Modeling Biological Activities of Chemical Compounds: Kernel Methods for Structure Activity Relationship

Jean-Luc Perret[1] luc@kuicr.kyoto-u.ac.jp
Yoshinobu Igarashi[1] igarashi@kuicr.kyoto-u.ac.jp
Minoru Kanehisa[1] kanehisa@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Keywords: chemoinformatics, SAR, molecule similarity, ADMETOX, kernel methods, graph kernel, SVM, KPCA

1 Introduction

The function of enzymes as well as the function of proteins involved in regulatory pathways often implies interactions with small chemical compounds. To understand the function of these proteins as well as for applications such as predicting activity or adverse effects of potential drugs we try here to compute the similarity between chemical compounds using a new similarity function based on the 2D structure of chemical compounds.

Small chemical compounds can be represented as undirected labeled graphs (2D structures). Although not perfect, this model representation of molecules has been shown to be related to the biological activity of chemical compounds[3].

Traditionally molecular similarity was computed in two different ways. One way is to find maximal of frequent common subgraphs[1]. This approached generally suffer from its computational complexity. Another approach was to transform graphs into vectors of molecular descriptors and then compare these vectors. This requires the expert choice of adequate molecular descriptors[3].

Here we explore another way to compare chemical compounds using the recently published kernel function for graphs by Kashima et al.[2]. This algorithm allows the direct computation of a similarity value between two graphs as a weighted sum of all common paths (sequences of atoms and bonds) in the two graphs. The path weights are choosen given a random walk probability model on a graph representation of molecules were undirected bonds are replaced by bidirectional edges. The kernel does not require an explicit vectorial representation of molecules. Kernel functions are particularly interesting because they allow the direct computation of multivariate statistical methods like kernel principal component analysis (KPCA), and support vector machines (SVM) on the similarity matrices.

Here we apply the graph kernel to two chemoinformatic tasks. Supervised learning for predicting an adverse effect of chemical compounds, and unsupervised learning for compound classification.

2 Method and Results

2.1 Graph Kernel

We implemented the graph kernel proposed by Kashima et al.[2] as a C++ library. To evaluate the appropriateness of the graph kernel similarity value, we tried to use it to predict the mutagenicity of a benchmark dataset of chemical compounds.

2.2 Supervised learning

We trained a Support Vector Machine to recognise mutagenic compounds among 230 aromatic and heteroaromatic nitro compounds. Prediction accuracy (estimated by leave-one-out error) reached 90% for homogeneous subsets of compounds, and compares favorably with state of the art methods using molecular descriptors. Similarly to these methods however, prediction accuracy diminished as the diversity of compounds in the dataset increased, indicating that active molecules might not be completely linearly separable in the feature space implicitely defined by the kernel function.

2.3 Classification

We tried to classify the dataset for mutagenicity using hierarchical clustering and projected the results on the first two axis of a kernel principal component analysis. The results demonstrate that mutagenic compounds belong to several clusters. Such classification might help resolve the mode of action of compounds, or help select diverse molecules for biological screening.

It was shown by toxicologists that the number of aromatic rings is a key feature for mutagenicity. This parameter is set as an additional categorical descriptor in the vectorial representation of compounds, and helps classify compounds. As the kernel does not use specific ring information, we verified if the graph kernel is able to distinguish compounds based on the number of aromatic rings. To do so we classified compounds with an identical number of heavy atoms (non hydrogens) but with 1, 2 or 3 aromatic rings. The resulting classification succeeds in grouping compounds according to the number of aromatic rings.

3 Discussion

The kernel on graphs used here is a new measurement of similarity between chemical compounds. The results obtained here indicate that it is useful for predicting the mutagenicity of compounds. These results demonstrate that the graph kernel function is a useful new tool for bio and chemoinformatics. Further developments on the kernel function for graphs are possible and may allow a better separation of compounds for their biological activity.

References

[1] Hattori M., Okuno Y., Goto S. and Kanehisa M. Development of a Chemical Structure Comparison Method for Integrated Analysis of Chemical and Genomic Information in the Metabolic Pathways *J. Am. Chem. Soc.*, 125:11853-11865, 2003.

[2] Kashima H., Tsuda K. and Inokuchi A. Marginalized Kernels Between Labeled Graphs *Proc. 20th Int. Conf. Machine Learn.*, 2003.

[3] Martin Y.C., Kofron J.L. and Traphagen L.M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.*, 45:4350-4358, 2002.