Report from Dagstuhl Seminar 13212

Computational Methods Aiding Early-Stage Drug Design

Edited by

Andreas Bender[1], Hinrich Göhlmann[2], Sepp Hochreiter[3], and Ziv Shkedy[4]

1   **University of Cambridge, GB, ab454@cam.ac.uk**
2   **Janssen Pharmaceuticals – Beerse, BE**
3   **University of Linz, AT, hochreit@bioinf.jku.at**
4   **Hasselt University – Diepenbeek, BE, ziv.shkedy@uhasselt.be**

—— **Abstract** ——

This report documents the program and the outcomes of Dagstuhl Seminar 13212 "Computational Methods Aiding Early-Stage Drug Design". The aim of the seminar was to bring scientists working on various aspects of drug discovery, genomic technologies and computational science (e.g., bioinformatics, chemoinformatics, machine learning, and statistics) together to explore how high dimensional data sets created by genomic technologies can be integrated to identify functional manifestations of drug actions on living cells early in the drug discovery process.

1   Executive Summary

Andreas Bender
Hinrich Göhlmann
Sepp Hochreiter
Ziv Shkedy

Besides discussing scientific findings enabled by computational approaches, the seminar successfully stimulated discussions between scientists from different disciplines and provided an exceptional opportunity to create mutual understanding of the various challenges and opportunities. It created understanding for technical terms and concepts and served as a catalyst to explore new ideas.

As a concrete example, it challenged the feasibility of utilizing chemical structure information for identifying correlations with biological data. Rather than attempting to define a most suitable way of translating chemical structure information into computer understandable form (e.g., via fingerprinting algorithms such as ECFP), the notion of utilizing functional readouts such as gene expression profiles was favored for prioritizing candidate drugs that demonstrate a favorable balance of desired and undesired compound effects.

Computational Methods Aiding Early-Stage Drug Design, *Dagstuhl Reports*, Vol. 3, Issue 5, pp. 78–94
Editors: Andreas Bender, Hinrich Göhlmann, Sepp Hochreiter, and Ziv Shkedy
        DAGSTUHL   Dagstuhl Reports
        REPORTS   Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2   Table of Contents

3 Overview of Talks

3.1 Enriched methods for analysis of high-dimensional data

Dhammika Amaratunga (Johnson & Johnson, US)

High-dimensional data are characterized as having an enormous number of features and relatively few samples. Exploration and classification of such data is an important aspect of bioinformatics and cheminformatics work. One of the challenges presented by such situations is that only a very small percentage of the features actually carries classification information; the other features add noise or carry secondary signals that could seriously obscure the true signal. In such situations, it is helpful to use methods that highlight features of the data that are most likely to be informative. We refer to such methods as enriched methods. Enriched methods have been developed for single-run procedures, such as SVD, as well for multiple-run (ensemble) procedures, such as Random Forest. Here we will discuss many issues that arise in this context and demonstrate the value of enriched methods in analyzing high-dimensional data.

3.2 Similarity-Based Clustering of Compounds and its Application to Knowledge Discovery from Kernel-based QSAR Models

Ulrich Bodenhofer (University of Linz, AT)

Joint work of Bodenhofer, Ulrich; Klambauer, Günter; Palme, Johannes; Kothmeier, Andreas; Hochreiter, Sepp
Main reference U. Bodenhofer, A. Kothmeier, S. Hochreiter, "APCluster: an R package for affinity propagation clustering," Bioinformatics, 27:2463–2464, 2011.
URL http://dx.doi.org/10.1093/bioinformatics/btr406

Quantitative structure-activity relationships (QSAR) have become a standard methodology in computational pharmacology and computer-aided drug design. In recent years, kernel-based approaches have been established as an alternative to traditional feature-based approaches using chemical descriptors or structural descriptors like ECFP fingerprints. Kernels can incorporate virtually any kind of chemical or structural information, as far as 3D structures. Many common kernels even facilitate good model interpretability similarly to feature- based approaches. This can be accomplished by the extraction of explicit feature weights and the superimposition of feature weights on given chemical structures to highlight sub-structures that are particulary relevant for the given modeling task. The only painful drawback of kernel approaches is their poor ability to scale to larger data sets.

In this contribution, we advocate the use of affinity propagation (AP) clustering for selecting representative samples/compounds, the so-called exemplars, with the following two possible applications: (1) If a set of exemplars is available, the kernel matrix can be compacted by removing all columns that do not correspond to exemplars. The Potential Support Vector Machine (P-SVM), for example, can process such non-quadratic kernel matrices and, thereby, leverage the scaling problem mentioned above. (2) As mentioned above, the superimposition of feature weights on chemical structures to highlight relevant sub-structures is an excellent tool for knowledge acquisition, but it can only be applied to a very limited number of samples. It seems natural to priorize samples/compounds that are

known — on the basis of an objective procedure — to be most representative for the entire data set of compounds.

It is worth to note that the use of AP clustering is essential for these two applications. Firstly, AP is similarity-based, therefore, it is not limited to explicit feature representations, but can be used for all kinds of kernels as well. Secondly, AP is able to compute exemplars that are members of the original data sets (as opposed to hypothetical averages used by k-means clustering). Last but not least, AP is efficient and, with appropriate computational strategies, it scales to large data sets.

3.3 Protein family focused, structure enabled chemical probes, to accelerate the discovery of new targets

Chas Bountra (University of Oxford – Nuffield College, GB)

The discovery of "pioneer medicines" (i.e. those acting via novel molecular targets) has proven to be an immensely complex, long term, expensive and high risk endeavour. During my presentation, I will discuss

- our focus on novel human epigenetic protein families,
- the generation of freely available inhibitors, and
- the partnership with many academic and industrial labs

I will describe progress with novel inhibitors for bromodomain and demethylase proteins, and their use in identifying new targets in can cer, inflammatory and neuro-psychiatric diseases.

3.4 Combining transcriptomics, bioassays and chemistry to aid drug discovery

Hinrich Göhlmann (Janssen Pharmaceutica – Beerse, BE)

The joint talk introduced the participants of the seminar to the steps and challenges of the drug discovery and development process. Janssen has collaborated over the past years with several universities to explore how transcriptomic data can be used to aid and overcome these challenges. With the financial support of the flemish IWT we have initially developed algorithms and approaches for defining and prioritizing compound clusters of equipotent compounds that were active in a phenotypic screen. In the second research and development project (QSTAR) we are attempting to combine transcriptomic data with data of bioassays and chemical structure information. Modelling either two of the three data types or jointly modelling all data we investigate what correlations are present in the data. As an essential tool for facilitating the collaboration between the different partners we have developed data processing pipelines that generate robust data structures that link all three data types via unified compound identifiers (InChIKey). By jointly creating R packages for data access as well as data analysis we hope to create the foundation that will allow us and other interested research teams to investigate what connections are present in the data and how we can use the information to aid drug discovery in the future.

3.5 False discovery proportions of gene lists prioritized by the user

Jelle J. Goeman (Leiden University Medical Center, NL)

Motivated by the practice of exploratory research, we formulate an approach to multiple testing that reverses the conventional roles of the user and the multiple testing procedure. Traditionally, the user chooses the error criterion, and the procedure the resulting rejected set. Instead, we propose to let the user choose the rejected set freely, and to let the multiple testing procedure return a confidence statement on the number of false rejections incurred. In our approach, such confidence statements are simultaneous for all choices of the rejected set, so that post hoc selection of the rejected set does not compromise their validity. The proposed reversal of roles requires nothing more than a review of the familiar closed testing procedure, but with a focus on the non-consonant rejections that this procedure makes. We suggest several shortcuts to avoid the computational problems associated with closed testing.

3.6 Systematic mapping of synthetic genetic interactions with combinatorial RNAi

Wolfgang Huber (EMBL Heidelberg, DE)

Biological systems are able to buffer the effects of individual mutations, and disease outcomes often depend on the combination of multiple genetic variants. Genetic interactions have been systematically measured in yeast and enabled the placement of genes into functional modules and the delineation of networks between modules at unprecedented coverage. However, such approaches have not been feasible for higher organisms. In this talk, I will report on our high-resolution genetic interaction maps of chromatin-related genes in Drosophila and human cells, obtained by combinatorial perturbation via RNA interference and single-cell phenotyping by automated imaging. Genetic interaction profiles were obtained by measuring multiple, non-redundant cellular phenotypes. The analysis of profiles revealed functional modules, among them many conserved protein complexes. Comparison with yeast showed a consistent, evolutionarily conserved pattern of genetic interactions for the substructures of the mediator complex, but also revealed the functional divergence of the kinase module Cdk8 and CycC in Drosophila. Genetic epistasis is an unresolved frontier of cancer genetics, as sequencing projects found that combinations of multiple, partially alternative and individually rare mutations lead to equivalent phenotypes. To dissect such interdependencies, we mapped recently reported recurrent cancer mutations onto our network and grouped them into clusters of putatively equivalent network functions.

3.7   Drug-induced transcriptional modules in mammalian biology: implications for drug repositioning and resistance

Murat Iskar (EMBL Heidelberg, DE)

In recent years, the publicly available data on small molecules has increased dramatically. Integrative analysis of these heterogeneous resources enables us to gain a better understanding of drug action in biological systems. Genome-wide expression profiling of cells treated with drugs summarizes the pharmacological and toxicological effects of these perturbations at the molecular level and further help us to bridge between the molecular basis of drug action and their phenotypic consequences. To systematically explore the biological responses of mammalian cells to a diverse set of chemical perturbations, we generated a comprehensive collection of drug-induced transcriptional modules from existing microarray data on drug-treated human cell lines and rat liver. More than 70% of these modules were identified in multiple human cell lines and 15% were conserved across organisms of human and rat, representing a lower limit. We systematically characterized these modules and could link antipsychotic drugs to sterol and cholesterol biosynthesis, providing an explanation for the metabolic side effects reported for these drugs. Moreover, we could identify novel functional roles for hypothetical genes, e.g. ten new modulators of cellular cholesterol levels and novel therapeutic roles for several drugs, e.g. new cell cycle blockers and modulators of alpha-adrenergic, PPAR and estrogen receptors. Our work not only quantifies the conservation of transcriptional responses across biological systems, but also identifies novel associations between drug-induced transcriptional modules, drug targets and side effects.

3.8   Semi-supervised investigation of association of gene expression with structural fingerprints of chemical compounds

Adetayo Kasim (Durham University, GB)

Exploring the relationship between a chemical structure and its biological function is of great importance for drug discovery. Whilst many studies attempt to introduce transcriptomics data into chemical function, little effort has been made to link structural fingerprints of compounds with defined intracellular functions such as target related pathways or expression of particular set of genes. Li *et al.* (2011) propose an approach to associate structural differences between compounds with the expression level of a defined set of genes by performing clustering on chemical structures to find differentially expressed genes between adjacent clusters of compounds from the same node. The identified set of genes were further subjected

to compounds re-classification to evaluate the accuracy of the prediction based on gene expression and chemical structures.

We propose a semi-supervised approach for investigation of association of gene expression with structural finger prints. Our approach starts with unsupervised biclustering of gene expression to identify subset of genes and compound with target related pathways. The expression levels of the relevant genes are used to weight structural fingerprints of chemical compounds to obtain clusters of compounds with a common set of structural fingerprints and similar level of gene expression. A similar approach was also applied to identify clusters of compounds with a common set of fingerprints and similar bioassay level.

3.9 Multi-view learning for drug sensitivity prediction

Samuel Kaski (Aalto University, FI)

License ⓒ Creative Commons BY 3.0 Unported license
© Samuel Kaski

We are developing machine learning methods for integrating multiple high- dimensional data sources. In the unsupervised task of decomposing the sources into shared and source-specific components, the new Bayesian Canonical Correlation Analyses and Group Factor Analyses can be applied to study omics- wide effects of chemical structures. The supervised personalized medicine task of predicting drug sensitivity based on multiple genomic measurement sources, can be addressed by a combination of multi-view and multi-task learning. This is joint work with several people from my group, and collaborators from Institute for Molecular Medicine Finland FIMM. For more details and code see http://research.ics.aalto.fi/mi.

3.10 Detecting differentially expressed genes in RNA-Seq drug design studies

Guenter Klambauer (University of Linz, AT)

License ⓒ Creative Commons BY 3.0 Unported license
© Guenter Klambauer
Joint work of Klambauer, Guenter; Unterhiner, Thomas; Hochreiter, Sepp
URL http://www.bioinf.jku.at/software/dexus/

Detection of differential expression in RNA-Seq data is currently limited to studies in which two or more sample conditions are known a priori. However, these biological conditions are typically unknown in drug design studies. We present DEXUS for detecting differential expression in RNA-Seq data for which the sample conditions are unknown. DEXUS models read counts as a finite mixture of negative binomial distributions in which each mixture component corresponds to a condition. A transcript is considered differentially expressed if modeling of its read counts requires more than one condition. DEXUS decomposes read count variation into variation due to noise and variation due to differential expression. Evidence of differential expression is measured by the informative/non-informative (I/NI) value, which allows differentially expressed transcripts to be extracted at a desired specificity (significance level) or sensitivity (power). DEXUS performed excellently in identifying differentially expressed transcripts in data with unknown conditions. On 2,400 simulated data sets, I/NI value thresholds of 0.025, 0.05, and 0.1 yielded average specificities of 92%, 97%, and 99%

at sensitivities of 76%, 61%, and 38% respectively. On real-world data sets, DEXUS was able to detect differentially expressed transcripts related to sex, species, tissue, structural variants, or eQTLs.

3.11 Intestinal microbiota, individuality and health

Leo Lahti (Wageningen University, NL)

Diverse microbial communities inhabit the human gastrointestinal tract, where hundreds of distinct bacterial phylotypes and a trillion bacterial cells per gram in a healthy adult individual can be encountered. This ecosystem constitutes a virtual metabolic organ that has a central role in nutrition, immune system and other bodily functions, and a profound impact on our well-being.

Recent accumulation of high-throughput profiling data sets is now for the first time enabling global characterization of the overall composition and variability of this intestinal microbiota. Integration of phylogenetic profiling data of a thousand phylotypes across thousands of human individuals scales up the current analyses by an order of magnitude based on the Human Intestinal Tract chip (HITChip), a phylogenetic microarray has enabled standardized data collection of over one thousand gut-specific bacterial phylotypes including many less abundant species that cannot be cultivated in a laboratory and whose functional role is less well known.

The analysis reveals huge inter-individual variability in microbial diversity as well as alternative ecosystem states that are associated with personal environmental and phenotypic factors such as ageing, overweight, host metabolism, and health status. We will discuss how these recent observations provide new insights into the role of our co-evolved microbial partners in individual health and well-being, as well as guidance for the design and interpretation of future studies.

3.12 Library-Scale Gene-Expression Profiling and Digital Open Innovation

Justin Lamb (Genometry Inc – Cambridge, US)

The Broad Institute's Connectivity Map project (www.broadinstitute.org/cmap) has demonstrated the value of a database of gene-expression profiles derived from cultured cells treated with a large collection of bioactive small molecules for drug discovery and development applications. It has also shown how exposing these data and allied search algorithms to the global biomedical-research community through a simple self-service webtool can successfully digitize and democratize the small-molecule screening process. The talk will review this earlier work then describe our efforts to greatly expand the Connectivity Map using a novel high-throughput low-cost gene-expression profiling technology we have developed. The idea

that a comparable system populated with expression profiles of a pharmaceutical company's proprietary chemical matter could serve as an efficient open-innovation platform will also be discussed.

3.13 A Maximum Common Subgraph Kernel Method for Predicting the Chromosome Aberration Test

Johannes Mohr (TU Berlin, DE)

License	ⓒ Creative Commons BY 3.0 Unported license
	© Johannes Mohr
Joint work of	Mohr, Johannes; Jain, B. ; Sutter, A.; Ter Laak, A.; Steger-Hartmann, T.; Heinrich, H.; Obermayer, K.
Main reference	J. Mohr, B. Jain, A. Sutter, A. Ter Laak, T. Steger-Hartmann, N. Heinrich, K. Obermayer, "Maximum Common Subgraph Kernel Method for Predicting the Chromosome Aberration Test," J. Chem. Inf. Modeling, 50(10), pp. 1821–1838, 2010.
URL	http://dx.doi.org/10.1021/ci900367j

The chromosome aberration test is frequently used for the assessment of the potential of chemicals and drugs to elicit genetic damage in mammalian cells in vitro. Due to the limitations of experimental genotoxicity testing in early drug discovery phases, a model to predict the chromosome aberration test yielding high accuracy and providing guidance for structure optimization is urgently needed. In this talk I will present a machine learning approach for predicting the outcome of this assay based on the structure of the investigated compound. It combines a maximum common subgraph kernel for measuring the similarity of two chemical graphs with the potential support vector machine for classification. The approach allows visualizing structural elements with high positive or negative contribution to the class decision.

3.14 Recursive Neural Networks for Undirected Graphs and Neural Network Pairwise Interaction Fields for annotating 2D and 3D small molecules

Gianluca Pollastri (University College Dublin, IE)

License	ⓒ Creative Commons BY 3.0 Unported license
	© Gianluca Pollastri
Joint work of	Pollastri, Gianluca; Lusci, Alessandro; Baldi, Pierre
Main reference	A.Lusci, G.Pollastri, P. Baldi, "Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules," Journal of Chemical Information and Modeling, 53(7), pp. 1563–1575, 2013
URL	http://pubs.acs.org/doi/abs/10.1021/ci400187y

I introduced two ways of "wiring" Artificial Neural Networks, that we have developed in my group, which can deal with structured data in the form of graphs. The first one, UG-RNN (Recursive Neural Networks for Undirected Graphs) factorises an undirected graph into a number of directed graphs that are used to transfer contextual information, and compresses this contextual information into a feature vector which can be mapped into a desired target property. The second model, NN-PIF (Neural Network Pairwise Interaction Field), subdivides a graph into all the pairwise interactions between its nodes (each, potentially, represented alongside a context of neighbours) and maps these pairwise interactions into a feature vector, which is then mapped into a target property. In both cases the feature vector is automatically

generated, and is effectively a fixed-size property-driven adaptive representation of the input. We have tested both models on a number of problems in the space of small molecules, in their 2D representation (in the case of UG-RNN) and 3D representation (for NN-PIF). I described the results we have obtained in these tests, which are generally comparable, and often superior, to those obtained by state of the art 2D and 3D kernels on the same sets. I also speculated on the feature vectors produced by both models, and on how they may be used for mapping the input space.

3.15   The nature of gene signature development

Willem Talloen (Janssen Pharmaceutica – Beerse, BE)

We now have entered the post-genomic era with much hope to harvest some of the fruits hidden in the genomic text. At the same time, the current difficulties faced by pharma research to discover generally applicable block-buster drugs have lead to think in terms of personalized medicine. Consequently, high hopes are on clinical opportunities for gene expression-based prediction of illness or drug response discovered using high-content technologies such as microarrays or RNAseq.

The 'omics revolution was also warmly welcomed by data analysts as its data properties imposed new and interesting statistical challenges. For example, the quest for biomarkers in the context of personalized medicine has made many statisticians and bioinformaticians think about classification models that are robust against overfitting for generation of molecular signatures.

Here, we will challenge that this enthusiasm made many researchers forget to think about the practical applicability and the biological nature, and hence clinical relevance of these developed classification algorithms. For example, the rationale behind signatures consisting out of many genes is generally overlooked. How these genes should be aggregated into one composite index (i.e., the marker) so as to reflect the underlying biology as well as to remain generalizable will be discussed in this presentation.

3.16   Scaling bioinformatics algorithms

Oswaldo Trelles (University of Malaga, ES)

Large scale genomics projects exploiting high throughput leading technology have produced and continue to produce massive data sets with exponential growing rates. So far, only a small part of this data can be abstracted, managed and processed, giving an incomplete understanding of the biological process being observed. The lack of processing power is a bottle neck in acquiring results.

A promising approach to address the processing of such massive data sets is the creation of new computer software that makes effective use of parallel and cloud computing.

Comparative genomics is a good example since it includes all the ingredients: huge and ever growing datasets, complex applications that demands large computational resources and new mathematical and statistical models for analysing and synthetizing genomic information.

This talk will provide an overview of cloud computing -from the user perspective- and the ways to exploit it with a real implementation in the framework of the Mr.SYmbiomath project

3.17 Minor Variant Detection In Virology with Model Based Clustering

Bie Verbist (Ghent University, BE)

Deep-sequencing is one of the applications of the new massively parallel sequencing (MPS) technologies allowing for an in-depth characterization of sequence variation in more complex populations, including low-frequency viral strains. However, MPS technology-associated errors in the resulting DNA sequences may occur up to equal or even higher frequency than the truly present mutations in the biological sample, impeding a powerful assessment of low- frequency virus mutations. As there are no obvious solutions to reduce the technical noise by further improvements of the technology platform, we believe that the search for statistical algorithms that can better correct the technical noise can be pivotal. Therefore algorithms that increase detection power in presence of technical noise and quantify base-call reliability are required. Phred-like quality scores, provided with the base-calls are such a quantification of the base-call reliability. These quality scores together with other covariates determine the multinomial model structure in a model-based clustering approach which will allow identification of viral quasi-species. This research program was granted by IWT, a governmental agency for Innovation by Science and Technology in Flanders, Belgium.

4   Scientific Background

(Written by Andreas Bender, Hinrich Göhlmann, Sepp Hochreiter, Ziv Shkedy)

4.1   Motivation

The efficiency and effectiveness of drug discovery has been challenged over the past years as increasing numbers of drug candidates failed to reach the market and patients. Accordingly, many efforts are underway to increase the productivity of the R&D process and avoid expensive late-stage clinical failures. A key concept is to de-risk drug candidates during the early preclinical stages. Toward this end, it is essential to reduce the time gap between the selection of promising candidate compounds (chemotypes) and the identification of potential side effects in later toxicity studies. In other words, relevant biologically data on the various desired and undesired effects of compounds need to be acquired early on in the research process.

At the same time various modern molecular biology technologies (e.g., next-generation sequencing, microarrays, high content screening) have advanced our understanding of the molecular basis of diseases and drug actions. One approach to increase the productivity

of drug discovery is to complement traditional pharmacology approaches by using modern molecular biology technologies together with computational techniques, e.g., studying the effects of drugs on a cell line. These new opportunities, in particular, the integration of gene expression data on a large scale, complement the established methods in computational chemistry that are focused on the chemical structures. Consequently, drug designers have to become able to interrelate and interpret transciptomic, genetic, proteomic, metabolomic, and assay data.

The aim of the seminar "Computational Methods Aiding Early-Stage Drug Design" was to bring scientists working on various aspects of drug discovery, genomic technologies and computational science (e.g., bioinformatics, chemoinformatics, machine learning, and statistics) together to explore how high dimensional data sets created by genomic technologies can be integrated to identify functional manifestations of drug actions on living cells early in the drug discovery process.

## 4.2   Previous Work and State-of-the-Art

The focus of the seminar is on utilizing computational methods in early-stage drug design. Pharmaceutical companies have large libraries of compounds at their disposal which they investigate for their potential medical applicability. In the early stage, compounds that seem promising to become drugs are selected for subsequent drug development phases.

Currently, two main branches to drug design are employed. On the one hand, structure-based drug design is based on predicting which compound binds to a specific target under investigation. Standard methods for target-compound docking are only feasible for screening a small set of compounds and include docking algorithms, molecular dynamics (force fields), and even simulations of quantum mechanics effects, which are all based on the 3D structure of the target. For screening many compounds simultaneously, specific features of the 3D structure of the target are incorporated into computational methods that predict whether a compound binds to the target.

On the other hand, the current practice in many pharmaceutical companies is to apply ligand-based drug design, i.e. to screen their huge compound libraries for possible drug candidates. Ligand-based drug design is often based on "analoging", where the activity of compounds that are similar to known active compounds is predicted by means of computational models. These analoging models are generated on the basis of similarity of compounds whose biological activity or target specificity has been verified experimentally. Analoging is mostly based on pharmacophores (target-specific functional groups in the compounds) and structure-activity relationship (QSAR) models which are used to represent the relationship between compound properties and the biological activity of the compound.

Ligand-based drug design typically consists of the following steps (see Fig. 1):

1. **Compound filtering**: From a large library of chemical compounds, a sub-selection of feasible compounds is chosen. Criteria for this filtering include, but need not be limited to, the following: solubility, permeability, (non-)toxicity, chemical feasibility, costs, and patent issues.
2. **Compound selection**: From the set of feasible compounds, those are chosen that seem to have the desired effect related to the disease under investigation. This selection may be based on phenotypic screens, disease-specific bioassays, or other techniques. Analoging is an approach to identify new active (having the desired effect) candidates, which are found by their similarity to existing active compounds.
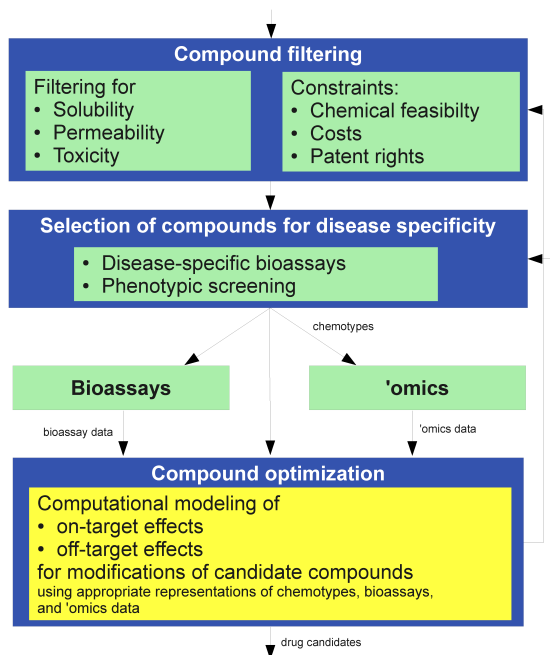
**Figure 1** Overview of the steps during ligand-based drug design.

3. **Bioassay measurements**: A large set of bioassays and various 'omics technologies (e.g. gene expression measurements) are used to assess on-target and off-target effects of the selected compounds.

4. **Compound optimization**: a small set of so-called lead compounds is selected; potentially promising modifications of these lead compounds are considered in order to maximize target effects and minimize off-target effects. The most promising lead compounds are fed back into steps 1, 2, or 3 to verify their expected properties experimentally.

That only a limited number of modifications can be studied experimentally is the central bottleneck of this procedure. Computational methods provide a powerful tool set to avoid this bottleneck. They are predominantly applied in the design-make-test cycle for compound optimization to screen large sets of compound modifications for the most promising candidates (analoging) that can then be tested experimentally. Secondly, computational methods are also highly relevant for predicting the solubility, permeability and toxicity of the compounds which further reduces the number of compounds which need to be tested experimentally.

For compound optimization, ideally, three kinds of data sets are available for each compound (see Fig. 2):

1. **chemical structure and properties**
2. **'omics data**, e.g. transcriptome (in one condition/cell line or in several conditions/cell lines)
3. **phenotypic data** (biological assays)

Using these data, compounds are optimized for being more effective (on-target) and, at the same time, for having less side effects (off-target). On-target and off-target effects are determined by bioassays and by 'omics technologies.
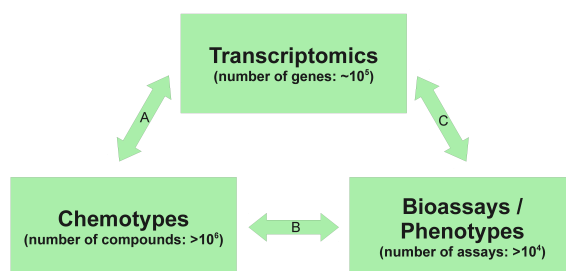
Figure 2 Available data for compound optimization.

The activity of a compound related to a (desired or undesired) target is predicted by classification or regression models (see arrow B in Fig. 2), where structural descriptions of compounds or, alternatively, similarities between compounds are used to predict the outcome of a phenotypic screen or a specific bioassay. Using these classification and regression models, new compounds can be found or existing compounds can be modified such that off-target effects are minimized while on-target effects are maximized. In practice, an objective has to be defined which trades on-target against off-target effects (it is common to speak of "Multifactorial Compound Optimization").

Biological activity including on-target and off-target effects can be measured by 'omics technologies. Again classification and regression models are constructed which now predict biological activity given by 'omics data from the structural description or similarities of compounds (see arrow A in Fig. 2).

To summarize, by interrelating chemistry, phenotype, and 'omics data, on-target and off-target effects can be predicted for a large set of candidate compounds in the compound optimization step, of which only the most promising ones need to undergo experimental validation.

All the data mentioned above (see Fig. 2) are typically high-dimensional, noisy, and technically biased, while only few samples or cell lines are available. These properties of the data entail the demand for advanced machine learning techniques and cutting-edge statistics. Data analysis starts with quality control and preprocessing. Then filtering techniques are needed to reduce dimensionality and finally structures in the data have to be identified. In a next level, these different data sources have to be combined and dependencies have to be recognized.

Summarizing, new computer science tools are required to tackle the computational challenges in early drug design. The more advanced those methods are, the more they are robust to data deficiencies, and the better the interplay between the different steps, the more helpful the results will be for early-stage drug development.

4.3   Conclusions

Combining high dimensional data from genomic technologies with chemical information of structures and classical measurements of compound effects in biological assays (e.g., biochemical or phenotypic assays) the seminar participants discussed various ways of how these data could complement established methods in computational chemistry. See Fig. 3 for topics covered in the seminar.

Figure 3 Summary of the topics covered in the seminar.

Being able to discover and utilize connections between the three data types has also been the focus of the QSTAR project (IWT-funded project of Janssen and academia). Some success stories of using transcriptomics in early drug design were presented by members of the QSTAR consortium. Examples of methodology being explored has been presented were transcriptomics was related to chemistry and to biological assays.

References

1 Mahrenholz C., Abfalter I., Bodenhofer U., Volkmer R., Hochreiter S.: Complex networks govern coiled coil oligomerization – Predicting and profiling by means of a machine learning approach, Mol Cell Proteomics, 2011 (doi: 10.1074/mcp.M110.004994)
2 Hochreiter S., Bodenhofer U., Heusel M., Mayr A., Mitterecker A., Kasim A., Khamiakova T., Van Sanden S., Lin D., Talloen W., Bijnens L., Göhlmann H., Shkedy Z., Clevert D.: FABIA: Factor Analysis for Bicluster Acquisition, Bioinformatics, 26(12): 1520–1527, 2010 (doi: 10.1093/bioinformatics/btq227)
3 Talloen W., Hochreiter S., Bijnens L., Kasim A., Shkedy Z., Amaratunga D., Göhlmann, H.: Filtering data from high-throughput experiments based on measurement reliability, National Academy of Sciences of the United States of America, 107(46): E173–E174, 2010 (doi: 10.1073/pnas.1010604107)
4 Hochreiter S., Clevert D., Obermayer K.: A new summarization method for affymetrix probe level data, Bioinformatics, 22(8): 943–949, 2006 (doi: 10.1093/bioinformatics/btl033)

Participants

- Dhammika Amaratunga
Johnson & Johnson, US
- Andreas Bender
University of Cambridge, GB
- Ulrich Bodenhofer
University of Linz, AT
- Chas Bountra
University of Oxford, GB
- Javier Cabrera
Rutgers Univ. – Piscataway, US
- Aakash Chavan Ravindranath
University of Cambridge, GB
- Hinrich Göhlmann
Janssen Pharmaceutica –
Beerse, BE
- Jelle J. Goeman
Leiden University Medical
Center, NL

- Sepp Hochreiter
University of Linz, AT
- Wolfgang Huber
EMBL Heidelberg, DE
- Murat Iskar
EMBL Heidelberg, DE
- Adetayo Kasim
Durham University, GB
- Samuel Kaski
Aalto University, FI
- Günter Klambauer
University of Linz, AT
- Leo Lahti
Wageningen University, NL
- Justin Lamb
Genometry Inc – Cambridge, US

- Johannes Mohr
TU Berlin, DE
- Gianluca Pollastri
University College Dublin, IE
- Ziv Shkedy
Hasselt Univ. – Diepenbeek, BE
- Willem Talloen
Janssen Pharmaceutica –
Beerse, BE
- Oswaldo Trelles
University of Malaga, ES
- Bie Verbist
Ghent University, BE
- Jörg Kurt Wegner
Janssen Pharmaceutica –
Beerse, BE