Matthias Dehmer[*] / Andreas Holzinger[**] /
Frank Emmert-Streib[***]

Personalized Medicine by Means of Complex Networks – a Big Data Challenge

Inhaltsverzeichnis

This conceptional note deals with the problem of performing personalized medicine by employing network-based techniques. We demonstrate this by way of example when detecting complex diseases.

Acknowledgments: We thank Florent Thouvenin for stimulating discussions concerning this work.

Key words: Personalized Medicine, Complex Networks, Biomarker, Big Data

I. Introduction

The research on personalized medicine has been still intricate to explore medical decision models and products for individual patients[1]. To do so, several decision models by using statistical techniques have been used and optimized[2]. A major

[*] Ao. Professor, Division for Bioinformatics and Translational Research UMIT – The Health and Life Sciences University, Tirol.

[**] Assistenzprofessor, IBM Watson Think Group, Research Unit HCI, Institute for Medical Informatics, Medical University Graz.

[***] Ao. Professor, Computational Biology and Machine Learning Laboratory, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Faculty of Medicine, Health and Life Sciences, Queen's University Belfast.

[1] Shastry B. S., Pharmacogenetics and the concept of individualized medicine. Pharmacogenomics J., 6:16–21, 2006.

[2] Emmert-Streib F., The chronic fatigue syndrome: A comparative pathway analysis. Journal of Computational Biology, 14(7), 2007; Siebert U., Rochau U. / Claxton K., When is

application of personalized medicine has been cancer management and its stratification, i.e., to estimate the cancer risk for the clinical management. In this paper, we sketch a conceptional approach to perform personalized medicine by using complex network analysis[3] where the pathways (networks) are inferred from DNA microarray cancer data[4].

In order to start, we briefly outline the most recent developments in network biology[5]. Afterwards, we state the crucial idea how to represent biomarker as complex networks. Also, we briefly mention mathematical techniques to analyze pseudo pathways quantitatively. The paper ends with a discussion and outlook.

Due to the increasing trend towards personalized, molecular and precision medicine (P4 medicine: Predictive, Preventive, Participatory, Personalized[6]), biomedical data today results from various sources in different structural dimensions, ranging from the microscopic world, and in particular from the omics world (e.g., from genomics, proteomics, metabolomics, lipidomics, transcriptomics, epigenetics, microbiomics, fluxomics, phenomics, etc.) to the macroscopic world (e.g., disease spreading data of populations in public health informatics[7]).

This «Big Data» is not a buzz word, actually we are at a paradigmatic shift of scientific work and medicine is turning into a data intensive science. Whilst, the traditional method of turning data into knowledge relied on manual analysis and interpretation by a medical doctor in order to find useful patterns in data for decision support, now The systematic exploration of this big data can be seen as a

---

enough evidence enough? using systematic decision analysis and value-of-information analysis to determine the need for further evidence. Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 107: 575–584, 2013.

3 Emmert-Streib F. / Dehmer M., (eds), Analysis of Microarray Data: A Network-based Approach. Wiley VCH Publishing, 2010; Emmert-Streib F. / Dehmer M., Networks for systems biology: Conceptual connection of data and function. IET Systems Biology, 5: 185–207, 2011; Junker B. H. / Schreiber F., Analysis of Biological Networks. Wiley Series in Bioinformatics. Wiley-Interscience, 2008.

4 Emmert-Streib F. / Dehmer M., Analysis of Microarray Data (Fn. 3).

5 Barabási A. L. / Oltvai Z. N., Network biology: Understanding the cell's functional organization. Nature Reviews. Genetics, 5(2):101–113, 2004; Emmert-Streib F. / Dehmer M., Networks for systems biology (Fn. 3).

6 Hood L. / Friend S. H., Predictive, personalized, preventive, participatory (P4) cancer medicine. Nature Reviews Clinical Oncology, 8, (3), 2011, 184–187.

7 Holzinger, A., Biomedical Informatics: Discovering Knowledge in Big Data, New York 2014, Springer; Holzinger A. / Stocker C. / Dehmer M., Big complex biomedical data: Towards a taxonomy of data, 2014. Springer; Communications in Computer and Information Science. Berlin, Heidelberg, New York 2014, Springer, pp. in print; Holzinger A. / Dehmer M. / Jurisica I., 2014, Knowledge Discovery and Interactive Data Mining in Bioinformatics – State-of-the-Art, Future challenges and Research Directions. BMC Bioinformatics, to appear.

completely new paradigm in the investigation of nature[8]: Machine learning approaches provide a mechanism for data driven hypotheses generation, optimized experiment planning, precision medicine and evidence-based medicine. The challenge is not only to extract meaningful information from this data, but to gain knowledge, to discover previously unknown insights, to look for patterns, and to make sense of the data[9]. Diverse approaches, including statistical and graph theoretical methods, data mining, and computational pattern recognition, have been applied to this task in the past[10] and although machine learning approaches are indispensable current trends are increasingly putting the human-into-the-loop[11], leaving the grand goals of the 1970ies artificial intelligence to make everything automatic. However, an increasingly important issue is the limited time which medical doctors have in their daily clinical routine: in average a medical doctor in a public hospital has only five minutes to make a decision[12], hence interactive tools for decision support are a necessity, which calls for powerful grid-based computing approaches. Inferring that the future of personalized medicine will be based on mobile technologies and cloud-based technologies, one of the most important future aspects are in privacy, data protection, data security and data safety

---

8 Bell G./Hey T./Szalay A., Beyond the data deluge. Science, 323, (5919), 2009, 1297–1298.

9 Holzinger A./Stocker C./Bruschi M./Auinger A./Silva H./Fred A., On Applying Approximate Entropy to ECG Signals for Knowledge Discovery on the Example of Big Sensor Data. In: R. Huang, E. a. E. (ed.), Active Media Technologies AMT 2012, LNCS 7669, Macau 2012, Springer, pp. 646–657; Holzinger A./Scherer R./Seeber M./Wagner J./Müller-Putz G., Computational Sensemaking on Examples of Knowledge Discovery from Neuroscience Data: Towards Enhancing Stroke Rehabilitation. In: Böhm C./Khuri S./Lhotská L./Renda M. (eds.), Information Technology in Bio- and Medical Informatics, Lecture Notes in Computer Science, LNCS 7451. Heidelberg, New York 2012, Springer, pp. 166–168.

10 Raymer M. L./Doom T. E./Kuhn L. A./Punch W. F, Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm. IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 33, (5), 2003, 802–813; Jurisica I./Mylopoulos J./Glasgow J./Shapiro H./Casper R. F., Case-based reasoning in IVF: prediction and knowledge mining. Artificial intelligence in medicine, 12, (1), 1998,1–24.

11 Holzinger A., Human–Computer Interaction & Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In: Cuzzocrea A./Dimitris C. K./Simos E./Weippl E./Xu L. (ed.) Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127. Heidelberg, Berlin, New York 2013, Springer, pp. 319–328.

12 Gigerenzer G., Gut Feelings: Short Cuts to Better Decision Making London, Penguin 2008; Gigerenzer G./Gaissmaier W., Heuristic Decision Making. In: Fiske S. T./Schacter, D. L./Taylor S. E. (eds.), Annual Review of Psychology, Vol. 62., 2011, Palo Alto: Annual Reviews, pp. 451–482.

and the fair use of data[13]. This calls for data accessibility, legal restrictions, confidentiality and data provenance. Consequently, the field of knowledge discovery and data mining is challenged by a number of non-trivial future aspects to properly address these complex restrictions in dealing with big data.

II. Network Biology

Network-based techniques have been investigated in various disciplines such as chemistry[14], biology[15], ecology[16] and finance[17]. In particular, network biology[18] deals with analyzing biological data sets in the context of systems biology[19]. In systems biology, the underlying biological systems are usually understood as dynamical systems which are investigated holistically[20].

Towards network analysis, two main categories of network analysis may be distinguished: One may use methods from so-called descriptive graph theory[21] or approaches from quantitative graph theory[22]. Methods which belong to the first category describe structural features of a network deterministically. The second category is concerned with developing techniques to quantify structural information by using a measurement approach[23]. In addition, so-called machine learning

---

13 Weippl E. / Holzinger A. / Tjoa A. M., Security aspects of ubiquitous computing in health care. Springer Elektrotechnik & Informationstechnik, e&i, 123, (4), 2006, 156–162.

14 D. Bonchev, Information Theoretic Indices for Characterization of Chemical Structures. Research Studies Press, Chichester, 1983.

15 Diudea M. V. / Gutman I. / Jäntschi L., Molecular Topology. Nova Publishing, 2001. New York, NY, USA; Junker B. H. / Schreiber F., Analysis of Biological Networks. Wiley Series in Bioinformatics. Wiley-Interscience, 2008.

16 Ulanowicz R. E., Quantitative methods for ecological network analysis. Computational Biology and Chemistry, 28: 321–339, 2004.

17 Boginski V. / Butenko S. / Pardalos P., Statistical analysis of financial networks. Computational Statistics and Data Analysis, 48(2): 431–443, 2005.

18 Barab'asi A. L. / Oltvai Z. N., Network biology: Understanding the cell's functional organization. Nature Reviews. Genetics, 5(2): 101–113, 2004.

19 Palsson B. O., Systems Biology: Properties of Reconstructed Networks. Cambridge University Press, 2006.

20 Emmert-Streib F. / Dehmer M., Networks for systems biology (Fn. 3).

21 Harary F., Graph Theory. Addison Wesley Publishing Company, 1969, Reading, MA, USA.

22 Dehmer M. / Emmert-Streib F., Quantitative Graph Theory. Theory and Applications. CRC Press, 2014. to appear.

23 Dehmer M. / Emmert-Streib F. / Mehler M. (eds.), Towards an Information Theory of Complex Networks: Statistical Methods and Applications. Birkhäuser Publishing, 2011; Dehmer M. / Emmert-Streib F. (Fn. 22); Emmert-Streib F. / Dehmer M., Analysis of Microarray Data (Fn. 3).

methods have been used to analyze network data efficiently[24]. On the one hand, network-based techniques have been successfully applied to tackle various questions in systems biology[25]. On the other hand, network-based methods have been proven amazingly efficient and flexible to examine problems related to the function and structure of biological systems in the postgenomics era[26].

In the following, we list some still outstanding and intricate problems in network biology and related areas in brief:.

- Network Biology for Medicine
  - Inferring complex networks from High-Throughput data sets, e.g. cancer[27]
  - Structural Analysis of biological networks[28]
  - Developing information-theoretic and statistical techniques for complex network analysis[29]
- Structure-oriented Drug Design
  - Representing molecule structures as graphs[30]
  - Quantifying the structure of molecules using so-called topological descriptors[31]
  - Predicting chemical or biological activities of the molecules using graph-theoretical and statistical techniques[32]

---

24 Gärtner T./Flach P. A./Wrobel S., On graph kernels: Hardness results and efficient alternatives. In Colt, pages 129–143, 2003; Müller L. A. J./Kugler K. G./Graber A./Dehmer M., A networkbased approach to classify the three domains of life. Biology Direct, 6:140–141, 2011.

25 Emmert-Streib F./Dehmer M., Networks for systems biology (Fn. 3).

26 Emmert-Streib F./Dehmer M., Networks for systems biology (Fn. 3).

27 Emmert-Streib F./Dehmer M., Analysis of Microarray Data (Fn. 3); Emmert-Streib F. (Fn. 2); Emmert-Streib F./Dehmer M., Networks for systems biology (Fn. 3).

28 Emmert-Streib F./Dehmer M., Analysis of Microarray Data, (Fn. 3); Emmert-Streib F. (Fn. 2).

29 Dehmer M., Information processing in complex networks: Graph entropy and information functionals. Appl. Math. Comput., 201: 82–94, 2008; Dehmer M./Emmert-Streib F./Mehler M. (Fn. 23).

30 Diudea M. V./Gutman I./Jäntschi L. (Fn. 15).

31 Dehmer M./Varmuza K./Borgert S./Emmert-Streib F., On entropy-based molecular descriptors: Statistical analysis of real and synthetic chemical structures. J. Chem. Inf. Model., 49: 1655–1663, 2009; Diudea M. V./Gutman I./Jäntschi L. (Fn. 15); Todeschini R., Consonni V., Mannhold R., Handbook of Molecular Descriptors. Wiley-VCH, Weinheim 2002.

32 Basak S. C./Magnuson V. R., Molecular topology and narcosis. Arzeim.-Forsch./Drug Design, 33(I): 501–503, 1983; Dehmer M./Varmuza K./Bonchev D. (eds.), Statistical Modelling of Molecular Descriptors in QSAR/QSPR. Quantitative and Network Biology. Wiley-Blackwell, 2012; Todeschini R./Consonni V./Mannhold R. (Fn. 31).

III. Personalized Medicine by Using Network-Based Approaches

In this section, we describe the problem of performing personalized medicine by using cancer data and complex networks by way of example. More precisely, we consider so-called complex diseases which have been defined due to Strohman[33] for inferring cancer networks.

Definition 3.1 A complex disease can not be defined by a single biomarker (gene) but a group of interacting genes (pathways).

For instance, the Coronary heart disease, Hypertension, Diabetes, Obesity, Various cancers, Alzheimer's disease, and Parkinson's disease are known to be complex diseases. Non-complex diseases, that is Mendelian diseases, are for example Huntington (D), BRCA (breastcancer, D), Cysticfibrosis, and Thrombophilia.

Now we briefly recall the classical view to describe Mendelian diseases on a gene level. If the expression of a gene changes, then the function of the gene changes. Also if the function of a gene changes, the phenotype of the organism changes accordingly. Consequently, we obtain a genotype-phenotype mapping in the Mendelian sense and the phenotype corresponds to the disease. Finally, we obtain that differentially expressed genes are markers of diseases which corresponds to the classical view of biological data analysis.

The new view is based on the hypothesis that biological processes in a cell form dynamical systems that maintain biological functions. That means the biological function is in general a collective effort. This implies that groups of genes participating in a biological process are important and groups of genes interact with each other. Otherwise we were unable to infer any dynamical system.

Ideally, this aims to infer pathways from an individual patient by using high-throughput data sets. Therefore this leads to a procedure to perform personalized medicine. In the following, we give a concrete example to do so when detecting complex diseases[34] and consider Figure 1. We see that two groups of patients exist, here sick versus non-sick. Then pseudo path- ways are inferred from the underlying DNA microarray data, e.g., cancer. Finally Figure 1 says that one can detect complex diseases by comparing the inferred pseudo pathways quantitatively. For instance, we could pairwisely compute the structural similarity between the pathways and then calculate the distance between the obtained similarity distributions.

---

[33] Strohman R., Maneuvering in the complex path from genotype to phenotype. Science, 296(5568): 701–703, 2002.

[34] Emmert-Streib F. / Dehmer M., Analysis of Microarray Data (Fn. 3).

Figure 1: Detecting complex diseases by means of complex networks and comparative network analysis[35].

Key Challenges along with big data include Data in the Cloud, mobile solutions, the trend towards software-as-a-service, and the massive increase in the amount of data, consequently lot of future effort must be spent in Privacy, Data Protection, Security and Safety. The challenges of data integration, data fusion and the increased use of data for secondary use put these issues from a «nice-to-have» into the key interest; just an example: In January 2013, the US Department of Health released the so-called Omnibus Final Rule, which significantly modified the privacy and security standards under the Health Insurance Portability and Accountability Act (HIPAA). These new regulations were driven by a need to ensure the confidentiality, integrity, and security of patients' protected health information (PHI) in electronic health records (EHRs) and addresses these concerns by expanding the scope of regulations and increasing penalties for health information violations[36].

A further big issue – particularly in the context of P4 medicine – is the secondary use of data, providing patient data for clinical and/or medical research. For most secondary data use, it is possible to use de-identified data, but for the remaining data protection issues are very important[37]. The secondary use of data involves the linkage of data sets to bring different modalities of data together, which raises more concerns over the privacy of the data. A classic example is the publication of the Human Genome, which raised new ways of finding relationships between clinical disease and human genetics. The increasing use and storage of P4 relevant data also impacts the use of familial records, since the information about the patient also provides information on the patient's relatives, which can indeed be very useful. Another big issue is the production of anonymized open data sets to support international joint research efforts.

IV. Discussion and Outlook

Inferring and analyzing networks by using DNA microarray data has been proven useful for at least a decade. Consequently, genome-wide high-through-put tech-

35 Emmert-Streib F./Dehmer M., Analysis of Microarray Data (Fn. 3).

36 Wang C./Huang D. J., The HIPAA conundrum in the era of mobile health and communications. JAMA, 310, (11), 2013, 1121–1122.

37 Safran, C./Bloomrosen, M./Hammond W. E./Labkoff S./Markel-Fox S./Tang P. C./Detmer D. E., Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. Journal of the American Medical Informatics Association, 14, (1), 2007, 1–9.

nologies have been proven useful for performing personalized medicine. In fact, applying the underlying mathematical methods bears enormous potential to perform personalized medicine in the future. As huge data sets for inferring and analyzing pathways exist, the issue of data protection is crucial too. This encompasses the whole range of issues from privacy, data protection, safety and security which always must be taken into account, see e.g. hci4all.at[38].

Also, it has been shown that not only the analysis of next-generation sequencing technologies to generate, e.g., DNA sequence data has been useful[39]. In this context, Emmert-Streib and Dehmer[40] recently explored the hypothesis that dynOmics data can be important for enabling personalized medicine by complementing genotype data[41].

A further grand challenge and big issue in research is the interaction with large graphs for graph-based data mining and knowlege discovery, where Holzinger and Dehmer do some pioneering future work[42]. During the last decades, many biological data sets have been growing exponentially in their size. Also, the complexity of those data sets (e.g., high- throughput data) bears potential problems when analyzing the data. For example, the networks might not be deterministically inferrable and, hence, one needs statistical techniques to cope structural errors. This example gives an idea about the complexity of the problem of selecting the right method.

---

38 Kieseberg, P. / Hobel, H. / Schrittwieser, S. / Weippl, E. / Holzinger, A. 2014. Protecting Anonymity in the Data-Driven Medical Sciences. *In:* Holzinger, A. / Jurisica, I. (eds.) *Interactive Knowledge Discovery and Data Mining: State-of-the-Art and Future Challenges in Biomedical Informatics, Springer Lecture Notes in Computer Science LNCS 8401.* Berlin, Heidelberg: Springer, pp. in print.

39 Emmert-Streib F. / Dehmer M., Enhancing systems medicine beyond genotype data by dynamic patient signatures: Having information and using it too. Frontiers in Genetics, 4(241), 2013.

40 Emmert-Streib F. / Dehmer M. (Fn. 38).

41 Emmert-Streib F. / Dehmer M. (Fn. 38).

42 Holzinger, A. / Ofner, B. / Stocker, C. / Valdez, A. C. / Schaar, A. K. / Ziefle, M. / Dehmer, M. 2013. On Graph Entropy Measures for Knowledge Discovery from Publication Network Data. *In:* Cuzzocrea, A. / Kittl, C. / Simos, D. E. / Weippl, E. / Xu, L. (eds.) *Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127.* Heidelberg, Berlin: Springer, pp. 354–362. Holzinger, A. / Ofner, B. / Dehmer, M. 2014. Multi-touch Graph-Based Interaction for Knowledge Discovery on Mobile Devices: State-of-the-Art and Future Challenges. *In:* Andreas Holzinger, I. J. (ed.) *Interactive Knowledge Discovery and Data Mining: State-of-the-Art and Future Challenges in Biomedical Informatics, Springer Lecture Notes in Computer Science LNCS 8401.* Berlin, Heidelberg: Springer, pp. in print.