)}) \tag{12}$$

I.e. we compute the optimal assignment of the atoms of both molecules while taking into account the similarity of their neighborhood structure. Since we proved that assignment kernels are positive definite in the previous section, we can conclude that k_{asn} is a valid positive definite kernel.

By looking at the computed optimal assignment $\hat{\pi}$ this kernel has the advantage of being transparent, because one can manually comprehend why a certain pair of molecules is given a higher similarity than another pair. This gives us the opportunity to actually interpret the kernel in a chemical context.

Instead of computing the optimal matching between both molecules one could also simply compute the expected match, i.e.

$$k_{em}(m, m') = \sum_{h,h'} k'_{base}(a_h, a'_{h'}) \tag{13}$$

The *expected match kernel* can be seen as a speed-up version of the optimal assignment kernel. However, it looses the nice feature of transparency.

Finally, in order to prevent that larger molecules automatically achieve a higher kernel value, we should normalize the kernel [12], i.e.

$$k(m, m') \leftarrow \frac{k(m, m')}{\sqrt{k(m, m)k(m', m')}} \tag{14}$$

where k is either k_{asn} or k_{em}.

B. Efficient Computation

We now turn to the question, how computations can be carried out efficiently. The first thing to realize is, that the number of neighbors of each atom in a molecule can be upper bounded by a small constant (usually 4). Hence, (9) can be computed in $\mathcal{O}(1)$ for each pair of atoms. The extended base kernel (11) can be rewritten as:

$$k'_{base}(a, a') = k_{base}(a, a') + \sum_{l=1}^{L} \gamma(l) R_l(a, a') \tag{15}$$

$$R_l(a, a') = \frac{1}{|N(a)||N(a)|} \sum_{h,h'} R_{l-1}(n_h(a), n_{h'}(a')) \tag{16}$$

That means we can compute k'_{base} by means of k_{base} and the recursive update formula (16). Let $n = \max(|m|, |m'|)$ then the complexity for the computation of (15) for all pairs of atoms is $\mathcal{O}(n^2)$. The optimal assignment between atoms in (12) can be computed efficiently by means of the classical Kuhn-Munkres algorithm (also known as the *Hungarian Method* [10]) in $\mathcal{O}(n^3)$. Although this seems to be a drawback compared to 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 n, there might be no real difference in computation time. For the expected match kernel (13) the overall complexity is just $\mathcal{O}(n^2)$ and hence the same as for the marginalized graph kernels.

TABLE I

ATOM AND BOND FEATURES CHOSEN IN OUR EXPERIMENTS

features	nominal	real valued
atom	type, valence, in donor, in acceptor, in donor or acceptor [1], in terminal carbon, in aromatic system [2], negative/positive, in ring [3]	electro-topological state, conjugated topological distance [14], partial charge [4], mass
bond	order, in ring [3], is aromatic [2], is rotor, is up/down, is in carbonyl/amide/primary amide group	—

TABLE II

5-FOLD STRATIFIED CROSS-VALIDATION ACCURACY ON DIFFERENT DATA (%) $\pm$ STD. ERROR (%)

data set	MG-Kernel	**OA-Kernel**	**EM-Kernel**
FM	64.76 ± 1.15 $p_t = 0.3$ $\sigma = 2^{-2}$	65.33 ± 0.94 $L = 1$ $\sigma = 2^{-4}$	64.47 ± 1.2 $L = 9$ $\sigma = 2^{-2}$
MM	69.05 ± 1.45 $p_t = 0.6$ $\sigma = 2^{-4}$	67.87 ± 1.7 $L = 6$ $\sigma = 2^{-4}$	66.97 ± 1.07 $L = 3$ $\sigma = 2^{-4}$
FR	70.09 ± 0.59 $p_t = 0.6$ $\sigma = 2^{0}$	70.37 ± 1.07 $L = 1$ $\sigma = 2^{-4}$	68.95 ± 0.7 $L = 3$ $\sigma = 2^{0}$
MR	62.5 ± 1.23 $p_t = 0.3$ $\sigma = 2^{0}$	63.39 ± 2.06 $L = 3$ $\sigma = 2^{-4}$	60.84 ± 1.67 $L = 9$ $\sigma = 2^{-4}$
HIA	84.75 ± 2.54 $p_t = 0.1$ $\sigma = 2^{-2}$	85.99 ± 1.19 $L = 3$ $\sigma = 2^{-4}$	84.17 ± 1.07 $L = 9$ $\sigma = 2^{0}$

IV. EXPERIMENTS

A. Datasets

We used the PTC dataset [7], 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 [8], 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 feature "atom type" for each remaining atom - see table I), 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 [16]. 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.

B. Experimental Setting and Results

We compare the optimal assignment kernel (OA) against the expected match kernel (EM) and the marginalized graph kernel (MG) using the same atom and bond features. Thereby for each atom we computed 9 nominal and 5 real valued features, and for each bond we selected 8 nominal features (table I). All features were computed by means of the open source software JOELIB[1] developed in our group. The kernels k_{atom} and k_{bond}, which compare atom and bond features, were the same for the OA, the EM and the MG kernel.

All real valued features were normalized to mean 0 and standard deviation 1 over the whole dataset. We explicitly set the value k_{nom} to 0, if for two atoms the atom type or for two bonds the bond type was not identical (this corresponds to a multiplication with a δ-kernel). This reflects the fact that computing a similarity for atoms of different type or bonds of different type is quite senseless. For the real valued attributes we chose a RBF-kernel of width $\sigma = 2^{-8}, 2^{-6}, ..., 2^{4}$. The decay parameter γ was set to $\gamma(l) = p_1(l)p_2(l)$ with $p_i(l) = 1 - \frac{1}{L}l, i = 1, 2$, and $l = 1, ..., L$, and for L we chose values $1, 3, 6, 9$ (for $L = 1$ only direct neighbors are considered).

For the marginalized graph kernel we tested the termination probabilities $p_t = 0.1, 0.3, 0.6, 0.9$. We used a SVM as the classification algorithm. The classification accuracy was evaluated by 5-fold stratified cross-validation, and on the training folds the parameter C was chosen via an extra 5-fold stratified cross-validation from the interval $2^{-2}, 2^{0}, ..., 2^{12}$. We trained the SVM with an asymmetric soft margin penalty $C_+ = w_+ \cdot C$ and $C_- = w_- \cdot C$, where $w_+ = 1$ and $w_- = \#negatives/\#positives$ in the dataset. Table II shows the best classification results we obtained over choices of kernel parameters σ, and p_t or L, respectively.

Our optimal assignment kernel gives almost identical results to the marginalized graph kernel. Our expected match kernel in 2 cases performs slightly worse than the optimal assignment kernels, but the difference is not significant. Hence, for large molecules the expected match kernel could be seen as an alternative to the optimal assignment kernel. Comparing computation times, we could not find any significant differences between the methods. Using our JAVA implementation one kernel function evaluation on our Pentium IV 3GHz desktop PC on average took around 20ms on the HIA and 5ms on the PTC dataset. However, we see the biggest advantage of our method that it better reflects a chemists' intuition on the similarity of molecules.

V. CONCLUSION

We introduced a new kernel for chemical compounds, which is based on the idea of computing optimal assignments between atoms of two different molecules including their

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

neighborhood structure. This lead to a new class of kernel functions, so called assignment kernels. We showed how the optimal assignment kernel between two molecules can be computed efficiently by means of an recursive update equation, even if not only the direct neighbors are considered. The *Hungarian method* was used to compute the final assignment. Alternatively, one can just compute the expected match, which is asymptotically faster and gives results close the solution obtained by the Hungarian method. Comparisons to the marginalized graph kernels by Kashima et al. showed an almost identical performance. However, we see an advantage of our approach that it better reflects a chemists' intuition on the similarity of molecules, because of its transparency and easy interpretability. The optimal assignment could also give the opportunity to look for *pharmacophores* in future research. To further increase the classification accuracy it may help to incorporate certain molecular properties known to be relevant for the problem at hand. For large molecules it would be beneficial not to use a molecule representation based on single atoms as it was used here, but to use a reduced representation based on certain motifs like rings, donors, acceptors, etc. This is subject to future research.

ACKNOWLEDGMENT

Our special thank goes to Jean-Yves Audibert, CERTIS, France, for the helpful discussion.

REFERENCES

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

[2] D. Bonchev and D. H. Rouvray, editors. *Chemical Graph Theory: Introduction and Fundamentals*, volume 1 of *Mathematical Chemistry Series*. Gordon and Breach Science Publishers, London, UK, 1990.

[3] J. Figueras. Ring Perception Using Breadth–First Search. *J. Chem. Inf. Comput. Sci.*, 36:986–991, 1996.

[4] J. Gasteiger and M. Marsili. A New Model for Calculating Atomic Charges in Molecules. *Tetrahedron Lett.*, 34:3181–3184, 1978.

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

[6] D. Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, University of California Santa Cruz, 1999.

[7] C. Helma, R. King, and S. Kramer. The predictive toxicology challenge 2000-2001. *Bioinformatics*, 17:107 – 108, 2001.

[8] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *Proc. 20th Int. Conf. on Machine Learning*, 2003.

[9] H. Kubinyi. From Narcosis to Hyperspace: The History of QSAR. *Quant. Struct. Act. Relat.*, 21:348–356, 2002.

[10] H. Kuhn. The hungarian method for the assignment problem. *Naval Res. Logist. Quart.*, 2:83 – 97, 1955.

[11] L. De Raedt and S. Kramer. Feature construction with version spaces for biochemical application. In *Proc. 18th Int. Conf. on Machine Learning*, pages 258 – 265, 2001.

[12] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[13] J. Shawe-Tylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[14] R. Todeschini and V. Consonni, editors. *Handbook of Molecular Descriptors*. Wiley–VCH, Weinheim, 2000.

[15] T. Washio and H. Motoda. State of the art of graph-based data mining. *SIGKDD Explorations Special Issue on Multi-Relational Data Mining*, 5, 2003.

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