Stuctural Analysis of Molecular Networks: AMES Mutagenicity

Laurin AJ Mueller and Karl G Kugler and Matthias Dehmer
laurin.mueller@umit.at and karl.kugler@umit.at and matthias.dehmer@umit.at

Institute of Bioinformatics and Translational Research
University for Health Sciences, Medical Informatics and Technology (UMIT)
Eduard Wallnöfer-Zentrum 1
6060 Hall in Tyrol, Austria

***Abstract*—The characterization of chemical compounds based on their molecular graphs is an important task for identifying properties such as toxicity or mutagenicity. We used different groups of topological descriptors using the AMES mutagenicity data. Instead of optimizing the classification performance, the aim of this study is to perform a structural analysis of the underlying set of molecular graphs to gain better insights of the data set.**

The structural analysis identifies two groups of molecular networks. One group contains graphs with linear patterns (outliers), and the other group contains graphs that exhibit patterns of regular graphs (remainders). We show that the set of used topological descriptors chosen for this study cannot capture enough group-specific structural information within the remainders group to achieve the discrimination ability of the outliers group. Finally, this leads us to the conclusion that it is necessary to identify existing or develop new descriptors that capturing specific structural information to achieve better discrimination ability.

Keywords: Topological network descriptors, network biology, drug design, machine learning

I. Background

The classification of drug-like compounds by using structural information of their underlying molecular graphs is an important task to identify chemical properties (e.g. toxicity or mutagenicity) [Feng et al., 2003], [Votano et al., 2004]. In general, graph classification is a challenging problem and has been tackled by using different methods [Cook and Holder, 2007], [Dehmer and Mehler, 2007], [Deshpande et al., 2003]. Note that classical work relates to applying methods from exact and inexact graph matching [Cook and Holder, 2007], [Dehmer and Mehler, 2007]. In a more biologically motivated work performed by Li et al., graph kernels to predict gene functions have been utilized [Li et al., 2007]. Chuang et al. used subnetworks to train a classifier for the detection of breast cancer metastasis [Chuang et al., 2007].

For our investigation we use the Ames mutagenicity data set, that is a benchmark set to classify graphs [Hansen et al., 2009]. It consists of 6512 graphs, that represent compounds that are categorized as Ames positive ($AMES^+$) or negative ($AMES^-$) by the Ames test [Ames et al., 1973]. Hansen et al. [Hansen et al., 2009] used the commercial software tool Dragon [Todeschini et al., 2003] to calculate a large set of molecular network descriptors to classify the Ames mutagenicity data set.

Dehmer et al. [Dehmer et al., 2010] used entropy-based descriptors [Dehmer and Mowshowitz, 2011] for weighted chemical structures to classify the AMES data set. After removing the isomorphic graphs, they showed that it is possible to classify the remaining graphs with a reasonable classification performance, by only using a set of seven descriptors.

For our analysis we modify this set of graphs, as we only consider the structural skeletons of the molecules. We construct a structural skeleton by using unlabeled nodes and unweighted edges. The main contribution of this paper is to identify discriminatory features of the AMES graphs to classify the structures properly. For this, we calculate the descriptors using the freely available R-package `QuACN` [Mueller et al., 2010b] and selected groups of measures from Dragon [Todeschini et al., 2003] on the resulting set of molecular skeletons.

Note, the classification of Ames mutagenicity by only using structural properties without labels is surely a critical undertaking. The aim of this study is not to increase or optimize the classification performance for this data set but rather to investigate the structural information of molecular networks.

This paper is structured as follows: The Material and Methods section describes the data set of molecular networks that we analyze and gives a brief overview about the used methods. The results section lists the results of the initial classification that motivates the structural analysis. It also contains the results of the structural analysis of the data. In chapter IV we summarize and discuss the results. Section V concludes the paper and provides an outlook on further investigation steps.

II. Material and Methods

The modified AMES Mutagenicity Set for Molecular Networks

The initial data set of Ames mutagenicity [Hansen et al., 2009] was designed to benchmark the

classification performance of different kind of graph classification strategies. It contains 6512 molecular compounds that were categorized positive or negative by the Ames test [Ames et al., 1973] for mutagenicity. Hansen et al. [Hansen et al., 2009] used six different public available data sets and studies to create this benchmark data set. This data set contains $n_+ = 3503$ AMES positive ($AMES^+$) and $n_- = 3009$ AMES negative ($AMES^-$) molecular networks. We used the data set of Dehmer et al. [Dehmer et al., 2010] were isomorphic graphs were removed and modified this set, as we only took the structural skeletons of the molecules. This means that each atom is represented by an unlabeled vertex. Moreover, we represent each kind of bond with an undirected edge. This results in a data set of $n = 3947$ skeletons of molecular networks with $n_+ = 2179$ AMES positive and $n_- = 1768$ AMES negative graphs. This set of molecular networks was used for further analysis.

Topological Network Descriptors

After modifying the AMES data set we calculate different groups of topological descriptors. Topological network descriptors are numerical graph invariants that quantitatively characterize the structure of the underlying network [Emmert-Streib and Dehmer, 2011] We calculated the entropy-based descriptors available in QuACN [Mueller et al., 2010b] and six groups of descriptors offered by the commercial software tool Dragon [Todeschini et al., 2003]. Table I gives an overview about the calculated descriptors.

Each descriptor in Table I results in a single value that characterizes the structure of the underlying molecular network in a certain way. The calculated descriptors can be treated like features and then be used for machine learning [Mueller et al., 2010a], [Mueller et al., 2011].

Also, we will not describe the descriptors in detail. For a better understanding of the selected measures see corresponding literature (e.g. [Bonchev, 1983], [Dehmer et al., 2010], [Todeschini and Consonni, 2009], [Mowshowitz, 1968]). Dehmer and Mowshowitz [Dehmer et al., 2010] discuss entropy-based descriptors, Todeschini et al. [Todeschini and Consonni, 2009] describes the descriptors implemented in Dragon.

Supervised Machine Learning

To classify the molecular networks between $AMES^+$ and $AMES^-$ we treat every topological descriptor as feature. We use support vector machines (SVM) [Vapnik and Lerner, 1963] with a radial basis function kernel.

To compare the results of the support vector machines we use Random Forest (RF) [Svetnik et al., 2003]. After optimizing the parameters and the classification with the mentioned algorithms we calculate the area under the ROC-curve (AOC), the accuracy and the f-score of the results. For each classification we perform a 10-fold cross validation.

To select the best set of topological network descriptors we use the feature selection algorithm information gain [Quinlan, 1993]. The best features of each group were combined to a so called superindex that is defined as follows [Bonchev et al., 1981], [Dehmer et al., 2010].

Definition 1. *Let $I_1, \ldots, I_j$ be topological network descriptors. The superindex of these measures is defined as $SI := \{I_1, \ldots, I_j\}$.*

III. Results

Supervised Machine Learning

The performance of the classification with support vector machines is shown in Table II. The corresponding ROC curves are shown in Fig. 1. It can be seen that the different groups lead to divergent results. The group of vertex degree-based topological descriptors (Dragon 3) achieves the best accuracy with 73.04%. Four groups (Dragon 1, 2, 3, and 5) achieve similar AUCs with about 72%. The groups Dragon 1 and Dragon 3 achieve the best f-scores of about 67%, for details see Table II. It can be summarized that the best classification performances (accuracy, AOC and f-score) is achieved by using the group Dragon 3, containing vertex degree-based topological descriptors.

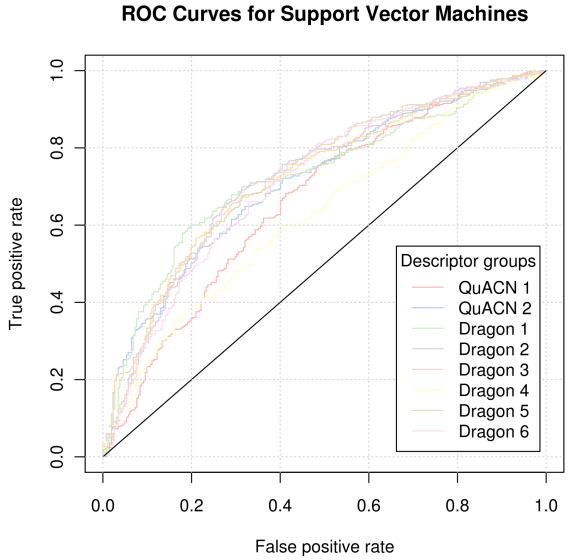

Fig. 1. This figure shows the ROC curves for each descriptor group for the classification using support vector machines.

To evaluate the performance of the support vector machine we use RF to classify the same groups again. Fig. 2 shows the ROC curves for the different groups of descriptors. The results in Table III show that the best performance is achieved by the groups Dragon 1, 2 and 3. The group called Dragon 1 has the highest AUC of 74.89%, Dragon 2 the highest accuracy of 71.55% and Dragon 3 achieved the highest f-score of 67.34%.

This result shows that the different groups of topological network descriptors perform similar using SVM and RF.

TABLE I
OVERVIEW OF THE USED INFORMATION-THEORETIC TOPOLOGICAL DESCRIPTORS.

Group name	Group	Subgroup	No. Descriptors
QuACN 1	Entropy based	Partition based and parametric graph entropy	9
QuACN 2	Polynomial based	–	50
Dragon 1	Walk and path counts	–	46
Dragon 2	Connectivity indices	–	37
Dragon 3	Topological indices	Vertex degree-based	26
Dragon 4	Topological indices	Distance-based indices	13
Dragon 5	Information indices	Basic descriptors	17
Dragon 6	Information indices	Indices of neighborhood symmetry	30

TABLE II
CLASSIFICATION PERFORMANCE FOR EACH GROUP USING SVM

	Precision	Recall	Specificity	Sensitivity	Accuracy	AUC	F-Score
QuACN 1	0.4785	0.6395	0.6486	0.6395	0.6456	0.6625	0.5474
QuACN 2	0.5803	0.6854	0.6971	0.6854	0.6927	0.7120	0.6285
Dragon 1	0.6476	0.7152	0.7344	0.7152	0.7266	0.7216	0.6797
Dragon 2	0.6041	0.7427	0.7210	0.7427	0.7289	0.7220	0.6663
Dragon 3	0.6357	0.7280	0.7320	0.7280	0.7304	0.7223	0.6787
Dragon 4	0.4649	0.6089	0.6357	0.6089	0.6266	0.6265	0.5273
Dragon 5	0.5288	0.7203	0.6855	0.7203	0.6970	0.7243	0.6099
Dragon 6	0.5696	0.6638	0.6868	0.6638	0.6780	0.7088	0.6131

TABLE III
CLASSIFICATION PERFORMANCE FOR EACH GROUP USING RANDOM FOREST

	Precision	Recall	Specificity	Sensitivity	Accuracy	AUC	F-Score
QuACN 1	0.5215	0.5895	0.6450	0.5895	0.6230	0.6281	0.5534
QuACN 2	0.5724	0.6216	0.6740	0.6216	0.6524	0.7102	0.5960
Dragon 1	0.6663	0.6747	0.7319	0.6747	0.7066	0.7489	0.6705
Dragon 2	0.6369	0.7007	0.7256	0.7007	0.7155	0.7406	0.6673
Dragon 3	0.6578	0.6898	0.7324	0.6898	0.7142	0.7363	0.6734
Dragon 4	0.6222	0.6599	0.7070	0.6599	0.6871	0.6022	0.6405
Dragon 5	0.6227	0.6613	0.7077	0.6613	0.6881	0.7223	0.6414
Dragon 6	0.5339	0.5885	0.6483	0.5885	0.6240	0.7166	0.5599

Moreover, the classification with RF achieves a slightly higher performance. However, it can be seen that the groups Dragon 1-3 are qualified best to discriminate between $AMES^+$ and $AMES^-$ for this set of molecular networks.

To study the classification performance we perform a feature selection with information gain for each group and selected the best three descriptors of each group to create a superindex. Classifying by applying the superindex leads to the results shown in Table IV. The ROC curves are shown in Fig. 3.

The performance of SVM and RF are similar but RF performs better with an accuracy of 74.21% and AUC of 76.62 and an f-score of 70.80%.

To evaluate the stability of the results we randomly select 1000 molecular networks and classify them using the superindex and RF. We repeat this procedure 1000 times. This results in a mean f-score of 64% with a standard deviation of 2%. This small standard deviation indicates that the classification performance is stable.

In order to analyze the classification performance we investigate the structural information of the set of the molecular networks. Therefore, we calculate a set of distance-based descriptors [Skorobogatov and Dobrynin, 1988] to explore basic structural properties.

Exemplarily, we use the average path length to outline a prototype of the structural analysis. Fig. 4 shows the average path length (APL) for all molecular networks. One function represents the graphs that are grouped as $AMES^+$ the other one shows the graphs that are $AMES^-$. The vertical lines represent the mean and the standard deviation (dashed) for each group. Fig. 4 shows, in a descriptive way, that the distribution of the average path length of the two groups ($AMES^+$ and $AMES^-$) is largely overlapping. This can also be observed for the other distance-based descriptors.

We hypothesize that the outliers are more discriminative than the remaining graphs. We define outliers as at least one standard deviation away from the mean of each group (see Fig. 4. Using this criteria we split the molecular networks into two groups (outliers and remainders) with $n_{outliers} = 1102$ and $n_{rest} = 2623$ graphs. We then classifying this two group separately, using the superindex and random forest. This results in an f-score for the outliers of 72.73%. The performance of the classification for the remainders obtained an f-score of 66.63%.

To identify structural information of the different groups we look at single graphs in the two groups ($AMES^+$ and $AMES^-$). Fig. 5 exemplary shows two graphs of each group. Fig. 5(a) and 5(b) show two outliers, and Fig. 5(c) and 5(d) represent two networks of the remainders. It can be seen that the outliers possess linear patterns, in contrast the remainders show regular patters. A regular graph is a graph where each

TABLE IV
CLASSIFICATION PERFORMANCE OF THE SUPERINDEX

	Precision	Recall	Specificity	Sensitivity	Accuracy	AUC	F-Score
Support vector machine	0.6561	0.7374	0.7439	0.7374	0.7413	0.7367	0.6944
Random forest	0.6980	0.7183	0.7604	0.7183	0.7421	0.7662	0.7080

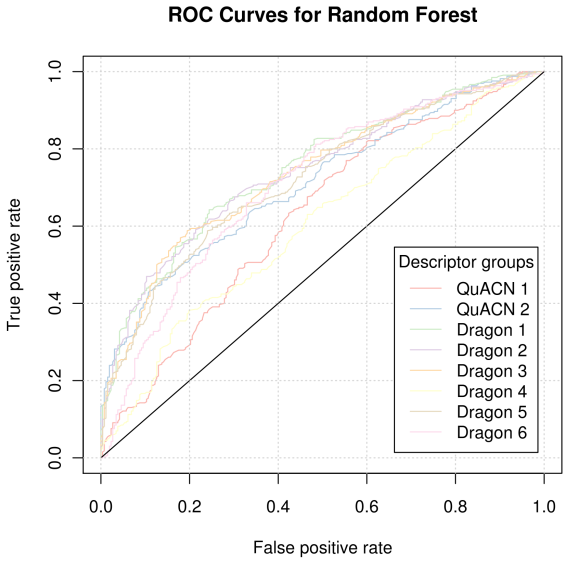

Fig. 2. This figure shows the ROC curves for each descriptor group for the classification using RF.

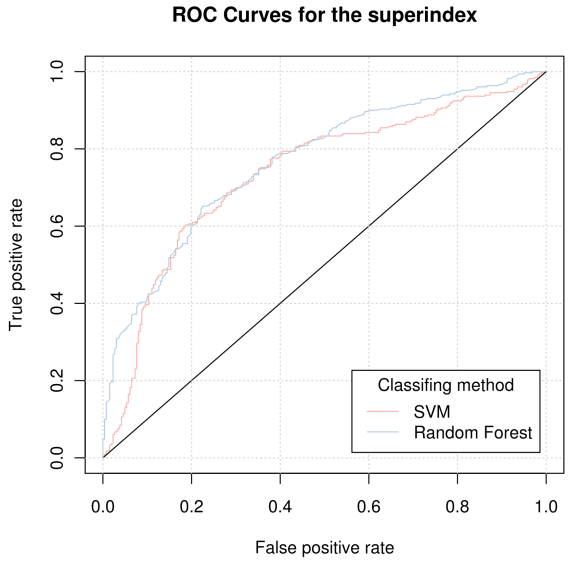

Fig. 3. This figure shows the ROC curves for the classification with SVM and RF using the superindex.

vertex has the same degree. These characteristics can also be observed for other graphs of the corresponding groups.

Repeating this kind of outlier analysis with different distance-based descriptors (i.e. eccentricity or average distance) leads to similar results. In summary we see that the outliers possess linear patterns. This is in contrast to graphs that are close to the mean of the corresponding descriptor, which exhibit rather regulatority.

IV. SUMMARY AND DISCUSSION

The AMES mutagenicity set of molecular networks is a benchmark set to evaluate the performance of graph classification algorithms. By only using the underlying skeletons, the classification of this AMES mutagenicity is a difficult and complex endeavor. It becomes even harder, when removing isomorphic graphs, the information of node labels and edge properties. To classify the remaining network skeletons we used different groups of topological network descriptors and constructed a so called superindex by selecting the best features of each group with the feature selection method information gain. We used support vector machines and random forest to perform the classification.

The group of vertex degree-based indices achieved the best results, what indicates that the degree has a hight discrimination ability within this set of molecular networks. Different groups of topological network descriptors capture different structural information, what led to a higher discrimination ability by combining them to a superindex.

Hansen et al. [Hansen et al., 2009] achieved an AUC of 86%. One can see that our classification performance is less than 10% lower. Considering the fact, that Hansen et al. also used groups of descriptors that take information about the atoms (e.g.: atom type, atom weights) and different binding types into account, and we reduced the information in the molecular network by reducing them to their structural skeletons, we achieve fairly acceptable results. Moreover, the removal of the isomorphic graphs can be a reason for the lower discrimination ability. Imagine that if a graph is correctly classified, all isomorphic graphs would also be correctly classified, what would increase the overall performance of the classification. By using molecular skeletons it can happen that two molecules are reduced to the same skeleton and then can be found in the $AMES^+$ and in the $AMES^-$ group. That can also be a possible reason for a lower classification power.

Comparing our results with Dehmer et al. [Dehmer et al., 2010], they achieved 71.4% including label information, shows that our best classification performance by using the superindex and random forest, is only about 0.6% lower. Compared to their results when using unlabeled graphs, the difference is even smaller. Note, that the fact that the results are fairly the same, strengthens our hypothesis that the classification performance cannot be increased dramatically,

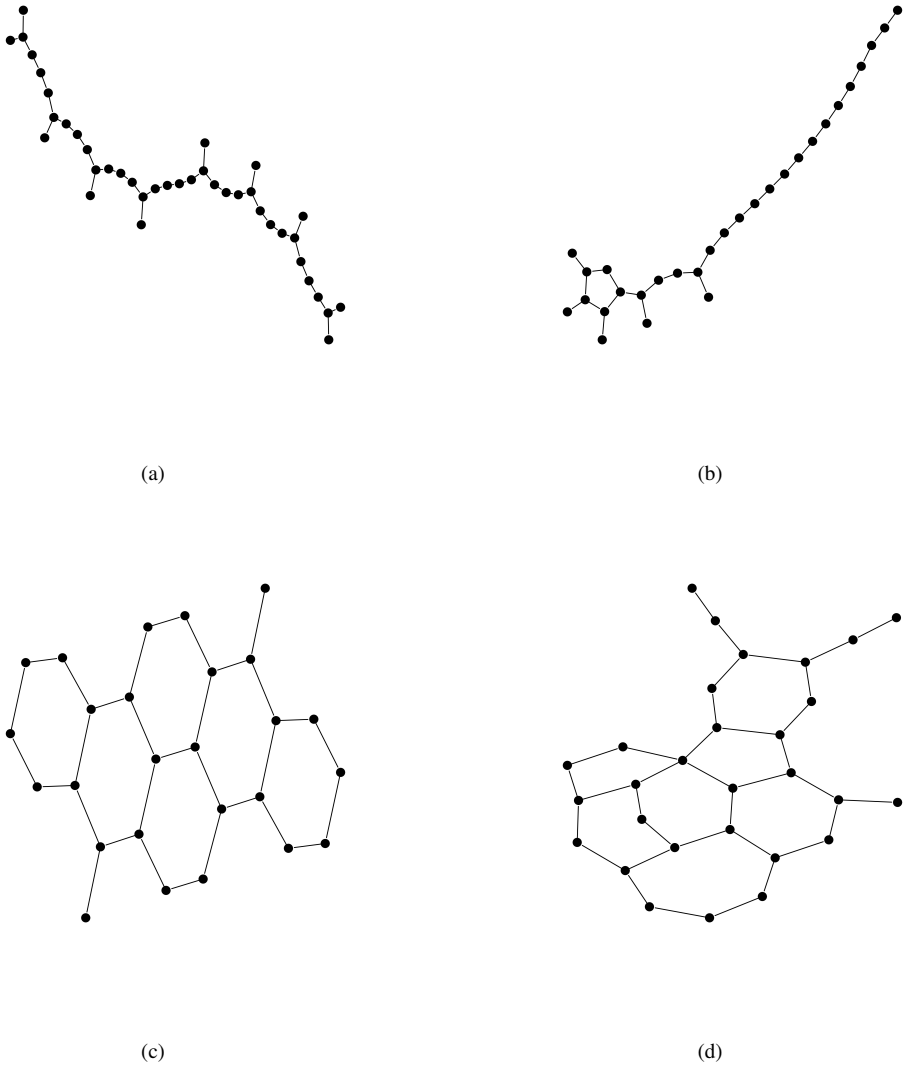

Fig. 5. This figure shows exemplary two graphs of the two groups, split by the value of the average path length. One group ((a) and (b)) represent outliers that are at least one standard deviation away from the mean (see Fig. 4. (c) and (d) represent the groups of the remaining molecular networks.

by only using the this set of molecular descriptors.

In order to increase the classification performance in further studies we analyzed the set of molecular networks structurally. Therefore, we applied a set of distance-based descriptors to them, and analyzed the structure of the outliers. An interesting finding is that the outliers show linear patterns, compared to the remaining graphs that show properties of regular graphs. Moreover, as these regular graphs show more equal vertex degrees than the linear ones, this assumption matches with observation that the group of vertex degree-based descriptors has the highest classification performance of all selected groups of topological network descriptors. An other interesting finding is that the remaining regular graphs contain ring-like structures.

V. Conclusion and Outlook

This study deals with the structural analysis of the AMES mutagenicity data set. It turned out that vertex degree-based descriptors led to a good classification performance. Combining different groups of descriptors to a superindex is promising as it increased the classification performance.

The major challenge of this study was to explore the selected topological network descriptors. They failed to capture enough structural information that would have been needed for achieving a better discrimination ability. The structural analysis showed that there is a set of graphs possessing linear patterns and a set of graphs showing regular characteristics. For future work it is necessary either to identify existing descriptors or develop new descriptors that can better discriminate between these graphs. Therefore, a thorough analysis of

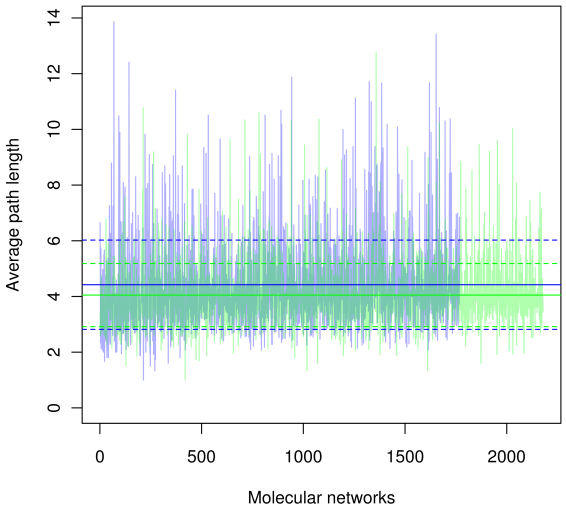

Fig. 4. This figure shows the average path length for each group: $AMES^+$ (green) and $AMES^-$ (blue). The vertical lines represent the mean and the standard deviation (dashed) for each group.

the data set is needed.

Moreover, defining superindices to combine different descriptors, which capture different kinds of structural properties can be a promising strategy. The combination of different superindices could lead to an approach that can capture group-specific combinations of different structural information to distinguish between AMES positive and AMES negative tested chemical compounds.

VI. Acknowledgement

This work was funded by the Tiroler Wissenschafts Fonds (Project CoNAn - Phase II) and the Tiroler Zukunftsstiftung.

This work was supported by the COMET Center ONCOTYROL and funded by the Federal Ministry for Transport Innovation and Technology (BMVIT) and the Federal Ministry of Economics and Labour/the Federal Ministry of Economy, Family and Youth (BMWA/BMWFJ), the Tiroler Zukunftsstiftung (TZS) and the State of Styria represented by the Styrian Business Promotion Agency (SFG).

We are grateful to Kurt Varmuza for calculating the topological network descriptors by using Dragon.

References

[Ames et al., 1973] Ames, B. N., Lee, F. D., and Durston, W. E. (1973). An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proceedings of the National Academy of Sciences of the United States of America*, 70(3):782–6.

[Bonchev, 1983] Bonchev, D. (1983). *Information theoretic indices for characterization of chemical structures*. Chemometrics research studies series. Research Studies Press.

[Bonchev et al., 1981] Bonchev, D., Mekenyan, O., and Trinajsiteć, N. (1981). Isomer discrimination by topological information approach. *Journal of Computational Chemistry*, 2(2):127–148.

[Chuang et al., 2007] Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140.

[Cook and Holder, 2007] Cook, D. and Holder, L. B. (2007). *Mining Graph Data*. Wiley-Interscience.

[Dehmer et al., 2010] Dehmer, M., Barbarini, N., Varmuza, K., and Graber, A. (2010). Novel topological descriptors for analyzing biological networks. *BMC Structural Biology*, 10(1):18.

[Dehmer and Mehler, 2007] Dehmer, M. and Mehler, A. (2007). A new method of measuring similarity for a special class of directed graphs. *Tatra Mountains Mathematical Publications*, 36:39–59.

[Dehmer and Mowshowitz, 2011] Dehmer, M. and Mowshowitz, A. (2011). A history of graph entropy measures. *Information Sciences*, 181(1):57–78.

[Deshpande et al., 2003] Deshpande, M., Kuramochi, M., and Karypis, G. (2003). Automated approaches for classifying structures. In *Proceedings of the 3-rd IEEE International Conference of Data Mining*, pages 35–42.

[Emmert-Streib and Dehmer, 2011] Emmert-Streib, F. and Dehmer, M. (2011). Networks for Systems Biology: Conceptual Connection of Data and Function. *IET Syst Biol*.

[Feng et al., 2003] Feng, J., Lurati, L., Ouyang, H., Robinson, T., Wang, Y., Yuan, S., and Young, S. S. (2003). Predictive Toxicology: Benchmarking Molecular Descriptors and Statistical Methods. *J. Chem. Inf. Comput. Sci.*, 43(5):1463–1470.

[Hansen et al., 2009] Hansen, K., Mika, S., Schroeter, T., Sutter, A., ter Laak, A., Steger-Hartmann, T., Heinrich, N., and Müller, K.-R. (2009). Benchmark data set for in silico prediction of Ames mutagenicity. *Journal of chemical information and modeling*, 49(9):2077–81.

[Li et al., 2007] Li, X., Zhang, Z., Chen, H., and Li, J. (2007). Graph Kernel-Based Learning for Gene Function Prediction from Gene Interaction Network. In *Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine*.

[Mowshowitz, 1968] Mowshowitz, A. (1968). Entropy and the complexity of the graphs i: An index of the relative complexity of a graph. *Bull Math Biophys*, 30:175–204.

[Mueller et al., 2010a] Mueller, L. A., Kugler, K. G., Dander, A., Graber, A., and Dehmer, M. (2010a). Network-based approach to classify disease stages of prostate cancer using quantitative network measures. *Conference on Bioinformatics & Computational Biology (BIOCOMP'10), Las Vegas/USA*, I:55–61.

[Mueller et al., 2010b] Mueller, L. A., Kugler, K. G., Dander, A., Graber, A., and Dehmer, M. (2010b). QuACN - An R Package for Analyzing Complex Biological Networks Quantitatively. *Bioinformatics*, submitted.

[Mueller et al., 2011] Mueller, L. A., Kugler, K. G., Netzer, M., Graber, A., and Dehmer, M. (2011). Distinguishing between the three domains of life using topological characteristics of their underlying metabolic networks. *submitted*.

[Quinlan, 1993] Quinlan, R. J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, CA, USA.

[Skorobogatov and Dobrynin, 1988] Skorobogatov, V. A. and Dobrynin, A. A. (1988). Metrical analysis of graphs. *Commun. Math. Comp. Chem.*, 23:105–155.

[Svetnik et al., 2003] Svetnik, V., Liaw, A., Tong, C., Culberson, J., Sheridan, R., and Feuston, B. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci*, 43(6):1947–1958.

[Todeschini and Consonni, 2009] Todeschini, R. and Consonni, V. (2009). *Molecular descriptors for chemoinformatics*. Vch Pub.

[Todeschini et al., 2003] Todeschini, R., Consonni, V., Mauri, A., and Pavan, M. (2003). Software dragon: Calculation of molecular descriptors, department of environmental sciences. Talete.

[Vapnik and Lerner, 1963] Vapnik, V. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24.

[Votano et al., 2004] Votano, J. R., Parham, M., Hall, L. H., and Kier, L. B. (2004). New predictors for several ADME/Tox properties: aqueous solubility, human oral absorption, and Ames genotoxicity using topological descriptors. *Mol Divers*, 8(4):379–391.