# Characterizing compounds and therapeutic targets with novel data engineering techniques

Ph.D. thesis

**Gergely Botond Temesi**

Semmelweis University
Molecular Medicine Doctoral School



Supervisor:
Dr. Csaba Szalai, Ph.D., D.Sc.

Opponents:
Dr. Attila Gulyás-Kovács, Ph.D.
Dr. Zsolt Rónai, M.D., Ph.D.

Final exam chairman:
Dr. Barna Vásárhelyi, M.D., Ph.D., D.Sc.

Final exam members:
Dr. Miklós Cserző, Ph.D.
Dr. Orsolya Szakács, Ph.D.

Budapest
2014

# Introduction

The new millennium has opened new horizons in several science, technology and economic areas. Healthcare spending related to aging in western societies has been posing an increasing economical challenge for decades. The pharmaceutical industry has experienced growing R&D spending every year in the last decade, while the number of approved new molecular entities is decreasing. These trends are neither sustainable on the level of the national economies nor on the level of the industrial sectors. Meanwhile, life sciences and primarily molecular life sciences have gone through enormous development in the last 50 years. Translating achievements of genomics to clinical practices seems to be slower than expected, but high-throughput measurement techniques provide more data about biological systems than ever before. Breakthroughs in computers science have made an even bigger impact, digital data and information volumes have been exponentially growing since the '60s ("Big Data").

An increased utilization of the information technologies could be a key asset to break the unsustainable business models of the healthcare and pharmaceutical industry. Based on these findings and on my previous MSc Computer Science and MSc Biomedical Engineering studies my goal in this PhD research was to contribute to the imminent transformation of the healthcare system and the pharmaceutical industry with information technology advancements. I was focusing on two parallel topics.

In my first topic ("asthma genetics") I studied the pathomechanism of asthma and atopy where I mainly focused on the genetic background. In this case my goal was to perform a complete biomedical discovery study: from the study design based on former animal model studies, through building a human biobank collection and multi-level omics measurements to evaluation with modern information technology tools. In this case information technology methods were tools used for the study design and systems based analysis of the data, and I primarily used the results of our former software development projects.

The focus of my second topic ("enrichment analysis") was the development of a new information technology methodology for pharmaceutical applications which I also tested on a domain specific problem. The technique is capable of predicting different properties (e.g. possible indications) of new molecular entities based on the heterogeneous information and measurement data of well known compounds. Retrospective testing was carried out with amantadine which is a compound known for its multiple indications so we could investigate the benefits and limits of the method.

**Asthma genetics**

Asthma is a chronic, complex multifactorial disease of the airways; both genetic and environmental factors are responsible for its occurrence. More people suffer from allergies and asthma than from all other chronic diseases. The symptoms and the therapeutic response are very heterogeneous, several endophenotypes can be distinguished. The significant molecular biology advancements of last decade made it possible to investigate several associated genes and pathomechanisms but in spite of the growing knowledge base the real causes are still unclear. Over 120 genes were found to be associated with asthma or atopy but only 10 were verified by more than ten studies. The state of the art considering most of the multifactorial diseases and traits are very similar. These difficulties give birth to several new hypotheses and new research strategies ("missing heritability", "dark matter of genetics").

Based on the technical capabilities of the age hypothesis driven discovery approaches were used almost exclusively till the end of the '80s: the expensive measurement techniques yielded only small amount of data, which were evaluated by the so called frequentist statistical hypothesis testing. Starting from the '90s the high-throughput measurement and information technology advancement made it possible to measure complete omics levels (e.g. gene expression with microarrays) and the emphasis has shifted towards hypothesis free study design and evaluation. The dimension of the data provided by the new methods is difficult to handle with the traditional statistical methods, thus it is advisable to apply Bayesian inference, an alternative normative framework. Our research group investigated, implemented and applied the methodology of Bayesian network based Bayesian multilevel analysis of relevance (BN-BMLA) in several studies. The technique is capable of providing a more specific characterization of the statistical associations and indicating weak signals, thus describing multifactorial diseases in a more detailed way.

**Enrichment analysis**

Recent developments indicate that the pharmaceutical industry might be able to avoid the disaster of the unsustainable business model ("patent cliff disaster"), but the companies are still intensively searching for new ways to improve efficiency. One of these innovative initiatives is aiming to extend or find new indications for known compounds. The so called drug repositioning is a much faster and cheaper way compared to the original research, additionally the number of shelved compounds at pharmaceutical companies is estimated to be between 2000 and 30 000. Repositioning is possible in two separate ways based on two scientific concepts. The first is that most of the compounds act on multiple molecular targets, so it is often worth to investigate these

off-target repositioning options. The second concept is that most of the molecular targets are involved in multiple biological pathways, so focusing on new signaling pathways could lead to new on-target repositioning options.

The improvements of computational methods in chemistry followed a very similar path to computational biology. The early simulation intensive computational chemistry was dominating the field until the '90s but as high-throughput measurement methods were spreading the focus shifted towards data management and integration techniques ("chemoinformatics"). Based on known molecular entities it is possible to predict the biological activity of a new molecule and in some cases also targets, side-effects or indications. Computational compound similarity searching and prioritization is a well researched area and there are also examples of its applications for computational drug repositioning. The different methods have their own benefits and drawbacks; the most important issue now is to appropriately integrate the results and to merge the new techniques into the pharmaceutical practice.

## Goals

**Asthma genetics**

The goal of my research was to identify new asthma associated genes, hopefully new therapeutic targets. I designed a candidate gene association study based on a former whole genome expression study of a mouse model of ovalbumin-induced asthma. My investigations have focused on three separate questions:

1. Is it possible to build a computational system that accelerates genetic association study design, especially the selection of the optimal set of gene polymorphisms with respect to measurement constraints?
2. What kind of new genetic variants, causal factors and therapeutic targets can be identified with a new asthma study supported by modern study design tools and multilevel validation techniques?
3. How can a systems based modeling tool, like the BN-BMLA (Bayesian network based Bayesian multilevel analysis of relevance) support frequentist statistics to reveal causal factors?

**Enrichment Analysis**

There is a huge amount of heterogeneous data available about compounds used in pharmacology. I investigated a proven data analysis method originating from molecular biology

(Set Enrichment Analysis) to be used in a pharmaceutical setting. My study focused on three separate questions:

1. How can Set Enrichment Analysis be used to integrate pharmaceutical data, to indicate weak signals, and what are the boundaries of the new application?
2. How can this technique support the development of a new drug candidate and how is it possible to merge the technique into the standard drug development workflow?

# Methods

**Asthma genetics**

My experiments were based on a former study of our research group which was carried out with an ovalbumin-induced mouse model of asthma. We used the results of a whole genome expression study to select 60 human orthologous gene candidates with a potential role in asthma and atopy.

The high-throughput genotyping of the polymorphisms was carried out with a primer extension technique in several studies of our group. The measurement technique has some technical constraints so it is impossible to measure polymorphisms with certain properties (e.g. short repeats in the PCR primer template region can lead to mismatches), or to measure some polymorphisms in the same reaction space together (e.g. the melting temperature of the primers has to be nearly identical). These somewhat contradictory requirements make it crucial to plan a multiplex measurement experiment carefully. Therefore, I implemented and described a software system and methodology (TIGER Study Design System) which integrates most of the databases needed for this process and automates several tasks.

During the study design process we selected 90 polymorphisms in 60 candidate genes based on functional, linkage and measurement technique considerations. The study population involved in genotype analysis comprised of 671 unrelated individuals (311 asthmatic children and 360 healthy controls) of the Hungarian (Caucasian) population about which we collected rich phenotypic data as well. Additionally, 34 adults were enrolled to an induced sputum gene expression experiment based on the genotyping results.

The Hardy-Weinberg equilibrium (HWE) was tested using the chi-square goodness-of-fit test, missing genotype data was imputed by sampling from the univariate distribution of the genotype data. I used HaploView and the IBM SPSS Statistics V20 software to perform Pearson's chi-squared test to assess statistical associations to asthma and the related odds-ratios for single markers, dominant-recessive models, and also haplotypes. Multivariate logistic regression analysis was also

performed with IBM SPSS Statistics V20, age and gender was included in the model as a covariate to apply sufficient statistical adjustment for the differences between cases and controls. To account for multiple hypothesis testing we applied permutation testing and Bonferroni correction.

I also analyzed the genotyping data with systems based modeling: I applied Bayesian network based Bayesian multilevel analysis of relevance (BN-BMLA) which was originally developed by our research group. The BN-BMLA method is capable of analyzing the Bayesian network graphs representing the causal relationships of the data, so it is possible to calculate the posterior probability of certain graph features, e.g. it can be predicted if the statistical association is due to a direct, transitive or confounded relationship. The high performance computing tasks were carried on the 512 CPU SGI Altix ICE supercomputer of the GenaGrid Consortium, the results were presented with the BayesEye client software.

For the analysis of the RT PCR measurements, I used the delta-delta-CT algorithm to normalize CT values for housekeeping genes and for the control samples. I used Student's t-test on the normalized fold change values to assess statistical significance.

**Enrichment analysis**

The simplest use case of a computational drug repositioning tool is when someone is searching for new indications to a given compound, so the input of the system is a compound under investigation. Our research group has developed computational drug repositioning methodology called QDF$^2$ which uses publicly available data of known compounds. The processed data can be noisy or incomplete, the system is capable of working with a wide range of data sources, e.g. from physical and chemical descriptors to any kind of quantitative measurements and clinical properties. The QDF$^2$ ranks the compounds of the database with a kernel fusion method based on their similarity to the compound under investigation. The pharmacological interpretation of the ranked list might bring up some new hypothesizes about the properties of the compound (e.g. indications, effected targets or pathways). There are several methods to analyze the ranked list (filtering, network analysis etc.), but these techniques require deep background knowledge and the results are difficult to merge with the pharmaceutical knowledge. The goal of my research was to give a possible solution for this problem.

Gene Set Enrichment Analysis (GSEA) is a well-known technique for analyzing tissue gene expression data and extracting biological insight. The method was developed for the interpretation of large ranked lists of entities on an abstract level of terms, thus the method with the appropriate modifications also can be applied for analyzing compound databases. During my work I developed a modified version of set enrichment analysis to be applied for ranked lists of compounds instead of

gene lists (CSEA or Compound Set Enrichment Analysis). This is a novel application of the proven Enrichment Set Analysis technique on this field, which would allow us to make a statement like the following: "if we rank all compounds on the basis of their similarity to our investigational compound, the dopaminergic agonists tend to be higher ranked, thus a dopaminergic agonist effect can be hypothesized".

We have built a reference database from publicly available data about well known compounds. We used three properties to describe the chemical profile (Molconn-Z™, MACCS keys and 3D pharmacophores) which were exported and computed from Schrödinger 2012 Suite software. The protein target profiles of the compounds were extracted from the DrugBank database. The unified side-effect profiles were computed with text-mining techniques based on the drug labels exported from the DailyMed database. We used the CMAP gene expression database to obtain expression profiles to the compounds. Eventually, we have built a reference database consisting 1730 compounds which is over 70% of the FDA approved compounds.

# Results

**Asthma genetics**

Collecting an optimal and measurable set of polymorphisms is a difficult and lengthy task which can take several months. In the TIGER Study Design System the users can define weighted filters for polymorphisms (e.g. genome locus, gene, functional region, allele types, MAF); the system controls their suitability for measurement; recommends possible substitutions if needed; and suggests an optimal set of polymorphisms that can be measured together in the same reaction space. The TIGER Study Design System was capable of accelerating the manual study design process by nearly an order of magnitude in several of our studies while the risks of the study design mistakes were also significantly reduced.

In our asthma study several of the 90 measured polymorphisms have shown different genotype distributions between asthma cases and controls. Among these 4 SNPs of 2 genes differed significantly in all of the applied statistical tests (the most significant polymorphisms in the genes: *SCIN* gene rs2240572, p=0,00007, OR=0,637; *PPARGC1B* gene rs32588, p=0,00012, OR=0,563). The polymorphisms of 6 other genes (*ITLN1, FABP3, MAT1A, OSGIN, LY9, LGMN*) showed statistically borderline differences with the frequentist evaluation. Further analysis has shown that the protecting effects of the *SCIN* and *PPARGC1* variants are significantly stronger among women,

whereas *ITLN1* in the atopic group and *LGMN* in dominant-recessive models show statistically significant association.

I analyzed the genotyping results also with the Bayesian network based Bayesian multilevel analysis of relevance (BN-BMLA) method to uncover the relationships between the polymorphisms and the endotypes. The BN-BMLA method confirmed the strong association of *SCIN* and *PPARG1B* with asthma. Additionally, the method also identified an interaction between *SCIN* and *TFF1* genes, and a stronger effect of the *PPARGC1B* in the exercise-induced asthmatic subgroup. Furthermore, the method found that the effect of *ITLN1* was only remarkable in the infection-induced asthmatic subgroup, but the association of *LGMN* with asthma (which was barely visible with the frequentist methods) was found to be particularly strong, especially in the atopic subgroup. The genetic background of the intrinsic asthma endotype differed significantly from all other endotypes, thus a different pathomechanism can be hypothesized. In addition, the method has also shown that the influence of the genetic background within the intrinsic group is much stronger than the age and gender variables, while these latter variables have significant effects in other asthma endotypes (especially atopic asthma). The asthma severity (GINA) has not shown any relationship with the genetic background.

Expressions of 3 genes (*SCIN, PPARGC1B, ITLN1*) were analyzed from the induced sputum samples of asthmatic subjects and healthy cases. The gene expression levels for all of the three genes were significantly lower in the case samples. The results were somewhat unexpected because the expression of all 3 genes changed in the opposite direction compared to our former mouse model of asthma. We compared publicly available data of asthma microarray studies from the NCBI GEO to our results and these confirmed our conclusions. We reviewed the literature to search for other examples of discordant human-mouse expression pairs and the phenomenon is well known, several systematic studies addressed the phenomenon and reported similar findings. There are several possible explanations. First, in the mouse model we measured the gene expression levels in whole lungs, while in humans in induced sputum, thus their different cell compositions could cause different gene expression averages. Second, it is known that asthma in humans is a chronic disease while mouse allergen-induced airway inflammation is an acute process and we measured the dynamic gene expression changes during this process. Nevertheless, differential expression of a gene in a process provides clear evidence for involvement in that process.

**Enrichment analysis**

I have built a software test system which is capable of characterizing a new compound with pharmaceutical terms (e.g. with potential indications or side-effects) based on the data of approved drugs. Instead of the functional gene sets used in GSEA I choose the Anatomical Therapeutic Chemical (ATC) Classification System, Level 4 compound sets to define the enrichment signals (compound sets for different indications). I used the basic cosine function to calculate the distances for side-effects and protein target descriptors, and Tanimoto distance function in case of the chemical descriptors. For a gene expression profile distance metric I used the default solution of the CMAP database.

To test the CSEA method I choose a compound and ranked all the other compounds in the reference database based on the different similarity metrics (chemical descriptors, target, expression profile, side-effects). In the next step I applied a rank fusion algorithm, SumScore, which is a simple yet mathematically unbiased method to merge the ranked list. Finally, I analyzed whether the distributions of the predefined compound sets in the ranked list are random. To calculate the enrichment scores of the sets (indications in this case) I applied the SaddleSum algorithm, which calculates the p-values with a special approximation technique (Lugananni-Rice formula).

To demonstrate the whole process I simulated a drug discovery workflow: at the first step the only available information about the compound is its chemical profile and at later stages more and more data becomes available. I choose amantadine for a retrospective demonstration, which is a known repositioned compound, originally it was developed for the treatment of influenza but later it received FDA approval also for the therapy of Parkinson's. First, I ranked all of the compounds in the database based on their similarity to amantadine with the different similarity metrics. Then I fused the ranked lists stepwise, I utilized more data sources in every step and analyzed the resulting ranked lists representing the combined effect of the related data sources. Dopaminergic agonists, one of the main pillars of the therapy of Parkinson's disease, consistently gained high ranks, also exhibiting a positive trend with the inclusion of more information sources. The joint application of the rank fusion method and the CSEA can cope with the problems of noisy data, but a very important confounding factor also stems from the similarity-based nature of the algorithm: the result indicated a number of groups containing dopaminergic antagonists. This may be explained by the similar chemical structures and target profiles of dopaminergic agonists and antagonists; similarity-based methods are inherently incapable of distinguishing such entities. This anomaly can be countered by employing other information sources which are less prone to this type of error, such as side-effect or expression-based ones. Using only these two information sources dopaminergic antagonists are not present on the list whereas dopaminergic agonists get a high rank.

Analogously (but the inverted case), chemically dissimilar entities with common biological function can be also connected.

Knowledge recycling is an information management approach that fosters the systematic reuse of accumulated information wherever possible and plausible, irrespectively of its origin and original purpose. This approach was also the motivation behind data warehousing technologies implemented in several industry sectors. In the current practice of drug development data utilization is usually finished altogether with the termination of the drug development process. The CSEA approach in a broader sense increases drug development efficiency with the tools of knowledge recycling. The case study has shown that the method can be easily integrated into the information workflow of the drug development process. Furthermore, it gives a computationally scalable, statistically robust and interpretation-friendly post hoc analysis technique which is capable of indicating important issues on the level of pharmaceutical terms.

# Conclusions

**1. Study Design System**

Genetic association study design is a difficult task which can take months. This lengthy, human resource intensive pre-optimization process can decrease the throughput of the genotyping process and the whole workflow by multiple orders of magnitude. The implemented TIGER Study Design System automates important phases of this preparation process which can increase its speed by nearly an order of magnitude.

**2. *SCIN***

This study is the first to investigate and show the association of *SCIN* (Scinderin) with asthma in a human population. Three polymorphisms of *SCIN*, namely on the exon 1 the minor allele of rs2240572 (H61R), and on the promoter region the major allele of rs2240571 and the minor allele of rs3735222 had statistically highly significant protective effects against asthma. I identified one potential causal variant with frequentist statistical methods which proved to be even more significant among women. The BN-BMLA method has confirmed these results and also indicated a modifying effect of the *TFF1* gene. This observation is in line with the known function of *SCIN*, both genes play an important role in mucus production.

**3. *PPARGC1B***

This was the first study reporting a relationship between the rs32588 (L42L) polymorphism on exon 2 of *PPARGC1B* (Peroxisome Proliferator-Activated Receptor Gamma Coactivator 1-Beta) and asthma. Frequentist methods indicated a statistically highly significant protecting effect of the minor allele which proved to be even more significant among women. The BN-BMLA confirmed the former findings and also pointed to an additional weak interaction with the exercise-induced endotype.

**4. *ITLN1***

My study has identified a potential protecting effect of the rs4656958 polymorphism of *ITLN*1 (Intelectin-1) related to asthma. The association was measurable with every method, but its statistical significance was in the borderline zone with frequentist methods. The BN-BMLA method has shown that the effect is much stronger in the infection-induced subgroup which is in line with the known function of the gene: the gene product plays an important role in recognizing bacterial components.

**5. *LGMN***

My research was the first to identify a potential role of *LGMN* (Legumain) in the pathomechanism of asthma. The effect proved to be week when assessed in a single allele model, statistically borderline significant as dominant-recessive model, whereas the BN-BMLA method identified a strong relationship especially in the atopic subgroup.

**6. Human asthma and mouse model gene expression**

We measured the elevated gene expression of *Scin*, *Ppargc1b* and *Itlna* in an ovalbumin-induced mouse model of asthma, whilst the expression of the homologue genes in the sputum samples of human chronic asthmatic patients was reduced. The effect was statistically significant for all of the three genes. Thus, their role in the pathomechanism received multiple confirmations but the opposite direction of the expression has shown the different dynamic properties of the two models and so the limits of a mouse model of asthma.

**7. Compound Set Enrichment Analysis**

This study was the first to successfully apply the proven mathematical method from molecular biology, the enrichment analysis framework for integrating wide range of pharmaceutical data. The new methodology significantly extended the former specific applications by involving

heterogeneous measurement data of arbitrary compound libraries and it is capable of indicating high level pharmaceutical properties in every stage of the drug development process.

**8. Information recycling with the CSEA method**

The CSEA method proved to be a statistically robust and computationally scalable solution for the continuous recycling of the pharmaceutical knowledge which is a somewhat unnoticed approach in the pharmaceutical industry with great opportunities. The method can be easily integrated in the drug development workflow in an iterative manner; it is capable of drawing conclusion based on significantly different compounds and distinguishing weak signals from the statistical noise by integrating heterogeneous information sources.

# Own publications

**Publications related to the subject of the thesis**

Temesi G, Bolgár B, Arany Á, Szalai C, Antal P, Mátyus P: Early repositioning through compound set enrichment analysis: a knowledge-recycling strategy. Future Med Chem 6(5), 563-575 (2014), IF: 4.000*

Temesi G, Virág V, Hadadi É, Ungvári I, Fodor LE, Bikov A, Nagy A, Galffy G, Tamási L, Horváth I, Kiss A, Hullám G, Gézsi A, Sárközy P, Antal P, Buzás E, Szalai C: Novel genes in Human Asthma Based on a Mouse Model of Allergic Airway Inflammation and Human Investigations. Allergy, asthma & immunology research 6(6), 496-503 (2014), IF: 3.084*

Ungvári I, Hullám G, Antal P, Kiszel Sz P, Gézsi A, Hadadi É, Virág V, Hajós G, Millinghoffer A, Nagy A, Kiss A, Semsei F Á, Temesi G, Melegh B, Kisfali P, Széll M, Bikov A, Gálffy G, Tamási L, Falus A, Szalai C: Evaluation of a partial genome screening of two asthma susceptibility regions using bayesian network based bayesian multilevel analysis of relevance. PloS one 7(3), e33573 (2012), IF: 3.730

Bolgár B, Arany Á, Temesi G, Balogh B, Antal P, Mátyus P: Drug repositioning for treatment of movement disorders: from serendipity to rational discovery strategies. Current topics in medicinal chemistry 13(18), 2337-2363 (2013), IF: 3.453

**Other publications**

Gabriella Jobbágy-Óvári, Csilla Páska, Péter Stiedl, Bálint Trimmel, Dorina Hontvári, Borbála Soós, Péter Hermann, Zsuzsanna Tóth, Bernadette Kerekes-Máthé, Dávid Nagy, Ildikó Szántó, Ákos Nagy, Mihály Martonosi, Katalin Nagy, Éva Hadadi, Csaba Szalai, Gábor Hullám, Gergely Temesi, Péter Antal, Gábor Varga, Ildikó Tarján: Complex analysis of multiple single nucleotide polymorphisms as putative risk factors of tooth agenesis in the Hungarian population. ACTA ODONTOLOGICA SCANDINAVICA 72:(3) pp. 216-227. (2014), IF: 1.309*