IEEE Computer Society 2006, ISBN 0-7695-2702-7, pp. 136-141. (http://dx.doi.org/10.1109/ICDMW.2006.51)

# Context-Aware Visual Exploration of Molecular Databases

Giuseppe Di Fatta[1] Antonino Fiannaca[2] Riccardo Rizzo[2] Alfonso Urso[2]
Michael R. Berthold[3] Salvatore Gaglio[24]

[1]School of Systems Engineering, The University of Reading
Whiteknights, Reading, Berkshire, RG6 6AY, United Kingdom

[2]ICAR-CNR, National Research Council, 90128 Palermo, Italy

[3]Department of Computer and Information Science, The University of Konstanz,
78457 Konstanz, Germany

[4]DINFO, The University of Palermo, 90128 Palermo, Italy

## Abstract

*Facilitating the visual exploration of scientific data has received increasing attention in the past decade or so. Especially in life science related application areas the amount of available data has grown at a breath taking pace. In this paper we describe an approach that allows for visual inspection of large collections of molecular compounds. In contrast to classical visualizations of such spaces we incorporate a specific focus of analysis, for example the outcome of a biological experiment such as high throughout screening results. The presented method uses this experimental data to select molecular fragments of the underlying molecules that have interesting properties and uses the resulting space to generate a two dimensional map based on a singular value decomposition algorithm and a self-organizing map. Experiments on real datasets show that the resulting visual landscape groups molecules of similar chemical properties in densely connected regions.*

## 1. Introduction

A crucial step in drug discovery remains the so-called High Throughput Screening (HTS) and the subsequent analysis of the generated data. In this screening, hundreds of thousands of potential drug candidates are automatically tested for a desired activity, such as blocking a specific binding site or attachment to a particular protein. This activity is believed to be connected to, for example, the inhibition of a specific disease. Once all these compounds have been automatically screened, a large amount of data have to be analyzed and explored in order to select a few hundred promising candidates for further, more careful and cost-intensive analysis. This step is critical for the success of the entire drug discovery process.

Recent approaches based on data mining techniques focus on the analysis of the molecular structure and the extraction of pieces of these molecules that are correlated with activity. Such fragments can be used to directly identify groups of promising molecules (clustering). They can also be used to predict activity in other compounds (classification) [6] and to guide the synthesis of new ones.

The number of relevant molecular fragments is often very large and they cannot be directly visualized nor exhaustively explored by biochemists. However, these fragments can be used to identify groups of molecules with similar characteristics. In this context, visualization and indexing techniques for large data spaces can provide a powerful tool for the overall HTS analysis process.

The present work describes a data mining process to generate a similarity map of molecular compounds. The distance and relative position of compounds in the map are representative of the similarity of their chemical structure. The generation and selection of the appropriate features is critical for the quality of the map. We, then, apply a dimensional scaling to the feature space and a subsequent projection in a 2D gray scale bitmap.

The unsupervised learning process has been applied to a well-known set of real molecular compounds, the NCI HIV screen dataset. The resulting map has produced distinct clusters of similar compounds. In order to verify the quality of the map we have exploited a publicly available classification of a small subset of the compounds.

The proposed approach can be used to explore a large set of molecules. The user (a biochemist) can interact with the map by selecting and visualizing neighboring compounds. The proposed technique can be used to generate a molecular compounds atlas from datasets with known chemical properties and classes. It can provide an overall view of the known compounds and an unknown compound can be projected on the map to have a first hint on its chemical class and properties.

In the next section, we briefly discuss some related works in the field of the visual exploration of scientific data and, in particular, of biochemical data. In section 3, we introduce the overall framework for the generation of maps of molecular compounds. Then, we present the components of the proposed framework, which are the frequent subgraph mining algorithm, the multi-dimensional scaling technique, and the self-organizing maps, respectively, in sections 4, 5, and 6. In section 7, we present and discuss the experimental results. Finally, we provide conclusive remarks.

## 2. Related Work

Studies in Quantitative Structure-Activity Relationships (QSAR) focus on the determination of proper features to analyze and classify molecular compounds. This is based on the assumption that structurally similar compounds will have similar physico-chemical properties, such as molecular weight, dipole moment, energies and other electronic and spectroscopic properties. For example, new compounds are selected for the synthesis when they maximize the presence of functional groups and desired features.

In our approach we do not use any of these features; we take directly the topological structure of the molecules into account. A molecular compound can be represented as an attributed graph and chemical substructures can be used to discriminate among different classes. We directly adopt the frequent subgraphs as features for our knowledge discovery process. While this approach seems quite straightforward, its computational complexity and the very large number of frequent fragments have limited its adoption so far.

Self-Organizing Maps have been used extensively in chemistry [8, 15] and biology applications, and in particular in the field of QSAR [3, 19, 22]. These studies have adopted the SOM to perform the QSAR analysis; while in other cases (e.g., [7]), the SOM have been used to select the best subset of features to carry out a subsequent QSAR analysis.

In some application related to the gene expression clusterization and visualization ([18, 21]) the SOM neural network was used as a tool for clustering and visualization in order to obtain an *"executive summary"* of a massive gene expression data set. In [18] the U-Matrix tool [23] for SOM visualization was used. In other works similar neural maps derived from the original SOM algorithm were used, e.g. the DSOM model in [26].

In [17] an online service provides a self-organizing map of 65 x 50 clusters of the compounds tested in the NCI anti-HIV Screen. The map contains about 42,000 AIDS-screened compounds, both active and inactive, clustered by structure similarity.

## 3. Framework for the generation of molecular compounds maps

The overall process of knowledge discovery for the visual exploration of molecular compounds is shown in Figure 1. Input data, undirected labeled graphs representing the molecular compounds, are first processed to extract and select features that might be relevant to the application domain. In our approach we adopt the frequent molecular fragments (subgraphs) as features of the molecules because they might be directly correlated with the activity and can be used to directly identify groups of similar compounds.

However, even for relatively small datasets the set of all frequent fragments is enormous: a single molecule of average size can already contain in the order of hundreds of thousands of different fragments. Hence, we adopt the closed frequent subgraphs which carry equivalent information more efficiently than the frequent ones.

Nevertheless, the set of features is still very large and defines a high-dimensional space of the input data, making the analysis very difficult. Techniques from statistics, machine learning and data mining can be applied to provide a model of the data that can be more conveniently interpreted by the user. We adopt a multi-dimensional scaling technique, the Singular Value Decomposition (SVD), to cope with such a problem. Our ultimate goal is a visual summary of the set of compounds that provides information on their similarity. Thus, the dimensionally-reduced space is further projected into a two-dimensional gray scale bitmap by means of the Self-Organizing Maps.

In the next section, we introduce the subgraph mining step, which determines the high-dimensional feature space.

## 4. Frequent subgraph mining

The problem of selecting molecular fragments in a set of molecules can be formulated in terms of Frequent Subgraph Mining (FSM) in a set of graphs, in analogy to the Association Rule Mining (ARM) problem [1, 29]. While in ARM the main structure of the data is a list of items and the basic operation is the subset test, FSM relies on graph and subgraph isomorphism.

Molecules are represented by attributed graphs, in which each vertex represents an atom and each edge a bond between atoms. Each vertex carries attributes that indicate the

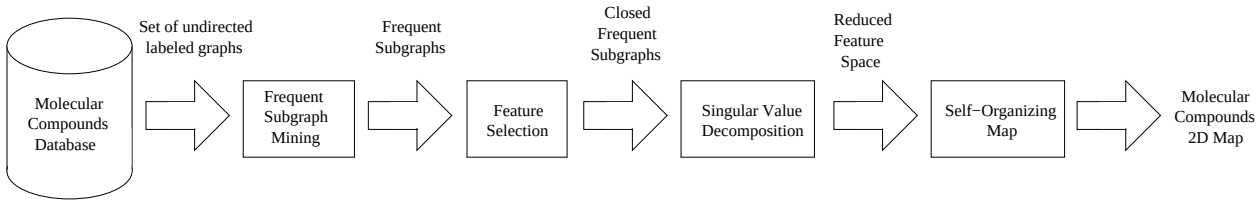



**Figure 1. Framework for the visual analysis of molecular compounds**

atom type (i.e., the chemical element), a possible charge, and whether it is part of an aromatic ring. Each edge carries an attribute that indicates the bond type (single, double, triple, or aromatic). Frequent molecular fragments are subgraphs that have a certain minimum support in a given set of graphs, i.e., are part of at least a certain percentage ($minSupp$) of the molecules. These topological fragments carry important information and may be representative of those functional components in the compounds that are responsible for a positive behavior. However, the high number of frequent fragments that can be found in a large dataset suggests the adoption of the closed frequent subgraphs for a more efficient definition of the feature space. A closed frequent subgraph is a frequent subgraph whose support is higher than the support of all its proper supergraphs.

Existing subgraph mining methods attempt to implicitly organize the space of all possible subgraphs in a lattice, which models subgraph relationships. The search then reduces to traversing this lattice, according to either a depth-first or a breadth-first strategy, and reporting all subgraphs that fulfill the desired criteria (feature selection). A number of approaches to find frequent molecular fragments have been published [4, 5, 13, 28]. The algorithm [4] that we have adopted, is based on a depth-first search strategy with advanced pruning techniques in order to reduce the generation of redundant search nodes.

## 5. Multi-dimensional scaling

The *feature selection* phase described in the previous section suffers the *curse of dimensionality* problem [2]. This drawback can be solved using a factorial analysis technique, in order to reduce the multidimensional vector space. The multidimensional input pattern set can be represented by a reduced number of *latent factors* that are able to describe the information contained in the original set. Given *m* compounds and *n* frequent fragments, it is possible to define a matrix *A*(*m*,*n*) where the element (*i, j*) is set to 1 when fragment *j* is contained in compound *i*. The *Singular Value Decomposition* (SVD) [9] of matrix *A* defines a mapping between the *m* compounds and *n* frequent fragments into a vector space *S*, in which both compounds and fragments are represented. The SVD is defined by

$$A = U\Sigma V^T, \tag{1}$$

where *U*(*m*,*n*) and *V*(*n*,*n*) are orthonormal matrices and $\Sigma$(*n*,*n*) is a diagonal matrix whose elements, $(\sigma_1, \sigma_2, ..., \sigma_n)$ with $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n \geq 0$, are called the *singular values* of *A*. The columns of *U* and the columns of *V* are called, respectively, *left* and *right singular vectors* of *A*.
The vector space reduction is performed using the *Truncated Singular Value Decomposition* (TSVD) [10]. Given a positive integer *r* with $r < n$, the TSVD is defined by

$$\widetilde{A} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T \tag{2}$$

where the matrix $\widetilde{U}$ is obtained from the matrix *U* by suppressing the last $n-r$ columns, the matrix $\widetilde{V}$ is obtained from the matrix *V* by suppressing the last $n - r$ columns, and matrix $\widetilde{\Sigma}$ is obtained from matrix $\Sigma$ by suppressing the last $n - r$ rows and the last $n - r$ columns. The molecular compounds represented in the reduced order vector space, are used for training the Self-Organizing Map, as described in the next section.

## 6. Self-Organizing Maps for Clustering and Visualization

Self-Organizing Maps [20] are neural networks inspired by some structure in biological brain. These neural structures (both biological and artificial) are capable of building maps of the input data, input stimuli in biological case, that preserve their neighborhood relationships. These maps can be used to build a visualization tool that allows to identify clusters or structures in input data [11].

The SOM algorithm principle defines a sort of elastic lattice of simple processing units that are organized in order to fit a set of points in a high-dimensional input space and to approximate their density function. For details of the learning algorithm we refer to [20].

The main application of the SOM is the visualization of high-dimensional data in a two dimensional manner and the creation of abstractions like in many clustering techniques [12]. After the learning stage, the inputs can be associated to the nearest network unit. When the surface is visualized, the inputs are distributed as landmarks on the map.

### 6.1 Map dimensions

The number of units in a SOM map can widely vary depending on different constrains. Usually it is related to the number of the available training patterns, according to some heuristics (e.g. the one used in the Matlab SOMToolbox [25]), it can also be related to the specific clustering problem or to the visualization needs.

In [24] the authors distinguish the *k-means SOMs* (the ones that have a number of units that is a fraction of the number of training patterns) and the Emergent SOM, were the map has a large number of neural units, that is not related to the clustering problem. The k-means SOMs are the large majority and widely used, but the use of a "large" SOM map can produce interesting results. Using a large SOM map it is possible to obtain a fine grain coverage of the input manifold, of course many of the units will be dead-units (i.e. they are not labeled with any input pattern) but they cover the space between clusters and, when the pattern clusters are isolated, constitutes a sort of "don't know" area for patterns that are added after the learning stage.

Large SOM maps can be visualized as images using the U-Matrix technique explained below.

### 6.2 The U-MATRIX visualization

The U-Matrix is a standard method for SOM visualization and is suitable for the visualization of large SOM maps [23]. The U-Matrix is a gray scale image matrix that is build "over" the two-dimensional SOM lattice and has the same dimensions. The gray level of the $g_i$ element of the U-Matrix is calculated using this equation:

$$g_i = \sum_{j \in N_i} d(w_i, w_j), \tag{3}$$

where $N_i$ is the set of the neighborhood units of the unit $i$ and $d(w_i, w_j)$ is the distance between the unit $i$ and the unit $j$ in the input space.

## 7. Experimental results

The validation of the proposed framework for the visual analysis of molecular compounds has been carried out on a set of real molecular compounds. We used the publicly available DTP AIDS Antiviral Screen dataset [16] of the National Cancer Institute. The screen measures the protection of human CEM cells from HIV-1 infection [27]. In particular, we used the 325 compounds belonging to the confirmed active (CA) class.

The feature vectors are obtained by means of the frequent subgraph mining process described in section 4 with a minimum support of 10%. The feature generation and selection phase obtains 640 molecular fragments. Consequently, each of the 325 compounds is represented by means of a vector of 640 binary components.

In order to obtain a data model that can be better interpreted and exploited, the multidimensional scaling technique described in section 5 has been applied. The original vector space has been reduced, using the *Truncated Singular Value Decomposition*, into the vector space defined by means of the most significant 20 singular values over the total 640.

The SOM used for visualization is a $100 \times 100$ square lattice. The training phase according to [20] is split into gross learning with $\alpha_i = 0.3$ $\alpha_f = 0.001$ and $\sigma_i = 20$, $\sigma_f = 1$ and the refinement with $\alpha_i = 0.1$ $\alpha_f = 0.001$ and $\sigma_i = 5$, $\sigma_f = 1$.

The overall data pipeline has been implemented in the data mining workflow management system KNIME [14]. The user can configure and execute the workflow, which produces an interactive map for the exploration and analysis of the set of compounds. The result is shown in Figure 2 using the U-Matrix representation.

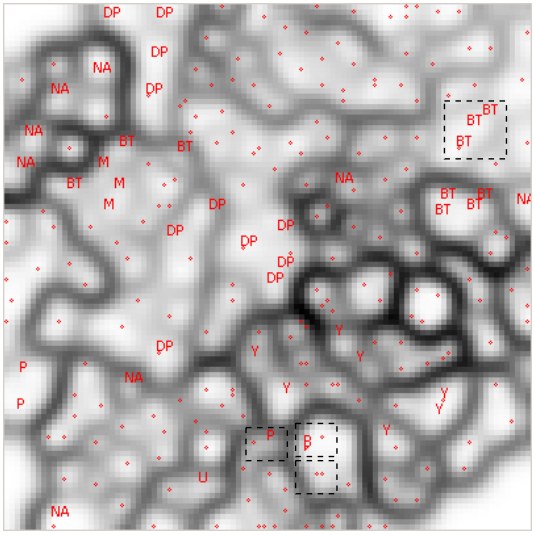
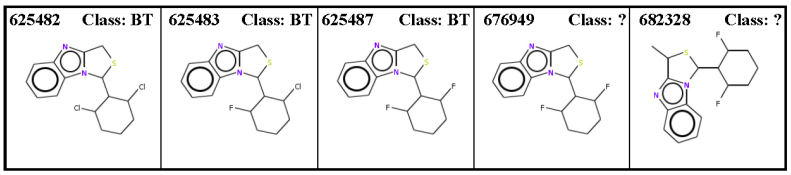


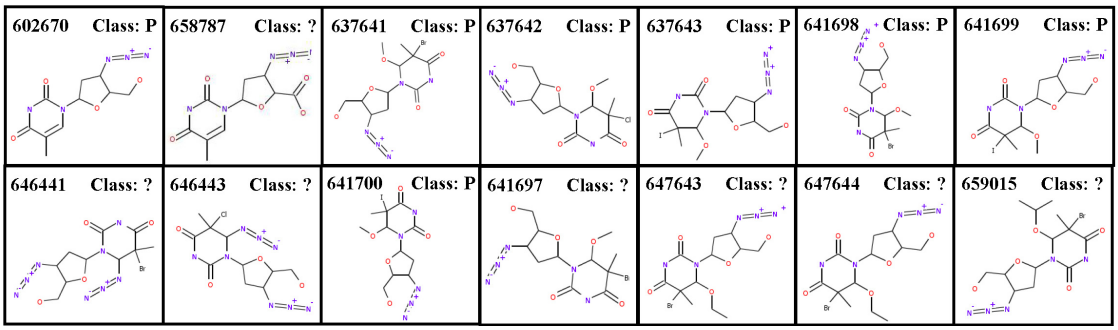

**Figure 2. UMatrix representation of 325 HIV-active molecular compounds**

A partial classification of the active compounds in the dataset is available at [16]. The classification identifies 7 chemical groups (table 1) for only 57 compounds of the 325 used in the training. In order to verify the correctness of the unsupervised learning process, we have marked those 57 compounds with their known class (Figure 2). Several clusters of compounds with the same label can be easily identified.

Finally, we have selected different areas and verified that similar compounds lie close on the map. In the top right




(a) 3 of 5 compounds have class label $BT$




(b) 7 of 14 compounds have class label $P$

**Figure 3. Two neighborhoods of molecular compounds**

area of Figure 2, three distinct labels BT appear. In the area of their close neighborhood, a total of 5 compounds are located (Figure 3(a)). Three correspond to compounds classified as BT, while two have no a priori classification. It should be noticed that the compounds 625487 and 676949 overlap in the map and actually have the same chemical structure. Apparently, the same compound has been stored in the NCI database with two different NSC numbers.

In the bottom center-right area of Figure 2, three labels $P$ appear. In the their neighborhood, there is a total of 14 compounds (Figure 3(b)), some of which overlap. Only 7 of the 14 compounds have been assigned to the class of Azido Pyrimidines (label P). The others have no a priori classification, while it is evident from their chemical structure that they do belong to this class. In particular, the compound 602670 is the Zidovudine, a.k.a. the azidothymidine (AZT) group, which is a well known antiretroviral drug, the first one approved for treatment of HIV.

## 8. Conclusions

We have presented an architecture for the visual exploration of molecular compounds. The overall knowledge discovery process is based on the generation of features, multidimensional scaling and self-organizing maps to generate a visual and interactive representation of the feature space and the similarity among molecules.

Two main underlying choices determine the focus of the analysis that can be performed with such maps, namely the set of the input molecules and the set of features. Different criterions can be adopted to select the features for a given dataset, which might be representative of different chemical classes (e.g., functional groups) or other properties, e.g. tested activity for the inhibition of different diseases. In our tests, we have adopted the closed frequent subgraphs for their topological information that is directly connected to chemical properties.

SOM and UMatrix visualization allow inspecting the relationships among the molecules by means of both inter and intra-cluster distances. This is exactly in the aim of building an atlas of drug candidates.

We have applied the method to the set of active compounds of the NCI AIDS antiviral screen dataset. A qualitative performance analysis has confirmed the validity of the proposed approach. Topologically similar compounds appear close on the map, while different ones are clearly separated by dark borders.

In future research activity we intend to perform a quantitative analysis of the results and a comparison with other techniques for the projection and the visualization of high-dimensional spaces. Moreover, clustering techniques can be applied to the generated 2D map and directly to the feature space. Further research efforts will also focus on the comparison and integration of different analysis contexts.

## References

[1] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. *Proc. of*

**Table 1. Compounds classification used to test the SOM training**

| Azido Pyrimidines (P) | Natural Products or Antibiotics (NA) | Benzodiazepines, Thazolobenzimidazoles and related compounds (BT) | Pyrimidine Nucleosides (Y) | Dyes and Polyanions (DP) | Heavy Metal Compounds (M) | Purine Nucleosides (U) |
|---|---|---|---|---|---|---|
| 10 compounds | 8 compounds | 10 compounds | 11 compounds | 13 compounds | 3 compounds | 2 compounds |


*the 1993 ACM SIGMOD Int. Conf. on Management of Data*, pages 207–216, May 26–28, 1993.

[2] R. Bellman. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ, 1961.

[3] Bienfait, B. Applications of high reolution self organizing maps to retrosynthetic and qsar analysis. *J. Chem. Inf. Comput. Sci. 34*, pages 890–898, 1994.

[4] C. Borgelt and M. R. Berthold. Mining molecular fragments: Finding relevant substructures of molecules. *Proc. of the IEEE Int. Conf. on Data Mining (ICDM 2002)*, pages 51–58, Dec. 9–12, 2002.

[5] M. Deshpande, M. Kuramochi, and G. Karypis. Automated approaches for classifying structures. *Proc. of Workshop on Data Mining in Bioinformatics (BioKDD)*, pages 11–18, 2002.

[6] M. Deshpande, M. Kuramochi, and G. Karypis. Frequent sub-structure-based approaches for classifying chemical compounds. *Proc. of IEEE Int. Conf. on Data Mining (ICDM'03)*, Nov. 19–22, 2003.

[7] Espinosa, G.; Arenas, A.; Giralt, F. An integrated som fuzzy artmap neural system for the evaluation of toxicity. *J. Chem. Inf. Comput. Sci. 42*, pages 343–359, 2002.

[8] Gasteiger, J.; Zupan, J. Neural networks in chemistry. *Angew. Chem. Int. Ed. 32*, pages 503–527, 1993.

[9] G.H. Golub, C.F. Van Loan. *Matrix Computation, Third edition*. The Johns Hopkins University Press, Baltimore and London, 1996.

[10] J. K. Cullum, R. A. Willoughby. *Real rectangular matrices, in Lancozos algorithms for large symmetric eigenvalue computations, Vol. 1 Theory*. Brikhauser, Boston, 1985.

[11] Kaski, S., Kangas, J., & Kohonen, T. Bibliography of self-organizing map (som) papers: 1981–1997. neural computing surveys, 1(3&4), 1-176.

[12] Kohonen, T., Kaski, S., Lagus, K., and Honkela, T. Very large two-level som for the browsing of the newsgroups. *Proc. of ICANN'96, Springer Verlag (1996) 269-274,*, 1996.

[13] S. Kramer, L. de Raedt, and C. Helma. Molecular feature mining in hiv data. *Proc. of 7th Int. Conf. on Knowledge Discovery and Data Mining, (KDD'01)*, pages 136–143, 2001.

[14] M. R. Berthold, N. Cebron, F. Dill, G. Di Fatta, T. Gabriel, F. Georg, T. Meinl, P. Ohl, C. Sieb, B. Wiswedel. Knime: the konstanz information miner. *Proc. of the 4th Annual Industrial Simulation Conf. (ISC 2006), Workshop on Multi-Agent Systems and Simulation (MAS&S)*, pages 58–61, June, 5–7 2006.

[15] Manallack, D.; Livingstone, D. Neural networks in drug discovery: Have they lived upto their promise? *Eur. J. Med. Chem. 34*, pages 195–208, 1999.

[16] National Cancer Institute. DTP AIDS Antiviral Screen Dataset. http://dtp.nci.nih.gov/docs/aids/aids/data.html.

[17] National Cancer Institute. Self-organized map (som) of compounds tested in the nci anti-hiv screen [online]. http://cactus.nci.nih.gov/services/som_qsar/.

[18] Nikkila J., and Toronen P. and Kaski S. and Venna J. and Castren E. and Wong G. Analysis and visualization of gene expression data using self-organizing maps. *Neural Networks*, **15**.

[19] Rose, V.; Croall, I.; Macfie, H. An application of unsupervised neural network methodology kohonen topology-preserving mapping to qsar analysis. *Quant. Struct.-Act. Relat. 10*, pages 6–15, 1991.

[20] T. Kohonen. *Self-Organizing Maps*. Springer Verlag, Berlin, 1995.

[21] Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E. S., Golub T. R. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. of the National Academy Science of USA* **96**, pages 2907–2912, 1999.

[22] Tetko, I.; Kovalishyn, V.; Livingstone, D. Volume learning algorithm artificial neural networks for 3d qsar studies. *J. Med. Chem. 44*, pages 2411–2420, 2001.

[23] Ultsch, A. Self-organizing neural networks for visualization and classification. *In O. Opitz, B. Lausen, & R. Klar (Eds.), Information and Classification*.

[24] Ultsch A. and Morchen F. Esom-maps: tools for clustering, visualization, and classification with emergent som. *Technical Report 46, CS Department, Philipps-University Marburg*, 2005.

[25] Vesanto J. Self organizing map in matlab: the som toolbox. *Proc. of the Matlab DSP Conf.*, pages 35–40, nov 1999.

[26] Wang D., Ressom H., Musavi M., Domnison C. Double self-organizing maps to cluster gene expression data. *Proc. of ESANN 2002*, 24–26 April, 2002.

[27] O. Weislow, R. Kiser, D. Fine, J. Bader, R. Shoemaker, and M. Boyd. New soluble formazan assay for hiv-1 cytopathic effects: Application to high flux screening of synthetic and natural products for aids antiviral activity. *Journal of the National Cancer Institute, University Press, Oxford, United Kingdom*, 81:577–586, 1989.

[28] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. *Proc. of the IEEE Int. Conf. on Data Mining (ICDM'02)*, 2002.

[29] M. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. *Proc. of 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD'97)*, pages 283–296, 1997.