GRAPH WAVELET ALIGNMENT KERNELS FOR DRUG VIRTUAL SCREENING

Aaron Smalter[1], Jun Huan[1], and Gerald Lushington[2]
[1] *Department of Electrical Engineering and Computer Science*
[2] *Molecular Graphics and Modeling Laboratory*
University of Kansas
Email :{*asmalter, jhuan, glushington*}*@ku.edu*

In this paper we introduce a novel graph classification algorithm and demonstrate its efficacy in drug design. In our method, we use graphs to model chemical structures and apply a wavelet analysis of graphs to create features capturing graph local topology. We design a novel graph kernel function to utilize the created feature to build predictive models for chemicals. We call the new graph kernel a graph *wavelet-alignment kernel*.

We have evaluated the efficacy of the wavelet-alignment kernel using a set of chemical structure-activity prediction benchmarks. Our results indicate that the use of the kernel function yields performance profiles comparable to, and sometimes exceeding that of the existing state-of-the-art chemical classification approaches. In addition, our results also show that the use of wavelet functions significantly decreases the computational costs for graph kernel computation with more than 10 fold speed up.

1. INTRODUCTION

The fast accumulation of data describing chemical compound structures and biological activity calls for the development of efficient informatics tools. *Cheminformatics* is a rapidly emerging research discipline that employs a wide array of statistical, data mining, and machine learning techniques with the goal of establishing robust relationships between chemical structures and their biological properties. Cheminformatics hence is an important component on the application side of applying informatics approach to life science problems. It has a broad range of applications in chemistry and biology; arguably the most commonly known roles are in the area of drug discovery where cheminformatics tools play a central role in the analysis and interpretation of structure-activity data collected by various means of modern high throughput screening technology. Traditionally the analysis of large chemical structure-activity databases was done only within pharmaceutical companies and up until recently the academic community has had only limited access to such databases. This situation, however, has changed dramatically in very recent years.

In 2002, the National Cancer Institute created the Initiative for Chemical Genetics (ICG) with the goal of offering to the academic research community a large database of chemicals with their roles in cancer research [18]. Two years later, the National Health Institute (NIH) launched a Molecular Libraries Initiative (MLI) that included the formation of the national Molecular Library Screening Centers Network (MLSCN). MLSCN is a consortium of 10 high-throughput screening centers for screening large chemical libraries [1]. Collectively, ICG and MLSCN aim to offer to the academic research community the results of testing about a million compounds against hundreds of biological targets. To organize this data and to provide public access to the results, the PubChem database and the Chembank database have been developed as the central repository for chemical structure-activity data. These databases currently contain more than 18 million chemical compound records, more than 1000 bioassay results, and links from chemicals to bioassay description, literature, references, and assay data for each entry.

These publicly-available large-scale chemical compound databases have offered tremendous opportunities for designing highly efficient *in silico* drug design method. Many machine learning and data mining algorithms have been applied to study the structure-activity relationship of chemicals. For example, Xue *et al.* reported promising results of applying five different machine learning algorithms: logistic regression, C4.5 decision tree, k-nearest neighbor, probabilistic neural network, and support vector machines to predicting the toxicity of chemicals against an organism of Tetrahymena pyriformis [21].

Advanced techniques, such as random forest and MARS (Multivariate Adaptive Regression Splines) have also been applied to cheminformatics applications [15, 17].

Recently Support Vector Machines (SVM) have gained popularity in drug design. Support vector machines work by constructing a hyperplane in a high dimensional feature space. Two key insights of SVM are the utilization of kernel functions (i.e. inner product of two points in a Hilbert Space) to transform a non-linear classification to a linear classification and the utilization of a large margin classifier to separate points with different class labels. Large margin classifiers have low chance of over fitting and works efficiently in high dimensional feature spaces.

Support vector machines have been widely utilized in cheminformatics study. Traditional way of applying SVM to cheminformatics is to first create a set of features (or *descriptors* in many quantitative structure-properity relationship studies) and then use SVM to train a predictive model [6, 16]. Recently using graphs to model chemical structures and using data mining approach to obtain high quality, task-relevant features gain popularity in cheminformatics [16]. In this paper, we report a novel application graph wavelet analysis in creating high quality localized structure features for cheminformatics. Specifically, in our method, we model a chemical as a graph where a *node* represents an *atom* and an *edge* represents an *chemical bond* in the chemical. We leverage wavelet functions applied to graph-structured data in order to construct a graph kernel function. The wavelet functions are used to condense neighborhood information about an atom into a single feature of that atom, rather than features spread over it's neighboring atoms. By doing so, we extract (local) features with various topological scales about chemical structures and use these wavelet features to compute an alignment of two chemical graphs.

We have applied our graph kernel methods to several chemical structure-activity benchmarks. Our results indicate that our approaches yields performance profiles at least competitive with, and sometimes exceeding that of current state-of-the-art approaches. In addition, the identification of highly discriminative patterns for chemical activity classification provides evidence that our methods can make generalizations about chemical function given molecular structure. More over, our results also show that the use of wavelet functions significantly decreases the computational costs for graph kernel computation.

The rest of the paper is organized in the following way. Section 2 presents an overview of related work on quantitative chemical structure-property relationship study. In Section 3, we present background information about graph representation of chemical structures, graph database mining, and graph kernel function. Section 4 discusses the algorithmic details of our work, and in Section 5 we examine an empirical study of the proposed algorithm using several chemical structure benchmarks. We conclude with a short discussion of the pros and cons of our proposed methods.

2. RELATED WORK

A *target property* of the chemical compound is a measurable quantity of the compound. There are two categories of target properties: continuous (e.g., binding affinities to a protein) and discrete target properties (e.g. active compounds vs. inactive compounds).

The relationship between a chemical compound and its target property is typically investigated through a quantitative structure-property relationship (*QSPR*) [a]. Abstractly, any QSPR method may be generally defined as a function that maps a chemical space to a property space in the form of

$$P = \hat{k}(D) \tag{1}$$

where D is a chemical structure, P is a property, and the function $\hat{k}$ is an estimated mapping from a chemical space to a property space.

Different QSPR methodologies can be understood in terms of the types of target property values (continuous or discrete), types of features, and algorithms that map descriptors to target properties. Many classification methods has been applied to build QSPR models and recent ones include Decision Trees, Classification based on association [2], and

[a] Such study also known as a quantitative structure-activity relationship (QSAR) but *property* refers to a broader range of applications than activity.

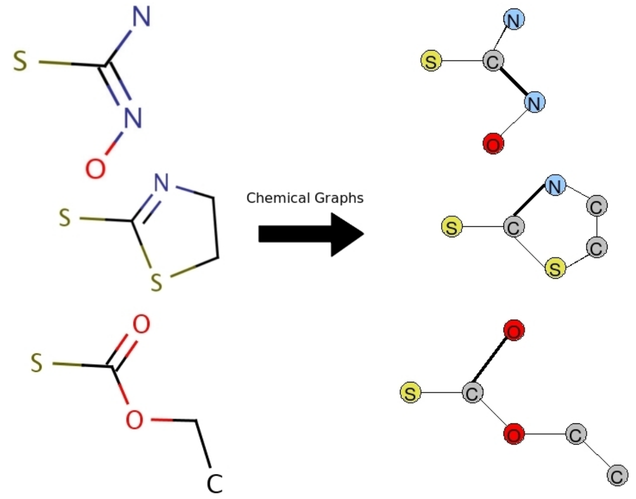

Fig. 1. Left: three sample chemical structures. Right: Graph representations of the three chemical structures.

Random Forest among many others. In our subsequent investigation, we focus on graph representation of chemical structures, graph wavelet analysis, and graph kernel methods that work well in high and even infinite dimensional feature space with low chance of over-fitting.

3. BACKGROUND

Before we proceed to algorithmic details, we present some general background regarding a computational analysis of chemical structure-property relationship which includes (i) a graph representation of chemical structures, (ii) graph kernel functions, and (iii) graph wavelet analysis.

3.1. Chemical Structure and Graph Modeling of Chemical Structures

Chemical compounds are organic molecules that are easily modeled by a graph representation. In our representation, we use *nodes* in a graph to model *atoms* in a chemical structure and *edges* in the graph to model chemical *bonds* in the chemical structure. In our representation, nodes are labeled with the atom element type, and edges are labeled with the bond type (single, double, and aromatic bond). The edges in the graph are undirected, since there is no directionality associated with chemical bonds. Figure 1 shows three sample chemical structures and their graph representation. Since we always use graphs to model chemical structures, in the following discussion, we make little distinction about graphs and chemicals, nodes and atoms, and edges and chemical bonds.

3.2. Graph Kernel Functions

The term *kernel function* in our context refers to an operation of computing the inner product between two objects (e.g. graphs) in a feature space, thus avoiding the explicit computation of coordinates in that feature space. Depends on the dimensionality of the feature space, we divide the current graph kernel function into two groups.

The first group works in a finite dimensional feature space [16]. Algorithms in the group first compute a set of features and performs subsequent classification in this feature space. Many existing application of machine learning algorithms to cheminformatics problems follow into this category.

The second group works in an infinite dimensional feature space. Example of this group include kernels that work on paths [12], on cyclic graphs [10]. The kernel computation in infinite dimensional feature space is usually challenging. To ease the prohibitive computational cost, Kashima et al[12] developed a Markov model to randomly generate walks of a labeled graph. The random walks are created using a transition probability matrix combined with a walk termination probability. These collections of random walks are then compared and the number of shared sequences is used to determine the overall similarity between two molecules. Recently frequent pattern based kernels are gaining popularity.

In this paper, we investigate a new way to create features of chemical graph structures. We also present an efficient computational way to compute such kernel called wavelet-alignment graph kernel. Our experimental study has demonstrated the efficiency and efficacy of our method.

3.3. Graph Wavelets Analysis

Wavelet functions are commonly used as a means for decomposing and representing a function or signal as its constituent parts, across various resolutions or scales. Wavelets are usually applied to numerically valued data such as communication signals or mathematical functions, as well as to some regularly structured numeric data such as matrices and images. Graphs, however, are arbitrarily structured and may represent innumerable relationships and topologies between data elements. Recent work has established the successful application of wavelet functions to graphs for multi-resolution analysis. Two examples of wavelet functions, the Haar and the mexican hat, are depicted in Figure 2.

Crovella et al. [4] have developed a multi-scale method for network traffic data analysis. For this application, they are attempting to determine the scale at which certain traffic phenomena occur. They represent traffic networks as graphs labeled with some measurement such as bytes carried per unit time. In their method, they use the *hop* distance between vertices in a graph, defined as the length of the shortest path between them, and apply a weighted average function to compute the difference between the average of measurements close to a vertex and measurements that are far, up to a certain distance. This process produces a new measurement for a specific vertex that captures and condenses information about the vertex neighborhood. Figure 3 shows a diagram of wavelet function weights overlayed on a chemical structure.

Maggioni et al. [13] demonstrate a general-purpose biorthogonal wavelet for graph analysis. In their method, they use the dyadic powers of an diffusion operator to induce a multiresolution analysis. While their method applies to a large class of spaces, such as manifolds and graphs, the applicability of their method to attributed chemical structures is not clear. The major technical difficulty is how to incorporate node labels in a multiresolution analysis.

4. ALGORITHM DESIGN

In the following sections we outline the algorithms that drive our experimental method. In short, we measure the similarity of graph structures whose nodes and edges have been labeled with various features. These features represent different kinds of chemical structure information including atoms and chemical bonds types among others. To compute the similarity of two graphs, the nodes of one graph are aligned with the nodes of the second graph, such that the total overall similarity is maximized with respect to all possible alignments. Vertex similarity is measured by comparing vertex descriptors, and is computed recursively so that when comparing two nodes, we also compare the immediate neighbors of those nodes, the neighbors of immediate neighbors, and so on so forth.

4.1. Graph Alignment Kernel

An *alignment* of two graphs G and G' (assuming $|V[G] \leq |V[G']|$) is a 1-1 mapping $\pi : V[G] \rightarrow V[G']$. Given an alignment π, we define the similarity between two graphs, as measured by a kernel function k_A, below:

$$\begin{aligned} k_A(G, G') := \max_\pi \textstyle\sum_{v \in V[G]} k_n(v, \pi(v)) + \\ \textstyle\sum_{u,v} k_e((u, v), (\pi(u), \pi(v))) \end{aligned} \tag{2}$$

The function k_n is a kernel function to measure the similarity of nodes and the function k_e is a kernel function to measure the similarity of edges. Intuitively in Equation 2 we use an additive model to compute the similarity between two graphs by computing the sum of the similarity of nodes and the similarity of edges. The maximal similarity among all possible alignments is defined as the similarity between two graphs.

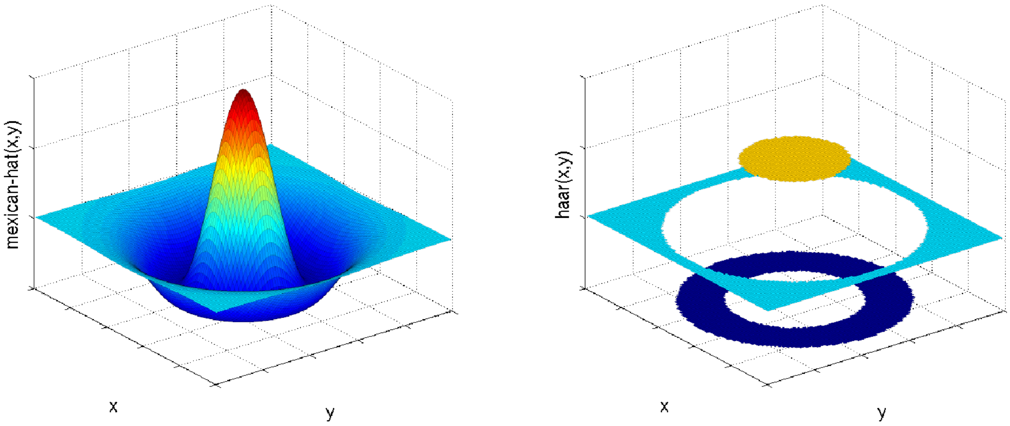

Fig. 2. Two examples of wavelet functions in 3 dimensions, the mexican hat on the left, and the Haar on the right.

4.2. Simplified Graph Alignment Kernel

A direct computation of the graph alignment kernel is computationally intensive and is unlikely scalable to large graphs. With no surprise, the graph alignment kernel computation is no easier than the subgraph isomorphism problem, a known NP-hard problem [b]. To derive efficient algorithm scalable to large graphs, we simplify the graph kernel function with the following formula:

$$k_M(G, G') = \max_{\pi} \sum_{v \in V[G]} k_a(f(v), f(\pi(v))) \qquad (3)$$

Where $\pi : V[G] \to V[G']$ denotes an alignment of graph G and G'. $f(v)$ is a set of features associated with a node that not only include node features but also include information about topology of the graph where v belongs to.

By equation 3, we are trying to compute a maximal weighted bipartite graph, which has an efficient solution known as the Hungarian algorithm. The complexities of the algorithm is $O(|V[G]|^3)$. See [7] for further details.

Below we provide an efficient method, based on graph wavelet analysis, to crate features to capture the topological structure of a graph.

4.3. Graph Wavelet Analysis

Originally proposed to analyze time series signals, wavelet analysis transforms a series of signals to a set of summaries with different scale. Two of the key insights of wavelet analysis of signals are (i) using localized basis functions and (ii) analysis with different scales. Wavelet analysis offers efficient tools to decompose and represent a function with arbitrary shape [5, 8]. Since invented, wavelet analysis has quickly gained popularity in a wide range of applications outside time series data, such as image analysis and geography data analysis. In all these applications, the level of detail, or *scale* is considered as an important factor in data comparison and compression. We show two examples of wavelet functions in a 3D space in Figure 2.

Our Intuition. With wavelet analysis as applied to graph represented chemical structure, for each atom, we may collect features about the atom and its local environment with different scales. For example, we may collect information about the average charge of an atom and atoms surrounding the atom and assign the average value as a feature to the atom. We may also collect information about whether an atom belongs to a nearby functional group, whether the surrounding atoms of a particular atom belong to a nearby functional group, and the local topology of

[b] Formally, we need to show a reduction from the graph alignment kernel to the subgraph isomorphism problem. The details of such reduction are omitted due to their loose connection to the main theme of the current paper, which is advanced data mining approach as applied to cheminformatics applications

an atom to its nearby functional groups.

In summary, conceptually we may gain the following two types of insights about the chemicals after applying wavelet analysis to graph represented chemical structure:

- *Analysis with varying levels of scale.* Intuitively, at the finest level, we compare two chemical structures by comparing the atoms and chemical bonds in the two structures. At the next level, we perform comparison of two regions (e.g. chemical functional groups) of two chemicals. At an even coarser level, small regions may be grouped into larger ones (e.g. pharmacophore), and we compare two chemicals by comparing the large regions and the connections among large regions.
- *Non-local connection.* In a chemical structure, two atoms that are not directly connected by a chemical bond may still have some kind of interaction. Therefore when comparing two graphs and their vertices, we cannot depend only on the *local environment* immediately surrounding an atom, but rather must consider both local and non-local environment.

Though conceptually appealing, current wavelet analysis is often limited to numerical data with regularly structures such as matrices and images. Graphs, however, are arbitrarily structured and may represent innumerable relationships and topologies between data elements. In order to define a reasonable graph wavelet functions, we have introduced the following two important concepts:

- h-hop neighborhood
- Discrete wavelet functions

The former, h-hop neighborhood, is essentially used to project graphs from a high dimensional space with arbitrary topology into a Euclidean space suitable for operation with wavelets. The h-hop measure defines a distance metric between vertices that is based on the shortest path between them. The discrete wavelet function then operates on a graph projection in the h-hop Euclidean space to compactly represent the information about the local topology of a graph. It is the use of this compact wavelet representation in vertex comparison that underlies the complexity reduction achieved by our method. Based on the h-hop neighborhood, we use a discrete wavelet function to summarize information in a local region of a graph and create features based on the summarization. These two concepts are discussed in detail below.

4.3.1. *h-Hop Neighborhood*

We introduce the following definitions.

Definition 4.1. Given a node v in a graph G the **h-hope neighborhood** of v, denoted by $N_h(v)$, is the set of nodes that are (according to the shortest path) exactly h hops away from v.

For example if $h = 0$, we have $N_0(v) = v$ and if $h = 1$, we have $N_1(v) = \{u|(u,v) \in E[G]\}$.

We use f_v denotes the feature vector associated with a node v in a graph G. $|f|$ is the feature vector length (number of features in the feature vector). The average feature measurement, denoted by $\overline{f}_j(v)$ for nodes in $N_j(v)$ is

$$\overline{f}_j(v) = \frac{1}{|N_j(v)|} \sum_{u \in N_j(v)} f_u \tag{4}$$

Example 4.1. The left part of the Figure 3 shows a chemical graph. Given a node v in the graph G, we label the shortest distance of nodes to v in the G. In this case $N_0(v) = v$ and $N_1(v) = \{t, u\}$. If our feature vector contains a single feature of atomic number, $\overline{f}_1(v)$ is the average atomic number of atoms that are at most 1-hop away from v. In our case, since $N_1(v) = \{t, u\}$ and $\{t, u\}$ are both carbon with atomic number equal to eight, then $\overline{f}_1(v)$ is equal to eight as well.

4.3.2. *Discrete Wavelet Functions*

In order to adapt a wavelet function to discrete structure such as graphs, we convert a wavelet function $\psi(x)$ to apply to the h-hop neighborhood. Towards that end, we scale a wavelet function $\psi(x)$ (such as the Haar, or Mexican Hat) to have support on the domain $[0, 1)$, with integral 0, and partition the function into $h+1$ intervals. We then compute the aver-

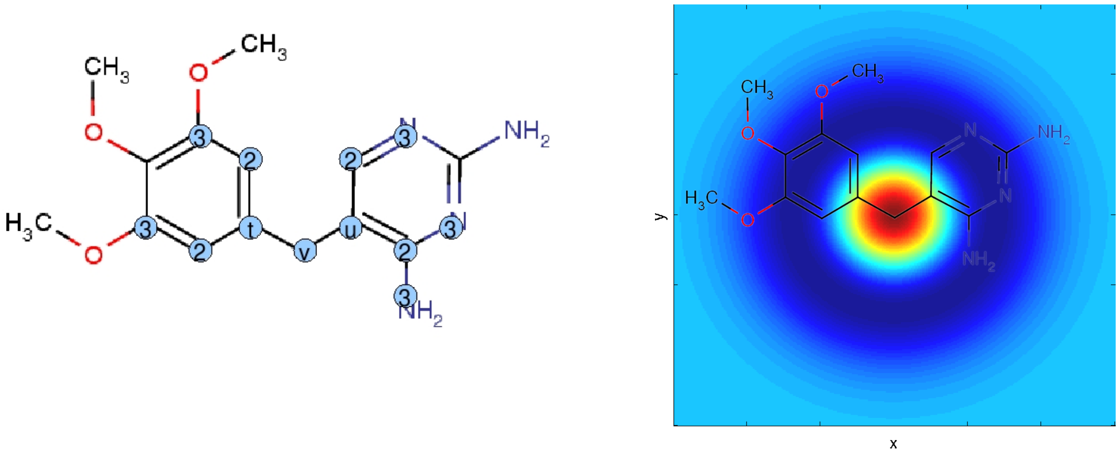

Fig. 3. Left: chemical graph centered on vertex v, with adjacent vertices t and u. Vertices more than one hop away are labeled with the hop number, up to hop distance three. Right: Superposition of a wavelet function on the chemical graph. Note here we can see the intensity of the wavelet function corresponds to the hop distance from the central vertex. Also this represents an idealized case where the hop distance between vertices corresponds roughly to their spatial distance. Unlabeled vertices correspond to carbon (C); hydrogens are shown without explicit bonds (edges).

age, $\psi_{j,h}$, as the average of $\psi(x)$ over the jth interval, $0 \le j \le h$ as below.

$$\psi_{j,h} \equiv \frac{1}{h+1} \int_{j/(h+1)}^{(j+1)/(h+1)} \psi(x)dx \qquad (5)$$

With neighborhood and discrete wavelet functions, we are ready to apply a wavelet analysis to graphs. We call our analysis *wavelet measurements*, denoted by $\Gamma_h(v)$, for a node v in a graph G at scale up to $h > 0$.

$$\Gamma_h(v) = C_{h,v} * \sum_{j=0}^{h} \psi_{j,h} * \overline{f}_j(v) \qquad (6)$$

where $C_{h,v}$ is a normalization factor with $C(h,v) = (\sum_{j=0}^{h} \frac{\psi_{j,h}^2}{|N_k(v)|})^{-1/2}$

We define $\Gamma^h(v)$ as the sequence of wavelet measurements as applied to a node v with scale value up to h. That is $\Gamma^h(v) = \{\Gamma_1(v), \Gamma_2(v), \ldots, \Gamma_h(v)\}$. We call $\Gamma^h(v)$ the *wavelet measurement vector* of node v. Finally we plug the wavelet measurement vector into the alignment kernel with the following formula.

$$k_\Gamma(G, G') = \max_\pi \sum_{v \in V[G]} k_a(\Gamma^h(v), \Gamma^h(\pi(v))) \qquad (7)$$

where $k_a(\Gamma^h(v), \Gamma^h(\pi(v))$ is a kernel function defined on vectors. Two popular choices are linear kernel and radius based function kernel.

Example 4.2. The right part of Figure 3 shows a chemical graph overlayed with a wavelet function centered on a specific vertex. We can see how the wavelet is most intense at the central vertex, hop distance of zero, corresponding to a strongly positive region of the wavelet function. As the hop distance increases the wavelet function becomes strongly negative, as we can see roughly at hop distances of one and two. At hop distance greater than two, the wavelet function returns to zero intensity, indicating negligible contribution from vertices at this distance.

5. EXPERIMENTAL STUDY

We have conducted classification experiments on five different biological activity data sets, and measured support vector machine (SVM) classifier prediction accuracy for several different feature generation methods. We describe the data sets and classification methods in more detail in the following subsections, along with the associated results. Figure 4 gives a graphical overview of the process.

We performed all of our experiments on a desktop computer with a 3Ghz Pertium 4 processor and

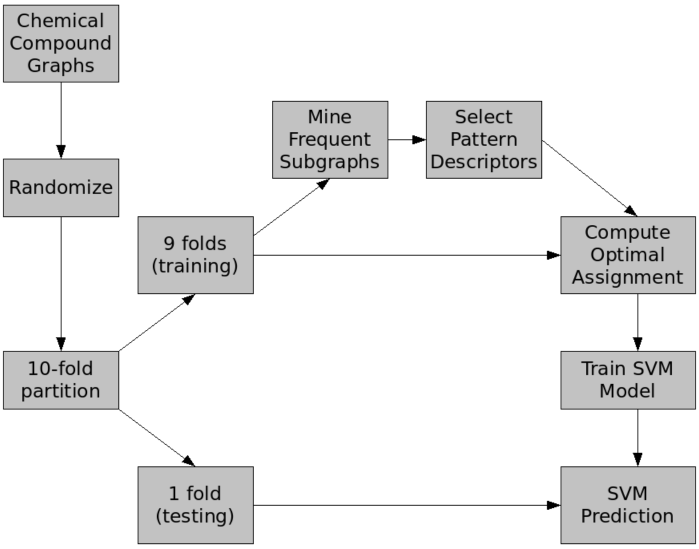

Fig. 4. Experimental workflow for a single cross-validation trial.

1 GB of RAM.

5.1. Data Sets

We have selected five data sets representing typical chemical benchmarks in drug design to evaluate our classifier performance. The Predictive Toxicology Challenge data set, discussed by Helma et al[9], contains a set of chemical compounds classified according to their toxicity in male rats (PTC-MR), female rats (PTC-FR), male mice (PTC-MM), and female mice (PTC-FM). The Human Intestinal Absorption (HIA) data set (Wessel et al.[19]) contains chemical compounds classified by intestinal absorption activity. The remaining data set (MD) is from Patterson et al[14], and was used to validate certain molecule descriptors. Various statistics for these data sets can be found in Table 1.

All of these data sets exist natively as binary classification problems, therefore in the case of the MD and HIA data sets, some preprocessing is required to transform them into regression and multi-class problems. For regression, this is a straightforward process of using the compound activity directly as the regression target. In the case of multi-class problems the transformation is not as direct. We chose to use a histogram of compound activity values to visualize which areas of the activity space are more dense, allowing natural and intuitive placement of class separation thresholds.

5.2. Methods

We evaluated the performance of the SVM classifier trained with different methods. The first two methods (WA-linear, WA-RBF) are both computed using our wavelet-alignment kernel, but use different functions for computing atom-atom similarity; we tested both a linear and RBF function here. In our experimental study, we experimentally evaluated different hop distance threshold and fixed $h = 3$ in all experimental study.

The method optimal alignment (OA) consists of the similarity values computed by the optimal assignment kernel, as proposed by Frölich et al[7]. There are several reasons that we consider OA as the current-state-of-the-art graph based chemical structure classification method. First, the OA method is developed specifically for chemical graph classification. Second the OA method contains a large library to compute different features for chemical structures. Third, the OA method has developed a sophisticated kernel function to compute the similarity between two chemical structures. Our experimental study shows that with the wavelet analysis we obtain performance profiles comparable to, and sometimes ex-

Table 1. Data set and class statistics.

Dataset	# Graphs	Class	Labels	Count
HIA	86	regression	0 - 100	86
		binary	0	39
			1	47
		multi-class	1	21
			2	18
			3	21
			4	26
MD	310	regression	0 - 7000	310
		binary	0	162
			1	148
		multi-class	1	46
			2	32
			3	37
			4	35
PTC-MR	344	binary	0	192
			1	152
PTC-MM	336	binary	0	207
			1	129
PTC-FR	351	binary	0	230
			1	121
PTC-FM	349	binary	0	206
			1	143

ceeding that of the existing state-of-the-art chemical classification approaches. In addition, we achieve a significant computational time reduction by using the wavelet analysis. The details of the experimental study are shown below.

In our experiments, we used the support vector machine (SVM) classifier in order to generate activity predictions. We used the LibSVM classifier implemented by Chang et al[3] as included in the Weka data-mining software package by Witten et al. [20]. The SVM parameters were fixed across all methods, and we use a linear kernel. For (binary) classification we used nu-SVC for multi-class classification with nu = 0.5. We used the Haar wavelet function in our WA experiments. Classifier performance was averaged over a 10-fold cross-validation set.

We developed and tested most of our algorithms under the MATLAB programming environment. The OA software was provided by [7] as part of their JOELib software, a computational chemistry library implemented in java. [11]

5.3. Results

Below we report our experimental study of the wavelet-alignment kernel with two focuses: (i) classification accuracy and (ii) computational efficiency.

5.3.1. *Classification Accuracy*

Table 2 reports the average and standard deviation of the prediction results over 10 trials. For classification problems, results are in prediction accuracy, and for regression problems they are in mean squared error (MSE) per sample. From the table, we observe that for the HIA data set, WA-RBF kernel significantly outperforms OA for both binary and multi-class classification. For MD data set, OA does best for both classification sets, but WA-linear is best for regression. For the PTC binary data, the WA-linear method outperforms the others in 3 of the 4 sets.

5.3.2. *Computational Efficiency*

In Table 3, we document the kernel computation time for both OA and WA methods using 6 different data sets. The runtime advantage of our WA algorithm over OA is clear, showing improved computation efficiency by factors of over 10 fold for the WA-linear kernel and over 5 fold for the WA-RBF kernel.

Figure 5 shows the kernel computation time across a range of dataset sizes, with chemical compounds sampled from the HIA data set. Using simple random sampling with replacement, we create data sets sized from 50 to 500. We did not try to run OA

Table 2. Prediction results of cross-validation experiments, averaged 10 randomized trials, with standard deviation in parentheses. For regression data sets labeled with real values, result is mean squared error (lower is better); for classification the result is prediction accuracy (higher is better). The best result for each data and label set is marked with an asterisk.

Dataset	Labels	OA	WA-RBF	WA-linear
HIA	real	979.82(32.48)*	989.72(33.60)	989.31(24.62)
	binary	51.86(3.73)	61.39(2.77)*	57.67(3.54)
	multi-class	29.30(2.23)	39.06(0.63)*	29.76(5.73)
MD	real	3436395(1280)	3436214(1209)*	3440415(1510)
	binary	67.16(0.86)*	52.51(3.34)	65.41(0.42)
	multi-class	39.54(1.65)*	33.35(3.83)	33.93(1.87)
PTC-FM	binary	58.56(1.53)*	51.46(3.45)	55.81(1.31)
PTC-FR	binary	58.57(2.11)	52.87(2.65)	59.31(1.95)*
PTC-MM	binary	58.23(1.25)	52.36(0.93)	58.91(2.078)*
PTC-MR	binary	51.51(1.20)	52.38(3.48)	52.09(2.61)*

Table 3. Running time for the computation of OA, WA-linear, and WA-RBF) kernels, in seconds. Speedup is computed as the ratio between the OA processing time and that of WA.

Dataset	Kernel	Time	Speedup
HIA	OA	75.87	-
	WA-RBF	13.76	5.51
	WA-linear	4.91	15.45
MD	OA	350.58	-
	WA-RBF	50.85	6.89
	WA-linear	26.82	13.07
PTC-FM	OA	633.13	-
	WA-RBF	103.95	6.09
	WA-linear	44.87	14.11
PTC-FR	OA	665.95	-
	WA-RBF	116.89	5.68
	WA-linear	54.64	12.17
PTC-MM	OA	550.41	-
	WA-RBF	99.39	5.53
	WA-linear	47.51	11.57
PTC-MR	OA	586.12	-
	WA-RBF	101.68	5.80
	WA-linear	45.93	12.73

on even larger data set since the experimental results clearly demonstrate the efficiency of the WA kernel already.

What these run time results do not demonstrate is the *even greater* computational efficiency afforded by our WA algorithm when operating on general, non-chemical graph data. As noted at the end of section four, chemical graphs have some restrictions on their general structure. Specifically, the number of atom neighbors is bound by a small constant (4 or so). Since the OA computation time is much more dependent on the number of neighbors, we can see that WA is even more advantageous in these circumstances. Unfortunately, since the OA software is designed as part of the JOELib chemoinformatics library specifically for use with chemical graphs, it will not accept generalized graphs as input, and hence we could not empirically demonstrate this aspect of our algorithm.

6. CONCLUSIONS

Graph structures are a powerful and expressive representation for chemical compounds. In this paper we present a new method *wavelet-assignment*, for computing the similarity of chemical compounds, based on the use of an optimal assignment graph

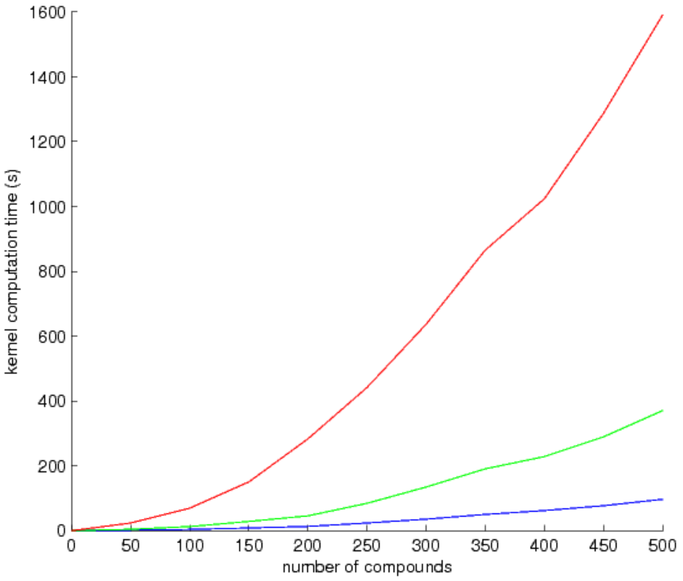

Fig. 5. Comparison of computation times between OA (top line), WA-RBF (middle line) and WA-linear (bottom line) kernels.

kernel function augmented with pattern and wavelet based descriptors. Our experimental study demonstrates that our wavelet-based method deliver an improved classification model, along with an order of magnitude speedup in kernel computation time. For high-volume, real world data sets, this algorithm is able to handle a much greater number of graph objects, demonstrating it's potential for processing both chemical and non-chemical data in large amounts. In our present study, we only used limited number of atom features. In the future, we plan to involve domain experts to evaluate the performance of our algorithms, including the prediction accuracy and the capability for identifying important features in diverse chemical structure data sets.

Acknowledgments

This work has been supported by the Kansas IDeA Network for Biomedical Research Excellence (NIH/NCRR award #P20 RR016475), the KU Center of Excellence for Chemical Methodology and Library Development (NIH/NIGM award #P50 GM069663), and NIH grant #R01 GM868665.

References

1. CP Austin, LS Brady, TR Insel, and FS Collins. Nih molec-
ular libraries initiative. *Science*, 306(5699):1138–9, 2004.
2. Yiming Ma Bing Liu, Wynne Hsu. Integrating classification and association rule mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 1998.
3. C. Chang and C. Lin. Libsvm: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.
4. M. Crovella and E. Kolaczyk. Graph wavelets for spatial traffic analysis. *Infocom*, 3:1848–1857, 2003.
5. Antonios Deligiannakis and Nick Roussopoulos. Extended wavelets for multiple measures. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, 2003.
6. M. Deshpande, M. Kuramochi, and G. Karypis. Frequent sub-structure-based approaches for classifying chemical compounds. *IEEE Transactions on Knowledge and Data Engineering*, 2005.
7. Fröohlich, J. Wegner, F. Sieker, and A. Zell. Kernel functions for attributed molecular graphs - a new similarity-based approach to adme prediction in classification. QSAR & Combinatorial Science, 2006.
8. M. Garofalakis and Amit Kumar. Wavelet synopses for general error metrics. *ACM Transactions on Database Systems (TODS)*, 30(4):888–928, 2005.
9. C. Helma, R. King, and S. Kramer. The predictive toxicology challenge 2000-2001. *Bioinformatics*,

17(1):107–108, 2001.

10. Tamas Horvath, Thomas Gartner, and Stefan Wrobel. Cyclic pattern kernels for predictive graph mining. *SIGKDD*, 2004.
11. Jun Huan, Wei Wang, and Jan Prins. Efficient mining of frequent subgraph in the presence of isomorphism. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM)*, pages 549–552, 2003.
12. H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *Proc. of the Twentieth Int. Conf. on Machine Learning (ICML)*, 2003.
13. M. Maggioni, J. Bremer Jr, R. Coifman, and A. Szlam. Biorthogonal diffusion wavelets for multiscale representations on manifolds and graphs. In *Proc. SPIE Wavelet XI*, volume 5914, 2005.
14. D. Patterson, R. Cramer, A. Ferguson, R. Clark, and L. Weinberger. *Neighbourhood Behaviour: A Useful Concept for Validation of "Molecular Diversity" Descriptors*, 39:3049–3059, 1996.
15. Put R, Xu QS, Massart DL, and Vander Heyden Y. Multivariate adaptive regression splines (mars) in chromatographic quantitative structure-retention relationship studies. *J Chromatogr A.*, 1055(1-2), 2004.
16. Aaron Smalter, Jun Huan, and Gerald Lushington. Structure-based pattern mining for chemical compound classification. *Proceedings of the 6th Asia Pacific Bioinformatics Conference*, 2008.
17. V. Svetnik, C. Tong A. Liaw, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43, 2003.
18. Nicola Tolliday, Paul A. Clemons, Paul Ferraiolo, Angela N. Koehler, Timothy A. Lewis, Xiaohua Li, Stuart L. Schreiber, Daniela S. Gerhard, and Scott Eliasof. Small molecules, big players: the national cancer institute's initiative for chemical genetics. *Cancer Research*, 66:8935–42, 2006.
19. M. Wessel, P. Jurs, J. Tolan, and S. Muskal. Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.*, 38(4):726–735, 1998.
20. I. Witten and E. Frank. Morgan Kaufmann, San Francisco, CA, 2nd edition edition, 2005.
21. Y. Xue, H. Li, C. Y. Ung, C. W. Yap, and Y. Z. Chen. Classification of a diverse set of tetrahymena pyriformis toxicity chemical compounds from molecular descriptors by statistical learning methods. *Chem. Res. Toxicol.*, 19 (8), 2006.