Analysis of Metabolic Networks: On the Similarity of the three Domains of Life

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

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

***Abstract*—Metabolic networks summarize and represent anabolic and catabolic processes that are driven by the enzymes in every organism. It has been shown that the metabolic networks of the three domains of life (Archae, Bacteria, and Eukaryota) have certain properties in common. However, we could previously demonstrate that it is still possible to find domain-specific attributes in the corresponding networks, that allow for a good inter-domain classification performance. In this paper we aim at finding domain dependent differences based on distances between vertices in the networks. We apply three different distance-based topological network descriptors using Shannon's Entropy. Our results show that a clear distinction between the three domains of life fails when using the employed network descriptors. This indicates that certain distance-related properties are common to all organisms in this study. We expect this to be a sign of the evolutionary optimization of the information spread within these networks.**

Keywords: Network biology, metabolic pathways, topological network descriptors, machine learning

I. Background

Catabolic and anabolic processes can be represented by metabolic networks, as they represent the interlinkage of metabolic processes that make up the human metabolism [Alberts et al., 2007]. By studying how these processes are organized in pathways it is possible to derive knowledge about the underlying functions. Jeong et al. systematically investigated the organization and structure of metabolic networks from 43 organisms that were representing the three domains of life [Jeong et al., 2000]. One of their main results was, that despite the evolutionary distance, properties related to the network diameter were found to be highly conserved [Jeong et al., 2000]. However, in recent work we demonstrated that it is possible to still discriminate between the three domains of life [Mueller et al., 2011]. The main goal of this paper is to analyze path distance-based properties in the networks by Jeong et al., in order to detect domain-specific effects. We aim at detecting distance-based effects, that may hint at evolutionary differences between the domains of life. To tackle this problem we utilize entropy-based topological network descriptors [Dehmer and Mowshowitz, 2011]. The structure of a network also reflects its function [Strogatz, 2001]. Thus, applying topological network descriptors might be useful for the analysis of complex networks, as they allow transforming structural information about a graph into a numeric value [Emmert-Streib and Dehmer, 2011]. Topological network descriptors have been employed in chemoinformatics, e.g. for predicting toxicity [Feng et al., 2003] or mutagenicity [Votano et al., 2004]. Recently, they have also been proven useful for analyzing biological networks [Mueller et al., 2010], [Emmert-Streib and Dehmer, 2011].

Jeong et al. first described the relatedness of metabolic networks, when they explored degree distributions and average path lengths [Jeong et al., 2000]. Later, Wagner and Fell found the metabolic network of E. coli to exhibit the small-world property for a slightly different set-up [Wagner and Fell, 2001]. Ma and Zeng analyzed the core networks and clustering properties of different organisms in their work [Ma and Zeng, 2003]. They also investigated the average path lengths of the largest subnetwork and the whole network for 65 organisms [Ma and Zeng, 2003]. A set of topological network descriptors was employed by Zhu and Qin to find differences in various single cell organisms [Zhu and Qin, 2005]. They found the average clustering coefficient and the average betweenness to differ between six Bacteria and four Arachaea [Zhu and Qin, 2005]. In the current paper we focus on topological network descriptors that are based on inferring distances between vertices. We hypothesize that the analysis of these distances might reveal knowledge about the spread of information within these networks.

This paper is structured as follows: After providing background information in this Section, we describe the employed data set and the methods in Section II. Thereafter, we illustrate the results in Section III, which are discussed in IV. This paper finishes with a final summary and conclusion in section V.

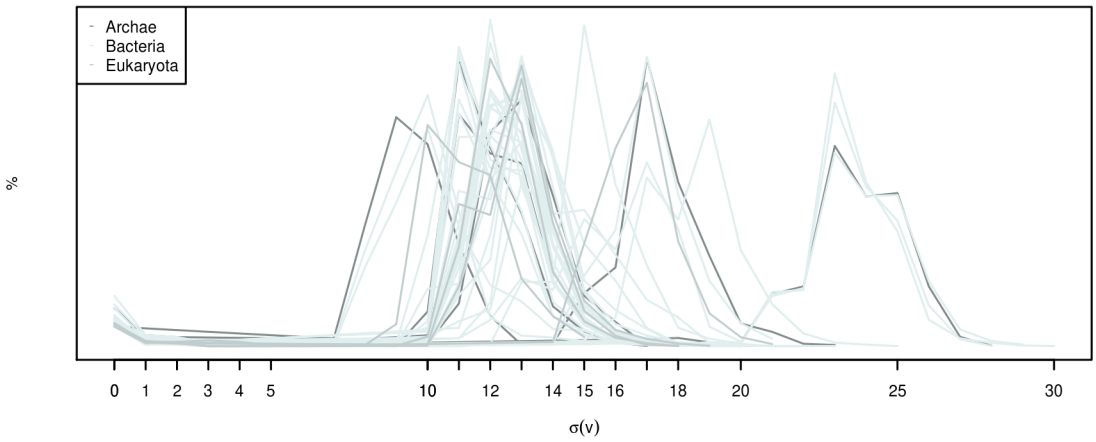

Fig. 1. The distribution of $\sigma(v)$ for the 43 organisms. Each of the three domains is depicted in a different color.

II. Material and Methods

Metabolic Networks

For the analysis of domain-specific effects we re-analyze the metabolic networks that have originally been studied by Jeong et al. [Jeong et al., 2000]. In their study they analyzed 43 organisms from the three domains of life ($n_{Archaea} = 6$, $n_{Bacteria} = 32$, and $n_{Eukaryota} = 5$). After we construct the networks, we extract the largest connected component, which represents the largest connected subgraph, for each organism. This results in a network G for every organism, where V is the set of labeled vertices and E is the set of directed edges. Overall, we then have 43 labeled and directed networks for the further analysis.
The eccentricity $\sigma(v)$ of a vertex v is an important feature within a network [Hage and Harary, 1995]. It gives the maximum of the distances from one vertex to all other connected vertices. In biological networks, small distances may indicate short communication processes, which allow for an organism to rapidly react to disturbances. To illustrate the distribution of $\sigma(v)$ for each of the 43 networks we plot it in Fig. 1.

Network Descriptors using Distances

Topological network descriptors represent the complexity of a network by a numeric value [Emmert-Streib and Dehmer, 2011]. Early applications of network descriptors date back to the work of Wiener [Wiener, 1947]. He utilized the sum of the distance matrix for predicting paraffin boiling points. Other well-known indices are the Balaban J index [Balaban, 1982], the Zagreb group indices [Diudea et al., 2001] or the Randić connectivity index [Li and Gutman, 2006]. Later, methods for quantifying the information content of a network were established [Bonchev and Rouvray, 2005], [Mowshowitz, 1968], [Rashewsky, 1955], [Trucco, 1956]. Note, that many of these descriptors are correlated. Bonchev and Trinajstić introduced an information index that captures molecular branching [Bonchev and Trinajstic, 1977]. Many other real-world applications are also based on problems of relational structures, e.g. transportation or communication networks [Kolaczyk, 2009]. Networks and topological descriptors have been extensively used in the social sciences [Wasserman and Faust, 1994], e.g. for identifying opinion-leaders or the spread of information in societies.

In the present work we put an emphasis on i) descriptors, that can be used to evaluate the information spread in a network. ii) Descriptors that calculate the information-content of a network. We select entropy-based network descriptors since they were shown in previous work to posses good classification performance when capturing domain-specific effects [Mueller et al., 2011]. For a comprehensive overview on entropy-based network descriptors see e.g. [Dehmer and Mowshowitz, 2011]. We focus on studying the information-spread as we are interested in finding structural differences that present themselves in the way information is spread within the metabolic networks. Our hypothesis is, that we might find structural differences that can be clearly linked to a domain-specific origin.

It has been shown that it is possible to quantify the information-content of a network by applying special functionals to the vertices of the network and using Shannon's Entropy [Dehmer, 2008]. Dehmer presented a vertex functional that is based on the j-spheres [Dehmer, 2008]:

$$\begin{aligned} f^V(v_i) := c_1|S_1(v_i,G)| + c_2|S_2(v_i,G)| + \\ \cdots + c_{\rho(G)}|S_{\rho(G)}(v_i,G)|, \\ c_k > 0, 1 \leq k \leq \rho(G). \end{aligned} \tag{1}$$

$S_j(v_i, G)$ is the set of vertices with distance j from vertex $v_i \in V$. Note, that c_k represents a weighting factor. In our case, we modeled it to follow a exponential function. So $c_k = \rho(G)e^k$ for $k = 0, 1, \ldots, \rho(G) - 1$. This allows emphasizing on vertices that are close to v_i. The structural information content of a graph G with respect to $f^V(v_i)$ is then defined by [Dehmer, 2008]:

$$I_{f^V}(G) = \sum_{i=1}^{|V|} \frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)} \log_2 \frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)}. \quad (2)$$

$f^V(v_i)$ can be seen as a function that represents the spread of information from v_i, so $I_f(G)$ is a model for the information spread in G [Dehmer, 2008].

Bonchev et al. introduced a descriptor that is based on the eccentricity $\sigma(v_i)$ and the mean information content [Bonchev et al., 1980]. The radial centric information index is defined by [Bonchev et al., 1980]:

$$\bar{I}_C^V(G) = \sum_{j=1}^{|V|} \frac{n_j}{|V|} \log_2 \frac{n_j}{|V|}. \quad (3)$$

Here, n_j gives the number of vertices with eccentricity $\sigma(v_i) = j$. It is common to assume that small $\sigma(v_i), v_i \in V$ indicate the possibility to spread information rapidly within G. So, $\bar{I}_C^V(G)$ should give an insight into how information is spread in G. If our hypothesis holds, organisms from different domains may exhibit systematic differences with respect to $\bar{I}_C^V(G)$.

It is possible to define information measures using local features of graphs, e.g. by quantifying the entropy of single vertices [Dehmer and Mowshowitz, 2011]. Konstantinova and Paleev introduced a measure that represents the vertex complexity by [Konstantinova and Paleev, 1990]:

$$I_D(v_i) = -\sum_{j=1}^{|V|} \frac{d(v_i, v_j)}{d(v_i)} \log_2 \left(\frac{d(v_i, v_j)}{d(v_i)} \right), \quad (4)$$

where $d(v_i)$ gives the sum of distances from vertex v_i to all other vertices in G. The entropy of G is then given as [Konstantinova and Paleev, 1990]:

$$I_D(G) = \sum_{i=1}^{|V|} I_D(v_i). \quad (5)$$

Here, we use $I_D(G)$ to model the heterogeneity of the vertices of a graph G with respect to the distances between the vertices. Based on our hypothesis we should see a domain-specific effect in this heterogeneity.

Univariate Analysis

After we calculate the three topological network descriptors for each of the 43 metabolic networks we proceed with the succeeding data analysis. First, we test for the presence of a domain-specific effect in at least one group by performing a one way ANOVA [Chambers and Hastie, 1991].

Unsupervised Machine Learning

We use hierarchical clustering in order to explore the groups that are formed by the employed distance measures. Our clustering is based on the Euclidean distance between features [Murtagh, 1985].

Supervised Machine Learning

For supervised machine learning we make use of support vector machines [Vapnik and Lerner, 1963], with a radial basis kernel. To optimize the outcome we set the cost parameter to 100. We then calculate the accuracy and the f-score for the classification of the domains of life based on our set of topological network descriptors.

III. Results

Distance-Based Network Descriptors

For each of the 43 species we calculate the three presented descriptors with the programming language R (http://www.r-project.org). The results are listed in Table I and illustrated as boxplots in Fig. 2.

Univariate Analysis

The results for the ANOVA are listed in Table II. To adjust for multiple testing we correct with the method by Bonferroni. However, even before the multiple testing correction no descriptor detects a significant effect in a single domain.

Unsupervised Machine Learning

The heatmap in Fig. 3 illustrates the results of the hierarchical clustering. The rows contain the 43 organisms and the columns represent the three employed topological network descriptors. We mark each domain in a specific color. We observe no meaningful clustering with respect to the three domains of life.

Supervised Machine Learning

The classification accuracy is 63%. While we reach a precision of 0.86 we only score a recall of 0.33. This leads to an overall f-score of 0.48.

IV. Discussion

In the present study our goal was to detect differences and characteristics for the three domains of life by making use of their metabolic networks. Therefore, we reused a set of 43 organisms that have originally been investigated by Jeong et al. [Jeong et al., 2000]. Here, we focused on analyzing potential differences in the distances between vertices in the metabolic networks. We employed a broad range of different approaches to this problem, which all failed to detect any domain-specific effects.

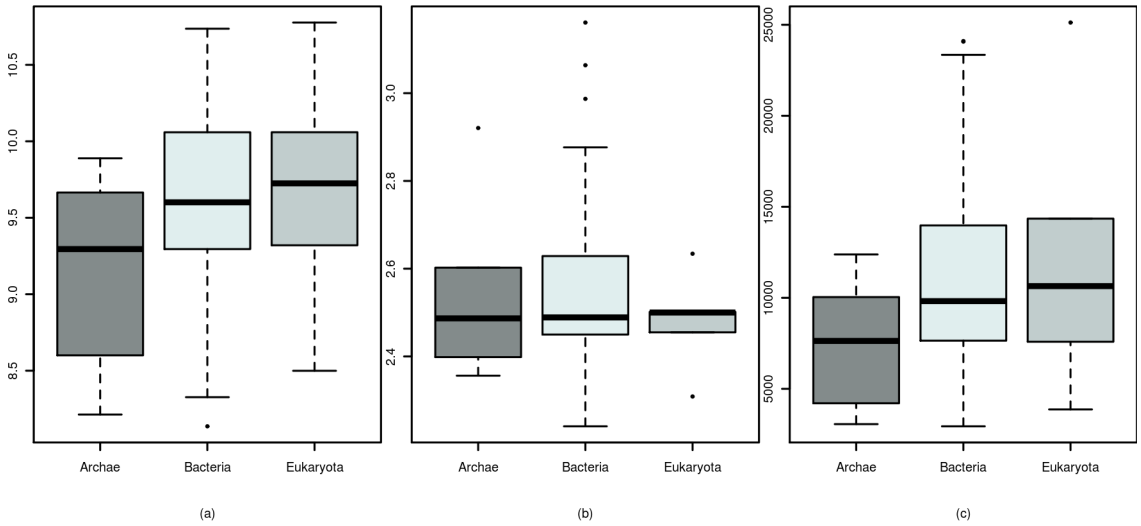

Fig. 2. For each of the 43 species we calculate three entropy-based network descriptors: (a) I_{fV}, (b) $\bar{I}_C^V$, (c) I_D.

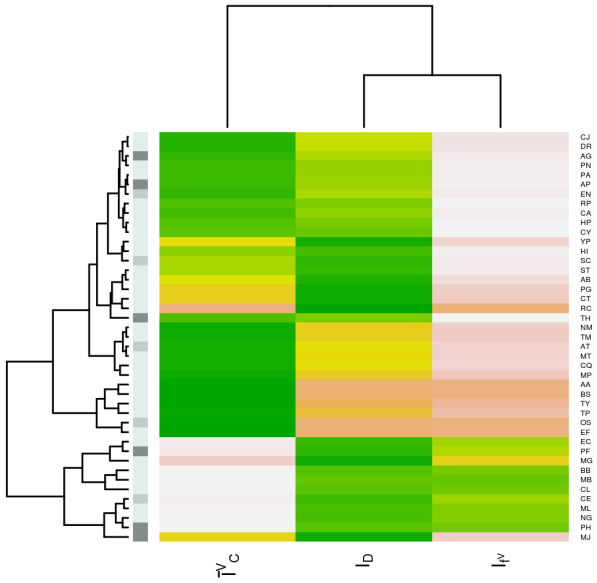

Fig. 3. We perform a hierarchical clustering for the 43 metabolic networks (rows) and the three employed network descriptors (columns). The three domains of live are depicted in three different colors.

In the original work by Jeong et al. they discovered several interesting aspects that were common to all networks. All the degree distributions of the networks were found to be scale-free and follow a power-law distribution [Jeong et al., 2000]. Moreover, the network diameters $\rho(G)$ were found to be relatively constant across all three domains of life [Jeong et al., 2000]. The similarity in the large-scale organization of the metabolic networks is also discussed in [Podani et al., 2001]. These observations indicate that core properties of the metabolic processes are common to all species and are to a certain degree not influenced by evolutionary processes. However, in recent work we could demonstrate that it is still possible to distinguish between Archaea, Bacteria, and Eukaryota based on topological properties of their metabolic networks [Mueller et al., 2011]. In that previous study we applied a set of supervised machine learning algorithms to 33 network descriptors that were calculated for the same data, and came up with a reasonable classification performance (Accuracy: 88.4%, weighted F-score: 0.88). Such a result has not been reached in the present study. However, in contrast to this previous work we now considered directed graphs for our analysis. This hardens a direct comparison of the previous results with the current ones. Interestingly, when we ignored the directional information, two measures that are related to path length and the spheres turned out to be significantly different in at least one group [Mueller et al., 2011]. This is a striking observation that will need to be verified and interpreted in future studies.

Considering that Jeong et al. observed highly conserved distance properties in their original study and that we focused our analysis on these network invariants the observed results come to no surprise. We conclude that this highlights the fact that metabolic networks are likely to have evolved in a way that allows spreading information efficiently, and that this design is common to most organisms in the present set of networks. Our results are to a certain degree coherent with other, related observations. In a similar study, clear differences between Bacteria and Archaea were found for the average clustering coefficient and the average betweenness, but not so much for the average path length and the diameter [Zhu and Qin, 2005]. These latter two are mainly related to distance between vertices, which was also the graph invariant

TABLE I
HERE, WE LIST THE RESULTS FOR THE 43 ORGANISMS AND THE THREE EMPLOYED TOPOLOGICAL NETWORK DESCRIPTORS.

Organism	I_f	$\bar{I}_C^V$	I_D	Domain
AP	9.665084	1.704296	10040.587666	Archaea
AG	9.564147	1.662454	9225.599594	Archae
TH	9.888839	1.803817	12382.120096	Archae
MJ	8.600314	1.63313	4206.884867	Archae
PF	9.025969	2.02442	6037.409112	Archae
PH	8.2133	1.74309	3060.56127	Archae
AA	10.658253	1.762958	23348.495827	Bacteria
CQ	9.938537	1.675467	12776.953312	Bacteria
CT	9.787773	1.887457	11276.072509	Bacteria
CY	9.505232	1.715802	8762.273987	Bacteria
PG	9.435009	1.804038	8489.037187	Bacteria
MB	8.136314	2.191041	2940.192524	Bacteria
ML	8.362968	1.837814	3543.299666	Bacteria
MT	10.070713	1.717556	14174.900512	Bacteria
BS	10.700696	1.678981	24103.108553	Bacteria
EF	10.736119	1.553148	24075.776966	Bacteria
CA	9.586382	1.701106	9517.179245	Bacteria
MG	9.416086	1.974329	8301.183302	Bacteria
MP	10.047548	1.615207	13779.564647	Bacteria
PN	9.523945	1.673842	8863.919547	Bacteria
ST	9.580282	1.792999	9745.236084	Bacteria
CL	8.350742	1.993873	3805.079281	Bacteria
RC	9.669622	1.924546	10291.659579	Bacteria
RP	9.615696	1.730073	9890.367254	Bacteria
NG	8.326707	1.783241	3343.943011	Bacteria
NM	10.142384	1.720369	15237.759713	Bacteria
CJ	9.657014	1.650279	9902.713247	Bacteria
HP	9.556632	1.711279	9098.264346	Bacteria
EC	8.964342	2.070491	5795.615593	Bacteria
TY	10.491932	1.71284	20399.468024	Bacteria
YP	9.154839	1.735033	6777.224163	Bacteria
AB	9.524078	1.806462	9364.071385	Bacteria
HI	9.174159	1.695162	6988.915709	Bacteria
PA	9.678677	1.709062	10192.080149	Bacteria
TP	10.341138	1.751982	18038.2337	Bacteria
BB	8.446315	2.123614	3699.92756	Bacteria
TM	10.171566	1.705844	15249.575415	Bacteria
DR	9.640121	1.647044	9901.92116	Bacteria
EN	9.724915	1.701616	10642.447537	Eukaryota
SC	9.320128	1.732907	7586.420683	Eukaryota
CE	8.499299	1.825795	3867.729448	Eukaryota
OS	10.776499	1.60026	25122.661425	Eukaryota
AT	10.059327	1.733995	14349.913609	Eukaryota

TABLE II
THE RESULTS FOR THE ANOVA TESTING. p_{Bonf} IS THE P-VALUES AFTER THE BONFERRONI CORRECTION.

	I_f	$\bar{I}_C^V$	$I_D(v_i)$
p	0.384	0.638	0.342
p_{Bonf}	1.000	1.000	1.000

of interest in our study. All three utilized topological network descriptors were quantifying the information-content of the networks. In recent work we were able to demonstrate that this family of descriptors is powerful for detecting differences related to the three domains of life when using this set of data [Mueller et al., 2011]. In the present work, the low power in finding domain-specific differences is caused by the underlying graph invariant. We hypothesize that in order to find domain-specific differences in the topology of the networks in this set it is better to focus on other graph invariants, e.g. vertex degrees or centralities.

V. SUMMARY AND CONCLUSION

Finding specific properties in groups of biological networks is a major goal in network analysis. Here, we wanted to detect topological properties responsible for the spread of information within a network and specific for the three domains of life (Archaea, Bacteria, and Eukaryota). Therefore, we employed a set of three network descriptors that capture properties related to distances within a graph. We calculated each of these three descriptors on a set of 43 metabolic networks from different organisms. To analyze the according data we utilized univariate methods as well as supervised and unsupervised machine learning procedures. However, with none of the applied approaches could we detect any meaningful discrimination or characterization of the three domains of life. Since we could demonstrate in previous work that is possible to discriminate between the three domains of life based on the present data, we conclude that the information-spread as captured by the employed measured fails to capture domain-specific properties for this set of directed networks. It will be part of future work to analyze what groups of topological network descriptors are best fitted to solve this undertaking. This could then give insights into evolutionary differences between the domains.

VI. ACKNOWLEDGEMENT

This work was funded by 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 thank Armin Graber and Frank Emmert-Streib for fruitful discussions. Frank Emmert-Streib also brought our attention to the analyzed data set.

REFERENCES

[Alberts et al., 2007] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2007). *Molecular Biology of the Cell*. Garland Science, 5 edition.

[Balaban, 1982] Balaban, A. (1982). Highly discriminating distance-based topological index. *Chemical Physics Letters*, 89(5):399–404.

[Bonchev et al., 1980] Bonchev, D., Balaban, A. T., and Mekenyan, O. (1980). Generalization of the Graph Center Concept, and Derived Topological Centric Indexes. *Journal of Chemical Information and Computer Sciences*, 20(2):106–113.

[Bonchev and Rouvray, 2005] Bonchev, D. and Rouvray, D. H. (2005). *Complexity in Chemistry, Biology, and Ecology*. Springer, New York.

[Bonchev and Trinajstic, 1977] Bonchev, D. and Trinajstic, N. (1977). Information theory, distance matrix, and molecular branching. *Journal of Chemical Physics*, 67:4517–4533.

[Chambers and Hastie, 1991] Chambers, M. and Hastie, T. (1991). *Statistical Models in S*. Chapman and Hall/CRC.

[Dehmer, 2008] Dehmer, M. (2008). Information processing in complex networks: Graph entropy and information functionals. *Applied Mathematics and Computation*, 201(1-2):82–94.

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

[Diudea et al., 2001] Diudea, M., Gutman, I., and Jäntschi, L. (2001). *Molecular topology*. Nova Science Pub Inc.

[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*. in press.

[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.

[Hage and Harary, 1995] Hage, P. and Harary, F. (1995). Eccentricity and centrality in networks. *Social Networks*, 17(1):57 – 63.

[Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654.

[Kolaczyk, 2009] Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated.

[Konstantinova and Paleev, 1990] Konstantinova, E. V. and Paleev, A. A. (1990). Sensitivity of topological indices of polycyclic graphs. *Vychisl. Sistemy*, 136:38–48. (in Russian).

[Li and Gutman, 2006] Li, X. and Gutman, I. (2006). Mathematical aspects of Randić-type molecular structure descriptors. *Math Chemistry Monographs*.

[Ma and Zeng, 2003] Ma, H.-W. and Zeng, A.-P. (2003). The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11):1423–1430.

[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., 2011] Mueller, L., Kugler, K., Netzer, M., Graber, A., and Dehmer, M. (2011). Distinguishing between the Three Domains of Life using Topological Characteristics of their underlying Metabolic Networks. submitted.

[Mueller et al., 2010] Mueller, L. A., Kugler, K. G., Dander, A., Graber, A., and Dehmer, M. (2010). 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.

[Murtagh, 1985] Murtagh, F. (1985). Multidimensional clustering algorithms. *Compstat Lectures, Vienna: Physika Verlag, 1985*, 1.

[Podani et al., 2001] Podani, J., Oltvai, Z. N., Jeong, H., Tombor, B., Barabási, A. L., and Szathmáry, E. (2001). Comparable system-level organization of Archaea and Eukaryotes. *Nat Genet*, 29(1):54–56.

[Rashewsky, 1955] Rashewsky, N. (1955). Life, Information Theory, and Topology. *Bull Math Biophys*, 17:229–235.

[Strogatz, 2001] Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410(6825):268–276.

[Trucco, 1956] Trucco, E. (1956). A note on the information content of graphs. *Bull. Math. Biol.*, 18(2):129–135.

[Vapnik and Lerner, 1963] Vapnik, V. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774—780.

[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.

[Wagner and Fell, 2001] Wagner, A. and Fell, D. (2001). The small world inside large metabolic networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1478):1803.

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge Univ Pr.

[Wiener, 1947] Wiener, H. (1947). Structural determination of paraffin boiling points. *J. Amer. Chem. Soc.*, 69:17–20.

[Zhu and Qin, 2005] Zhu, D. and Qin, Z. S. (2005). Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics*, 6:8.