

# Integrating Medical Databases, the Interactome, and the Environment

-PhD Thesis-

Francisco Simões Roque

Center for Biological Sequence Analysis  
Department of Systems Biology  
The Technical University of Denmark

November 26, 2010





# Preface





This work is submitted as the requirement for obtaining the PhD degree at the Department of Systems Biology of the Technical University of Denmark. Funding supporting this work was received from the Villum Kann Rasmussen Foundation. The work was carried at the Center for Biological Sequences Analysis (CBS), Department of Systems Biology, under the supervision of Professor Søren Brunak and at the Center for Pediatric Research, Massachusetts General Hospital, Boston, under the supervision of Dr. Patricia Donahoe.



# Contents

<b>Preface</b>	<b>iii</b>
<b>Contents</b>	<b>vi</b>
Abstract . . . . .	ix
Dansk resumé . . . . .	xi
Acknowledgments . . . . .	xiii
Papers included in this thesis . . . . .	xv
Paper not included in this thesis . . . . .	xv
<b>I Introduction</b>	<b>I</b>
<b>2 Electronic health care</b>	<b>5</b>
2.1 Electronic health records . . . . .	5
Implementation and usage . . . . .	5
Data types . . . . .	6
Secondary use . . . . .	6
2.2 Extracting information from electronic health records . . . . .	8
Structured data . . . . .	9
Text mining . . . . .	9
The Sct. Hans database . . . . .	10
2.3 Paper I . . . . .	13
2.4 Paper II . . . . .	26
<b>3 Disease gene finding</b>	<b>37</b>
3.1 The human interactome . . . . .	38
Estimating the size of the interactome . . . . .	38
Protein interaction data . . . . .	38
The InWeb . . . . .	39
3.2 Paper III . . . . .	41
3.3 Paper IV . . . . .	53
<b>4 From chemicals to disease</b>	<b>67</b>
4.1 Chemicals and disease . . . . .	68

	Susceptibility and exposure . . . . .	68
	Polypharmacology and side-effects . . . . .	69
	Systems chemical toxicology . . . . .	69
4.2	Paper V . . . . .	71
4.3	Paper VI . . . . .	88
<b>5</b>	<b>Epilogue</b>	<b>97</b>
	<b>Bibliography</b>	<b>99</b>
	<b>Appendices</b>	<b>117</b>
	<b>Supplementary Information to Paper I</b>	<b>119</b>



## Abstract

THE past decades have seen an exponential growth of biomedical data that remains unexplored for research purposes. The increase in available unidentifiable patient record data, high-throughput proteomics results, and chemical pharmacology data, provides the opportunity of studying diseases as complex phenomena, triggered by a series of factors. In the present thesis, these disparate data types are combined in order to gain further insight into human disease etiology.

Electronic Health Records (EHR) contain coded and unstructured information, posing additional challenges in data extraction. In chapter 2, data and text mining techniques are used to retrieve medical terms from an electronic health record database, in the context of discovering disease correlations and stratifying patient cohorts. Furthermore, the re-use of electronic health record content in visualization systems is analyzed. Secondary use of EHRs is useful for comparing and detecting trends in the data, for both clinicians and researchers.

In the third chapter, protein-protein interaction data is explored for deducing functional relationships in human cellular systems in relation to health and disease. By combining disease gene information with protein data, disease related protein complexes can be identified, and then mapped to tissues to gain spatial resolution. With this, insight into the biological processes that prime certain tissues for developing tissue-specific disorders is gained. By integrating phenotypic data from mouse models with protein networks, the systems biology driving organ development can be discovered, paving the way for new approaches in treatment, diagnostics, and in regenerative medicine.

Finally, a network based approach is used to combine protein interaction data, disease information, functional annotation and chemical structures. By integrating the chemical biology knowledge with network biology, chemical exposures can be related to disease, and novel molecular targets for drugs can be identified. In the last article in this thesis, a new web server for exploring environmental chemicals, drugs, and natural products, based on their activity profile against biological targets and their adverse effects, is shown.

Collectively, this work demonstrates the power of integrating divergent data types in disease systems biology for understanding the complex nature of human biological systems.



## Dansk resumé

DE seneste årtier har der været en eksponentiel vækst i biomedicinske data, der stadig står tilbage som uanalyseret i forskningsformål. Stigningen i de disponible uidentificerbare patientjournal data samt high-throughput proteomics- og kemiske farmakologiske data, giver mulighed for at studere sygdomme som komplekse fænomener, udløst af en række faktorer. I den foreliggende afhandling er disse forskellige datatyper kombineret med henblik på at opnå yderligere indsigt i menneskelige sygdommes ætiologi.

Elektroniske Patient Journaler (EPJ) indeholder kodet og ustruktureret information, hvilket skaber yderligere udfordringer i forbindelse med dataudtrækning. Data- og tekst udvindelses teknikker bruges her for at hente medicinske termer fra en elektronisk patientjournalssystem database, med målet at opdage sygdoms korrelationer og stratificere patient kohorter. Desuden er genbrug af elektroniske patientjournalssystemers indhold på visualisering systemer analyseret. Sekundær anvendelse af EPJ er nyttig til at sammenligne og afsløre tendenser i data, både for klinikere og forskere. I det tredje kapitel bliver protein-protein interaktion data udforsket for at udlede funktionelle relationer i humane celler i relation til sundhed og sygdom. Ved at kombinere oplysninger fra sygdomsgener med protein data, er det muligt at identificere sygdomsrelaterede protein komplekser, og derefter kortlægge disse til væv for at få rummelig opløsning. Med dette, har denne øgede indsigt i de biologiske processor der primer bestemte vævs typer, givet en større indsigt i udviklingen af vævsspecifikke lidelser. Ved at integrere fænotypiske data fra musemodeller med protein netværk, kan denne system biologisk drevne organ udvikling blive opdaget, hvilket har banet vejen for nye tilgange i behandling, diagnostik og regenerativ medicin.

Endelig er en netværk baseret tilgang brugt, der anvendes til at kombinere protein interaktion data, sygdom information, funktionelle annotation og kemiske strukturer. Ved at integrere den kemisk biologiske viden med netværks biologi, kan kemiske eksponeringer relateres til sygdom, og nye molekyllære mål for lægemidler kan identificeres. I den sidste artikel i denne afhandling præsenteres en ny web-server der udforsker miljømæssige kemikalier, lægemidler og naturlige produkter, baseret på deres aktivitet profil mod biologiske mål og deres skadelige virkninger. Kollektivt, viser dette arbejde effekten af at integrere forskellige datatyper i sygdoms systembiologi for at forstå den komplekse karakter af de menneskelige biologiske systemer.





## Acknowledgments

Because no man is an island, I would like to acknowledge a number of people that have helped and supported me throughout my work on this thesis.

I have had the great pleasure to be staying at the Center for Biological Sequence Analysis (CBS), with its helpful staff and fantastic atmosphere. The center is led by Professor Søren Brunak who doubles as my supervisor in my PhD. Thank you for your reflected advices, your trust, and your inspiration.

A special thanks to Prof. Patricia Donahoe, for a profitable stay at the Massachusetts General Hospital in Boston, and for her continued support and interest.

My gratitude to Ian Donaldson, for offering me a desk in Oslo.

For having the best office space in CBS, thanks to Kristoffer Rapacki. I deeply cherish all the advice you gave me, all the buttons you pushed for helping me, and everything you have done for my success, both academic and personal.

To Kasper Lage, my special regards for making me believe in my work as meaningful research, for your continued advice and support, and for receiving me with open arms in Boston. And thanks for that U2 concert!

A big thanks to all the members of the ISB group. Nils Weinhold, Daniel Edsgård, Thomas Stranzl, Thomas Skøet Jensen, Ramneek Gupta, Tune Pers, Karine Audouze, Konrad Krysiak-Baltyn, Malene Larsen, Mette Larsen, and Chris Workman. Thanks for all the journal clubs, ISB meetings, helpful discussions, and fantastic social events. I am quite proud of winning the indian bow-and-arrow set!

For Rodrigo, thank you for launching me into the academic world, for working with me on InterMap3D, for speaking portuguese, and for giving me advice.

To Anders Gorm Petersen, Henrik Nielsen, and Rasmus Wernersson, thanks for introducing me to the evolutionary world, and for fantastic discussions.

A special place in my heart is reserved to the administration staff at CBS. Thank you Lone, Dorthe, Marlene and Annette, and all the office aids. You make the world a better place.

Of course a big thanks to the system administration crowd, Kristoffer, Peter, and Hans-Henrik. Without your work and help, my research would not be possible.

To the people involved in the InWeb project, my continuous gratitude for all the help. Thank you Olga for all the database work, Rasmus for the help, Chris for discussions and cinweb, Niclas for all the input, and Daniel Hansen and Grzegorz for the user scripts. Thank

you Kasper for giving me the big responsibility.

My appreciation to Massimo, Peter Bjødstrup Jensen, and Lars Juhl Jensen, for all your help with text mining the patient records.

A big thank you to all the social crowd at CBS, the beer and whisky club, the Predictors, Solmaz, Irene, Ilka, Nico, Bent, Kasper Nielsen, Nicolai and all the people that also know how to have fun.

My acknowledgements to the proofreaders Bo, Nils, Bent and Monica, for making sure there are (almost) no mistakes in the thesis, and to Nico, for his LaTeX expertise.

To Bo, Bent, Niko, Nico, Aron, a big thanks for making Denmark a friendly place, and to Rita for making me feel at home away from home.

And to my friends back home or somewhere else, my thanks for accepting me the way I am, and most of all, for being friends. Bruno, Tiago, Pico, for sticking around ever since I remember; the LEBM crowd, João Tiago, Ricardo, Becas, Nery, Guida, Barradas, Lili, Rebordão, Inês e David, you're cool guys!

To Mãe, Pai e Bea, for being close, for believing, and for everything. Thank you.

Obrigado avós, por tudo, e por serem amigos.

Thank you to Vanda, Nuno, Custódio, and João, and to my cousins, for being a great big family.

Monica Gjuvsland, my dearest fiancée, thank you for being always by my side. Thank you for your cheerful mood, patience, and understanding. Thank you for opening my eyes to other things than work. Thank you for fredagskos, and for all the koselig things you do. Thank you for being a partner.

## Papers included in this thesis

- ∞ *Using electronic patient records to discover disease correlations and stratify patient cohorts..*  
**Francisco S Roque**, Peter B Jensen, Henriette Schmock, Massimo Andreatta, Thomas Hansen, Søren Bredkjær, Anders Juhl, Thomas Werge, Lars J Jensen, and Søren Brunak. [manuscript in preparation]
- ∞ *A comparison of several key information visualization systems for secondary use of electronic health record content.*  
**Francisco S Roque**, Laura Slaughter, Aleksandr Tkatsenko. [accepted for publication in the proceedings of *Louhi 2010: Second Louhi Workshop on Text and Data Mining of Health Documents*].
- ∞ *A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes.*  
Kasper Lage, Niclas Tue Hansen, E Olof Karlberg, Aron C Eklund, **Francisco S Roque**, Patricia K Donahoe, Zoltan Szallasi, Thomas Skøt Jensen, Søren Brunak. *Proc Natl Acad Sci USA*, 2008 vol. 105 (52) pp. 20870-5.
- ∞ *Dissecting spatio-temporal protein networks driving human heart development and related disorders.*  
Kasper Lage, Kjeld Møllgård, Steven Greenway, Hiroko Wakimoto, Joshua M Gorham, Christopher T Workman, Eske Bendsen, Niclas T Hansen, Olga Rigina, **Francisco S Roque**, Cornelia Wiese, Vincent M Christoffels, Amy E Roberts, Leslie B Smoot, William T Pu, Patricia K Donahoe, Niels Tommerup, Søren Brunak, Christine E Seidman, Jonathan G Seidman, Lars A Larsen. *Mol Syst Biol*, 2010 vol. 6 pp. 381.
- ∞ *Deciphering diseases and biological targets for environmental chemicals using toxicogenomics networks.*  
Karine Audouze, Agnieszka Sierakowska Juncker, **Francisco S Roque**, Konrad Krysiak-Baltyn, Nils Weinhold, Olivier Taboureau, Thomas Skøt Jensen, Søren Brunak. *PLoS Comput Biol*, 2010 vol. 6 (5) pp. e1000788.
- ∞ *ChemProt: A Disease Chemical Biology Database.*  
Olivier Taboureau, Sonny Kim Nielsen, Karine Audouze, Nils Weinhold, Daniel Edsgard, **Francisco S Roque**, Irene Kouskoumvekaki, Alina Bora, Ramona Curpan, Thomas Skøt Jensen, Søren Brunak and Tudor Oprea. [in review]

## Paper not included in this thesis

- ∞ *InterMap3D: predicting and visualizing co-evolving protein residues.*  
**Francisco S Roque**, Rodrigo Gouveia-Oliveira, Rasmus Wernersson, Thomas Sicheritz-Ponten, Peter W Sackett, Anne Mølgaard, Anders G Pedersen. *Bioinformatics*, 2009 vol. 25 (15) pp. 1963-5.



# Introduction

SYSTEMS BIOLOGY and new technological advances supporting large scale biological studies have pushed us away from traditional studies of single genes, proteins or phenotypes towards a more integrative holistic approach to biological research. Complex processes, such as those triggering many common diseases like cancer, diabetes, hypertension, and psychiatric disorders, can be better understood by studying them as systems. Potential and emergent causes can then be revealed by analyzing the high order interactions between the component parts.

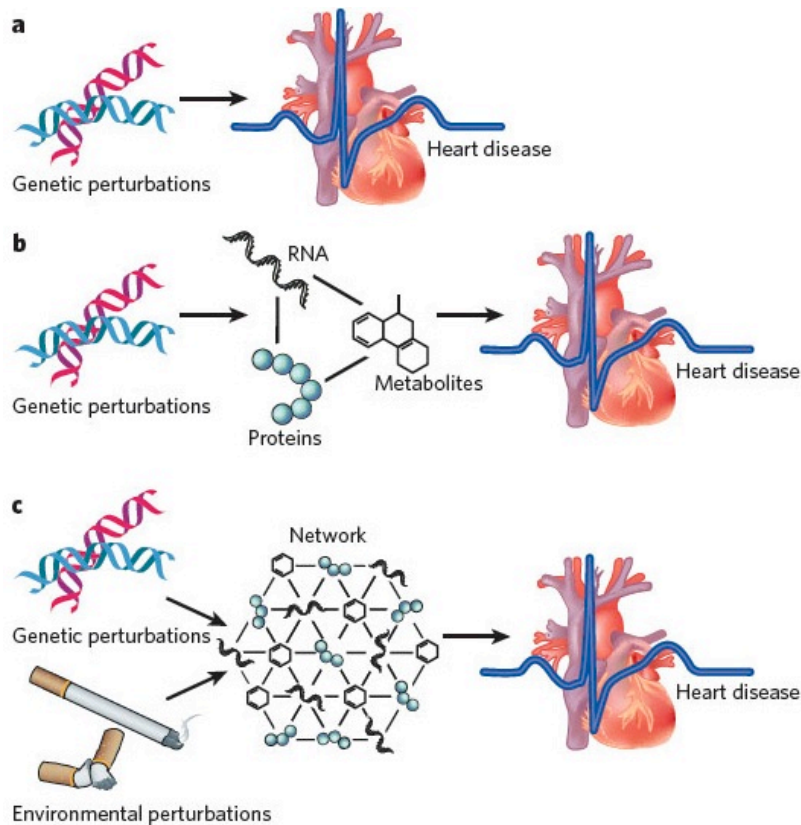
For decades, disease gene finding has been driven by single factorial perturbations leading to phenotypical changes. One affected gene leading to changes in abundance and activity of its transcribed protein, causing disease. This simple hypotheses lead to the discovery of thousands of genes for rare Mendelian diseases, but has been failing for the complex inherited phenotypes in the human population [1].

Recent advances in large-scale genomic studies revealed that common forms of disease might be caused by disruption of functional molecular networks. Disease states are then affected by a complex interaction of genetic and environmental factors (Figure 1.1). To understand the behavior of any one gene in the context of human disease, individual genes must be understood in the context of molecular networks that define the different disease states.

The new trends in biology, shifting the focus from singular events to large scale analysis of systems have given rise to continuous data generation, forcing the field of data analysis to move faster, in order to keep up the pace. To complicate things, new data acquisition methods are developed every day, enforcing the creation of new data types and adding to the existing, making it hard to come up with standardized ways of sharing and storing them.

The evolving of technology itself, with new and faster machines, and more storage space fuels this advancement, while data analysis lags slightly behind. New tools are continuously improved, and data is now being shared faster than before, generating scientific findings at an unprecedented pace.

Recent sequencing efforts have made available entire genome datasets. The availability of fast and cheap sequencing machines has dropped the cost of sequencing an entire human genome to a couple of thousand dollars. When Craig Venter's genome was sequenced, the cost was around \$70 million! Accompanying the evolution of high throughput sequencing, biological databases experienced extreme growth. Protein structure, expression array, and pathway databases have



**Figure 1.1:** Views on complex diseases [2]. **a**, the classic reductionist approach of identifying variation in DNA correlating with disease phenotypes, effective for simple Mendelian diseases. **b**, isolated DNA changes do not lead directly to disease, but instead lead to genetic changes in RNA, proteins and metabolites that then affect disease propensity. Intermediate phenotypes can describe early relationships between genes and disease. **c**, integrationist approach to disease exploration. A combination of genetic and environmental factors triggers disruptions in molecular networks, which then go on and trigger disease risk.

evolved side-by-side. At the same time manual curation lagged behind, and cannot keep up with the constant flow of information. This justifies the need of high performance computing, and automation strategies in biology.

Up to now I only referred to the more structured data, that is readily available and can be easily imported into tools and computer programs for analysis. In reality, however, the bulk of biological data is not in the form of structured and neatly arranged information, but rather in free text in scientific literature and medical databases. Just MEDLINE [3] alone contains over 18 million references to journal articles, and their searchable interface PubMed handles over 80 million queries per month, as seen in Figure 1.2. It's hard not to get lost, and even harder to keep up with all the new research being published, even on very specialized research fields. This vast amount of easily accessible, but poorly structured data calls for new strategies and methods to tackle the complex task of information extraction from free text.

Side-by-side with the publicly available data, there has been an increase in the amount of information collected for other purposes, remaining unused for research. For example, medical

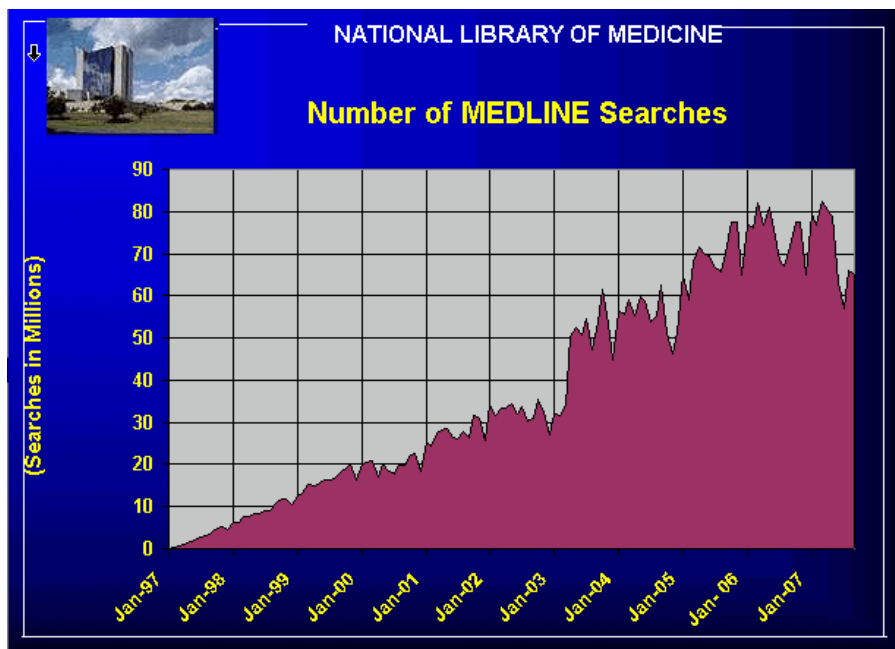


Figure 1.2: Number of MEDLINE searches from Jan-97 to Sep-07 [4].

databases from the public or private health systems are often subject to much more strict ethical and privacy laws than sequence data, and thus are not made readily available to the public. One other example is pharmaceutical databases, often kept in the secret of the companies which develop them. Allowing researchers to access those databases may lead to additional insights into drug usage and side-effects.

All these different pieces of information are collected and assembled independently, and even publicly available databases lack the amount of integration that would allow for large-scale analysis of complex biological systems. This thesis will focus on visualizing all those data as little pieces of a large, ever-growing puzzle. It covers fields that just started to attract scientific attention, such as integrating medical databases with other biological knowledge for disease gene discovery.

The structure of the thesis reflects the workflow I intend to show: Chapter 1 is this introduction. Chapter 2 describes the Electronic Health Records and the data rich medical databases, and the challenges working with such data poses. Paper I shows an integrative analysis of a medical database, and lays ground for exploring disease causality in the next chapters. Paper II is a brief review on secondary use of medical records. Chapter 3 describes the use of protein interaction data for creating the human interactome, and how it was put together with disease gene information, tissue resolution, and gene expression levels to understand the causes of tissue specificity of diseases (Paper III), and combined with spatio-temporal resolution of gene expression for understanding organ development (Paper IV). Chapter 4 dwells into external causes for disease, and explores chemical-protein and chemical-disease associations. Paper V integrates toxicogenomics data for drugs and environmental compounds, and protein interaction data, to identify new molecular targets for chemicals and linking them to disease. Paper

VI describes a new server, ChemProt, a compilation of chemical-protein association resources and protein-disease information.



# Electronic health care

## 2.1 Electronic health records

The use of health information systems has increased substantially in the past few years, and there has been a generalized adoption of electronic health records (EHR) in primary care. An EHR is a repository of patient data in digital form, stored and exchanged securely, and accessible by multiple authorized users. It contains retrospective, concurrent, and prospective information and its primary purpose is to support continuing, efficient and quality integrated health care [5]. The amount of knowledge contained in those records is potentially immense, and early adoption studies have shown that EHRs facilitate searching and retrieval of patient data [6]. With a multitude of data types aggregated under a single record, there is a current need to address ethical and permission issues, and to develop tools and methodologies for gaining insight into this wealth of information.

Implementation of these systems has posed challenges in developing the appropriate data warehouses to store them. Medical databases contain data in a variety of formats: images in the form of X-ray images or scans, phenotypical description of diseases, medical histories, nursing notes, EKG signals, drug administration information, billing and administrative codes, etc. Frequently, the data is not located in the same system, but is distributed amongst several servers and locations, depending on the origin and nature of it. Information retrieval from these systems is therefore a non-trivial task.

Even though EHR systems are designed to simplify the daily routines of clinicians, there are downsides to their implementation. Adoption of electronic patient systems takes time and challenges experienced professionals to sacrifice some flexibility, which is inherent to the manual track-keeping used for decades [7]. The information contained within EHRs draws attention towards new research areas, opening up new perspectives and directly supporting medical advances. Acknowledgement of these facts might allow the widespread use of electronic health records and apply them for strategic management decisions and clinical research.

### Implementation and usage

Health information systems' implementation has been a high priority in several countries for the past decade [8]. As numerous hospitals adopted such systems, there was an increased aware-

ness towards the lack of integration between the different systems implemented. Many vendors developed their own user interfaces and models for data storage, and used different information representations and technology platforms, creating additional challenges for data sharing and re-using information across different sources. Several groups have now started collaborative efforts towards data integration of electronic health records, such as CEN, HL7 and OpenEHR [9--11]. In Denmark, after a widespread adoption of the first EHR systems, there has been a continuous effort towards cross-platform integration [12].

Different systems are used depending on how the health system is organized within one country [13]. Generally these systems are implemented in primary care, the office of a general practitioner; secondary care, or specialist upon referral by the primary care practitioner; and tertiary care in a major hospital. Some systems are implemented for the patient's self-monitoring at home [14].

EHRs are used not only by doctors, but also by different health care professionals and administrative staff. Additionally, patients or their parents might use the system at some point for data entry [13]. Table 2.1 on the facing page describes them in detail and illustrates the different components of the system they use.

## Data types

As mentioned before, each EHR is an aggregate of different data sources and formats. EHRs include both unstructured free text and coded data; the later one using a number of different terminologies, such as the International Classification of Diseases (ICD)<sup>1</sup> codes for diagnoses, Anatomical Therapeutic Chemical Classification Index (ATC)<sup>2</sup> codes for medication, and the Systematized Nomenclature of Medicine (Snomed)<sup>3</sup> for coding pathological findings. Patient outcomes can also be described by statements such as descriptions of pain. Inclusion of other types of data is dependent on where the system is being developed, and who will be the main user of the system. For example, in a radiology ward it is useful to include X-ray pictures and radiology reports in the patient's electronic record, while these are not required for a psychiatry ward.

## Secondary use

Besides the obvious main use for EHRs, there are numerous other activities that have interest in using these records. Analysis, research, quality and safety measurement, public health, payment, provider certification and accreditation, marketing, and general business applications, might re-use the whole or parts of the patient's journals and extract information for their own purposes [15]. The issue of secondary use of EHRs is reviewed in detail in Paper II, *A Comparison of Several Key Information Visualization Systems for Secondary Use of Electronic Health Record Content*.

Figure 2.1 illustrates current and future usage of EHRs. Many existing systems have implemented tools for medical documentation, order entry and results review. Data extraction

---

<sup>1</sup><http://www.who.int/classifications/icd/en/>

<sup>2</sup><http://www.whocc.no/atc>

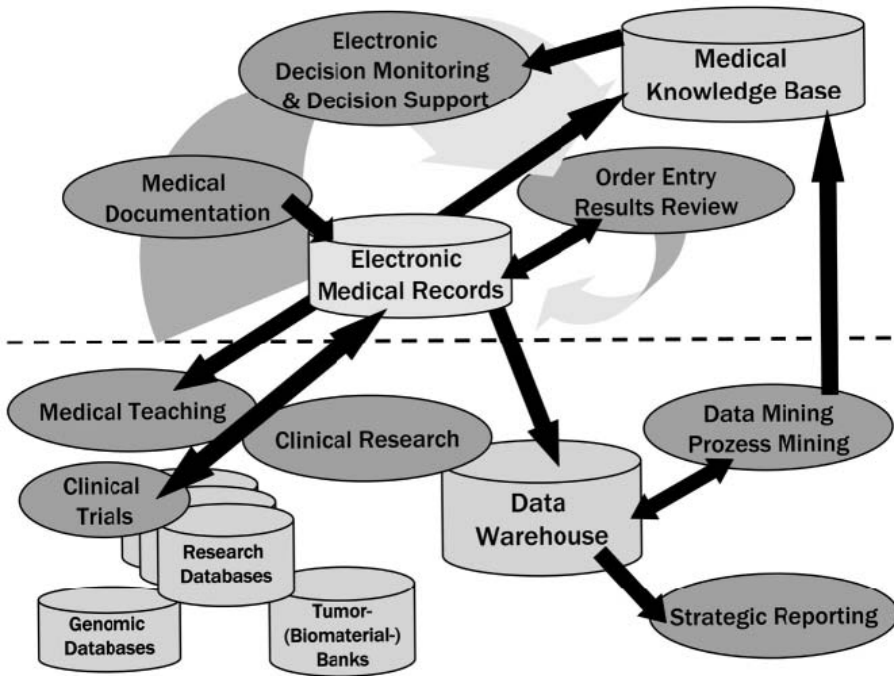
<sup>3</sup>[http://www.nlm.nih.gov/research/umls/Snomed/snomed\\_main.html](http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html)

User	Component of EHR
Nurse	Daily charting; medication administration; physical assessment; admission nursing note; nursing care plan
Physician	Referral; present complaint, e.g. symptoms; past medical history; life style; physical examination; diagnoses; tests; procedures; treatment; medication; discharge
Patient	History; diaries; test
Parents	History
Secretarial staff	Procedures; problems; diagnoses; findings; immunization
Pharmacists	Medication
Multiprofessional: nurse; physician; laboratory staff; radiology staff; clerk or administrative staff; pharmacy personnel; health care professionals	Referral; present complaint, e.g. symptoms; past medical history; life style; physical examination; diagnoses; tests; procedures; treatment; medication; discharge; administration of medication; admission nursing note; daily charting

**Table 2.1:** Users of EHR systems and used data components. Table adapted from *Hayrinen et al.* [13]

has been limited to building medical knowledge bases for decision monitoring and decision support (e.g. [17], and several others). Current trends are moving towards using the wealth of information present in these records to support clinical research and clinical trial support, and using data mining and complex data extraction procedures for further enhancing medical knowledge bases [16]. Several large projects have begun tackling translational research, combining many different data sources and further improving the EHRs richness (e.g. the i2b2 project [18]).

It is important to mention the protection of the privacy of patient data. When retrieving records from the protected clinical health care system for purposes of secondary use, we have to take into account that the data needs to go through an anonymization step, making sure most of the sensitive data is filtered out [19, 20].



**Figure 2.1:** Using EHRs: major areas of past (upper part) and future research areas (lower part) [16].

## 2.2 Extracting information from electronic health records

Health information systems contain many different data types, combining both structured information and unstructured free text, the latter one making up for the majority of the electronic records. Biomedical and clinical language pose additional challenges to working with literary texts. They frequently use domain specific terminology, acronyms, and polysemic words. Often these contain different spellings or typographical variants, and use different writing styles. Furthermore, clinical text adds extra complexity due to its narrative nature.

Clinical texts are those written by clinicians in the clinical setting. These can cover patients description, their diseases, examination findings, personal and social histories, and all types of comments and observations that are useful upon the examination procedure. Clinical text is mostly composed of narratives, ranging from very short (e.g. *"The patient ECG is normal."*), to very long and descriptive entries (e.g. a detailed medical examination). Frequently, these narratives are different from those occurring in biomedical literature, being mostly ungrammatical and composed of short, often abbreviated messages. Sometimes the texts are dictated into the computer, or written only for documentation purposes. Because the narratives are not intended to be reviewed or to be used outside of the setting of the clinical environment, they are often plagued with abbreviations, acronyms, and jargon, many of them specific to the doctor or practice in question. Ambiguity is often high, and there is a frequent lack of consistency on the wording used [21]. If the written patient documentation is not monitored by a built-in spellchecker, the resulting texts will contain a high amount of misspelling.

## Structured data

Extracting information from the structured parts of the EHR is not a trivial task; the records contain missing and incomplete data, and often have to be complemented with the unstructured text for establishing meaningful analyses. Structured data entry is a tedious process and often doctors only input what is required for billing purposes. Furthermore, this type of data lacks the expressiveness of natural language, and is often complemented by the use of unstructured data. What could otherwise be a content rich data source is often hindered by these drawbacks.

For the purpose of standardizing data input across clinical systems, and to map synonyms to a same term, a number of clinical vocabularies have been consistently used. These are mainly used for a number tasks, namely:

- Searching knowledge resources, and tagging;
- Supporting clinical practice analysis, quality measurement, and outcomes research;
- Providing data for clinical epidemiological analyses;
- Supporting payment processing and reimbursement;
- Identifying proper guidelines, paths, and trigger reminders in patient care.

When a clinician evaluates a patient, he or she starts the documentation using free text and unstructured information, such as history and physical findings. As the clinician's evaluation process continues, the unstructured data is transformed into more structured data, often linked to billing and reimbursement. These claims-related structured data sets are primarily used for structured billing, and may not be enough to capture clinical details. The most commonly used terminology is the International Classification of Diseases (ICD), both in its version 9, in the United States, and version 10, in Europe, Australia and New Zealand. For medication and drug usage, the Anatomical Therapeutic Chemical Classification Index (ATC) codes are widespread.

Some EHR systems are currently implementing the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT) ontology for recording test results and describing events that require a more fine-grained control over the nomenclature.

All of the terminologies used are organized in a somewhat hierarchical way, and can describe diseases, or drugs, that target a specific system in the human body. For this reason they can be extracted and used for epidemiology studies, or drug usage research. Paper I in Section 2.3 describes the use of ICD10 for stratifying patient cohorts and identifying disease correlations.

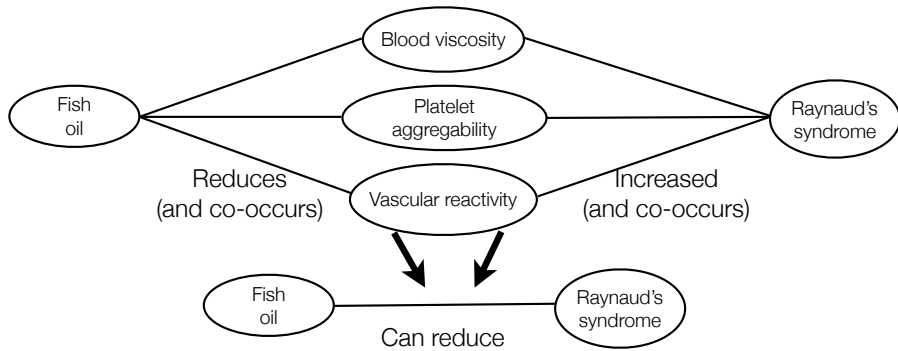
## Text mining

Searching within free text can be performed in many of different ways. Depending on the complexity of the query, and the type of information to be extracted, one can use from very simple string matching methods to very complex machine learning algorithms.

Text mining and information extraction (IE) differ from information retrieval (IR) in the way that the first two involve extracting predefined information from text, while the latter focus on finding documents on large units of data. Examples of IR systems are search engines such as

Google [22] or PubMed [23]. In simplified terms, Information Retrieval returns documents while Information Extraction returns information or facts.

IR can be used to select documents to be analyzed, thus resulting in new associations between documents. One such example of IR was described by *Swanson* [24], who found an association between fish oil and Raynaud's syndrome by manually searching the literature, illustrated in Figure 2.2. This hypothesis was later corroborated both experimentally and clinically.



**Figure 2.2:** Hypothesis advanced by *Swanson* [24], after IR from the literature. *Swanson* discovered that Raynaud's patients had altered blood properties. By further exploring the literature, this time for articles about factors influencing the characteristics of blood disorders common in this syndrome, he found that fish oil appeared to be an important topic. Later, his association was tested experimentally and in a clinical setting, and became the first trial-and-error strategy for linking two previously unknown subjects.

IE is a sub-domain of Natural Language Processing (NLP). NLP research focuses on building computational models to understand and interpret natural language. Named Entity Recognition (NER) is also important to mention, as a sub-field of information extraction, and refers to the task of recognizing entities such as drugs, diseases, protein names, etc. in free text documents. NER systems can be rule-based, or can use machine learning approaches, which requires large amounts of training corpora.

Text mining uses IE to discover and extract knowledge from unstructured data, and to derive relationships between entities. Usually text mining comprises of three steps: information retrieval, information extraction, and data mining (to find associations between the different extracted pieces of information).

The paper in Section 2.3 builds on the basic concepts of information retrieval, information extraction and text mining. Here I have covered only the key concepts of the technology. More details can be found in the works of *Cohen et al.* [25], *Jensen et al.* [26], and *Ananiadou et al.* [27], and a further review on information extraction for the medical domain from *Meystre et al.* [28].

## The Sct. Hans database

For the purpose of Paper I, we worked with a database from the Sct. Hans Mental Health Center, in Roskilde, Denmark. A total of 5543 patients were followed from 1998-2008, and their records stored in an EHR database. 70% (4822) patients are from the Copenhagen area,

61% of these are male. The average age is 30 years old. The records are a mix of structured diagnose assignments of ICD10 codes, ATC codes for medication usage, patient care notes from nurses and doctors, personal information, as well as admission and discharge summaries.

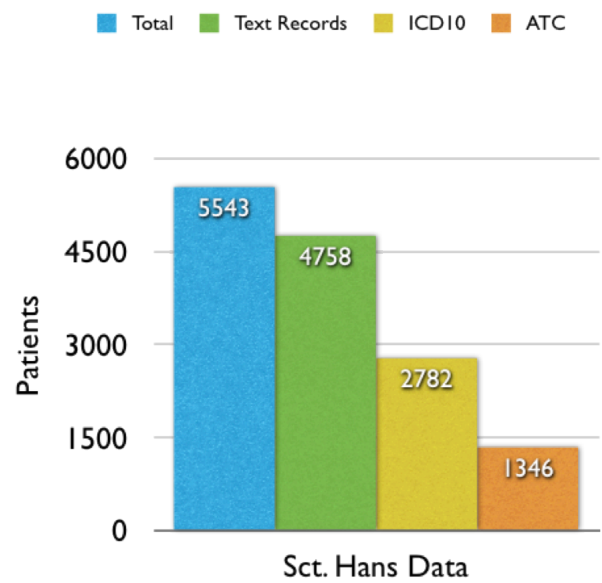
The screenshot shows the EHR system interface. On the top left, there is a list of patient notes with columns for date, time, and author. Below this is a table with tabs for 'Socialt', 'Baggrund', 'Recept', 'Læbliste', 'Udskriv', and 'Skriv'. The 'Læbliste' tab is currently selected. On the right side, there is a large text area for 'Behandlingsnotat' (Treatment Note). The note contains several paragraphs of text, including dates and times, and mentions of medication and patient status. The bottom of the interface shows a status bar with various icons and text.

**Figure 2.3:** Snapshot of the EHR system at the Sct. Hans Psychiatric Center, Roskilde, Denmark. On the top left are the different types of notes a patient has been given, and on the right the free text notes that were written in each entry.

Figure 2.3 illustrates a snapshot of the user interface, currently used by medical staff in this hospital to input patient data. Most systems are tailored to a specific application, or the medical speciality they are used in. This poses challenges to the integration and data extraction process, as no system is the same. In this particular case there is a much higher content of free text in the record, in detriment of structured information, due to the fact that psychiatric practice involves much longer sessions with patients. On average, the patients' internment is also much longer. Another noticeable aspect in this corpus are the discrepancies on the size of each entry, depending on the physician who wrote the report.

The system does not have any data export feature. For Paper I, *Using electronic patient records to discover disease correlations and stratify patient cohorts*, the extraction was made using a dump of the database's backend, and the fields were manually cross-referenced based on exploration of the front-end. Figure 2.4 shows the contents of the database. About 85% of the patients had textual records, but less than half had coded diagnosis. ATC codes were only present in 25% of the patients.

The lack of consistent structured data throughout the corpus motivated the use of text mining from unstructured free text, capturing additional medical terms to complement existing ones. In the next Section, the analyses conducted on this data for patient stratification and disease co-occurrence exploration are shown.



**Figure 2.4:** Contents of the Sct. Hans database. There are in total 5 543 patients. Of these, 4 758 have textual records, 2 782 have some kind of ICD10 code for diagnosis, and 1 346 have drug usage ATC codes.



## 2.3 Paper I

# Using electronic patient records to discover disease correlations and stratify patient cohorts

Francisco S Roque<sup>1†</sup>, Peter B Jensen<sup>2†</sup>, Henriette Schmock<sup>3</sup>, Massimo Andreatta<sup>1</sup> Thomas Hansen<sup>3</sup>, Karen Søebye<sup>4</sup>, Søren Bredkjær<sup>3,5</sup>, Anders Juhl<sup>6</sup>, Thomas Werge<sup>3</sup>, Lars J Jensen<sup>2</sup>, and Søren Brunak<sup>1,2\*</sup>

<sup>1</sup> Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark.

<sup>2</sup> Center for Protein Research, University of Copenhagen, Copenhagen, Denmark

<sup>3</sup> Institute of Biological Psychiatry, Mental Health Center Sct. Hans, Copenhagen University Hospital, Roskilde, Denmark

<sup>4</sup> Department of Clinical Biochemistry, Hvidovre University Hospital, Denmark

<sup>5</sup> Psychiatry Region Sealand, Ringsted, Denmark

<sup>6</sup> Copenhagen University Hospital, Denmark

\* To whom correspondence should be addressed. † These authors contributed equally to this work.

### Abstract

Electronic patient records remain a rather unexplored, but potentially rich data source for discovering correlations between diseases. We describe a general approach for gathering phenotypic descriptions of patients from medical records in a systematic and non-cohort dependent manner. By extracting phenotype information from the free-text in such records we demonstrate that we can extend the information contained in the structured record data, and use it for producing fine-grained patient stratification and disease co-occurrence statistics. The approach uses a dictionary based on the International Classification of Disease ontology and is therefore in principle language independent. As a use case we show how records from a Danish psychiatric hospital lead to the identification of disease correlations, which subsequently can be mapped to systems biology frameworks.

### Introduction

The most important prerequisites for personalized medicine are meaningful patient stratification and a detailed understanding of how diseases, clinical phenotypes, and genotypic variation correlate. Ultimately this depends on our ability to quantify and compare phenotypic descriptions of diseases and symptoms, taking past treatment and disease history into account. For that we need access to, and integration of clinical data, which is exactly what the recent introduction of Electronic Patient Records (EPR) in modern healthcare promises to deliver [29--33].

EPR systems document patient treatment and care over time. They comprise different types of structured and unstructured data, ranging from coded diagnoses, ordinary physiological measures, laboratory test results over medication prescriptions, and treatment plans, to free text notes about disease, treatment and care [13]. Structured EPR data (and registry data in general) have known biases, in part related to reimbursement practice (hospitals are reimbursed based on which diseases they report) and administrative tasks such as activity monitoring [34]. Incompleteness is another problem, since registry data will normally pertain strictly to procedures and diagnoses relevant to the current hospitalization. In contrast, free text notes contain much additional information, but in an inherently unstructured form [35]. In this paper we show that text mining can be used to augment the coded diagnoses and thus complement the information stored in structured formats. This approach provides the means for a more fine-grained phenotypic description of patients, which is comparable across cohorts, and goes far beyond what is normally stored in public registries.

The growth in EPR systems and health registries is changing the focus of health informatics towards the clinical research potential of the collected data [16]. Structured data from these sources have previously been used to uncover patterns of disease and comorbidity [36, 37], and for patient recruitment and monitoring in clinical trials [38]. In unstructured health data, such as in EPR texts, information extraction approaches, including Natural Language Processing (NLP), have been used for diagnosis detection [39--41], decision support [34], and medication surveillance [42--44]. These studies have been partially aided by tools like MetaMap [45] for mapping medical texts to controlled vocabularies such as the Unified Medical Language System (UMLS).

Independently of the research assisted by the information presented in the patient records, several approaches have been developed to discover novel disease associations, either based on

shared disease causing genes or on overlapping pathways [46--48]. Known disorder-gene associations, from available resources like OMIM, have been used to establish links between diseases, thus creating a network of disorders [46]. Common to many of these approaches is the extensive use of protein-protein interactions from large-scale proteomic studies. Linking disease-gene information with the growing data present in EPR systems will allow for a better understanding of disease etiology and mechanisms.

Here we describe a strategy for exploring data from EPR systems in the context of subsequent systems biology analysis. By mining the free-text parts of the EPR from a psychiatric hospital we are able to augment the disease information assigned in structured formats, such as ICD10 (Version 10 of the International Classification of Disease ontology) codes, and thus obtain a much richer phenotype profile of each patient. Treating these profiles as phenotype vectors [47] in the controlled vocabulary space of the ICD10 disease classification, we demonstrate how they can be used to investigate disease co-morbidity and patient stratification, paving the way for discovery of the underlying molecular level disease etiology in the form of overlapping genes and pathways. A longer-term perspective is to also include genetic profiles of the individuals in these data integration schemes, but this is not explored in the present paper.

## Results and Discussion

We based our study on a corpus of 5,543 patient records from the Sct. Hans Hospital (the largest Danish psychiatric hospital) collected in the period 1998 - 2008. For these records we extracted all assigned ICD10 codes from structured fields. Next we used a dictionary based on the Danish translation of the ICD10 classification to retrieve medical terms from the patients' free text entries in the corpus (see Materials and Methods). On average we found 9.5 ICD10 associations in addition to the 1.5 assigned codes (see Supplementary Information in the Appendix). Rounding ICD10 codes to the third level we found 351 different assigned ICD10 codes and 554 different mined codes. In total, 674 different ICD10 level 3 codes (see Materials and Methods) were represented in the corpus. Gathering all mined and assigned codes, we created a Patient-ICD10 association matrix, by assigning each Patient-ICD10 combination a binary value indicating whether or not a given code was associated with a given patient. The precision of our text mining was investigated by manually checking all 2724 mining hits for 48 patients (Table 2.2). The validation set covered 214 full level ICD10 codes, corresponding to 151 level 3 codes. A hit was considered correctly assigned when it was possible to infer a direct clinical link between the term and the patient from the record context. We defined precision in two ways: Incidence precision of all curated hits, and association precision, where an ICD10 code is considered correctly associated with a patient if it has at least one correct incidence. In both cases we considered how the precision was distributed among the different chapters (see SI). We found a total incidence precision of 87.78% and an association precision of 84.03%. The 333 false associations were further subdivided into categories with this distribution: Negations, 105; Wrong individual, 17; Delusion, 9; Putative, 40; Polysemic, 10; Patient information, 92; Other, 60 (see SI).

Chapter	Incidence precision (mining hits)			Association precision (ICD10 codes)		
	Correct	False	Precision	Correct	False	Precision
I	7	10	41.18%	7	6	53.85%
II	0	1	0.00%	0	1	0.00%
IV	30	4	88.24%	17	4	80.95%
V	486	20	96.05%	128	7	94.81%
VI	124	16	88.57%	46	9	83.64%
VII	19	13	59.38%	11	9	55.00%
IX	26	11	70.27%	13	5	72.22%
X	78	11	87.64%	36	4	90.00%
XI	67	12	84.81%	19	2	90.48%
XII	73	10	87.95%	29	9	76.32%
XIII	57	2	96.61%	17	2	89.47%
XIV	12	2	85.71%	6	1	85.71%
XVIII	1234	115	91.48%	252	53	82.62%
XIX	141	101	58.26%	36	8	81.82%
XX	4	0	100.00%	3	0	100.00%
XXI	33	5	86.84%	27	3	90.00%
All	2391	333	87.78%	647	123	84.03%

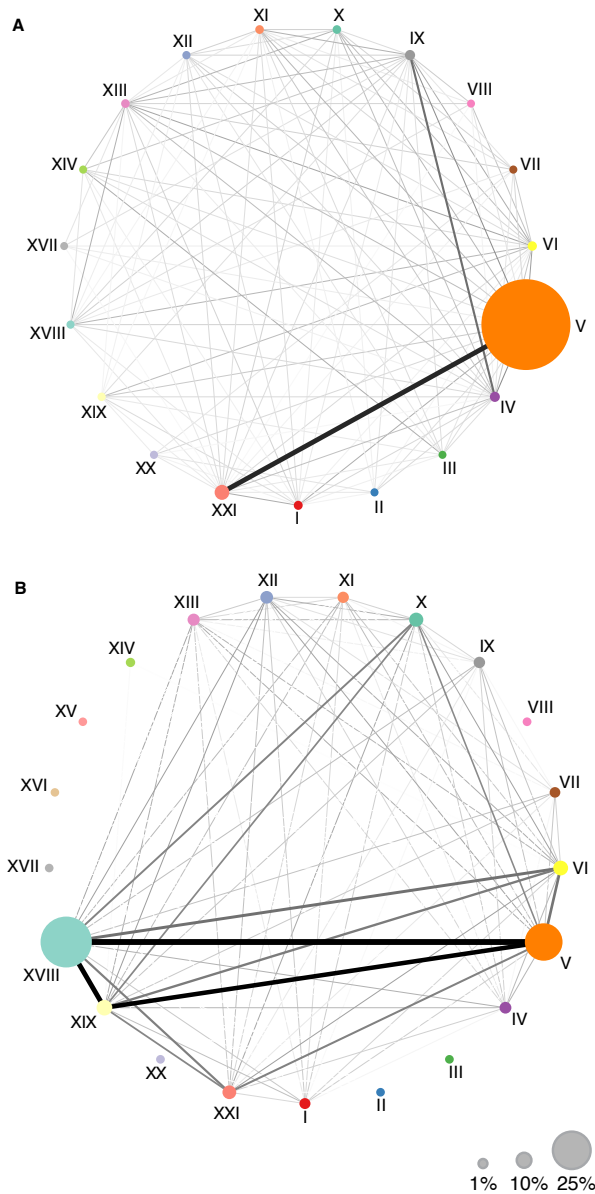
**Table 2.2:** Precision of text-mining associations. Precision is the number of true positives divided by the sum of true and false positives. Incidence precision distinguishes every individual mining hits as either correct or false. In association precision each ICD10 code is counted just once per patient and is considered correct if just one of the incidences of the code with this patient is correct. The final row contains the precision over all chapters.

## Co-morbidity

ICD10 is organized into 22 chapters according to disease areas (see Materials and Methods). To discover the degree of co-morbidity between chapters, we constructed an ICD10 chapter network (Figure 2.5a and b). Based on which diseases belonging to a specific chapter each patient has in the corpus, we calculated a similarity score between the different chapters, ranging between 0 (for the lowest co-morbidity), to 1 (highest co-morbidity), see Materials and Methods. Codes for chapter V ‘Mental and behavioral disorders’ account for over 80% of the assigned codes given by physicians at Sct. Hans Hospital, while codes for chapter XXI ‘Factors influencing health status and contact with health services’ have a frequency of around 7%. These are also the two most correlated chapters. The strong correlation between mental disorders of chapter V and the observational Z-diagnoses of chapter XXI is most likely explained by a large ward in the hospital for forensic psychiatry, where patients are frequently admitted for mental observation following a criminal offense.

When including both the assigned and the mined codes from the textual records we capture many symptomatic descriptions for diseases. As seen on Figure 1b, more than 35% of all codes are pertaining to chapter XVIII ‘Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified’, e.g. general medical complaints, edema, back pain, and elevated blood glucose. Chapter XIX ‘Injury, poisoning and certain other consequences of external causes’, as well as chapter XVIII, exhibit a high correlation with chapter V. Assigned codes are often restricted to the principal psychiatric illness and important for billing and social purposes, not necessarily reflecting the actual psychiatric treatment and care, nor the somatic disorders affecting the patient. For this reason, introducing the mined codes in the analysis allowed capturing correlations that were previously impossible to find.

In our attempt to identify pairs of interesting unexpected co-morbidities, as well as general trends of correlation, we investigated the ICD10 codes in patient space (columns in the patient-



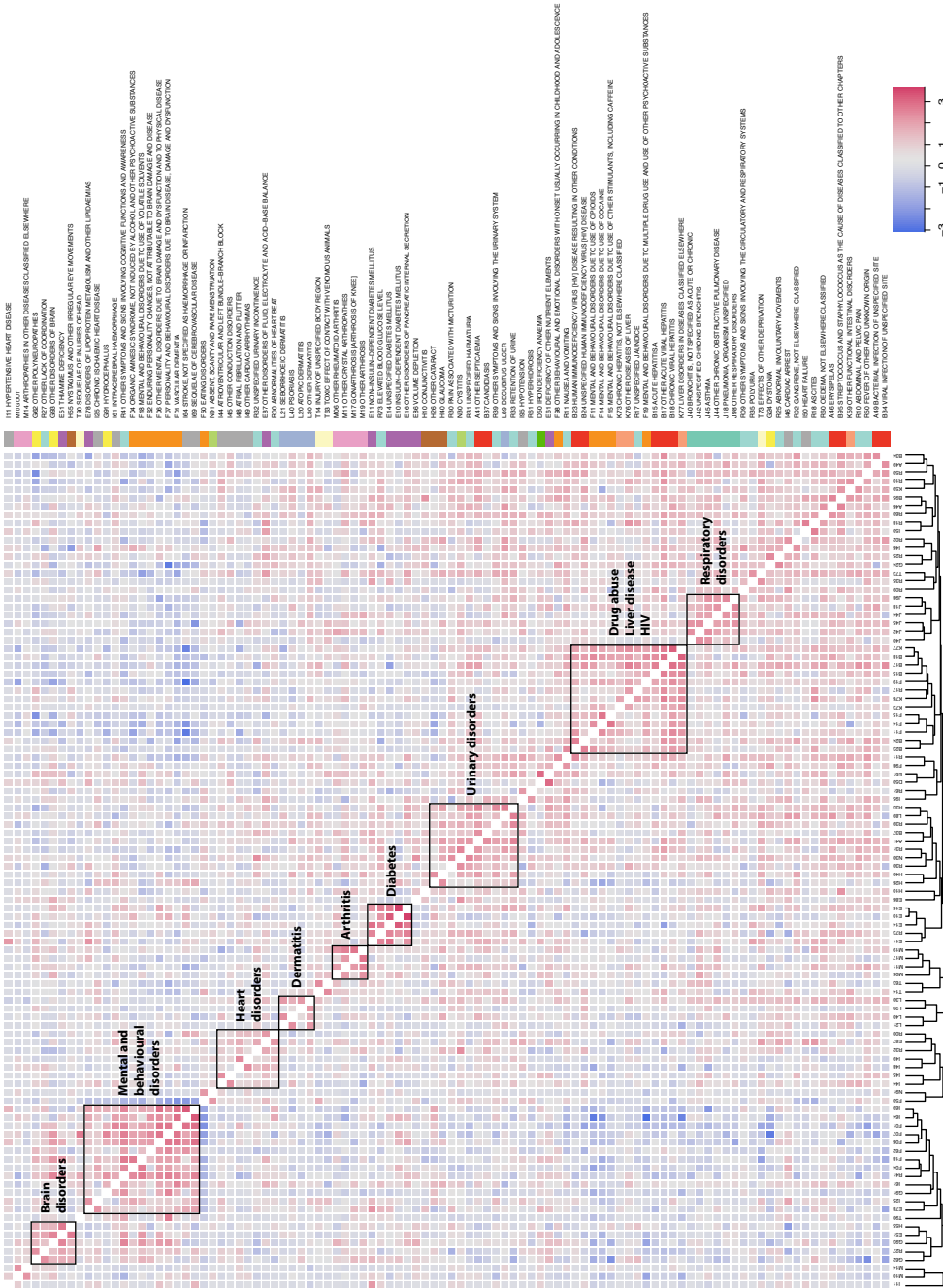
**Figure 2.5:** Disease Chapter Networks. ICD10 Chapters are nodes; links are correlations. Link weight represents correlation strength between two chapters; node area represents the proportion of codes from that chapter in the entire corpus. A. Network based only on the assigned codes for each patient. Most frequent chapter is chapter V 'Mental and behavioral disorders' with a frequency of 81%. The strongest correlation is between chapters V and XXI with a cosine similarity score of 0.45. Chapters IX, 'Diseases of the circulatory system' and IV 'Endocrine, nutritional and metabolic diseases' have a score of 0.3. B. Full network containing both the assigned and mined terms for all patients. Chapters V and XVIII have a frequency of 24% and 35% respectively, and have a score of 0.92. After mining, 'Diseases of the respiratory system' - chapter X, and 'Injury, poisoning and certain other consequences of external causes' - chapter XIX, now have a cosine similarity score of 0.6 and 0.78, respectively.

ICD10 association matrix). We used two measures to rank the 226,801 possible pairs of the 674 ICD10 codes, according to their co-association, compared to what would be randomly expected. Pairs were sorted based on p-values and a cut-off was imposed based on a co-morbidity score and a false discovery rate of 1% (see Materials and Methods). The result is a list of 802 candidate ICD10 diagnostic pairs that occur more than twice as often as expected by random, and that are statistically significant at a false discovery rate of 1% (Supplementary Table S1).

Using the co-morbidity score as a similarity measure we clustered all 674 ICD10 codes and created a corresponding heatmap of the co-morbidity scores for all ICD10 pairs. Figure 2.6 shows a truncated version of the entire heatmap, containing the scores of all the interactions for the top ranking 100 ICD10 codes; i.e. the top 100 codes found when sorting the list of 802 candidate pairs by their co-morbidity score. Figure 2.6 illustrates the general ability of our approach to capture correlations between different disorders. Several clusters of ICD10 codes relating to the same anatomical area or type of disorder can be identified along the diagonal of the heatmap. They range from trivial correlations, e.g. different arthritis disorders, to correlations of cause and effect codes, e.g. stroke and mental/behavioral disorders, to social and habitual correlations like drug abuse with liver diseases and HIV. Another interesting observation from figure 2 about the composition of the corpus is the lower than expected co-occurrence between the codes of the ‘mental and behavioral disorders’ cluster and the ‘drug abuse, liver disease, HIV’ cluster, as indicated by the blue areas in the upper and lower corners. These are very different groups of disorders that strongly stratify the patient corpus, and inspection of the specific diagnoses indicate that the correlation reflects two of the primary causes for admittance to Sct. Hans Hospital (i.e. two distinct clinical departments): psychiatric disorders caused by stroke or brain injury, and mental illness accompanied by drug abuse.

To discriminate potentially interesting, novel candidate co-morbidities from the many trivial ones, an experienced medical doctor manually inspected the candidate list of 802 pairs. Trivial pairs are e.g., between two codes for essentially the same disease (e.g. E11 ‘Non-insulin-dependent diabetes mellitus’ and R73 ‘Elevated blood glucose level’), or between trivial disease-symptom pairs (e.g. N30 ‘Cystitis’ and R30 ‘Pain associated with micturition’) or between pairs of well-established correlations (e.g. E51 ‘Thiamine deficiency’ and H55 ‘Nystagmus and other irregular eye movements’). Pairs with surprising correlations with or without possible hypothesis were flagged resulting in a list of 93 pairs. A full list of all the code-pairs analyzed can be seen in the Supplementary Table S2.

Disease correlations may or may not have genetic causes. To identify a possible molecular basis for the flagged pairs, we extracted genes implicated in those particular diseases when a good mapping from ICD10 to OMIM was possible (see Materials and Methods). We then created a protein-protein interaction network by determining the first order interactions of those genes in refined experimental proteomics data (see Materials and Methods). For each disease pair, we searched for shared first order interactions connecting the two networks. Despite the difficulty of mapping the different terminologies and genes with this approach [36], the analysis revealed several connected proteins which are novel in relation to the diseases used to generate the networks. For example, we narrowed down an interesting case story between Alopecia (i.e., hair loss, ICD10 L65) and Migraine (ICD10 G43). We found that THRA, thyroid hormone receptor, not previously associated with any of the two diseases, is a shared interaction partner of Protein Hairless (HR, a putative single zinc finger transcription factor protein) involved in



**Figure 2.6:** Heatmap of the first 100 ICD10 codes, based on ranking the list of 802 candidate pairs by their co-occurrence morbidity scores. Chapter colors are highlighted next to the ICD10 codes. Diseases that occur often together have red color in the heatmap, while those with lower than expected co-occurrence are colored blue. The color label shows the log2 change of co-morbidity between two diseases when compared to the expected.

alopecia [49], and the Estrogen Receptor 1 (ESR1) associated with migraine [50]. This may suggest that these two diseases share a similar molecular mechanism of action. Migraine and alopecia occurred in 12 patients with a co-morbidity score of 1.92 and a p-value of  $2.07 \times 10^{-6}$ . Most of the associations of these two codes come from mining, and in these cases a physician manually inspected their records to evaluate the clinical context. In two cases where the term ‘migraine’ was retrieved from a nursing note, a more correct clinical description would have been ‘headache’. One patient did not suffer from hair loss but from a somatic delusion thereof. Adjusting for those cases, the recalculated co-morbidity score is reduced to 0.4, while the recalculated p-value is  $2.81 \times 10^{-6}$ . Of the remaining 9 patients with migraine and alopecia, six are women aged 21-63 and three are men aged between 47 and 54.

The observed co-morbidity may reflect different side effects from medication [51--53]; most prominently seen with SSRIs (Selective Serotonin Re-uptake Inhibitors for treatment of depression) that have been associated with cutaneous reactions, including alopecia, and migraine [54]. Also, frequently prescribed oral contraceptives are associated with migraines (REF) [55]. In fact, inspection of the nine co-morbidity cases, revealed that three patients were being treated with SSRIs (with a possible link to hair loss mentioned in the medical notes), two patients were administered oral contraceptives and one patient was treated with calcium antagonists and anti-epileptic drugs

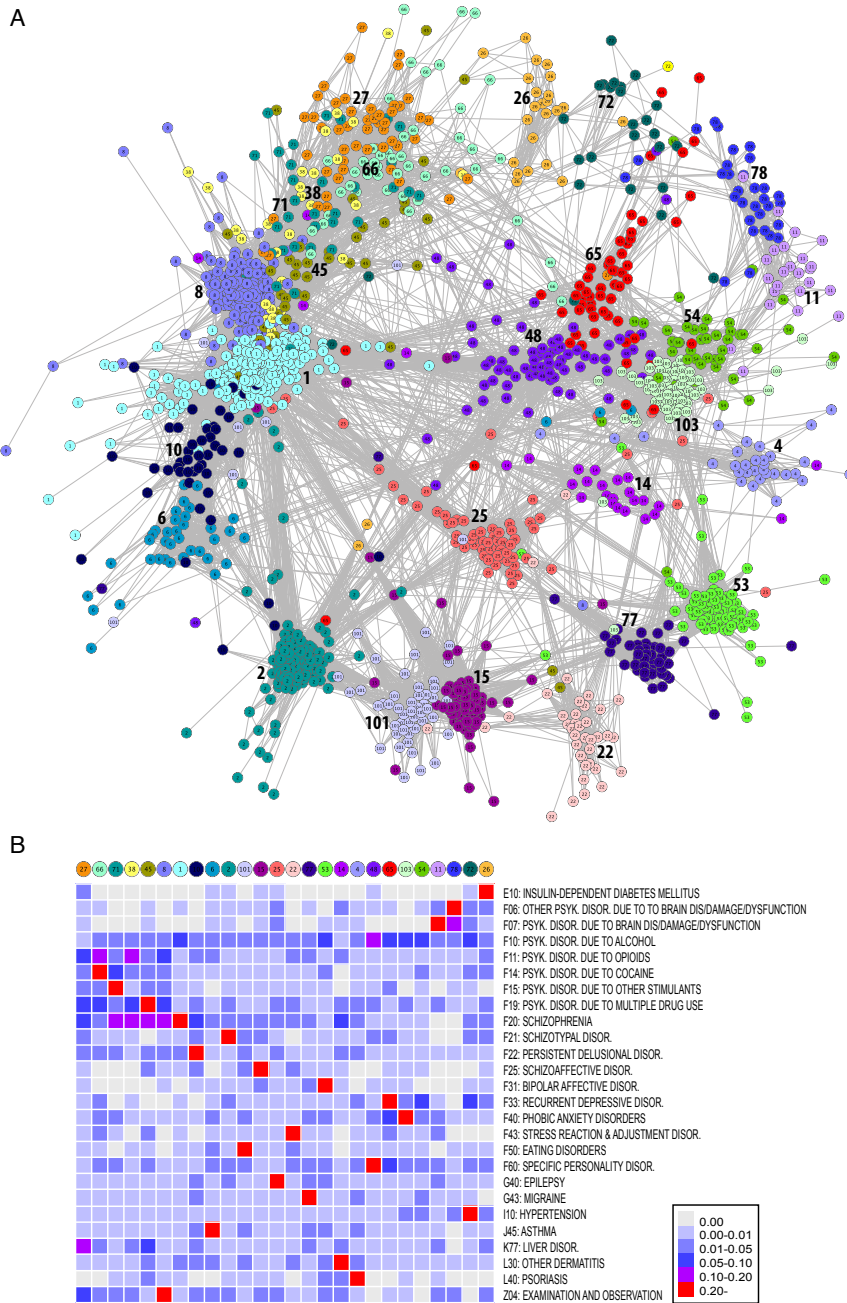
The co-morbidity may also have an etiological cause that relates to schizophrenia, the primary disease of the patients. It has previously been shown that schizophrenia is associated with coeliac disease, i.e. the highly under-diagnosed condition of gluten allergy [56], which in turn has been linked to both alopecia, and migraine; in fact the two latter conditions are now indications for diagnostic work-up for Coeliac disease according to the recent official US guidelines [57, 58].

### Patient stratification

In a specific hospital corpus of patients the most important level of stratification is generally based on the primary diagnosis, or inclusion, which dictates treatment and care. The stratification can also be very specific and based on lab results and tests for molecular markers, such as in the case of hormone receptor variants in breast cancer [59]. We were interested in determining if the combined mined data could lead to a richer structure in the patient population, spanning a wider range of phenotypes, not typically considered when stratifying a specific corpus by assigned codes.

In the patient-ICD10 association matrix each patient is represented as a vector of associated ICD10 codes in the space of all the 674 ICD10 codes. Based on these vectors we grouped patients into clusters according to the similarity of their phenotypes. Associations of ICD10 codes to patients were not treated as binary, but were weighted using TF-IDF (see Materials and Methods). Figure 2.7 shows members of the largest clusters resulting from a clustering based on Cosine-Similarity of patients ICD10 vectors (see Materials and Methods). In all but one cluster (54) one ICD10 code stands out as the most discriminating code. The TF-IDF value for this code constitutes up to 18-40% of the sum of all TF-IDF values in the vector. Furthermore, no two clusters share the same main code. The ICD10 characteristics of each cluster are shown in Figure 2.7b. From this figure, we see that Schizophrenia has a strong component in several clusters, primarily located in the top left of the network. As pictured, many of these clusters





**Figure 2.7:** Patient network. (A) Nodes represent 1497 patients from 26 clusters. Edges are correlations between patients. Node color denotes cluster membership. (B) Heatmap showing ICD10 composition of each cluster. Values are the fraction of the cluster ICD10 vector covered by this code. Shown are only the 26 ICD10 codes that are most distinguishing codes for a cluster. The heatmap columns match the network clusters in a counter clockwise direction starting at cluster 27.

are also characterized by various codes for alcohol/drug use, indicating the type of abuse as a good sub-stratification of schizophrenia. Similarly, alcohol seems to be a common denominator for clusters 48-54, which are primarily characterized by depressive disorders, anxiety disorders, and other personality disorders. What is also interesting is that many patients fall into clusters characterized by somatic codes like diabetes and psoriasis, which have certainly not been the initial reason for admittance to the hospital. This is largely attributable to data coming from text mining (see Supplementary Table S3 in the Appendix).

## Discussion

As EPR systems become the norm in modern health care, focus is naturally turned to exploring this treasure trove of data for improving health care and research [60]. Extracting the data is a first step, and as EPR systems maintain the use of free text to complement structured data, text-mining approaches are necessary for extracting data usable in further analyses.

The enrichment of existing structured patient data by text mining significantly expands phenotype profiles, both within the specific pathology of the corpus, but especially into other disease areas. We present one example of co-morbidity between two diseases that are very often not coded in the record by the physician, but show up written in the patient journal and are later picked up by mining. The enrichment from mining is also visible in our attempts to stratify patients, where it shows potential for uncovering additional layers of the population structure. More detailed stratification of patient cohorts could help improve population homogeneity and signal strength in Genome Wide Association Studies, leading to stronger results in smaller case-control studies.

The procedure described here represents, in our opinion, a practical non-hypothesis driven approach for extracting valuable information from patient records where manual inspection and ICD10 association would turn into an otherwise impossible task. Furthermore, we show how this information can be used in researching disease co-morbidity and patient stratification and how it can be mapped to the underlying systems biology revealing possible causes for the observed correlations.

## Materials and Methods

### Patient Corpus

The patient population data was from collected from the Sct. Hans Mental Health Centre, in Roskilde, Denmark. A total of 5543 patients, were followed from 1998-2008, and their records stored in an EPR database. 70% (4822) patients are from the Copenhagen area, 61% of these are male. The average age is 30 years old. The records are a mix of structured diagnose assignments of ICD10 codes, ATC<sup>4</sup> codes for medication usage, patient care notes from nurses and doctors, admission and personal information, etc. A corpus was created containing all text entries for each patient that were verified and signed by a physician. In total, the corpus contains text for 4765 patients. To each entry we assign an entry date, the note type, and the text. The note type identifies the type of text entry, such as the epicrisis, discharge note, treatment note,

---

<sup>4</sup><http://www.whocc.no/atc>

nursing note etc. Each record contained approximately 25,000 words in free text. In addition, we extracted all ICD10 codes assigned to patients that were stored in a structured format.

### Data extraction

ICD10 associations were extracted from the free text entries of the patient records. The texts were parsed (exact matching) against a dictionary based on the Danish translation of the WHO International Classification of Diseases (ICD10), downloaded from the Danish national board of health the 2<sup>nd</sup> Nov 2009. ICD10 is divided into 22 chapters, and has a hierarchical structure with increased specification of terms in each lower level. Each term is uniquely matched to code of between 3 and 5 characters. Our dictionary contains all the Danish ICD10 terms, plus a number of permutations of these, to reflect language usage rules [61]. The objective was to get as many terms as possible for each ICD10 code. In addition, a blacklist was also created to remove dubious and uninformative terms from the text. The final dictionary consisted of 53452 terms. During parsing, candidate terms were tested for preceding negations or mention of family members, which acted as disqualifiers. For the data analysis all codes were rounded up to the third level for consistency and increased precision. Further information about the dictionary generation rules, parsing and rounding is contained in the Supplementary Information text.

### Chapter networks

For each disease we created a vector mapping its presence or absence from a patient record. This resulted in 22 vectors for each disease chapter. The pair wise overlap between vectors was quantified by calculating the cosine of the angle between normalized vector pairs [47]. The result is a score between 0 and 1, mapping the co-morbidity value of each of the chapter pairs. We also calculated the frequency of each chapter in relation to the total number of chapter assignments. In Figure 2.5, the roman numerals represent the different ICD10 chapter numbers: I, Certain infectious and parasitic diseases; II, Neoplasms; III, Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism; IV, Endocrine, nutritional and metabolic diseases; V, Mental and behavioral disorders; VI, Diseases of the nervous system; VII, Diseases of the eye and adnexa; VIII, Diseases of the ear and mastoid process; IX, Diseases of the circulatory system; X, Diseases of the respiratory system; XI, Diseases of the digestive system; XII, Diseases of the skin and subcutaneous tissue; XIII, Diseases of the musculoskeletal system and connective tissue; XIV, Diseases of the genitourinary system; XV, Pregnancy, childbirth and the puerperium; XVI, Certain conditions originating in the perinatal period; XVII, Congenital malformations, deformations and chromosomal abnormalities; XVIII, Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; XIX, Injury, poisoning and certain other consequences of external causes; XX, External causes of morbidity and mortality; XXI, Factors influencing health status and contact with health services; XXII, Codes for special purposes.

### Co-morbidity ranking

For the purpose of exploring co-morbidity between ICD10 codes we used two measures to rank the 226801 possible  $((674 \times 674 - 674) / 2)$  pairs of different codes, according to how often

they come together in patients, compared to what would be randomly expected assuming no a-priori correlations. The two measures represent our desire to ensure statistical significance, while focusing on pairs with a noticeably increased co-association.

First, for each pair of ICD10 codes A and B, the patient corpus is divided and counted in the four categories: A & B, A NOT B, B NOT A and NOT A NOT B, according to their association to A and B. Using this p-values are calculated using Fishers exact test, and the pairs are sorted accordingly. We then filtered this list by imposing a cut-off value of 1.0 of a co-morbidity score between diseases A and B defined as:

$$CS_{AB} = \ln_2 \left( \frac{Obs + 1}{Expt + 1} \right), Expt = \frac{n_A \cdot n_B}{n_{tot}}$$

Where *Obs* is the observed number of ICD10 co-associations, and *Expt* is the expected number. Expected overlaps are calculated based on the prevalence of each disease in the actual corpus ( $n_A$  and  $n_B$ ). To make the tendency to favour pairs of low prevalence ICD10 codes less pronounced, a pseudo-count of 1 is added to nominator and denominator. Since we take  $\log_2$  of this ratio, a cut-off value of 1.0 means we restrict our focus to pairs with a higher than two fold (approximately) over co-association. This co-morbidity measure is very similar to the one used by *Hidalgo et al.* [62].

Finally we used a Benjamini-Hockberg false discovery rate method on the ranked list to correct for multiple testing. The p-values for all pairs are multiplied by the total number of pairs (226801) and divided by the rank of the pair in the sorted list. A cut-off is then imposed where the corrected p-value drops below 0.01. The result is a selection of 802 potentially interesting candidate pairs, with a false discovery rate of 1 percent, from the total of 226801 pairs.

### Creating gene lists from ICD10 codes

There is no direct mapping between ICD10 codes and OMIM [63] record entries. Furthermore the disease names used by ICD10 and OMIM are not identical, so there was a need to map OMIM disease names into ICD10 codes. Work has been done mapping the online database and ICD9 codes, a previous version of the ICD [36]. We used the ICD10 to ICD9 General Equivalence Mapping available online from CMS (<http://www.cms.gov/ICD10/>) to map the ICD codes to their previous version. With the mappings in place, OMIM was parsed for phenotypic descriptions of defects in genes, as described in *Lage et al.* [47]. From the OMIM records, the *clinical synopsis* field was extracted for retrieving phenotypic descriptions regarding a certain disease. Additional information was retrieved from the *morbidity map* tables, a map of disorders included in OMIM that have the syndrome name, chromosomal localization, and name of the disease causing gene. A manual curation step by a medical doctor ensured that each ICD10 code to be included in the analysis was assigned the correct OMIM entries.

### Genetic overlaps between ICD10 pairs

For each disease, a network was generated by taking the disease causing genes extracted from OMIM and determining their first order interactions in a human protein interaction network of refined experimental proteomics data. This procedure is described in detail by *Lage et al.* [47,

[64, 65]. For determining genetic overlaps between two ICD10 diseases, we take their networks and identify those genes which are shared and have first order interactions with the seed genes. After a round of automatic overlap detection, we manually curated the results of the different steps in the pipeline, in order to detect erroneous assignments of disease names or genes, and reran the overlap detection in those cases. For those pairs where overlapping protein-protein interaction networks indicate underlying biological evidence, a final round of validation was done by manually checking if the binary associations from text mining of patients to the ICD10 codes were correct. Based on the corrected data, new p-values were calculated by Fishers exact test, and it was controlled that the p-value remained lower than the lowest p-value of the list of 802 candidates. The candidate genes found to overlap in the two disease networks were scored using the enrichment of OMIM seed genes in their first order interaction network, in a similar procedure as the one used by *Lage et al., 2010* [64]. The score assigned to a candidate was the hyper geometric p value of observing the amount of interactions to the OMIM set out of all the interaction partners of the candidate. Our example of THRA has a total of seventeen interaction partners in the network, and two are with the input genes (HR and ESR1), having a p-value of  $1.17 \times 10^{-3}$ .

### Patient stratification

By looking at the Patient-ICD10 matrix by rows, or patient vectors in ICD10 space, we can stratify patients based on the codes they have. Instead of a binary association of a given code to a given patient, we weighed the significance of icd10 occurrences using the term frequency – inverse document frequency measure (TF-IDF). TF-IDF rewards high code frequency in the individual record, and penalizes high prevalence across the corpus. As a patient-patient stratification measure, we used the cosine similarity CS [47] to calculate the cosine of the angle between all pairs of vectors. We included only patients with at least 3 associated codes, and exclude a number of trivial/symptom codes (e.g. pain, coughing, itching). A total of 2584 patients were found to have at least 3 associated codes. We used 1-CS as a distance measure and calculated average linkage clustering to divide patients into clusters. Manual inspection of the clustering dendrogram led us to cut the tree at a CS value of 0.6, which created a total of 307 clusters. 26 clusters contained 25 or more members, accounting for a total of 1800 patients. Taking all edges with CS greater than 0.6 between these patients, the network in Figure 3a of 1497 patients was created. The network layout is based purely on an edge weighted layout algorithm. In order to investigate the clinical characteristics of each cluster, we concatenated the assigned and mined data for all members of a cluster, and calculated a new TF-IDF code vector for the entire cluster.

### Acknowledgements

The authors would like to thank Kasper Lage for feedback and critical discussions. The work carried out in this study was supported by the Villum Kann Rasmussen Foundation, the Novo Nordisk Foundation and the Danish Strategic Research Council. The project was approved by the National Board of Health, Denmark (No. J. nr. 7-604-04-2/33/EHE).

## 2.4 Paper II

# A Comparison of Several Key Information Visualization Systems for Secondary Use of Electronic Health Record Content

Accepted for publication in the of *Louhi 2010: Second Louhi Workshop on Text and Data  
Mining of Health Documents*

★Francisco S. Roque<sup>1</sup>, Laura Slaughter<sup>2,3</sup>, Alexandr Tkatšenko<sup>4,5</sup>

<sup>1</sup>Center for Biological Sequence Analysis, The Technical University of Denmark, Lyngby, Denmark

<sup>2</sup>The Interventional Center, Oslo University Hospital, Oslo, Norway

<sup>3</sup>Department of Computer and Information Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

<sup>4</sup>Institute of Computer Science, University of Tartu, Tartu, Estonia

<sup>5</sup>Software Technology and Applications Competence Center, Tartu, Estonia  
<http://dsv.su.se/hexanord>

★All three authors contributed equally to this work.

### Abstract

An overview is provided of six information visualization systems designed specifically for gaining an overview of electronic health records (EHR). The systems discussed all make use of timelines: Lifelines, Lifelines2, KNAVE II, CLEF Visual Navigator, Timeline, and AsbruView. With the exception of Lifelines2, the main user groups targeted are physicians involved in direct patient care. Little attention has been paid towards supporting true secondary use of EHR contents, for activities such as assessing quality of care, patient health and safety monitoring, and clinical trial recruitment. Future work on such systems needs to address the complexity of EHR data, missing and incomplete information, and difficulties in displaying data with differing levels of granularity.

### Introduction

This paper provides an overview of several information visualization (infovis) systems that have been built for exploring abstracted information from Electronic Health Records (EHR). EHRs are systems that are used to document care of patients. The records can include a wide range of data and information, including medications prescribed and administered, immunization history, laboratory test results, allergies, radiology images, treatment plans, and care notes. Currently, most EHR systems implemented are proprietary and highly customized when used by larger care institutions.

It is usually the case that only clinicians and other healthcare professionals with direct responsibility for providing care have access to patient data. The suggestion of secondary use of health data is not new and has been handled separately from the issue of creating user interfaces and visualizations. *Safran et al.* [15] discuss the purpose of clinical data repositories in their white paper and point towards the goal of a national framework for the secondary use of health data in the U.S. According to their definition, secondary use includes activities such as analysis, research, quality and safety measurement, public health, payment, provider certification and accreditation, marketing, and general business applications, while at the same time taking into account the ethical, political, technical and social implications of such re-use. *De Lusignan and van Weel* [66] highlight the challenges of making use of clinical data for research, stating, “The available research methods for working with large data sets are limited; it is difficult to infer meaning from data; there is a rapid pace of change in both medicine and technology; and integrating data without reliable unique identifiers is difficult.” *Prokosch and Ganslandt* [16] have recently summarized the latest advances in enabling clinical data re-use for research purposes. They identify as key challenges the establishment of comprehensive clinical data repositories, the establishment of professional IT infrastructure to support clinical data capture, and the integration of medical record systems and clinical trial databases. As discussed in these articles, aggregated, abstracted and manipulable information is underutilized and hard to come by.

The emerging field of *Visual Analytics* [67] is relevant to this review. This field is focusing on combining related research areas such as visualization, data mining and statistics to handle large and heterogeneous volumes of data, such as EHR. The systems we encountered are integrating human judgment with automated analysis, suggesting that future work will be related to handling massive amounts of data that contains missing elements --- including the results of textual analysis of records content.

### Purpose

Our motivation for creating this overview is to compare and discuss some of the available information visualization/visual analytics tools and how are these used for secondary, i.e. for purposes other than direct patient care. This is a first step towards infrastructure and coordinating efforts to produce systems that are based on standard input formats, and meet the needs of specifically defined users. The reader of this overview is most likely working on information extraction, temporal abstraction, and summarizing EHRs.

<i>Source</i>	<i>Search Keywords</i>
Pubmed	visualization health records Medical Records Systems, Computerized Computer Graphics User-Computer Interface
ACM DL	electronic health records or medical record information visualization or visualization healthcare or health care user interface
IEEE DL	visualization medical records
Google Scholar	electronic medical records or EHR information visualization visual analytics

**Table 2.3:** Keywords searched.

### Scope

The review is non-systematic. We didn't expect to find large numbers of articles, since this is a relatively narrow area of interest. The search was confined to user interfaces and visualizations for EHR data, we searched pubmed, ACM digital library, IEEE library, and Google Scholar, using basic keywords and checked references in found articles. We also looked for papers on work we had read or known about previously from conferences or other sources. The literature search covered articles in English only. Keywords used are listed in Table 2.3.

### Systems

In this section we give an overview of the state-of-the-art systems related to visualization of temporal information in EHRs. Our intention is to cover broad areas of application including representation of medical histories, visual data query and aggregation, generation of temporal abstractions and visualization of treatment plans. Due to the limitations in space, we focus



only on the most representative systems, which feature interesting and potentially reusable visualization techniques.

### **Lifelines**

LifeLines uses a timeline visualization technique to represent personal histories, medical records and other types on biographical data [68]. In LifeLines, horizontal bars are used to depict temporal duration and location of events on a horizontal time axis. Similar events are organized into facets, which can be expanded and collapsed to provide increasing or decreasing level of detail. Color notations and line thickness are used to indicate the importance and relationship of events. To handle regions with high data density, LifeLines provides zooming functionality allowing users to compress and stretch the time scale at any location. Additional content (e.g., multimedia) can be added in a linked fashion. Authors apply LifeLines in the analysis of complex patient medical records to visualize temporal relationships between treatments, consultations, disorders, prescriptions, hospitalizations and other events.

### **Lifelines2**

LifeLines2 [69] is an extension of LifeLines, allowing the user to analyze records from multiple patients at a time. The system facilitates comparative visualization of records by means of aligning, filtering and sorting operations. By aligning patient records on some common reference event (e.g., the first heart attack), users can easily spot co-occurring and neighboring events. Ranking and filtering operations complement alignment by interactively reordering or narrowing the set of records to suit a user's changing focus. The system proved to be particularly suitable for observational research, where researchers analyze data from different studies in order to better understand health problems or study the effect of treatments, and in finding patients for clinical trials. Evaluation studies showed that the system significantly simplifies typical analytical tasks and that medical specialists can quickly learn the interface. LifeLines2 is currently used to display EHR data provided by the Informatics for Integrating Biology & the Bedside (i2b2) Project [18].

While in LifeLines2 the main focus is on visualizing temporal ordering of events, *Wang et al.* [70] emphasizes practical need in viewing multiple records as an aggregate in order to study frequency of event data over time. For instance, a user might be interested to analyze blood pressure of all patients who have had an open-heart surgery within 3 months of their first heart attack. As a solution, authors complement LifeLines2 framework with a new visualization technique, called temporal summaries, which represents distributional trends of events over a set of records in a histogram-like chart. Furthermore, the system allows splitting the whole dataset of records into multiple subsets and use temporal summaries to compare event patterns between these groups.

### **CLEF**

Hallet [71] proposes a visualization architecture for browsing medical histories, which integrates visual navigation tools and automatically generated textual summaries. While the graphical interface facilitates interactive navigation, textual descriptions can, in addition, convey

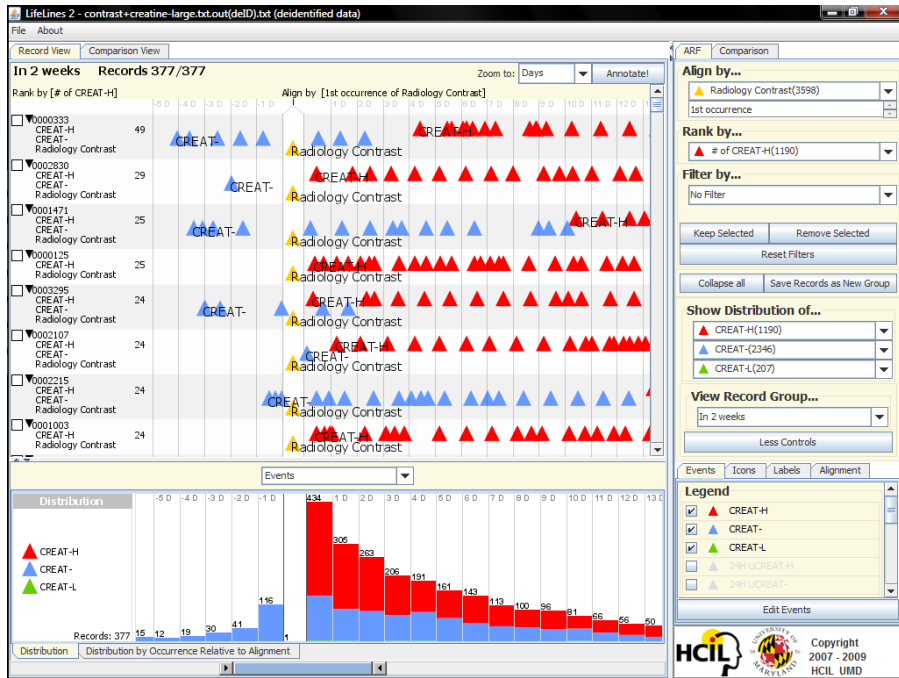


Figure 2.8: The Lifelines2 main window, with focus on timelines.

complex temporal information and display details that would otherwise be too complex for visualization components. Within the system, the patient's medical history is represented as a network of semantically and temporally organized events, which serves as an input for visualization and natural language generation components. The visual navigator depicts a high level overview of a patient's medical history by plotting events along three parallel timelines, corresponding to diagnoses, treatments and investigations. In addition to zooming time scale and detail-on-demand functionality, the navigator provides interactive visualization of semantical relationships between events (e.g., caused-by, has-locus, indicated-by, etc.). Having different features from the LifeLines interface, the navigator also allows the user to visualize numerical data (e.g., results of blood tests) by plotting results of measurements on separate line charts. Natural language generation is used for two purposes: 1) to create customized textual reports for printing or exchange purposes and 2) as a support tool for the visual navigator, to enable better description of complex events and relationships between them.

## KNAVE-II

KNAVE-II [72] is an interface enabling knowledge-based visualization and interactive exploration of time-oriented data at different levels of temporal abstractions (e.g., abstraction of periods of bone marrow toxicity from raw individual hematological data). Users can navigate through the links of a semantic network while simultaneously navigating visually through multiple degrees of temporal abstraction of the dataset under observation. The evaluation results have shown that users of KNAVE-II were able to perform queries both faster and more accurately than with other standard tools.

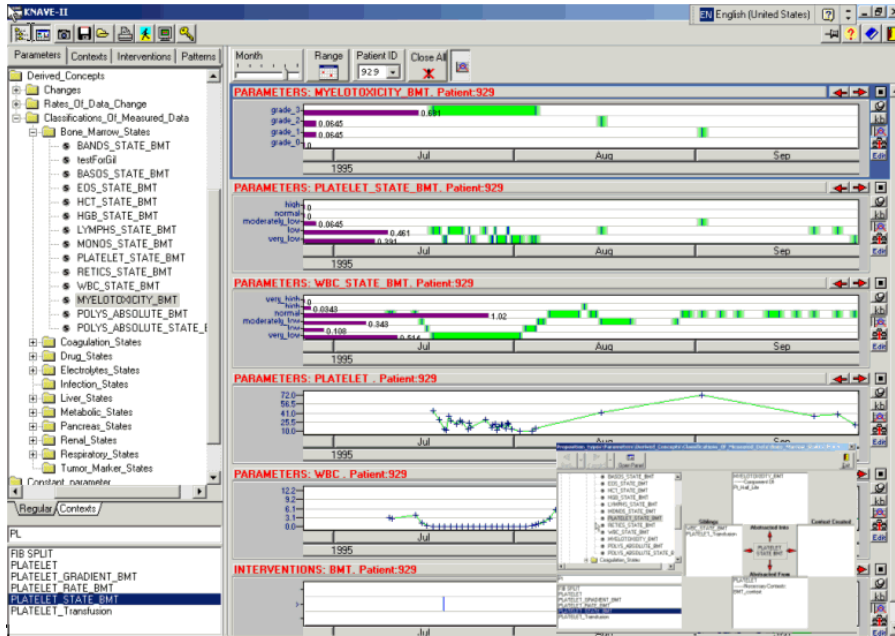


Figure 2.9: The Knaive-II system.

## TimeLine

The TimeLine system [73] is a problem-centric temporal visualization of patient records. The contents of the EHR are integrated, reorganized, and displayed within the user interface (UI) along a timeline. It is similar to Lifelines in the way that the different elements of the EHR are grouped along the y-axis: imaging, reports, lab tests, etc are collapsible categories. However, unlike Lifelines, the TimeLine system uses an XML data representation to handle data from distributed, heterogeneous medical databases. Data elements that are displayed in the UI are classified based on a knowledge base that guides both data inclusion rules and the visualization metaphors used to render the data.

## AsbruView

AsbruView [74] is a visualization and user interface on top of Asbru language [75] designed to represent treatment procedures as structured time-oriented plans. AsbruView represents hierarchical and temporal relationships between treatment plans using a 3D visualization perspective. Plans are aligned along the time axis and can be stacked on top of each other and laid out in different ways. To simplify the interface, all graphic elements are represented by well-known real world objects (e.g., track, traffic light, etc.). Also a 2D view is available which focuses on temporal aspects of plans in greater detail. To depict uncertainty of future events, AsbruView extends the timeline by using time annotation glyphs [76].

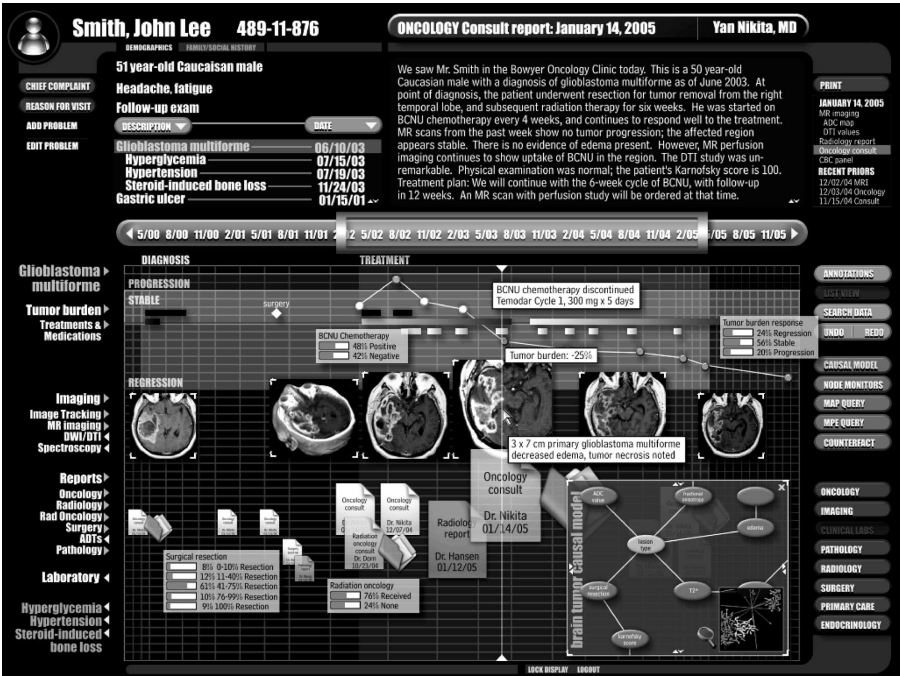


Figure 2.10: Timeline system.

### Comparisons

Infovis techniques are a way of augmenting human cognitive capabilities, to help humans find patterns in large volumes of data. The systems described above target specific user types that will benefit from the visualization methods. While some user interfaces were developed in close dialog with medical practitioners, like Lifelines2 and Knave-II, others, such as the first Lifelines, Clef and Asbruvview have had only minimal input from their intended audience.

### Users, Goals and Tasks

Most of the tools were directed at clinicians and clinical practice, although they were not always developed in close relation to them. Table 2 gives an overview of intended users for each of the named systems, and their proposed goals/tasks. From the user point of view, a number of tasks and goals can be defined for each tool. Some are very specific and tend to care for niche usages, while others provide more general visualization methods that can be applied to a number of situations.

These systems were designed with input from only a few medical personnel involved in the project. In general, articles we read concerning these systems that have a more guided development process, i.e. closely related with physicians, have more specific goals and tasks, because they were designed with these in mind. Visualizing data for decision-making and analyzing treatment outcome is often a general goal in many of the tools developed in interaction with medical staff [77--79]. There is an emphasis on pre-processed patient data, specifically numeric, such as lab tests, heart rate, and blood pressure. Systems mainly try to help physicians answer

<i>System</i>	<i>Users, Goals, Tasks</i>
Lifelines	Clinician Patient care Use EHR content in temporal time-based view
Lifelines2	Clinical researchers Research Compare patterns of events, detecting trends
CLEF	Clinician, Biomedical researcher Patient care Visualize timelines, use NLP to extract complex temporal data, aggregate numerical data
KNAVE-II	Clinician Patient care Generation and exploration of context sensitive abstractions of temporal data
TimeLine	Clinician Patient care Use EHR content in temporal time-based view, with additional filters on data based on NLP techniques
AsbruView	Clinician Patient care Medical therapy planning and execution

**Table 2.4:** Users, Goals, Tasks.

questions about correlations in the patient's data, and provide a means for supporting quick decision-making when combining several types of highly heterogeneous data. Physicians can follow a specific treatment plan and check the patient's physiological variables over time. This also enables the practitioner to check the influence of certain variables in the treatment process and change the protocol if needed. CLEF, for example, allows the physician to discover events during specific time spans, such as searching for past specific liver problems. Lifelines2 is specifically geared towards research uses and towards answering complex queries. In Lifelines2, a case study involved verifying the results of a clinical study with real-life EHR data to see if clinical care data differ from the study results.

The systems we discuss have conducted evaluation studies as a part of the end-stages of development. The Lifelines evaluations were conducted in another domain (use of pattern searching related to monitoring graduate student progress), with limited interviews and input from experts in the medical domain. The KNAVE system conducted a crossover study with doctors, comparing KNAVE with existing tools. TimeLine was evaluated following the development of the interface by five radiologists- focusing on questions related to data integration and the temporal display. AsbruView was evaluated using questionnaires sent to clinicians.

## Visualization Methods

The focus of this paper is on temporal visualization methods since this has been the primary visualization type studied for aiding humans in organizing and exploring patterns in abstracted EHR content. All the systems that are compared in this paper display some type of timeline with time running from the left part the screen to the right, time being on the x-axis, and categories of events along the y-axis. Various techniques for graphically representing specific events are used (e.g. icons, shapes), AsbruView makes use of 3D, while all the others are flat 2D.

Infovis has been the keyword used to describe these systems, with the idea of presenting a method for human users (most often stated as being clinicians), to recognize patterns and thereby *amplify cognition* [80]. Other methods for recognizing patterns in EHR for secondary use are purely automated and conducted through data mining techniques. *Bertini and Lalanne* [81] wrote about the complementary role of automatic data analysis and visualization in knowledge discovery. They discuss *visual analytics*, an outgrowth of infovis that can be seen as an integrated approach combining visualization, human factors, and data analysis. They suggest 4 categories for classifying approaches: Pure Visualization (VIS), Computationally-enhanced Visualization (V++), Visually enhanced Mining (M++), and Integrated Visualization and Mining (VM). In the systems we have compared, there is a spectrum of ideas about how to visualize EHR contents, including movements towards *enhanced* or *intelligence* in the processing of the underlying EHR data. In Lifelines2, the data visualized was obtained from anonymized EHRs through cooperation with the i2b2 Project [18]. The input form of the data is a simple 3-column table containing 'ID', 'Event Type', and 'Time'. Each ID can have multiple events happening at various times. Lifelines2 allows sorting of the data so that records with the most incidents of one type of event are shown at the top of the screen. This type of infovis relies on human pattern recognition only and would be considered as *VIS* by *Bertini and Lalanne* [81]. In the CLEF project, the CLEF Chronicle, which underlies the visualizations, is a semantic network modeling of what happened to the patient, why, and how. Semantic relations are: causality, reason, finding, and consequence. The types of events modeled are: problem, investigation, and treatment. The CLEF Visual Navigator might be considered as *V++*, computationally enhanced visualization because some sort of automated computation supports the visualization. In CLEF, the visual display is "enhanced with visual techniques for highlighting relationships between events on the timeline." None of the systems so far that we have seen, would qualify as *visually enhanced mining* or *integrated visualization and mining*. Table 2.5 provides a full overview for all systems reviewed.

The papers we have read that cover EHR visualization, as seen in the systems presented, express the complexity of abstracted EHR data. Missing and inconsistent data, dealing with hierarchical data, and problems with granularity are all concerns that become readily apparent through attempting to build infovis systems. *Wang* [69] summed it up best "Clinical data tend to be messy with aspects that become only obvious when the data is visualized. The same heart attack might be recorded three times in three days (by the emergency room physician, a cardiologist, and a clerk from the billing office) and it can be hard to differentiate it from 3 separate events. Even if medical event information is carefully recorded at the time of the doctor visit or during a hospitalization, the time stamp is usually inaccurate by nature." Future

<i>System</i>	<i>Category</i>	<i>Notes</i>
Lifelines	VIS	
Lifelines 2	VIS	
CLEF	V++	automated generation of summaries semantic network of EHR record events
KNAVE-II	V++	semantic (ontology-based) navigation and exploration of the data knowledge base is used to interpret raw data
TimeLine	V++	data mapping and reorganization content-based techniques to elucidate predominant sub- ject of reports for classification
AsbruView	VIS	

**Table 2.5:** Visual Analytics of Systems using the classification from *Bertini and Lalanne* [81].

work on visualizations needs to adequately address the complexity of the data rather than work with test data that is too simplistic.

### Text Mining Tasks

All mentioned systems, except the CLEF and TimeLine, operate with readily available lists of type- and time-tagged events. However, clinical records are often stored in textual form what makes them inaccessible for machine processing. Text mining techniques need to be applied to automatically transform textual data into structured, normalized form. Key tasks involve event extraction, classification and normalization.

The CLEF system uses an advanced information extraction engine to identify pre-defined classes of entities (e.g. diseases, investigations, problems, drugs, etc.) and semantic relationships between them (e.g. investigation indicates problem) in natural language texts. The information extraction process involves lexical and terminological analysis, syntactic and semantic analysis, and discourse analysis. To address the complexity of medical language, the system makes use of language resources including the Unified Medical Language System and the Gene Ontology. Extracted information is stored in templates, which can be queued or used to generate textual summaries. The TimeLine system makes use of both textual contents of the EHR as well as numerical data and codes. An NLP-based system is used in conjunction with the TimeLine UI, for example, performing section analysis in radiology reports to determine whether specific subsections exist within the reports that are related to certain medical problems [73].

### Conclusions

The infovis systems analyzed allow secondary use of EHR content data especially aimed at clinicians documenting patient care. All of them are focused on visualizing temporal data in a timeline, while displaying specific events from the patient data.

Although directed at medical practitioners in their daily patient care routine, they were not always developed with user feedback. Evaluation of the different tools was often based on situations outside of the clinical setting, and might not reflect reality. A more intimate dialog with clinicians would benefit the creation of targeted systems addressing specific needs of the medical community.

The overall goal of these tools is to present users temporal information contained in a record, improving their ability to recognize patterns for knowledge discovery and following treatment. They introduce simple visualization tools, but some include automated computational enhancements supporting it.

EHR contain missing and inconsistent data, which is in general messy. Due to the complexity of the underlying data, future work needs to address these intricacies rather than using simplistic approaches.

### **Acknowledgments**

We would like to thank Nordforsk and the Nordic Council of Ministers for the funding of our research network HEXAnord - Health text Analysis network in the Nordic and Baltic countries. This work was partially supported by a grant from the Villum Kann Rasmussen fund.



## Disease gene finding

**B**ASED on the hypothesis that diseases are an effect of disrupted functional networks which regulate the many systems in the human body, this chapter introduces some of the tools used to gain further insight into the underlying complex interactions of many disease etiologies.

Our work in many of the articles in this thesis is predicated on the following: if we consider a functional module (complex) of interacting proteins, the phenotypic effect of disrupting any single protein in the module will be very similar independently of which individual proteins are disrupted. If this assumption generally holds true, then protein interaction data can be used for discovering novel proteins related to a disease just by having a set of candidate proteins and reliable functional modules of the disease. For this purpose we have constructed an inferred human protein interaction network, described in Section 3.1. This network was used in most of the analyses in this thesis.

In the following sections I will briefly introduce the human interactome, and how it can be used for disease gene finding. The CBS in-house inferred human protein interaction network is described, and the data underlying explained.

In the paper entitled *A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes*, we analyze the tissue specificity of disease genes and complexes. In this integrative approach we systematically intersect pathologies, tissues, protein complexes and gene expression to interpret the underlying causes of tissue-specific pathology on several layers of cellular organization. We used text mining to map disease phenotypes to tissues, while OMIM [63] and Pubmed were used to map disease phenotypes to genes. Gene complexes were mapped to tissues using gene expression array data. The resulting list of generated disease complexes is also used in Paper VI, described in Section 4.3 of the next chapter. The next manuscript, entitled *Dissecting spatio-temporal protein networks driving human heart development and related disorders*, combines phenotypic information from specific knockout mutants in genes leading to heart pathologies, with protein interaction data, in order to explore the systems biology driving organ development.

### 3.1 The human interactome

Proteins do not act as isolated units in the human body. They interact with each other, and these physical interactions are a strong indication of functional association [82]. These interactions allow them to co-operate and perform vital cellular functions, working together in functional modules. Furthermore, transcripts from those genes associated with similar disorders, show a higher likelihood of having physical interactions between them [46]. If this holds true, then failure of a single component in the network, e.g. due to loss of functionality stemming from mutation, will render the entire module dysfunctional, and result in similar disease phenotypes.

Protein-protein interaction networks (PPIs), or interactomes, are then an interesting data type to be included in a workflow for disease gene finding, and have proved to be an important resource for prediction of disease genes [47, 83, 84]. Despite the advances of the technology driving large scale proteomics experiments, there are reports of a high number of false positives, and the majority of the experiments are performed in model organisms. In order to fully explore this data, we devised an approach to integrate PPIs across several organisms, and to remove false positives based on a confidence score.

#### Estimating the size of the interactome

The size of an organism's interactome has been suggested to correlate better with the biological complexity of the organism, rather than genome size [85]. Recent studies estimate that the human interactome contains approximately 130,000 interactions, most of them unmapped, and that the fraction of those identified up to now represents 8% of the full interactome [86]. This statement comes to contradict earlier estimates by *Stumpf et al.*, sizing the interactome with 650,000 interactions [85]. Similar estimates were reported during the ongoing efforts of the Human Genome Project, although after sequencing the number of genes turned out to be much lower than expected. *Venkatesan et al.* [86] suggest that the ambiguity concerning the size of the human interactome is due to unresolved differentiation between sets of protein pairs that can interact (biophysical interactions) and do interact (biological interactions). Furthermore, there are suggestions that high-throughput yeast two-hybrid (Y2H) interactions for human proteins are more precise than literature-curated interactions supported by a single publication [87].

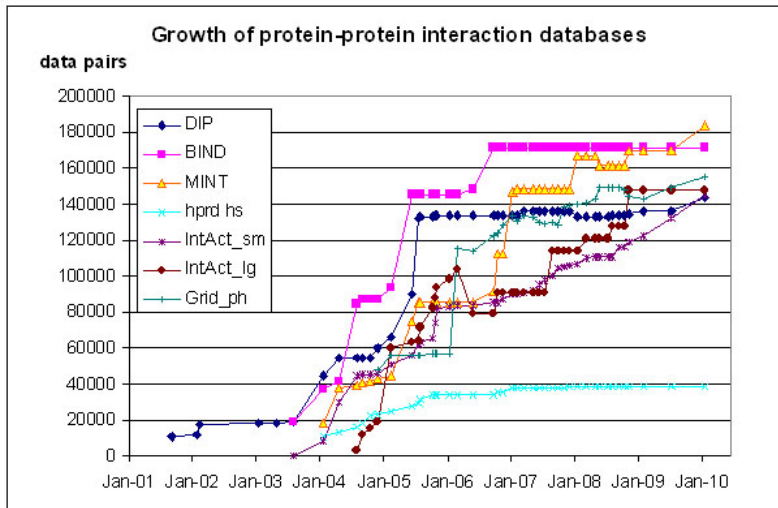
#### Protein interaction data

Physical protein interactions are mainly screened for in large scale experiments using one of two common methods: yeast-2-hybrid (Y2H) approaches, or affinity purification followed by mass spectrometry (AP/MS). Both are reported to be of high quality, but of different and complementary nature [88]. Y2H is based on the activation of downstream reporter genes when a transcription factor binds to an upstream activation site. If one protein is hybridized to the DNA binding domain, and another protein to the activation domain of the same transcription factor, physical interaction between these two proteins is detected by translational activity. AP/MS is a two-step process: during the first step, purification, bait proteins are tagged with a molecule or chemical, and then trapped in an affinity column together with all their interacting partners (in a complex). During the second step, identification, the trapped complexes are analyzed by a mass spectrometer, in order to identify all the components. Due to the nature of

this method, all the constituents of the complex are revealed, but it's hard to say which of the proteins in the complex actually interact physically.

## Databases

As experimental data continues to flow in, a number of databases have emerged, combining text mining and manual curation to extract the interactions from the literature. The main databases are MINT [89], BIND [90], Grid [91], IntAct [92], DIP [93], and hprd [94]. Their growth in the past years is shown in Figure 3.1. There is little overlap between the different databases, therefore in order to get the best coverage of PPI data, we need to integrate all of them.



**Figure 3.1:** Growth of the major protein-protein interaction databases based on the number of ppi pairs contained (statistics and figure by Olga Rigina, olga@cbs.dtu.dk).

## The InWeb

For the purpose of determining human interactions and maintaining high coverage of the human interactome, we devised a strategy to create an in-house human protein interaction network (InWeb). The database was set up so that it can be updated on a regular basis, retrieving data from the available protein-protein databases.

The publicly available data of human protein interactions is low compared to the large number of interaction determined in model organisms, but it has been shown that many interactions are conserved across different species [95]. Therefore it is a feasible strategy to import cross-species interaction data from model organisms using orthology-based mappings to the human interactome.

We downloaded and reformatted data from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [96--98], and MPact, in addition to the databases mentioned in Section 3.1. To map

orthologous proteins between species, we used the InParanoid database [99]. The resulting human inferred PPI network contains more than 510,000 unique interactions, between transcripts of 22523 human genes<sup>1</sup>.

A reliability score is computer for all the interactions in the network, in order to filter out spurious interactions, likely to be false positives. First, all the interactions are assigned a topology score based on the characteristics of the network surrounding the interaction, as reported by *de Lichtenberg et al.* [100]; and a supporting evidence metric, taking into account the size of the experiment that reported the interaction, and the number of separate experiments that confirm it. The raw score for each interaction is then a product of the topological strength and supporting evidence. Using the raw score, a calibration curve is fitted against the overlap with a gold standard set of human interactions<sup>2</sup>, for converting it to a probabilistic confidence score. A cutoff of 0.154 was applied to the network [47], resulting in 124,759 high confidence protein-protein interactions.

By combining multiple protein interaction resources, we were able to infer a human protein interaction network supporting more than 500,000 interactions. Furthermore, by applying a scoring scheme, we constructed a network of almost 125,000 high confidence interactions, coming close to the suggested size of the human interactome [86]. Integration from multiple databases, and orthology-based transfer from model organisms, provides a more complete picture of the human interactome, which can be used for subsequent analyses. In the following articles, the InWeb assumes a central role in establishing previously unknown correlations between proteins involved in different diseases.

---

<sup>1</sup>As of version 4.1, released July 2010.

<sup>2</sup>The gold standard is composed of trusted human protein-protein interactions from several sources: high confidence small scale data (less than 5 human interactions per study) from MINT, BIND and IntAct, KEGG enzymes involved in neighboring steps (ECrel) and KEGG annotated protein-protein interactions (PPrel), and interactions from protein complexes, indirect complex reactions, and neighboring reactions from Reactome [101]. The gold standard is composed of close to 45,000 non-redundant high-confidence interactions.

## 3.2 Paper III

# A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes

*Proc Natl Acad Sci USA*, 2008 vol. 105 (52) pp. 20870-5.

Kasper Lage<sup>1,2,3\*</sup>, Niclas Tue Hansen<sup>1\*</sup>, E. Olof Karlberg<sup>1,4</sup>, Aron C. Eklund<sup>1</sup>, Francisco S. Roque<sup>1</sup>, Olga Rigina<sup>1</sup>, Patricia K. Donahoe<sup>2,3</sup>, Zoltan Szallasi<sup>1,3,5</sup>, Thomas Skot Jensen<sup>1</sup>, Soren Brunak<sup>1</sup>

<sup>1</sup>Center for Biological Sequence Analysis, Technical University of Denmark, building 208, DK-2800 Lyngby, Denmark

<sup>2</sup>Pediatric Surgical Research Laboratories, MassGeneral Hospital for Children, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA

<sup>3</sup>Harvard Medical School, 25 Shattuck Street, Boston, MA 02115, USA

<sup>4</sup>Stockholm Bioinformatics Center, Albanova, Stockholm University, Roslagstullsbacken 35, SE-114 21 Stockholm, Sweden

<sup>5</sup>Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, 300 Longwood Avenue, Boston, MA 02115, USA

\*These authors contributed equally.

### Abstract

Heritable diseases are caused by germline mutations, which despite tissue-wide presence often lead to tissue-specific pathology. Here, we make a systematic analysis of the link between tissue-specific gene expression and pathological manifestations in many human diseases and cancers. Diseases were systematically mapped to tissues they affect from disease relevant literature in Medline to create a disease-tissue co-variation matrix of high-confidence associations of more than 1,000 diseases to 73 tissues. By retrieving more than 2,000 known disease genes, and generating 1,500 disease-associated protein complexes, we analyzed the differential expression of a gene or complex involved in a particular disease in the tissues affected by the disease, compared to non-affected tissues. When this analysis was scaled to all diseases in our data set there is a significant tendency for disease genes and complexes to be overexpressed in the normal tissues where defects cause pathology. In contrast, cancer genes and complexes were not overexpressed in the tissues from which the tumors emanate. We specifically identified a complex involved in XY sex reversal that is testis-specific and downregulated in ovaries. We also identified complexes in Parkinson disease, cardiomyopathies, and muscular dystrophy syndromes that are similarly tissue specific. Our method represents a conceptual scaffold for organism spanning analyses and reveals an extensive list of tissue-specific draft molecular pathways, both known and unexpected or novel, that might be disrupted in disease.

### Introduction

Pathology caused by defects in human genes is usually highly tissue-specific [46, 102--104]. In heritable diseases, this suggests that specific spatiotemporal functions of the implicated genes are disrupted due to germline mutations. Research on tissue specificity of human diseases has focused on the analysis of single disease genes in affected tissues [105, 106], and although it has been shown that disease genes generally tend to be expressed in a limited number of tissues [46], it is still unclear in many cases how the tissue-specific expression patterns of disease genes correlate with their pathological manifestations.

Proteomics approaches have established that most gene products exert their function as members of one or more protein complexes [107--111], and that mutations in different proteins participating in the same complex, such as cellular machines, rigid structures, dynamic signaling or metabolic networks, and post-translational modification systems, generally lead to similar phenotypes [47, 108, 112]. A next logical step is to model entire disease complexes and to analyze the link between tissue-specificity of the complexes and the pathological manifestations with which they are associated when defective. However, such efforts are hampered by the lack of adequate coverage on experimental proteomic data in humans and of strategies for systematically analyzing hundreds of diseases, and their related genes and protein complexes, across multiple tissues of the human organism.

Here, we describe a strategy (Figure 3.2) for systematically correlating pathological manifestations of diseases with expression patterns of implicated genes and protein complexes across many human tissues. For this analysis we created and validated a number of data sets including more than 1,500 disease-associated protein complexes and to these added tissue and sub-cellular localization. Then a method for systematically associating diseases to affected tissues was developed. Across all diseases in the Online Mendelian Inheritance in Man (OMIM) [63] database

to which causative genes could be mapped, we analyze the correlation between tissue-specific expression and pathological manifestation both at the cellular level of single disease genes and for entire disease-associated protein complexes. Finally, we systematically compared the tissue-specific pattern of expression and pathology in cancer initiating genes and complexes, causing familial cancers, to that of non-cancer disease genes and complexes.

## Results

### Systematic generation of an atlas of disease-associated protein complexes with tissue resolution

By mining the GeneCards [113] resource for genes associated with diseases, we generated a list of 2,227 unique disease-related proteins. Similar to the method that we reported earlier [47], an *in silico* approach for generating disease-associated protein complexes based on an inferred human protein-protein interaction network was used (see SI and Figure S1<sup>3</sup>). Following this strategy, we generated 1,524 raw complexes comprising 45,662 unique interactions between 5,202 unique proteins. The quality of the complexes was validated by measures identical to the ones reported in major experimental screens in *S. cerevisiae*, *E. coli* and *H. sapiens* [48, 107--111, 114], showing that the quality of our data matches the reproducibility, average probabilistic interaction scores, accuracy, and coverage reported in these studies (see SI and Figure S2). Finally, the complexes were mapped to tissues using the expression data from 73 non-diseased tissues from the Novartis Research Foundation Gene Expression Database (GNF) [115]. The expression level of a complex in a tissue was calculated by averaging over the expression levels of all genes represented in the complex.

### Mapping complexes to diseases

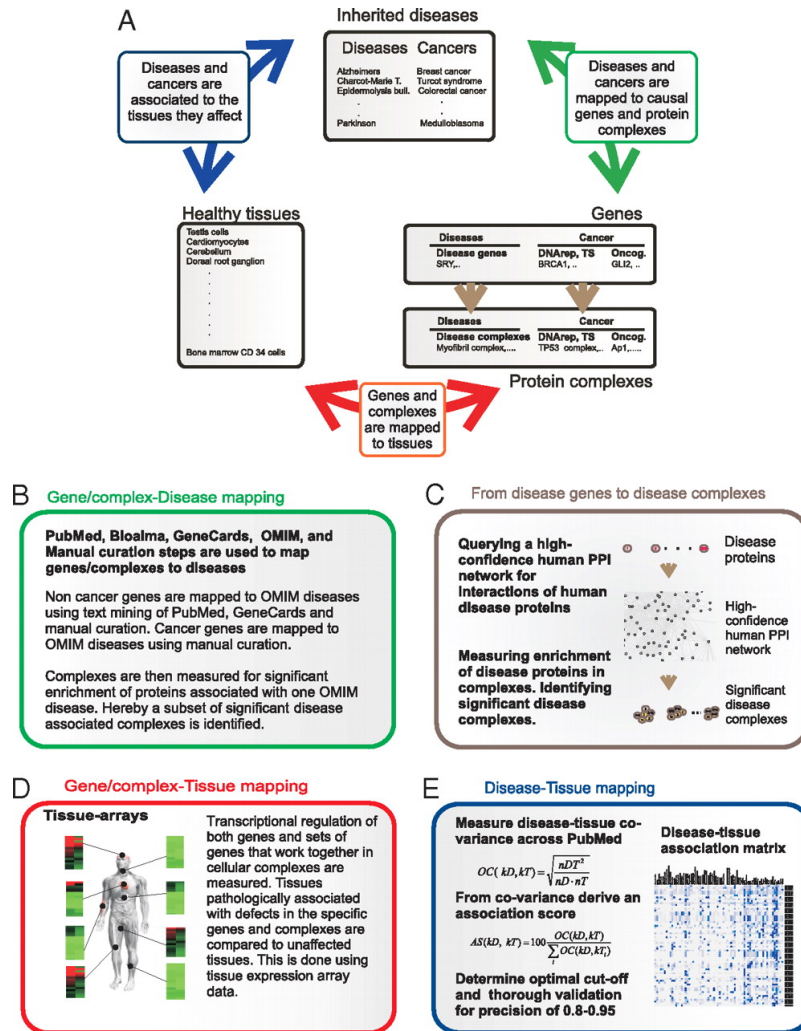
To map complexes to diseases we systematically identified the proteins that had been associated to each of the diseases mentioned in OMIM. This was done using the protein to OMIM mapping displayed in GeneCards<sup>4</sup> database. We then measured the overlap between proteins in complexes and proteins associated with the diseases and calculated the significance of this overlap. Because a number of complexes are known to be involved in different diseases we allowed for a complex to be associated with more than one disease. In total the 1,524 raw complexes were mapped to 1,054 OMIM diseases. In the further text we refer to these as disease complexes.

### Disease-tissue association matrix

To our knowledge there exists no systematic mapping of diseases to affected tissues. We determined the covariance of a disease with a tissue by identifying the number of publications co-mentioning the disease and tissue (and synonyms thereof), relative to the number of publications mentioning the disease or tissue alone [116]. We transformed the covariance into an association score between a tissue and a disease by calculating the fraction of covariance

<sup>3</sup>Online at <http://www.pnas.org/content/105/52/20870/suppl/DCSupplemental>.

<sup>4</sup><http://www.bimas.cit.nih.gov/cards/>



**Figure 3.2:** Overview of the study. (a) The different analyses and how they relate to each other. (b) 59 inherited cancers and more than 1,000 other Mendelian disorders are mapped to 1,265 causative genes and 1,524 complexes using a combination of automated parsing of OMIM and Pubmed. Genes and complexes are stratified into three major categories, non-cancer disease, cancer-gain of function, and cancer-loss of function. This stratification is done by a combination of manual curation and semi-automated steps. (c) A unique set of 1,524 protein complexes associated to disease are generated by querying the proteins of disease genes for direct interaction partners in a human protein interaction network followed by several quality control steps. (d) Transcriptional regulation of both genes and sets of genes that work together in cellular complexes are analyzed across tissues of the human organism. (e) Diseases are mapped to relevant tissues using association degree of particular diseases and tissues across MEDLINE. Steps are taken to reduce errors in word recognition and handle synonyms accurately. These steps are followed by determination of an optimal cut-off and rigorous quality control. Hereby we produced a matrix where diseases are mapped to tissues relevant to the pathology with a precision of more than 0.8. Cancers are mapped to tissues that are the primary origin of tumor formation with a precision over 0.95.



that a given tissue-disease pair constituted, of the total covariance for a given disease. Calculating an association score for the 73 tissues used in the GNF tissue atlas [115], versus 1,054 OMIM diseases yielded a disease-tissue association matrix (Figure 3.3). By manually validating the associations we determined a cut-off where tissues associated with the pathology of a given disease could be determined with a precision over 80% (see SI and Figure S3), meaning that above this threshold tissues relevant to the pathology of a given disease can be accurately identified amongst the GNF atlas tissues in more than 80% of the cases. Tissues associated with the pathology of a given diseases are in the further text defined as disease-tissue associations scoring above this cut-off.

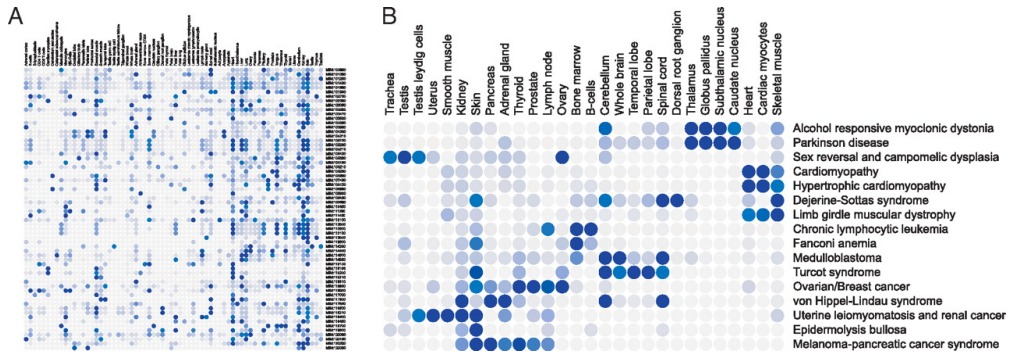
### **Mapping complexes to cancers**

A large number of genes have been associated with cancers, due to aberrant expression or somatic mutations in tumors. However, few of these genes have actually been proven to play a role in the initiation of the tumor. Hence, an automated mapping of cancer genes to complexes would include many genes that are mutated in tumors, but do not cause the cancer. As we are interested in studying the tissue distribution of disease initiating genes and complexes, we manually created an exhaustive list of heritable cancer genes that initiate tumors through germline mutations. These genes were mapped to OMIM diseases describing the cancers manually (Table S2<sup>5</sup>). For this subset of genes, there is compelling evidence that defects are the primary cause of the cancer. In total we extracted a subset of 51 genes in which mutations lead to heritable cancers and mapped them to 59 cancers. Since most cancer mutations are either loss or gain of function which could influence the mechanisms of disease progression and have bearing on the mechanisms of tissue-specificity, we further stratified the cancer genes into loss or gain of function as defined in Vogelstein et al. [103]. Examples of loss of function genes are tumor suppressor or DNA repair genes that become defective when mutated, and examples of gain of function are kinases which become constitutively activated by mutations (Table S3). Cancer associated complexes were identified as complexes enriched for this subset of genes. In the further text we refer to these as cancer complexes.

### **Generating a disease-tissue association matrix for cancers**

Cancer to tissue association mapping is not straightforward. In this study we were interested in exclusively studying the tissues in which tumors are initiated through germline mutations of particular genes. Since cancers generally affect many tissues through downstream effects such as metastases, associations to non-initiating tissues had to be filtered out. Furthermore, many cancer syndromes, arising from germline mutations in cancer genes, also include non-malignant pathology, for which disease-tissue association had to be disregarded in this analysis. For this reason, we manually analyzed the complete subset of tissues associated to heritable cancer syndromes resulting in a precision approximating 100% for the cancer-tissue associations (SI and Table S4).

<sup>5</sup>Online at <http://www.pnas.org/content/105/52/20870/suppl/DCSupplemental>.



**Figure 3.3:** Disease-tissue association matrix. The color range goes from light grey, which corresponds to no association of disease and tissue, to dark blue at 12% association. The percent association is the proportion of a disease's association to a particular tissue in the Novartis Research Foundation Gene Expression Database (GNF) atlas, out of the cumulative association to all tissue in the atlas. (a) The first 100 diseases mapped to the 73 tissues in the GNF atlas. (b) A subset of the disease-tissue associations.

### Correlation between pathology and tissue-specific expression

First, we analyzed the expression of disease genes in the tissue with the highest disease-association in the disease-tissue matrix (rank one). This analysis was repeated for the 2nd to 25th highest associated tissues (rank two to 25) and the average z score at each rank level was plotted as a curve (Figure 3.4a). For example myosin heavy chain 6 (MYH6) is involved in hypertrophic cardiomyopathy and the tissues from the GNF atlas ranked first and second in relation to hypertrophic cardiomyopathy are heart and cardiac myocytes. We determined the z-score of MYH6 in heart (tissue rank one), the average z-score of MYH6 in the two highest ranked tissues, heart and cardiac myocytes (tissue rank two). This procedure is repeated for ranks three to 25. This gives a set of rank dependent z-scores for MYH6. This procedure is repeated for every disease gene in every disease yielding rank dependent z-scores for every gene-disease combination, which is plotted in Figure 3.4a. This figure shows the clear tendency of overexpression for disease genes in tissues with the highest rank (blue curve). The curves for cancer genes show two different trends. While gain of function genes are overexpressed in tissues with the highest rank (red curve), loss of function genes are underexpressed (green curve).

To see if the observed expression trends were significant, we averaged the z scores in the tissues associated with the disease and compared to their expression levels in non-affected tissues (Figure 3.4b). For non-cancer disease genes we observed a significant tendency of overexpression ( $p < 1.0e-6$ ), which is also the case for gain of function cancer genes ( $p = 3.9e-2$ ), but with less significance. Loss of function cancer genes show the converse trend of underexpression, ( $p = 1.0e-2$ ).

We carried out the same analysis for the protein complexes which showed that the expression trend observed for disease genes is conserved at the level of disease protein complexes (see Figure 3.4d and 3.4c). These disease complexes display a significant tendency to be overexpressed in tissues where they are involved in pathology ( $p < 1.0e-6$ , blue curve). While protein complexes significantly enriched for gain of function cancer genes follow the tendency of over-

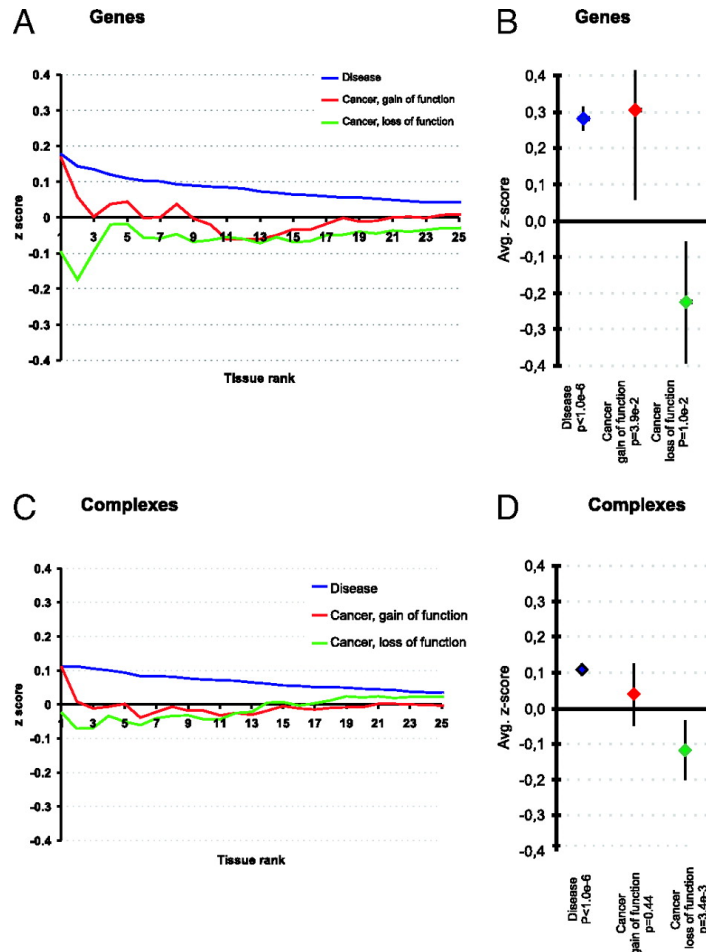
expression ( $p=0.44$ , red curve), and complexes enriched for loss of function cancer genes are underexpressed ( $p=3.4\text{e-}3$ , green curve).

As the z scores were lower for the cancer genes and complexes compared to the more robust values of the non-cancer disease genes and complexes, we tested if this result was influenced by the data set and normalization method. We replicated the analysis using a different and novel robust multi-array (RMA) based normalization scheme [117]. Expression data normalized with this algorithm still showed a significant overexpression of disease genes and complexes, but both the over and underexpression trends for the cancer genes and complexes decreased in significance. To test if a few diseases or tissues were driving the observed trend, we analyzed the expression trend broken down into single tissues (Figure S4) and by bootstrapping the dataset both on disease and tissue level. This analysis shows that that most tissues contribute to the observed results and they are robust to bootstrapping of the data set.

### Examples of disease complexes with tissue and phenotype correlation

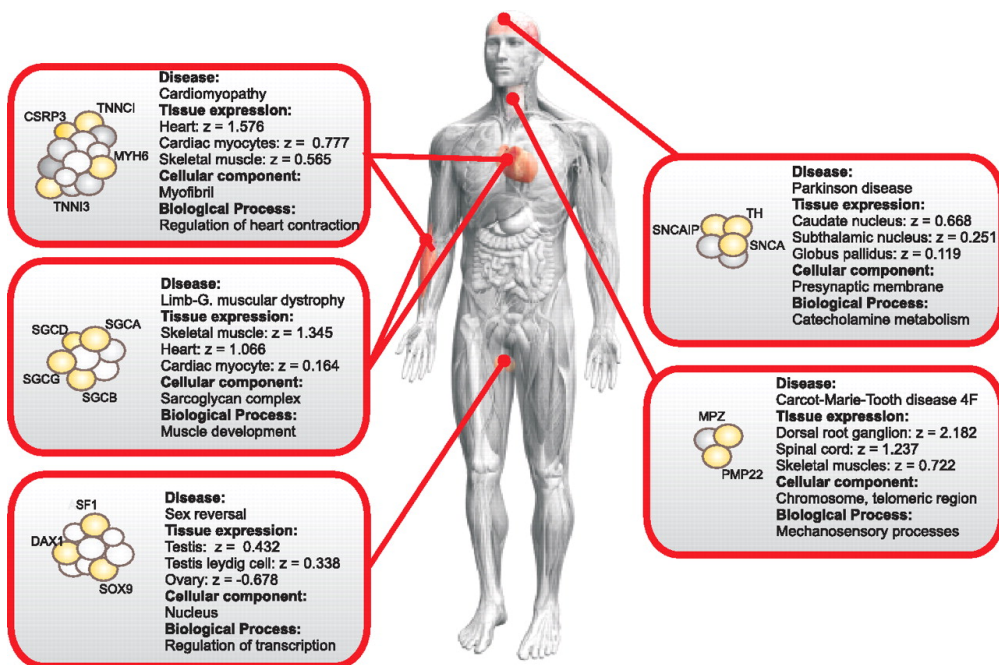
Examples of the correlations found between tissue expression and pathology or phenotype reported are provided in Figure 3.5. Also, the most significant gene ontology (GO) subcellular and functional categories for the complex in question are indicated followed by the significance with which the complex can be assigned to this GO category. Tissue names are as defined in the GNF atlas. The full sets of proteins in each complex can be seen in Figure S5 (Supplementary Material, online).

XY sex reversal can be caused by mutations in the transcription factors SRY (Sex determining Region Y) [118]. SOX 9 (the SRY sex determining region Y-box 9 gene) [119], NR5A1 (the nuclear receptor subfamily 5A1), more commonly known as SF1 [120, 121]; and NR0B1 (nuclear receptor subfamily 0B1), more commonly known as DAX1 [122]. Additionally SOX 9 is associated with campomelic dysplasia a bone disorder that leads to a number of associated skeletal and cartilaginous deformities [123]. SF1 is needed for gonad and adrenal differentiation [120, 124] and for proper steroidogenesis as well as for Mullerian Inhibiting Substance (MIS) ligand and MIS receptor expression [124, 125]. DAX1, which leads to XY sex reversal both when overexpressed, by inhibiting SF1 [122], and when inactivated, as it is required for testis differentiation by regulating expression of SOX9 [126]. While the activity of SF1, DAX1 and SOX9 is required for testis differentiation and development, none of these genes are essential for ovarian development [126--129]. Here we identify a transcriptional regulation complex (GO:0006355;  $p=1.9\text{e-}8$ ) containing DAX1, SF1, and SOX9 all of which are known to be associated with sex reversal ( $p=6.9\text{e-}6$ ). Furthermore, the complex contains SOX8 that is closely related to SOX9 and implicated in regulating the expression of testis-specific genes [130]. While, the complex is overexpressed in testis cells, it is underexpressed in ovaries (Figure 3.5 on page 49), which coincides with the known biology of the most well characterized of its components. Our method thus has predictive value as it can i. detect interactions between molecules which, by themselves, are known to be important in sex differentiation and determination by producing sex reversal, ii. validate these findings by demonstrating dimorphic tissue-specific expression that correlates with the pathology resulting from inactivation of several members of the complex, and iii. reveal the importance of new interactors worthy of further study.



**Figure 3.4:** Expression levels of disease genes and complexes in pathologically associated tissues. (a) The expression level of genes associated with diseases and cancers in the tissues most associated with the particular disease caused by the genes. Tissues are ranked with the most associated tissue at the intersubsection with the y-axis and in declining order from left to right. This plot shows the trend of overexpression for disease genes and gain of function cancer genes in tissues with the highest rank. Loss of function cancer genes are generally underexpressed in the tissues with the highest rank. (b) The average disease gene expression in associated tissues is shown. Disease genes are overexpressed with an average z-score of 0.28 ( $p < 1.0e-6$ ). The cancer associated genes show two different trends; gain of function follow the trend of all disease genes, with an average z-score of 0.30 ( $p = 3.9e-2$ ), but loss of function genes have a tendency to be underexpressed in the tissues associated with tumor formation, with an average z-score of -0.21 ( $p = 1.0e-2$ ). (c,d) The same analysis is shown at the level of protein complexes, where the trend is conserved.

Four other complexes, where tissue-specific overexpression correlates with pathological manifestations, are depicted in Figure 3.5, (see SI and Figure S5 for more details on these four complexes and for examples of cancer related complexes). These include i. a complex involved in Charcot-Marie-Tooth disease type 4F and overexpressed in spinal cord, dorsal root ganglion, and skeletal muscles; ii. a sarcoglycan complex involved in Limb-Girdle muscular dystrophy overexpressed in skeletal muscle, cardiac myocytes and heart; iii. a myofibril complex involved in familial cardiomyopathy overexpressed in several tissues associated with the disease such as heart and cardiac myocytes; iv. and a complex involved in catechol metabolism and Parkinson disease, overexpressed in a number of relevant brain tissues including the caudate nucleus, subthalamic nucleus, and globus pallidus. While the overexpression of the sarcoglycan and myofibril complex in skeletal tissues is well known, the ovarian-testes dimorphic expression pattern of the sex-reversal complex, and the overexpression of a Parkinson complex in several relevant brain tissues of the basal ganglia are suggestive of the underlying tissue-specific biology of these disorders. Across all examples the tissue-specific expression patterns correlate with the pathological changes observed when one or several members of the complex are defective.



**Figure 3.5:** Representative examples of disease complexes are displayed. Diseases are associated with tissues using our disease-tissue matrix, and expression data are from the GNF data set. The expression levels of complexes are shown as z-scores. If a disease is associated with more than three tissues, only the three most associated tissues are shown for clarity. In a given complex, proteins relevant to the disease in question are yellow. The figure shows the general tendency of overexpression of the complexes in the tissues in which they are involved in pathology compared to their expression level in other tissues.

## Discussion

The complex data set reported here is more than three times larger than our reported set of complexes [47] and contains approximately seven times more interactions compared to the only previously reported experimental screen for human complexes [131]. To our knowledge, this data set comprises the first set of systematically generated complexes with tissue, phenotype, and sub-cellular resolution in any mammalian organism. The entire atlas is made available online at <http://www.cbs.dtu.dk/suppl/dgf/>.

A theoretical limitation of our approach is that we use gene expression data to map complexes to tissues due to the lack of good coverage of quantitative proteomics expression data. Early studies of the relationships between mRNA expression and protein abundance levels have consistently reported modest correlations [132--134]. However, recent work, which uses a probabilistic framework to model the relationship between the experimentally recorded protein and mRNA patterns, has confirmed that in 75% of all genes tissue mRNA expression patterns linearly correlate with protein abundance, and this overall good correlation is shown for the dataset we use in this work [135]. However, to test how a lack of correlation for 25% of the genes affects our results, we randomized 25% of the data points and found that the results achieved for disease genes and complexes, and for loss of function cancer genes and complexes were robust ( $p < 1.0e-3$ , see SI). Furthermore, the tissue resolution of our complexes is supported by the observation that they are significantly enriched in proteins co-occurring in tissue samples that are analyzed using manually curated immunohistochemistry data (SI and Figure S2).

Our results support the notion that known disease genes generally are tissue specific [46, 104], by being selectively overexpressed in the tissues in which specific gene defects cause pathology. Alternatively high levels of gene expression may be needed for the functional activity of the tissue. Moreover, we show that this trend is conserved also at the level of the protein complexes in which the disease genes carry out their biological function.

Most known genes which initiate cancer are involved in ubiquitous processes such as DNA repair, cell cycle regulation and apoptosis [103, 136--138] and Table S3. And it remains a key puzzle in oncology to determine how germline mutations in general genes initiate tissue-specific tumors [137]. To investigate this contradiction, we also analyzed the expression patterns of cancer genes and complexes involved in heritable cancer syndromes. The gain of function cancer genes and complexes follow the trend of non cancer disease genes and are generally overexpressed in tissues where they initiate tumors, conversely complexes enriched for loss of function genes are underexpressed in the tissues where mutations cause neoplastic transformation. Our results for cancer genes and complexes were not robust when different algorithms were used to normalize the expression data. There could be a number of reasons for the lack of a tissue-specificity signal for the analyzed cancer genes and complexes. The current concepts of cancer indicate that some tumors are initiated by a small subset of stem cells [139] whose specific expression levels would be impossible to detect in tissue samples with the resolution used here. Another hypothesis is that tumor initiation is caused by a combination of mutations in a key gene, exposure to mutagenic substances or ionizing radiation, and high proliferation rates of specific cell populations in a tissue [137], a combination we do not analyze here. However, our results highlight the fundamental difference between the tissue specificity of cancers and other diseases, and shows that this difference is consistent on both gene and complex level.

Functional genomics and sequencing have been extremely useful tools for identifying the complete sets of genes in humans and model organisms, and deducing how disruption of different genes in a common molecular pathway can lead to similar phenotypic pathologies. These results indicate how the function of genes is organized in space and time. The next step is to model entire systems using data integration and systems biology. This has proven difficult in humans due to experimental limitations and ethical issues, suggesting that other strategies must be considered. We take a step towards this goal by creating a number of new data sets, in part by refining, re-analyzing, and integrating existing data to identify a comprehensive list of functional modules that are associated with pathological processes in humans. We analyze their spatial tissue-specific and sub-cellular patterns and correlate this information with the diseases that are the result of defects in the modules. As such, our data set and the scaffold of the analysis presented could be useful in disease systems biology of humans, and provides draft mechanistic pathways that can serve as potential molecular drug targets.

## **Materials and Methods**

### **Mapping genes and complexes to tissues**

We used the GNF tissue atlas [115] that includes reproduced RNA expression experiments from 79 human tissues. Six tissues were removed as they were derived from cancer tissues. We chose the GNF data set as it displays high reproducibility [140], and the transcript levels show generally a linear relationship with protein abundance [135]. We log-transformed hybridization levels and normalized within each tissue (to ensure equal weight), followed by a normalization across all tissues, thereby ensuring that expression levels represented the relative presence of a transcript in one tissue compared to the other 72 healthy tissues in the data set. For complexes, the normalized expression levels of all genes in a complex were averaged for each tissue. To test the effect of different normalization methods on our results, we prepared the same data set with Eklund and Szallasi's novel normalization method [117] and compared the results.

### **A curated set of genes in which mutations lead to tumor formation**

We curated a set of genes in which mutations had been shown to lead to heritable tumor formation and mapped them to OMIM diseases (see Table S3). By following the definitions introduced by Vogelstein et al. [103] we also noted whether the genes were oncogenes or non-oncogenes (such as tumor suppressors or DNA repair proteins) (see Table S4).

### **Mapping of complexes to OMIM diseases**

We calculated the enrichment of proteins involved in the same OMIM disease using the annotations in GeneCards, which has previously been shown to be an accurate way of mapping genes to diseases [47]. We calculated the significance of an enrichment using a hyper-geometric test.

### **Under- and overexpression significance**

We averaged the expression z score over all disease genes in the most disease-associated tissue as determined from the disease-tissue matrix. For each rank from 1 through 25, we calculated the average z score yielding a curve. In Figure 3.4 on page 48, this curve is plotted as the average z scores of all gene-disease pairs in tissues with a particular rank. This procedure was repeated for gain of function and loss of function cancer genes. Again this approach was repeated on a protein complex level. All reported significances are two-tailed using the Students t-test.

### **Disease-tissue association matrix.**

To identify the tissues most affected by diseases described in the OMIM database [63], we used co-mentioning of a given disease with a given tissue across MEDLINE [116]. The tissue names from the Novartis Research Foundation Gene Expression Database (GNF) [115] were manually curated and translated to corresponding medical subject heading (MeSH) terms (to reduce errors in word recognition and handle synonyms properly). Similarly, the disease names were determined using disease titles provided in OMIM. Also, these titles were manually curated and translated to the relevant MeSH terms. We used Ochiai's coefficient (OC) as a measure of similarity derived from the co-occurrences [141--143], and calculated an association score (see below), as the percentage of the total normalized co-occurrence of a given disease that could be attributed to a given tissue. Validation was carried out as described in the SI.

### **Acknowledgements**

The authors wish to thank Matthias Mann, Jiri Bartek, Gert-Jan B. van Ommen, Barbara Pober and Jonathan Rosand for valuable input on the manuscript and project. We further wish to thank the Villum Kann Rasmussen Foundation, the Simon Spies Foundation, the National Institute of Child Health and Development (NICHD) #CD RO1 HD0551-50 and the National Institute of Health (NIH) for financial support and Anders Lendager and Lene Hep from MAPT for help with the figures. Zenia Marian Størling assisted in the initial analyses. We thank Kasper Fugger and Christopher Workman for helpful discussions.



### 3.3 Paper IV

## Dissecting spatio-temporal protein networks driving human heart development and related disorders

*Mol Syst Biol*, 2010 vol. 6 pp. 381.

Kasper Lage<sup>1,2,3,4\*</sup>, Kjeld Møllgård<sup>5</sup>, Steven Greenway<sup>6</sup>, Hiroko Wakimoto<sup>6</sup>, Joshua M. Gorham<sup>6</sup>, Christopher T. Workman<sup>4</sup>, Eske Bendsen<sup>7</sup>, Niclas T. Hansen<sup>4</sup>, Olga Rigina<sup>4</sup>, Francisco S. Roque<sup>4</sup>, Cornelia Wiese<sup>8</sup>, Vincent M. Christoffels<sup>8</sup>, Amy E Roberts<sup>9,10</sup>, Leslie B. Smoot<sup>10</sup>, William T. Pu<sup>5,10,11</sup>, Patricia K. Donahoe<sup>1,2</sup>, Niels Tommerup<sup>12</sup>, Søren Brunak<sup>4,13</sup>, Christine Seidman<sup>6</sup>, Jonathan Seidman<sup>6</sup>, Lars A. Larsen<sup>12\*</sup>

<sup>1</sup> Pediatric Surgical Research Laboratories, MassGeneral Hospital for Children, Massachusetts General Hospital, Boston, Massachusetts, USA.

<sup>2</sup> Harvard Medical School, Boston, Massachusetts, USA.

<sup>3</sup> Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

<sup>4</sup> Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark.

<sup>5</sup> Developmental Biology Unit, Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark.

<sup>6</sup> Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA

<sup>7</sup> Fertility Clinic, Department of Obstetrics and Gynecology, University Hospital of Odense, Odense, Denmark

<sup>8</sup> Center for Heart Failure Research, Academic Medical Centre, Amsterdam, The Netherlands.

<sup>9</sup> Partners HealthCare Center for Genetics and Genomics, Boston, Massachusetts, USA

<sup>10</sup> Department of Cardiology, Children's Hospital, Boston, Massachusetts, USA

<sup>11</sup> Harvard Stem Cell Institute, Harvard University, Cambridge, Massachusetts, USA.

<sup>12</sup> Wilhelm Johanssen Centre for Functional Genome Research, Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark.

<sup>13</sup> Center for Protein Research, University of Copenhagen, Copenhagen, Denmark.

\* Correspondance to: Kasper Lage (lage.kasper@mgh.harvard.edu) and Lars A. Larsen (larsal@sund.ku.dk)

### Abstract

Aberrant organ development is associated with a wide spectrum of disorders, from schizophrenia to congenital heart disease, but systems-level insight into the underlying processes is very limited. Using heart morphogenesis as general model for dissecting the functional architecture of organ development, we combined detailed phenotype information from deleterious mutations in 255 genes with high-confidence experimental interactome data, and coupled the results to thorough experimental validation. Hereby, we made the first systematic analysis of spatio-temporal protein networks driving many stages of a developing organ identifying several novel signalling modules. Our results show that organ development relies on surprisingly few, extensively re-cycled, protein modules that integrate into complex higher-order networks. This design allows the formation of a complicated organ using simple building blocks, and suggests how mutations in the same genes can lead to diverse phenotypes. We observe a striking temporal correlation between organ complexity and the number of discrete functional modules coordinating morphogenesis. Our analysis elucidates the organization and composition of spatio-temporal protein networks that drive the formation of organs, which in the future may lay the foundation of novel approaches in treatments, diagnostics and regenerative medicine.

### Introduction

Insight into the biology of molecular networks driving organ development is an important and emerging field since aberrations in these systems underlie a wide spectrum of highly polygenic human disorders, ranging from schizophrenia [144], to congenital heart disease (CHD) [145]. Understanding the functional architecture of networks that orchestrate the development of organs may also lay the foundation of novel approaches in regenerative medicine, because manipulation of these systems will be necessary for the success of tissue-engineering technologies and stem cell therapy [146].

We used heart development and CHD as a general model for dissecting the functional protein networks underlying a developing organ and its related, genetically complex, human disorder. The heart is particularly suitable for such an analysis, because it is among the most studied of all organs, it is the organ most susceptible to disease and its developmental processes and genes are extraordinarily conserved enabling straight forward integration of data between humans and model organisms [147, 148]. Genetic studies in humans and model organisms have identified hundreds of genes involved in heart development. In mice, phenotypes caused by targeted mutations can be organized into hierarchical morphological subgroups, which point at the spatio-temporal role of the disrupted genes. These results have led to a hypothesis suggesting that during organ development autonomous anatomical sub-structures are coordinated by discrete protein complexes or pathways (i.e., functional modules) integrating into higher-order functional networks, and that evolutionary newer anatomical structures might re-cycle parts of the networks used in more ancient structures [149]. Although transcription factors have been identified as central players in these processes [148--158], we currently lack overviews of how most genes integrate into functional modules and networks during the different developmental stages. Our lack of understanding of this biological architecture is exemplified by the knowledge that genetic factors contribute significantly to CHD [145], but less than 5% of CHD patients

have mutations within the few identified causal human genes, suggesting that many genetic principles of the molecular networks driving heart development remain to be understood.

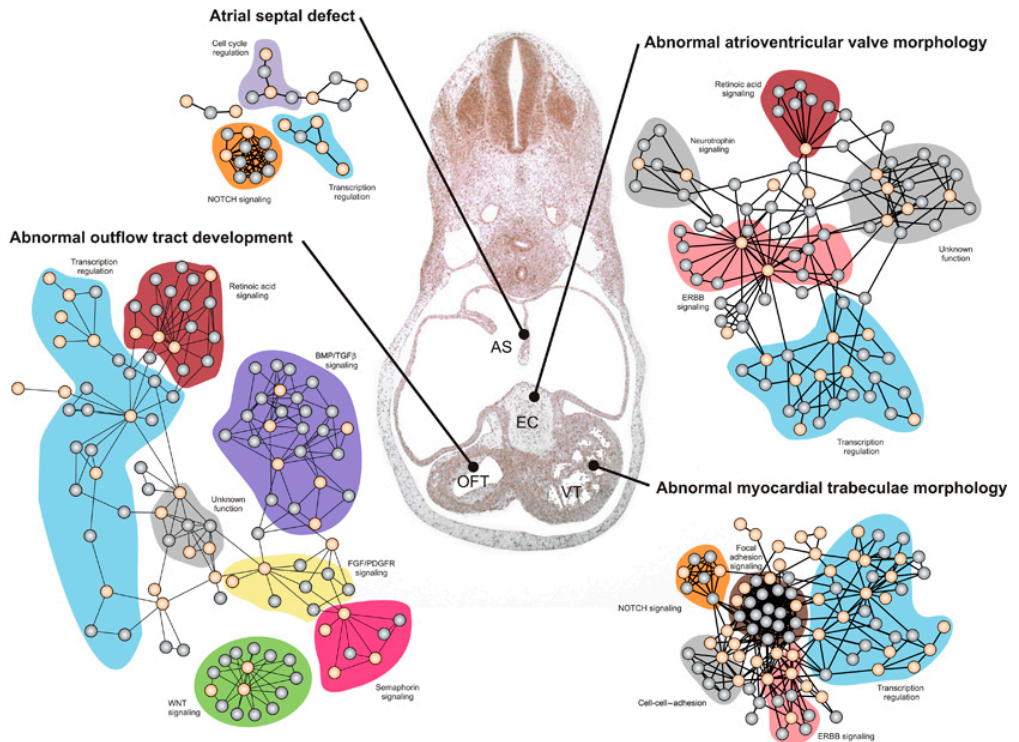
## Results and Discussion

First, we manually curated a set of 255 cardiac developmental genes, in which targeted mutation leads to heart phenotypes in mouse models, from the Mouse Genome Database ver. 3.44 [159]. We used the InParanoid orthology database [160] to find the orthologous 255 human genes, and then identified their corresponding human proteins. We used InParanoid since this method several times has been shown to be superior to other methods for mapping functional orthologs [161, 162]. We refer to this set of human proteins as cardiac developmental (CD) proteins. The 255 proteins are stratified into a total of 19 morphological subgroups reflecting the specific phenotype associated with their mutation (Table S1<sup>6</sup>), which can be used as an indicator for the spatio-temporal role of the individual genes. For each of the 19 sets of proteins we constructed functional networks (Figures S1-S4) using their interaction patterns in refined experimental proteomics data (Materials and Methods), and indeed several novel modules not previously associated with heart development emerged from our analysis (see below). Randomization tests of the resulting networks show that 18 of the 19 gene sets significantly interact at the protein level after adjustment for multiple testing using a Bonferroni correction (Materials and Methods and Figures S1-S4), indicating that the genes involved in each developmental stage have a strong tendency to directly interact at the protein level, or are part of connected pathways. In total the resulting interaction networks consists of 629 unique proteins, have both time and tissue resolution, and describe a wide variety of developmental stages and anatomical structures in the developing heart. These data represent a new framework for the study of organ development at the systems level, and they extend considerably our understanding of the highly polygenic nature of organ developmental processes, which has been shown previously at the level of gene expression [163].

We manually annotated the functional clusters in the networks by literature curation. We chose manual literature curation over automated gene ontology analyses, to exploit the considerable experience and expertise in our group on developmental programs. This analysis revealed several functional modules which are novel in relation to heart development, including focal adhesion signaling modules and a module of unknown function which include Sorting nexin 9 (SNX9, Supplementary Information, Figure S2C and S3A). The quality of the data was confirmed by the existence of many known functional modules in the networks (e.g. NOTCH signaling in development of the ventricular trabeculae [164]). Examples of four networks are shown in Figure 3.6 on the following page; the proteins involved in four phenotypes and their interaction partners fall into distinct modules, represented as highly interconnected subclusters in the networks. E.g., the data show that WNT, semaphorin, FGF/PDGFR, BMP/TGFbeta, and retinoic acid signaling are involved in development of the outflow tract and suggest that extensive communication takes place within and between modules.

To get insight into how the modularity of heart development is organized across spatio-temporal morphological stages, we created module maps of the different networks and grouped them according to temporal development (i.e., early, intermediate and late developmental

<sup>6</sup>Available online on [http://www.nature.com/msb/journal/v6/n1/supinfo/msb201036\\_S1.html](http://www.nature.com/msb/journal/v6/n1/supinfo/msb201036_S1.html)



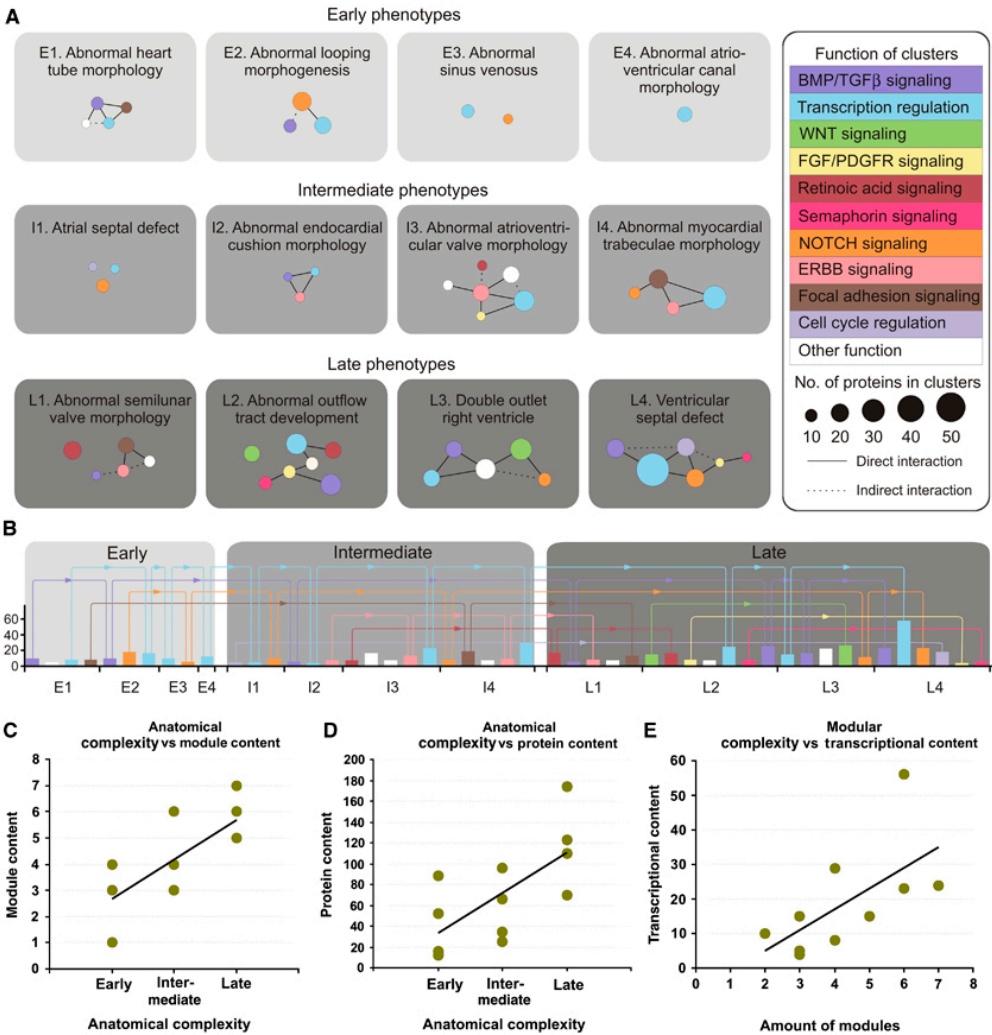
**Figure 3.6:** Examples of four functional networks driving the development of different anatomical structures in the human heart. These four networks constructed by analyzing the interaction patterns of four different sets of cardiac development (CD) proteins corresponding to the morphological groups 'atrial septal defects', 'abnormal atrioventricular valve morphology', 'abnormal myocardial trabeculae morphology', and 'abnormal outflow tract development' (Table S1 online). CD Proteins from the relevant groups are shown in orange and their interaction partners are grey. Functional modules annotated by literature curation are indicated with a coloured background. High-resolution figures (including protein names) can be seen in Figure S2A, S2C, S2D and S3B, respectively. Centrally in the figure is a haematoxylin-eosin stained frontal section of the heart from a 37 days human embryo, where tissues affected by the four networks are marked; AS (developing atrial septum), EC (endocardial cushions, which are anatomical precursors to the atrioventricular valves), VT (developing ventricular trabeculae) and OFT (developing outflow tract). The entire set of 19 networks is shown in detail in Figure S1-S4, and can be downloaded from <http://www.cbs.dtu.dk/suppl/dgf/>.

stages of organogenesis, Figure 3.7 on the next page). Here, the modular design of the functional networks becomes clear across developmental stages and anatomical structures. Surprisingly, although the networks in some instances contain hundreds of proteins, they consist of relatively few protein modules that are extensively recycled across developmental stages (Figure 2 3.7 A and B). Moreover, each network consists of a combinatorial unique module pattern. Although modularity is known to be a core feature of the organization of organisms [165], to our knowledge, this concept has not been shown at the level of protein networks in organ development before. This organizational concept allows for the formation of a very complex organ using relatively simple building blocks and suggests how mutations in the same genes and modules can lead to very diverse phenotypes. For example, NOTCH signalling modules are present in the networks representing atrial septal defects (ASD, Figure S2A) and double outlet right ventricle (DORV, Figure S3C) in line with the observation that mutations in NOTCH1 may lead to atrial septal defects in one individual, and double outlet right ventricle in another [166].

Development of the human heart starts approximately two weeks after fertilization with the formation of the cardiac crescent and the subsequent formation and looping of the primitive heart tube. At this stage the heart is an anatomically simple structure associated with the 'early phenotype' networks in Figure 3.7. Looping is followed by extensive tissue remodelling which includes septation of the atrium and ventricles, and development of trabeculae within the ventricles. Defects at this stage results in 'intermediate phenotypes'. The last stages of heart development include construction of the heart valves and separation of the outflow tract, as determined by 'late phenotypes'. Throughout this transformation the organ, along with the embryo, becomes an anatomically much more elaborate structure [148], which remarkably is mirrored in the complexity of the functional networks we have identified as drivers of these processes.

We have quantified network complexity based on i) the number of distinct functional modules present in each network, and ii) the total amount of proteins in each network. The amount of modules in networks associated with 'early phenotypes' is on average 2.5 which increases to an average of 5.8 for 'late phenotypes' (Figure 3.7C, Spearman  $\rho = 0.76$ ,  $p = 0.010$ ). A similar observation can be made for the total amount of proteins in each network that increases from an average of 42 to 119 (Figure 3.7D, Spearman  $\rho = 0.72$ ,  $p = 0.016$ ). Although the stages of heart development are broadly defined, there is a clear trend across all networks for later phenotypes to be associated with more complex and functionally diverse networks.

Thus, our analysis of the networks strongly suggest that increased morphological complexity of the heart is correlated with increased signalling complexity at the molecular level, which support a model predicting that the four chambered heart has evolved by addition of new autonomous anatomical structures [167, 168]. Analysis of functional data at the systems level suggests that this evolution in part relies on recycling and shuffling of existing functional modules, to create combinatorial unique functional networks that drive the formation of new anatomical structures. Furthermore, our results mirror the findings of evolution modelling and algorithms, which show that modularity in networks can spontaneously arise under changing environments [169, 170], a principle which allows for rapid organism adaptation to new demands [169]. Importantly, we couple these findings to the anatomical development of organs and together these studies give insight into the forces that advance structural simplicity in biological networks underlying organ development.



**Figure 3.7:** (A). An overview of the modular organization of heart development. Protein interaction networks are plotted at the resolution of functional modules. Each module is colour coded according to functional assignment as determined by literature curation. The amount of proteins in each module is proportional to the area of its corresponding node. Edges indicate direct (lines) or indirect (dotted lines) interactions between proteins from the relevant modules. (B) Recycling of functional modules during heart development. The bars represent functional modules and recycling is indicated by arrows. The bars follow the colour code of (A) and the height of the bars represent the number of proteins in each module, as shown on the Y-axis (left). (C-E) Correlations between anatomical, modular and transcriptional complexity in organ developmental networks. We plotted network complexity along an axis of increasing anatomical complexity as defined by the early, late and intermediate phenotypes (C and D), and observe a significant correlation. Also, modular and transcriptional complexity correlate significantly during the traversing of organ developmental programs and stages (E). In a given network, module content is the amount of modules, protein content the amount of proteins, and transcriptional content the amount of proteins directly involved in transcriptional activation.

Formation of organs depends on highly conserved sets of transcription factors [171] of which spatio-temporal regulation is critical to achieve correct patterning [156]. During development, series of transcription factors function hierarchically to regulate specific developmental programs [163, 172, 173]. These observations raise the question of how transcriptional programs are linked to the timing of developmental processes and the relationship between the transcriptional programs, cellular networks, diseases, and the developing organ.

*GATA4*, *NKX2-5*, and *TBX5* are known to be involved in many stages of heart development and defects in these genes have been established as the cause of familial CHD [150, 152, 157]. As expected, we observe these transcription factors participating in most of the networks and across almost all stages of heart development, stressing their importance (Figure S1-S4). In addition to *GATA4*, *NKX2-5*, and *TBX5*, the networks also contain a large amount of other proteins directly involved in transcriptional control either as transcription factors, or by participating in transcription initiation complexes and networks (Figure 3.7 on the facing page and Figures S1-S4).

Two transcriptional concepts have been observed in organ development. Some regulators are only active for a brief period of time and usually produce a uniform response in the expressing cells [155, 163, 173]. Other regulators such as *GATA4*, *NKX2-5*, and *TBX5* are continuously expressed, but activate different sets of genes at different developmental stages, suggesting they are parts of more heterogeneous and complex transcriptional programs [153, 154, 158, 174]. The latter type of regulators exert their specific function by exploiting promoter affinity gradients, and by through complicated patterns of promoter elements that scaffold sets of transcriptional proteins [153, 154]. Our data show that *GATA4*, *NKX2-5*, and *TBX5* participate in most of the transcriptional modules throughout heart development as expected (Figure S1-S4), but interestingly, the modules vary widely in complexity and in the specific composition of the participating proteins. Thus, on the level of transcriptional protein networks, we observe combinatorial regulation, which provides the organism with a high degree of flexibility for *GATA4*, *NKX2-5*, and *TBX5*, and enables them to play a broad role during heart development. This is consistent with the remarkable variability of phenotypic outcome that can be the result of mutations in each of these genes.

Interestingly, the amount of transcriptional proteins in the networks increases from an average of 11 in the networks associated with 'early phenotypes', to 32 in the networks associated with 'late phenotypes'. Moreover, there is a significant correlation between the amount of modules in each network and the amount of proteins directly involved in transcriptional control in the same networks (Figure 3.7E, Spearman  $\rho = 0.69$ ,  $p = 0.035$ ). Thus, our results show a direct relationship between anatomical, modular and transcriptional complexity during the traversing of organ developmental programs and support the concept of combinatorial regulation at the protein level.

To experimentally test the biological accuracy of the module maps, we systematically identified 49 novel heart developmental proteins from the modules (Table S2, Table S3). These candidates were interacting significantly with the CD set, but were not in the literature associated with heart developmental processes (the procedure for scoring and identifying the final 49 candidates is described in detail in Supplementary Information and Figures S5). Twelve of these candidates were selected for immunohistochemistry (IH) analyses to test if they were expressed at the time and place determined by the functional networks in which they participate and the

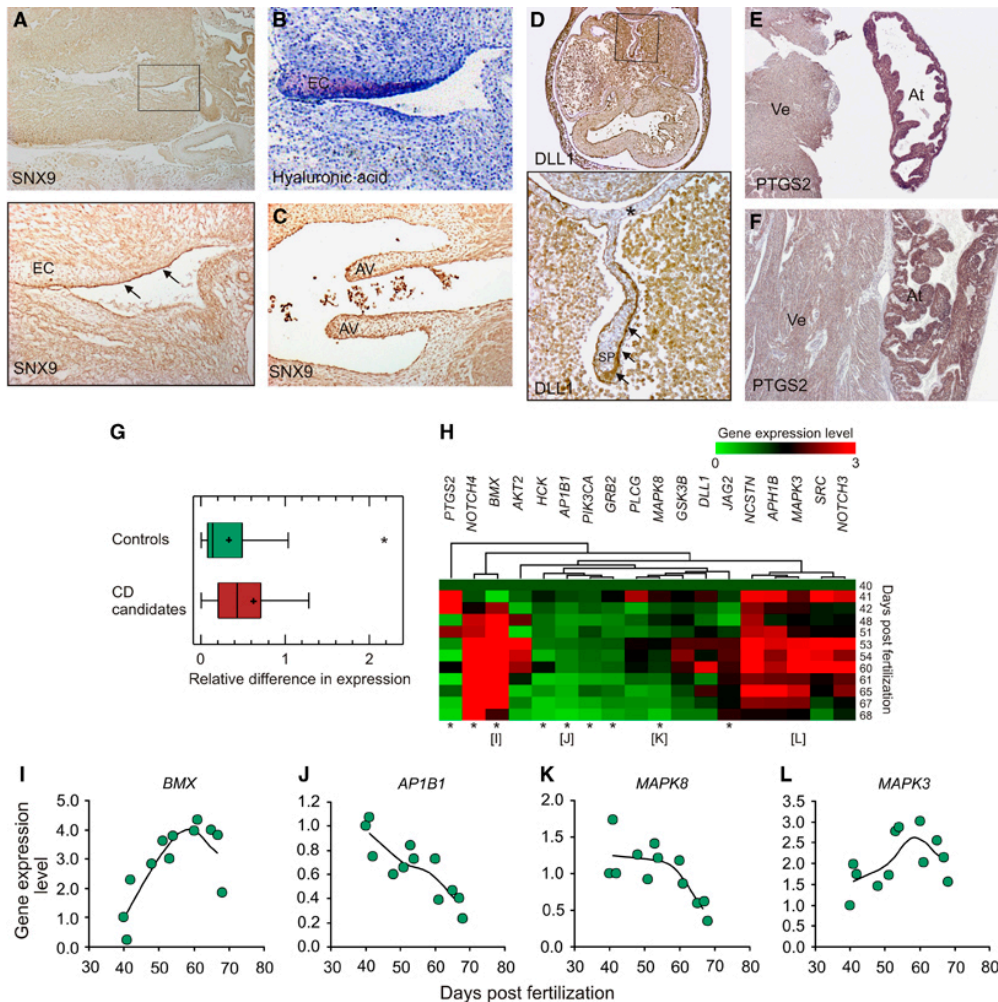
specific morphological groups associated with those networks (the procedure and criteria for choosing the tested proteins is described in detail in Supplementary Information and Figures S5, and S6). Immunohistochemical analyses were systematically performed on a total of 382 tissue sections from 19 developing human hearts in a blinded fashion, and a semi-quantitative measure for the expression level of each protein was determined in at least six anatomical structures or tissues, and across six developmental time points (Table S4 and Figures S7-S13).

For all twelve proteins we see strong evidence of heart developmental function (Table S4). Importantly, eleven of the proteins (SNX9, DLL1 (DELTA), NOTCH3, JAG2, PTGS2 (COX-2), CAV3, SRC, MAPK3 (ERK1), MAPK8 (JNK1), BMX and PTK2B), and are specifically expressed in the anatomical structures associated with the morphological grouping of the functional network in which they participate (Table S5 and S7-S13).

As a standardized estimate of the network-based prediction signal, we calculated the precision of our predictions as the true functional predictions amongst all functional predictions (or the amount of true positives amongst all positives). Using a conservative estimate of a true functional prediction, the precision is 0.72 (Supplementary Information and Table S5), meaning that the morphological groups associated networks in which a candidate emerges, correlate with the candidates being specifically, and highly, expressed in relevant tissues at relevant time points in 72% of the cases. For example, SNX9, which is not associated with heart development in the literature, emerges in several networks associated with valve development (Figure S2C and S3A). We confirmed this role of SNX9 by observing that it is highly expressed specifically in the cell populations driving the development of the endocardial cushions (EC), which are anatomical precursors of the heart valves, and aortic valves (AV, Figure 3.8 on the next page A-C). Analogously, DLL1 and PTGS2 are predicted to be involved in the development of the atrial septum (Figure S2A), and cardiac myocardium (Figure S4A-E), respectively. These predictions are confirmed by their specific expression patterns in the relevant structures of the developing heart at the correct developmental time points (Figure 3.8D-F). For enlargements of IH pictures, many more details, examples of antibody specificity, and a more thorough discussion of the functional roles of these candidates see Supplemental Information and Figure S7-S14.

To further validate the network data we carried out expression profiling of the larger set of 49 candidates across different developmental stages using quantitative real-time RT-PCR on RNA extracted from 14 embryonic human hearts (Figure 3.8G-L). The candidates were significantly differentially expressed during heart development, compared to a set of controls (Figure 3.8G, Mann-Whitney (Wilcoxon) W test,  $p < 0.006$ , Table S6), and significantly higher expressed in heart tissues than random controls ( $p = 0.016$ , Figure S15 and Supplementary Information). To investigate this trend in more detail, we analyzed the relative expression levels of a subset of the candidates in 12 additional hearts at 12 different time points between 40 and 67 days post fertilization (Table S7). This analysis showed that half of the candidates were significantly differentially expressed across these twelve time points further supporting their role in human heart development (Figure 3.8H-L and Table S7). Together with the IH results, these data strongly establish the biological signal in our network data, and the high accuracy of the module maps.





**Figure 3.8:** Examples of functional validations of candidates emerging from the networks. (A-F) The functional validations rely on a total of 382 tissue sections of which only a very small subset (six of 382 or less than 2%) are shown in this figure. (A) SNX9 is located in clusters of unknown function involved in development of the cardiac valves (Figure S2C and S3A), which correlates with its specific expression in endothelial cells of the endocardial cushions (EC, precursors of heart valves) in a nine week old human heart. Note that strong expression of SNX9 is confined to the endothelial cells lining the developing portion of the EC (section below (A)). (B) Toluidine blue staining of hyaluronic acid, a marker for epithelial-mesenchymal transformation in the endocardial cushions correlates with the expression pattern of SNX9, hereby confirming its expression specifically in the developing parts of the EC. (C) SNX9 expression in endothelial cells of the developing aortic valves (AV) in a nine week old human heart. Strong expression is shown in endothelial cells lining the valves and in cells within the valves. (D) DLL1 expression in the leading edge of septum primum (SP) in a six week old human heart. DLL1 is located in a NOTCH signaling cluster involved in development of the atrial septum (Figure S2A). Note the stronger expression in the migrating and developing part of the SP (arrows) compared to endothelial cells lining the inner surface of the atrial wall (asterisk). (E,F) Expression in cardiomyocytes of the ventricle and atrium of PTGS2 in an 18 week old human heart. PTGS2 is located in several clusters involved in myocardial growth and organisation (Figure S4A-E). Note the stronger expression of PTGS2 in the atrium (At) compared to the ventricle (Ve). (G-L) Validating a larger set of 49 candidates by real-time quantitative RT-PCR. (G) The gene expression level of the candidate genes and 29 control genes were measured in two hearts collected from embryos at ages 46 days and 67 days, respectively. Controls were genes corresponding to randomly chosen proteins that did not significantly interact with the 255 CD proteins, but were represented in our interaction data set. The data distribution is displayed by a box-and-whisker plot. A single outlier data point in the control group is shown with an asterisk. (H) Heatmap showing the relative level of gene expression of 18 representative candidates (level at day 40 = 1). The gene expression level of each of the 18 candidate genes was analyzed in hearts collected from 12 human embryos or fetuses of the indicated stages of development (between day 40---68 post fertilization). The data was sorted in four groups according to expression pattern of the genes using hierarchical clustering. Statistical significant correlation between expression value and days post fertilization is marked by an asterisk (I-L) Representative plots of gene expression within the four groups. A trend line representing the average value of the data in four groups of three data points is shown as a smoothed line. Expression levels were measured by QPCR and the data was normalized using the average value of six housekeeping genes (*GAPDH*, *COX4A*, *B2M*, *ATP6A*, *HPRT*, *RPL13*).

## Conclusion

We present a framework for gaining new insights into the systems biology of the protein networks driving organ development and related polygenic human disease phenotypes, exemplified here with heart development and CHD. Our analysis is the first example of large-scale integration of phenotypic data from targeted mice mutants with high-confidence experimental proteomics data and represents the most comprehensive characterization and analysis of the functional protein networks underlying the development of an organ system to date. A strength of our approach is that it immediately puts new candidates in the functional context of other, more well-characterized, network components.

We have shown that analysis of organ development at the systems level can be used to discover new developmental modules, gain insight into the evolution of organs, and understand the biology of highly polygenic disorders associated with aberrant organ development. A weakness of the method is the lack of cellular resolution of the networks due to use of macroscopic phenotypes as the starting point of the analysis. However, the morphological subgroups associated with the networks, and the IH data (which has the resolution of individual cells), strongly suggests in which cell populations the individual networks are active. The networks generated here can be used as a community resource for addressing major questions in developmental and cardiac biology, and we have made a database of the relevant network data available at <http://www.cbs.dtu.dk/suppl/dgf/>.

In principle, the framework can be applied to any organ, to widen our understanding of the functional architecture of protein networks that drive the formation of organs. Additionally, they can facilitate the evolution of novel approaches in regenerative medicine, because a thorough characterization and understanding of the genes, proteins, pathways and concepts underlying organ developmental programs will be necessary for the successful manipulation of these systems in tissue-engineering technologies and stem cell therapy. Finally, the networks can be used as a functional scaffold for understanding combinatorial effects of gene-gene and gene-environment interactions in complex heart phenotypes.

## Materials and Methods

### Generating a functional network

A network is generated by determining the first and second order interactions of CD proteins associated with a given morphological subgroup in a human protein interaction network consisting of refined experimental proteomics data. This network is described in high detail in (Lage et al, 2008; Lage et al, 2007)[47, 65], and online <http://www.cbs.dtu.dk/suppl/dgf/>. The full network (InWeb 29) can be downloaded from <http://www.cbs.dtu.dk/suppl/dgf/>. Interactions of the CD proteins are integrated into a network by always including direct interactions between CD proteins, and only including indirect interactions mediated through proteins with Q percent of its interactions to the CD set. Various thresholds for Q are iteratively tested and value of Q for the final network is chosen based on which value gives the optimal network significance, this procedure is described in detail in [175, 176]. The method for determining network significances can be seen below. Detailed views of the networks can be seen in Figures S1-S4.

### Determining network significances

The significance of each of the generated 19 networks was determined by randomization testing as described in detail previously [175, 176]. Specifically, for an input set of  $N$  input proteins yielding an interaction network (connected component) with  $G$  input proteins and  $T$  total proteins a network score ( $NS_{input}$ ) was determined. This network score is the fraction of input proteins of all proteins in the network ( $G/T$ ). We then determined the significance of the network score by empirically estimating the probability of observing a similar or better network score in networks generated by using 10,000 random input sets of size  $N_{input}$ . The random gene sets were chosen so the degree distribution of proteins in the random sets approximate the input set. As each query generates a varying number of networks (connected components) the probability estimates can be calculated from the total amount of networks produced by all 10,000 randomizations that have a network score  $> NS_{input}$ . For this reason network p-values can be lower than the amount of random queries. All network p-values can be seen in Figures S1-S4 below the title of the network. To rule out the chance of functional bias in the CD set, we analyzed the set for bias as discussed in Supplementary Information.

### Identifying candidates for IH and expression analyses

A set of raw candidates were determined by querying all proteins in our interaction network for the amount of interactions to the CD set and determining the hypergeometric probability of this interaction profile. Out of all proteins in the proteome, forty nine novel candidates had a significant interaction profile to the CD proteins after adjustment for multiple testing (described in detail in Supplementary Information and Figure S5). We then used the functional networks assigned to each morphological subgroup to determine the most likely developmental function of the candidates. This was done by identifying the specific sub-networks to which the interactions of the candidates were most significant (as described in Supplementary Information and Figure S6). 12 of the 49 candidates were chosen for IH analysis based on the overlap between morphological subgroups and the developmental stages present in our panel of embryonic hearts available for validation experiments.

### Human embryonic and fetal heart tissues

Human embryonic tissues were collected from legal abortions, according to the Helsinki Declaration II, and their use was approved by the local science ethics committee. Embryonic or fetal age was based on measurement of crown-rump length (CRL). Immediately after dissection the samples were snap frozen in liquid nitrogen or treated with RNA later according to manufacturer's instructions (Ambion, Austin, TX). Samples for immunohistochemistry were dissected into appropriate tissue blocks and fixed for 12-24 hours at 4°C in either 10% neutral buffered formalin, 4% Formol-Calcium, Lillie's or Bouin's fixatives. The specimens were dehydrated with graded alcohols, cleared in xylene and paraffin embedded. Serial sections, 3-5 (m thick, were cut in transverse, sagittal or horizontal planes and placed on silanized slides.

### Immunohistochemical analysis

Sections were deparaffinized and rehydrated in xylene followed by a series of graded alcohols according to established procedures. The sections were treated with a fresh 0.5% solution of hydrogen peroxide in methanol for 15 minutes to quench endogenous peroxidase and then rinsed in TRIS buffered saline (TBS, 5mM Tris-HCl, 146 mM NaCl, pH 7.6). Non-specific binding was inhibited by incubation for 30 minutes with blocking buffer (ChemMate antibody diluent S2022, DakoCytomation, Glostrup, Denmark) at room temperature. The sections were then incubated overnight at 4°C with the primary antibody in blocking buffer (ChemMate antibody diluent S2022, DakoCytomation). The sections were washed with TBS and then incubated for 30 minutes with a peroxidase-labelled secondary antibody. The sections were washed with TBS, followed by incubation for 10 min with 3,3'-diamino-benzidine chromogen solution. Positive staining was recognized as a brown color. The sections were dehydrated in graded alcohols followed by xylene and cover-slipped with DPX mounting media. Non-immune rabbit IgG1 (Xo936) was used as negative control. Specificity of the antibodies were determined by their ability to stain specific cell populations in the tissue sections (examples are shown in figure S7-S12).

The following primary antibodies were used: anti-PTGS2 (35-8200, Invitrogen), anti-MAPK8 (SC-6254, Santa Cruz Biotechnology, Santa Cruz, CA), anti-CAV3 (610421, BD transduction laboratories, Franklin Lakes, NJ), anti-MAPK3 (SC-7383, Santa Cruz Biotechnology), anti-SRC (AT-7016, MBL international, Woborn, MA), anti-JAG2 (SC-8157, Santa Cruz Biotechnology), anti-DLL1 (SC-9102, Santa Cruz Biotechnology), anti-NOTCH3 (SC-7474, Santa Cruz Biotechnology), anti-NOTCH4 (SC-5594, Santa Cruz Biotechnology), anti-BMX (ab73887, Abcam, Cambridge, UK), anti-PTK2B (ab78119, Abcam), anti-BMP4 (ab31165, Abcam), Anti-EGFR (#2232, Cell Signaling Technology, Boston, MA).

### Real-time quantitative RT-PCR

We chose quantitative real-time quantitative RT-PCR (QPCR) for this analysis because it is considered to be the most accurate and sensitive method for detecting RNA differences also at very small amounts [177]. Total RNA was isolated from tissues using TRIzol Reagent (Invitrogen, Taastrup, Denmark) and cDNA synthesized with SuperScript II (RNase H-) reverse transcriptase (Invitrogen) according to manufacturer's instructions. QPCR analysis was carried out on a ABI 7500 Fast real-time PCR system using a LightCycler FastStart DNA MasterPLUS SYBR GreenI kit (Roche, Hvidovre, Denmark). Primer sequences used for QPCR analysis are available on request. To exclude that polymorphic gene expression between the individual developing hearts could account for the observed differential expression trends reported by QPCR, we also used Polony Multiplex Analysis of Gene Expression (PMAGE) [178] to measure the expression of the 49 candidates in right ventricular outflow tract (RVOT) from TOF patients at the time of primary surgical repair and left ventricle (LV) collected from patients with either heart failure or diabetic cardiomyopathy. The expression levels of the candidates were compared to the expression levels of a different set of 49 randomly chosen controls after normalizing both gene sets against gene expression in glioblastoma tissue. Here, the heart developmental candidates were significantly higher expressed in heart tissue than the controls ( $p = 0.016$ ) (see Supplemental Information and Figure S13).

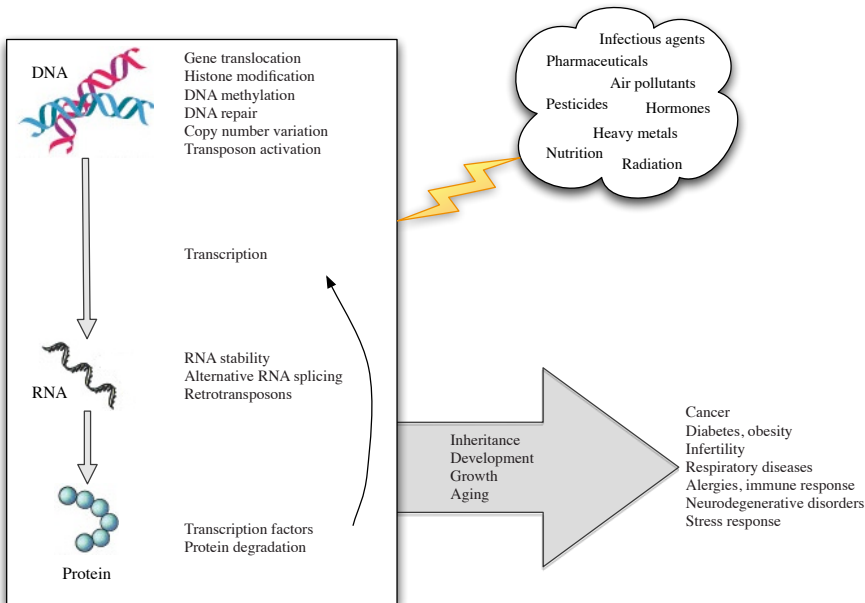
## Acknowledgements

We would like to thank Christine Kocks, Barbara Pober, Chris Cotsapas, Josh Korn, Soumya Raychaudhuri, Ben Voight, Gerasimos Sykiotis and Lars Juhl Jensen, for critical discussions. We are also grateful to the Charles Lee lab for discussions and input on the project and results. This work was supported by The Danish Heart Foundation, the Villum Kann Rasmussen Foundation and The Novo Nordisk Foundation. Wilhelm Johannsen Centre for Functional Genome Research is established by the Danish National Research Foundation. KL is supported by a grant from 'Forskningsrådet for Sundhed og Sygdom'; and KL and PKD are supported by NICHD RO1 grant HD055150-03. The authors would like to thank Lillian Rasmussen and Kirsten Winther for technical assistance.



## From chemicals to disease

**D**ISEASE development is not only attributable to gene polymorphism or heritable defects. In many instances, the actions of genes are known to be modified by environmental conditions, and human disease propensity is shaped by interactions between each individual's genes and the environment [179] (Figure 4.1). Environmental agents have been shown to influence chronic disease susceptibility, combined with an array of factors, such as genetic fit, age, and other predisposition conditions [180]. True understanding about diseases cannot come through, unless both the genetic and environmental contributions to its triggers are discovered.



**Figure 4.1:** Summary of gene regulatory mechanisms affected by exposure to external agents, with disease implications. Adapted from *Edwards et al.* [179].

When thinking about disease etiology, one should have in mind that disease risk is best predicted by considering genetic and environmental factors combined. There are numerous types of environmental agents, including infectious agents, chemicals, diet, and radiation (Figure 4.1). They affect genes in multiple ways, including DNA methylation and mutations, RNA stability, and influence gene expression and protein function. At the same time, naturally occurring polymorphisms in the population may in turn affect the chemical susceptibility and increase the disease predisposition [181].

For better understanding the effects of the environment in the human health, and decipher their complex mechanisms of action, we need to combine the already existing information about drugs, pesticides, and other environmental chemicals with gene and protein interaction data across species and how they relate to human diseases. Ultimately we would like to create a framework for accessing human disease risk in a community, or even better, at the level of the individual.

The purpose of this chapter is to frame drug targets in the context of cellular and disease networks. The article *Deciphering Diseases and Biological Targets for Environmental Chemicals using Toxicogenomics Networks*, in Section 4.2, describes a generic approach to understanding the underlying molecular mechanisms that regulate chemical activity in the human body, and discover which biological pathways they perturb. The method is based on the integration of toxicogenomics data, chemical structure, protein interaction data, and disease and functional annotation. The main task in the article was creating a protein-protein association network (P-PAN), where two proteins are associated if they are affected by the same chemicals. This network was benchmarked against refined experimental protein interaction databases (refer to Chapter 3).

In the last manuscript included in this thesis, entitled *ChemProt: A Disease Chemical Biology Database*, we construct a database (ChemProt)<sup>1</sup>, compiling several networks developed in previous work. This online database integrates chemical-protein annotation resources and the disease complexes from Paper III. ChemProt was designed to support *in silico* evaluation of chemicals and environmental compounds, and the selection of new compounds based on their activity profile against biological targets.

## 4.1 Chemicals and disease

### Susceptibility and exposure

As explored in the previous chapters, many common human diseases seem to be polygenic, where the incidence of defects on a single gene may not trigger the disease, but might become detrimental if multiple variants of susceptibility genes accumulate in a disease cluster. These susceptibility genes alone are not enough for causing disease, they rather increase or decrease risk in combination with other genes and with exposure to external agents [180].

*Cooper* [182] relates to the example of the African-American population in the United States, referring that this particular minority suffers from more cases of vascular disease and dementia than the caucasian population in the same area, but that both conditions are infrequent in West

---

<sup>1</sup>Available at <http://www.cbs.dtu.dk/services/ChemProt-1.0/>

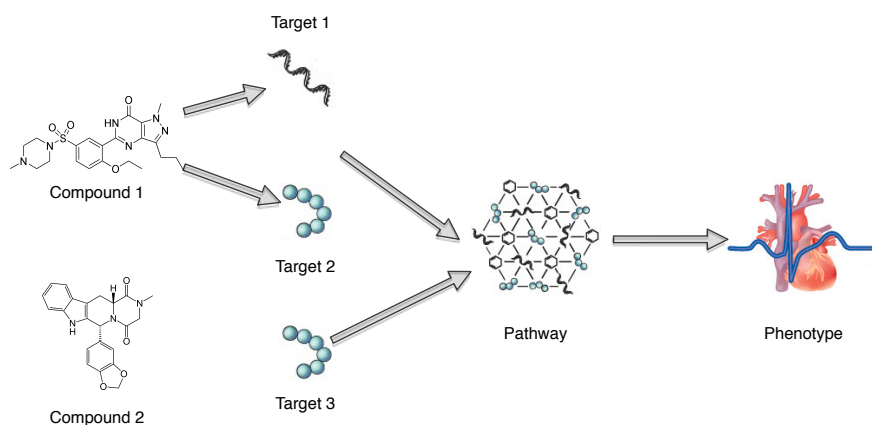


Africa. The explanation for the lower disease rates in the American versus African population can be pictured as an example of gene-environment interactions.

In the broader public aspect, common diseases result from common exposures to environmental conditions which we all are susceptible for, but in varying degrees. There is therefore an impact on both the genetic and the environmental factors on the distribution of phenotypes.

## Polypharmacology and side-effects

A few years back, drug design was based on the premise that drugs selectively interact with one or two molecules, and thus preventing or treating disease. This notion has seriously been put back, and now we know that most drugs interact with multiple targets (polypharmacology). For some drugs this effect is beneficial, and probably essential in psychiatric medication [183]. A simplistic workflow for beneficial polypharmacology can be seen in Figure 4.2, where one compound acts on two different targets that are part of the same pathway, therefore potentiating its action and triggering a bigger phenotypic response. Other therapeutic drugs have been withdrawn from the market due to serious adverse side-effects, because they interact with molecular targets other than those they were designed for.



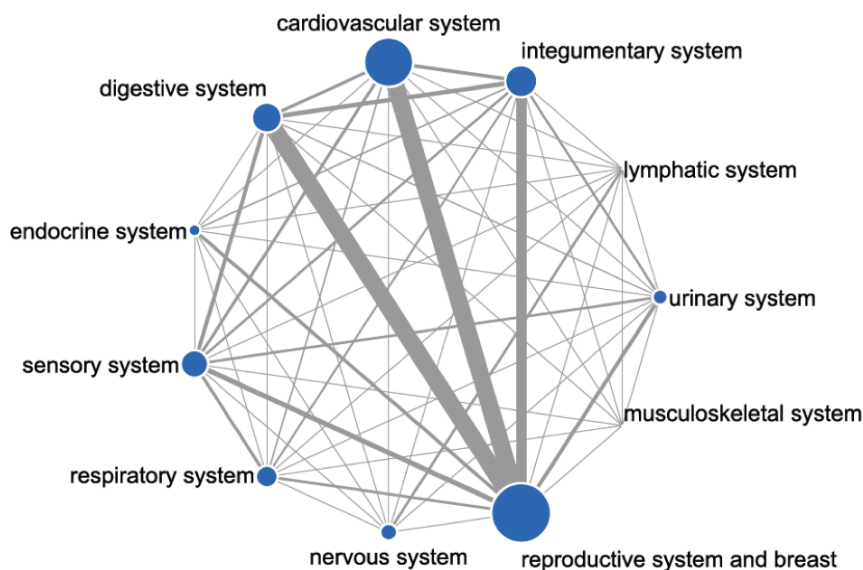
**Figure 4.2:** Compound 1 is a polypharmacology example: one drug affects two molecular targets from the same pathway, potentiating the phenotypic change. Compound 2, on its hand, targets a different molecule and is chemically dissimilar, but disturbs the same pathway as compound 1, resulting in a similar phenotype.

Side-effect similarity has also been used to infer drug targets, even on chemically different drugs and therapeutic indications [184]. Drug side-effects could be due to interaction with the primary or additional targets, downstream network perturbations, protein interaction, drug-drug interference, dosage effects, or problems in the metabolism of the drug.

## Systems chemical toxicology

Data integration at various levels can help predicting the adverse effects caused by drugs and environmental chemicals, and contributing for lowering drug development costs. By taking toxicology data and combining it with gene and tissue resolution, we can better pinpoint where

the molecular targets for each compound are, and contribute for an earlier detection of side-effects.



**Figure 4.3:** Overlap between chemicals in the different target systems of the human body. Systems are connected according to their common chemicals. Area of the circle is proportional to the number of chemicals targeting the system, and edge thickness represents the number of common chemicals between two systems. Figure by Audouze *et al.*, in preparation.

Figure 4.3 is an example of a such approach, where we took toxicology data for 25,656 chemicals annotated for 11 systems in the human body and checked which chemicals were targeting more than one system in the human body. The idea of the manuscript (Audouze *et al.*, in preparation) is to connect each system to its component tissues, and integrate them with gene and protein data, and chemicals, thus achieving tissue and protein resolution of chemical adverse effects.

## 4.2 Paper V

# Deciphering diseases and biological targets for environmental chemicals using toxicogenomics networks

*PLoS Comput Biol*, 2010 vol. 6 (5) pp. e1000788.

Karine Audouze<sup>1</sup>, Agnieszka Sierakowska Juncker<sup>1</sup>, Francisco S. Roque<sup>1</sup>, Konrad Krysiak-Baltyn<sup>1</sup>, Nils Weinhold<sup>1</sup>, Olivier Taboureau<sup>1</sup>, Thomas Skøt Jensen<sup>1</sup> & Søren Brunak<sup>1\*</sup>.

<sup>1</sup>Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark.

\*Correspondence should be addressed to brunak@cbs.dtu.dk, tel: +45 45 25 24 77, fax: +45 45 93 15 85

### Abstract

Exposure to environmental chemicals and drugs may have a negative effect on human health. A better understanding of the molecular mechanism of such compounds is needed to determine the risk. We present a high confidence human protein-protein association network built upon the integration of chemical toxicology and systems biology. This computational systems chemical biology model reveals uncharacterized connections between compounds and diseases, thus predicting which compounds may be risk factors for human health. Additionally, the network can be used to identify unexpected potential associations between chemicals and proteins. Examples are shown for chemicals associated with breast cancer, lung cancer and necrosis, and potential protein targets for di-ethylhexyl-phthalate, 2,3,7,8-tetrachlorodibenzo-p-dioxin, pirinixic acid and perme-thrine. The chemical-protein associations are supported through recent published studies, which illustrate the power of our approach that integrates toxicogenomics data with other data types.

### Introduction

Humans are daily exposed to diverse hazardous chemicals via skincare products, plastic cups, computers and pesticides to mention but a few sources. The potential effect of these environmental compounds on human health is a major concern [179, 185]. For example, chemicals such as phthalate plasticizers have been widely linked to allergies, reproductive disorders and neurological defects. Humans are intentionally exposed to drugs used for treatment and cure of diseases. Many drugs affect multiple targets and may interact or affect the same proteins as environmental chemicals [186--188]. The mechanism of action of these small molecules is often not completely understood and can be associated to adverse and toxic effects through for example drug-drug interactions [189]. There is thus a need to improve our understanding of the underlying mechanism of action of chemicals and the biological pathways they perturb to fully evaluate the impact of small molecules on human health.

An essential step towards deciphering the effect of chemicals on human health is to identify all possible molecular targets of a given chemical. Various network-oriented chemical pharmacology approaches have been published recently to identify novel protein candidates for drugs, using structural chemical similarity [183, 184, 190, 191]. For example *Keiser et al.* [190] applied network analysis to drugs and their targets. The authors identified unexpected molecular targets such as muscarinic acetylcholine receptor M<sub>3</sub>, alpha-2 adrenergic receptor and neurokinin NK<sub>2</sub> receptor for methadone, emetine and loperamide, respectively. Additionally, recent studies have demonstrated that chemicals could be classified based upon their effect on mRNA expression detected by microarrays [192, 193]. *Lamb et al.* [192] showed that genomic signatures could be used to recognize drugs with common mechanism of action allowing discovery of unknown modes of action. Despite the explosion of chemical-biological networks, the chemical toxicity remains a major issue in human health. Analysis of environmental chemicals with similar gene expression profiles is still lacking. With the recent advances in toxicogenomics, information on gene/protein activity in response to small molecule exposures becomes more available. This provide necessary data to develop computational systems biology models to predict both high level associations (linking chemical exposures to diseases) and more detailed associations (linking chemicals to proteins).

In this paper we present a method that can associate chemicals to disease and identify potential molecular targets based on the integration of toxicogenomics data, chemical structures, protein-protein interaction data, disease information and functional annotation. The core of our procedure is derived from the ‘target hopping’ concept defined previously [186]. But instead of considering only binding activity, we extended the concept to gene expression. If two proteins are affected with two chemicals, then both proteins are deemed associating in chemical space. Our approach is not only a statistical model but mimics the true biological system by constructing a network of associations between human proteins defined as Protein-Protein Association Network (P-PAN). We have validated our network by comparison with two high confidence protein-protein interaction (PPI) networks, and by assessing the functional enrichment of clusters in the network generated. The P-PAN revealed both known as well as many novel surprising connections between chemicals and diseases or proteins. We provide literature support for some of the unexpected associations, such as the connection between diethylhexylphthalate (DEHP) and gamma-aminobutyric acid A receptor beta target [194], as well as between apocarotenal, a chemical found in spinach, and necrosis. This illustrates the usefulness of an approach that integrates toxicogenomics data with other diverse data types.

## Results

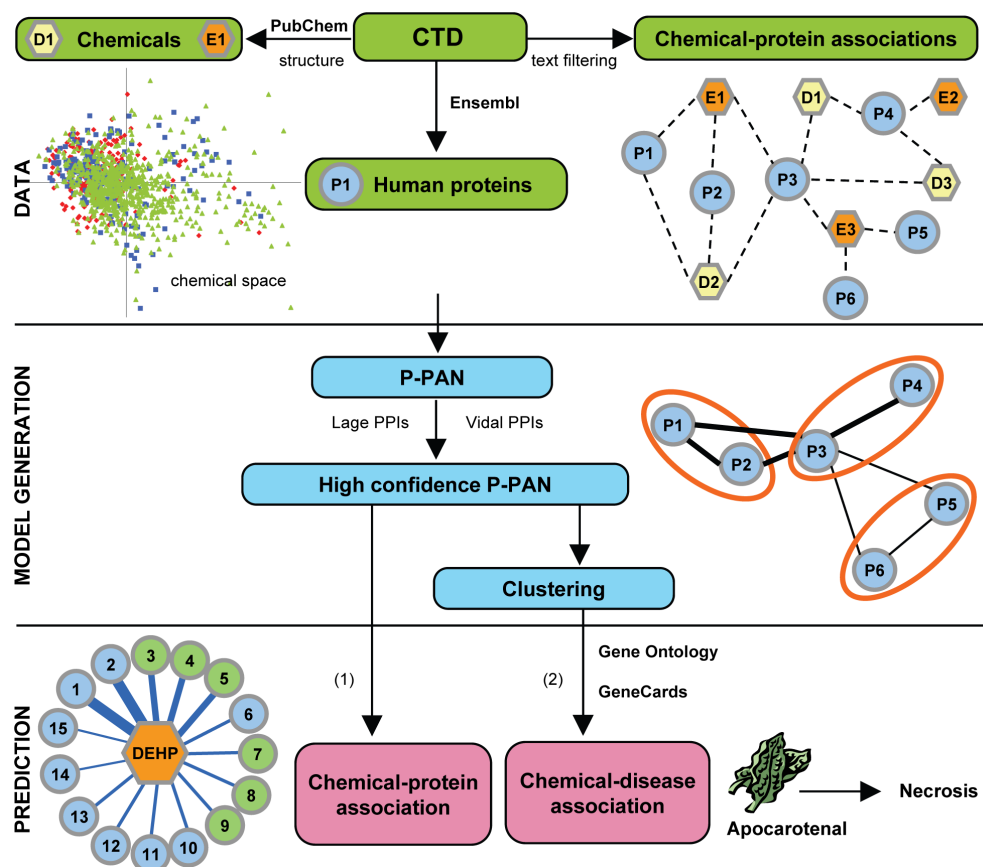
Based on the Comparative Toxicogenomics Database (CTD) [181], we constructed a human P-PAN. A workflow of the strategy is shown on Figure 4.4. We extracted 42,194 associations between 2,490 chemicals and 6,060 human proteins from the CTD. We mapped compounds to chemical structures from PubChem and extracted their indication of use from Medical Subject Headings (MeSH<sup>2</sup>) to classify them as either drugs (MeSH: ‘Pharmaceutical Actions’) or environmental chemicals (MeSH: ‘Toxic Actions’ and ‘Specialty Uses of Chemicals’).

In the CTD, drugs and environmental compounds are claimed to be associated with toxicologically important proteins. To estimate how much the information from the CTD differs from available data on pharmacological action of drugs, we compared the data shared between CTD and DrugBank, as of May 2009 [195]. DrugBank is a repository of pharmacological action for ‘Food and Drug Administration’ approved drugs. From the 1358 drugs gathered in DrugBank, 420 drugs matched in CTD. Interestingly, whereas 1403 proteins are associated to these drugs in DrugBank, only 194 proteins are found in both databases. For example, according to Drug Bank celecoxib, a known non-steroidal anti-inflammatory drug, is associated to two metabolizing enzymes: the Cytochrome P450 2C9 (CYP2C9) and the Cytochrome P450 2D6 (CYP2D6) and to two drug targets: the Prostaglandin G/H synthase 2 (COX-2) and the 3-phosphoinositide-dependent protein kinase 1 (PDK1). In the CTD, celecoxib is linked to 33 human proteins including CYP2C9 and COX-2. The toxicity information extracted from CTD is relatively different to the known pharmacological action of drugs and should be considered as a complementary source of information.

## Structure-target relationship

To investigate the assumption that two compounds sharing similar structure can potentially affect the same molecular targets, we compared chemical properties of the compounds collected

<sup>2</sup><http://www.nlm.nih.gov/mesh/MBrowser.html>



**Figure 4.4:** Workflow of the strategy for generating a human P-PAN and predicting novel associations. **DATA:** Extraction and filtering of human protein-chemical associations from CTD. The visualization of the chemical space by Principal Component Analysis projection confirms that drugs (D) and environmental chemicals (E) shared structural properties, and then may affect similar protein targets. The two first principal components, which explained about 44% of the variance on the calculated properties are shown (green: pharmaceutical actions, red: toxic actions and blue: specialty uses of chemical). All proteins (P) were mapped to Ensembl gene identifiers to facilitate further data integration. **MODEL GENERATION:** Construction of the P-PAN. The P-PAN was created from associations present in the CTD (dashed edge lines) between chemicals and proteins. In the P-PAN, two proteins are connected to each other (edge lines) if they share a common chemical. A weighted score, represented by the width of the black edges, was assigned to each protein-protein association. It represents the strength of the network between two proteins as defined by the number of shared compounds for both molecular targets. Selection of a scoring function and a high confidence P-PAN after overlaps comparison with two human interactomes (PPIs) based on experimental evidences. Clustering of the P-PAN and evaluation of the biological meaningful of the clusters using Gene Ontology annotations. **PREDICTION:** (1) Prediction of novel molecular targets for chemical using a neighbor protein procedure. DEHP (orange) is known to be connected with blue proteins and is predicted to be associated with green proteins. A confidence score was calculated for each protein, represented by the width of the edges; thick edge for high score to thin edge for low score. (2) Prediction of disease associated with chemical after integration of protein-disease information using GeneCards in clusters. As example, apocarotenal, a compound found in spinach is predicted to be link to necrosis.

from the CTD. The chemicals were characterized by 50 properties calculated from the structure, including the molecular mass and affinity for a lipid environment. The distribution of properties, as it appears in a multi-dimensional properties space, was projected and visualized in two dimensions using principal component analysis (PCA) (shown in Figure 4.4). There is substantial overlap in the PCA projections between environmental chemicals and drugs indicating that they can potentially affect the same protein targets. We also compared the oral bioavailability profiles of compounds based on standard Lipinski [196] and Veber [197] rules. Again, overlaps were observed, indicating that environmental chemicals mimic drug properties (see Figure S1<sup>3</sup>). These results confirm that it is reasonable to generate a network by integrating toxicogenomics knowledge from both drugs and environmental compounds, as they share many properties.

### Generating a high confidence human Protein-Protein Association Network

The human P-PAN was generated based on the assumption that if two proteins are biologically affected with the same chemicals (defined as shared chemicals), they are likely to be involved in a common mechanism of action of the chemicals. Then, two proteins are connected to each other if they are linked to the same chemical in the CTD. The resulting P-PAN consists of 2.44 million associations. To reduce noise and select the most significant associations, we assigned two reliability scores to each protein-protein association: a score based on hypergeometric calculation and a weighted score. The weighted score was calculated as the sum of weights for shared chemicals, where weights were inversely proportional to the number of associated proteins for a given compound.

We went one-step further and compared the P-PAN with two human PPI databases: (1) a high confidence set of experimental PPIs extracted from a compilation of diverse data sources [47] and (2) PPIs based on an internal consistent single data source [48]. Our P-PAN performed well compared to both PPIs. Based on the calibration curves (Figure S2), we considered a threshold that capture good overlaps between our P-PAN and the PPI networks for different reliability scores thus reducing our P-PAN to around 200,000 reliable associations. Using this approach, the molecular target predictions are limited to the 3,528 proteins present in the P-PAN. To confirm that biological information is not lost when selecting only 8% of the entire P-PAN, we compared functional enrichment for the complete network (6,060 proteins) and for the high confidence sub-network (3,528 proteins) using Gene Ontology (GO) [198]. For example cell proliferation (p-values of  $3.22\text{e-}36$  and  $1.46\text{e-}27$  for the large network and the sub-network, respectively) and protein binding (p-values of  $1.2\text{e-}72$  and  $4.13\text{e-}47$  for the large network and the sub-network, respectively) were the most overrepresented terms.

Since proteins tend to function in groups, or complexes, an important step has been to verify that our high confidence network mimics true biological organization. This task is commonly executed using graph clustering procedures, which aim at detecting densely connected regions within the interaction graph. Two clustering methods have been applied to our network. The molecular complex detection (MCODE) approach [199] that allows multiple clusters assignment for a protein, mimicking the reality as a protein can participate in several complexes

<sup>3</sup>Supplementary Information, online at <http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000788#s5>

simultaneously. On the other hand, the markov cluster algorithm (MCL) [200] which assign one protein to a unique cluster has been shown to be superior to other graph clustering methods in recent studies [201, 202]. Applied on our network, MCODE extracted few large core clusters and several tiny clusters (possibly singleton clusters). The MCODE approach results in a clustering arrangement with a weak cluster-wise separation. Compared to MCL, MCODE yielded a lower number of clusters, with a higher number of proteins per cluster. Only 35 clusters varying in size from five to 845 proteins were extracted. Using the MCL algorithm we obtained a more heterogeneous separation with 58 clusters varying in size from five to 462 proteins. Therefore, to identify the biological meaningfulness of our network, we used complexes extracted using the MCL method. Each cluster was then investigated for functional enrichment based on GO terms. To ensure the high quality of functional annotations we used only annotations experimentally supported or with traceable references. Hypergeometric testing was used to determine GO functional annotation overrepresented amongst each cluster. The two top scoring molecular functions found were heme binding (p-value of  $6.60\text{e-}25$ , cluster 4) and glucuronosyl transferase activity (p-value of  $2.34\text{e-}21$ , cluster 12). Regulation of apoptosis (p-value of  $1.67\text{e-}17$ , cluster 2) and oxidation reduction (p-value of  $6.67\text{e-}14$ , cluster 4) were the most highly enriched categories in the biological process branch of the GO. This analysis thus confirms that clusters in the network, and therefore the proteins associated with each other, are functionally coherent. This was further evidence that the organization of the network is meaningful.

### **Diseases associated to clusters**

In the clusters of the P-PAN, proteins are more connected with other proteins within the cluster than with the other targets in the network. As proteins are associated based on their shared relationship with chemicals, proteins within a given cluster tend to be more linked to specific compounds. It is thus possible to find associations between diseases and the chemicals that underlie the protein-protein associations within the cluster using protein-specific disease annotations. For each cluster, we investigated if specific disease annotation was found more frequently than expected by using protein-disease information [203]. We identified several diseases associated with specific clusters. These included the two most common types of cancer, breast cancer (cluster 1, p-value of  $9.67\text{e-}18$ ) and lung cancer (cluster 12, p-value of  $4.84\text{e-}12$ ), as well as necrosis (cluster 2, p-value of  $2.26\text{e-}12$ ), ichthyosis (a skin disorder associated to cluster 4, p-value of  $1.41\text{e-}5$ ), retinoblastoma (cluster 7, p-value of  $9.46\text{e-}8$ ) and inflammation (cluster 8, p-value of  $1.55\text{e-}5$ ).

### **Mining the network for chemicals associated with disease**

To predict which chemicals may affect human health, we then analyzed selected clusters to identify new chemical-disease associations (see Table 4.1). When linking diseases to compounds, it is important to keep in mind that there is no direction in the association, i.e. it is not possible from the network to separate positive from negative associations between a chemical and a disease. Discriminating between whether a compound prevents or causes disease requires manual interpretation of the association.



<i>Cluster ID</i>	<i>Disease</i>	<i>Chemical name</i>	<i>p-Value</i>
1 (462 proteins)	Breast cancer (128 proteins)	<i>estradiol</i>	7.68e-134
		<i>bisphenol A</i>	4.46e-92
		<i>PCBs</i>	1.15e-88
		<i>genistein</i>	2.20e-78
		<i>fulvestrant</i>	7.05e-63
12 (59 proteins)	Lung cancer (29 proteins)	<b>thimerosal</b> (10 proteins)	1.57e-26
		<b>DNCB</b> (12 proteins)	3.29e-22
		<i>styrene</i>	7.78e-06
		<i>arsenic disulfide</i>	4.76e-35
2 (433 proteins)	Necrosis (122 proteins)	<b>apocarotenal</b> (8 proteins)	1.63e-29
		<i>doxorubicin</i>	2.66e-26

**Table 4.1:** Mining the P-PAN for chemicals associated with breast cancer, lung cancer and necrosis, using a clustering procedure. Chemicals already known from the literature to be associated to disease are shown in italic. In bold are the chemicals significantly associated to disease, which are unknown to be disease-causing chemical from the literature. The number of proteins is shown in brackets for each cluster, disease and novel association. As example, among the 433 proteins associated to cluster 2, 122 are known to be linked to necrosis. Among these 122, 8 are connected to apocarotenal in CTD.

One of the clusters showed high enrichment for breast cancer. The most significantly associated chemicals are already known from the literature to be related to cancer, thus supporting the clustering quality of the P-PAN. Among the most significantly associated chemicals are the well-known polychlorinated biphenyls (PCBs). PCBs are used for a variety of applications i.e. flame retardants, paints and plasticizers. After being banned due to their toxicity, they still persist in the environment. Previous results suggest that specific PCBs may indeed be associated with breast cancer [204]. Several organizations (EPA, IARC) have classified PCBs as probable human carcinogens. When we inspected another cluster highly connected to lung cancer using our P-PAN method, thimerosal, dinitrochlorobenzene (DNCB) and styrene were significantly associated with this cluster. Thimerosal and DNCB are not known lung cancer-causing chemicals, while the last compound, styrene has been classified as a possible carcinogen. Thimerosal is an organomercury chemical widely used as preservative in health care products and in vaccines. It may have possible adverse health effects such as a role in autism and in nervous system disorders [205] as well as possible gene-toxic effects to human lymphocytes [206]. No study has previously related it to lung cancer. The second chemical DNCB is known to be a skin allergen that may cause dermatitis. Genes associated with allergies were shown to be up regulated in rat lung tissue after DNCB exposure [207], but no direct link to lung cancer has been demonstrated so far. Another interesting finding is the association between apocarotenal and necrosis. Apocarotenal, a natural carotenoid found in spinach and citrus, is used as a red-orange coloring agent (E160E) in foods, pharmaceuticals and cosmetics products. No direct evidence

has been found that links apocarotenal to necrosis. However, *in vitro* and *in vivo* studies [208] have suggested that spinach may be a good anti-cancer agent. This is in line with epidemiologic studies that have shown that those who consume higher dietary levels of fruits and vegetables have a lower risk of certain types of cancer [209] due to the presence of carotenoids. Furthermore, carotenoids have been defined as chemopreventive agents [210]. Studies have established associations between carotene and beta-carotene with reduced risk of prostate cancer [211] or breast cancer [212]. The prediction that apocarotenal is positively associated to necrosis and could prevent certain types of cancer is thus indirectly supported by other studies. The other chemicals significantly associated to disease (Table 4.1) are discussed in the supplementary text (see Text S1).

### Predicting novel molecular targets for chemicals

Besides revealing disease-chemical associations, the network can be used to predict novel targets for chemicals. It has been shown that many small molecules affect multiple proteins rather than a single target, and that proteins sharing an interaction with a chemical are targeted by the same chemicals [190]. Based on the CTD data available, strong promiscuities between some proteins exist. For example, more of 25% of chemicals annotated to estrogen receptor 1 (ESR1) affects also progesterone receptor (PGR). In the same order, cytochrome p450 2D6 (CYP2D6) and cytochrome p450 2C9 (CYP2C9) shared one-third of their respective associated compounds. By the term *affected*, we consider effects such as up regulated, down regulated, agonist, antagonist and inhibitor. Then, our network can not be used to identify chemical synergies or opposite effect on proteins. Thus, if two proteins are affected by two chemicals and one of the proteins is further deregulated by an additional chemical, then it might be that both proteins are in fact deregulated with the same three chemicals. Based on this assumption and in order to suggest novel associations between chemicals and proteins, a neighbor protein procedure was used which scored the association between each protein and each chemical (see Materials and Methods). Molecular targets known to be associated with a chemical were extracted from the CTD, and the P-PAN was scanned for proteins associated with a high score. The significance of enrichment was calculated by random testing (for the confidence scores see Text S2), and sub-networks were subsequently ordered according to their significance. Four examples of various chemicals are presented in Table 4.2 on page 81 (other case stories are shown in Table S1). To estimate the performance of our approach for approved drugs, we analyzed the level of recall and precision obtained for the 420 common drugs between DrugBank and CTD. We obtained a recall and a precision of 5.91% and 3.77% respectively, corresponding to the percentage of interactions in DrugBank retrieved and percentage of interactions in DrugBank from all interactions predicted obtained from CTD data and from the neighbor protein procedure. These values illustrate that information between the two data sources are relatively different.

### Examples of proteins associated to chemicals

Phthalates, mainly used as plasticizers, have received a lot of attention as environmental compounds because they are potential human carcinogens. As there are many phthalates, we focused on Di-EthylHexyl Phthalate (DEHP) that has been associated with more proteins compared to other phthalates such as additional information on kinases (e.g. mitogen-activated protein

kinase 1, and mitogen-activated protein kinase 3) [213]. DEHP is widely used due to its suitable properties and low cost, and is present in the general environment at high levels. Exposure to DEHP is of particular concern with regard to developing fetuses where it is believed to cause malformation of reproductive organs and neurological defects [214]. Using our approach, several proteins were identified as being associated with DEHP (Table 4.2 on page 81). Cysteine dioxygenase type I (CDO1) and peroxisome proliferator-activated receptor alpha (PPARA), the two top scoring proteins, are already known in the CTD and from the literature [215, 216] as molecular targets for DEHP. Six other high ranking proteins are new potential DEHP molecular targets which are not recorded in the CTD (thus not input data). Among them, four gamma-aminobutyric acid A (GABA) receptors were predicted as potential DEHP molecular targets. These associations are supported by a recent study showing that DEHP can modulate the function of ion channels as GABA receptors in a manner similar to volatile anesthetics in experiments on expressed receptors [194]. This makes sense because the GABA neurotransmitter system has been implicated in the pathogenesis of bipolar disorders (neurological disorders) via gamma-aminobutyric acid receptor subunit alpha-1 (GABA $\alpha$ 1) [217], and DEHP is also associated with neurological defects [214]. In addition to GABA receptors, we identified several other candidates including proopiomelanocortin (POMC) and a cytochrome P450 (CYP3A11). We looked at another environmental chemical, the 2,3,7,8-TetraChloroDibenzo-p-Dioxin (TCDD), which originates from burning or incineration of chlorinated industrial compounds. TCDD is believed to cause a wide variety of pathological alterations, with the most severe being progressive anorexia and body weight loss [218]. TCDD is also known to be a neurotoxin leading to neurodevelopmental and neurobehavioral deficits [219, 220], and accumulating in the brain as well as other organs [221]. We identified six proteins associated with TCDD that are not recorded in the CTD for human (Table 4.2). Among them five are supported by literature (see Text S2). This included protein kinase C epsilon (PRKCE), known to be involved in brain tumors [222], carnitine palmitoyltransferase I (CPT1A), 11 $\beta$ -hydroxysteroid dehydrogenase type 1 (HSD11B1) and apolipoprotein B (APOB) which are all linked to obesity [223--225]. Furthermore, we investigated in detail the drug pirinixic acid (PA) (also named WY14,643), which is a peroxisome proliferator-activated receptor (PPAR) agonist with strong hypolipidemic effects. PA was never approved for clinical use due to hepatocarcinogenesis adverse effect shown in animal studies [226]. To date there is no evidence that PA promotes carcinogenesis in humans [227], and this has spurred new studies for identifying cellular processes that are capable of responding to PA. Among 11 molecular targets identified and not recorded in the CTD (Table 4.2), only five are supported by the literature (see Text S2). For example the expression of the C3 protein, an acylation stimulating protein involved in necrosis and afibrinogenemia (blood disorders), has been shown to be affected by PA in rats [228]. Finally we studied proteins associated with permethrin in more detail. Permethrin is a widely used insecticide, acaricide and insect repellent, classified by the US EPA as a likely human carcinogen, but still used in healthcare for the treatment of lice infestations and scabies. Four proteins not recorded in the CTD were identified as associated with permethrin. Three of them are supported by literature (see Text S2 for details) including a cytochrome P450 (CYP2B1) [229, 230] and sex hormone-binding globulin (SHBG) [231], which are proteins linked to the endocrine system. These findings suggest a mechanism by which chronic expo-

sure of humans to pesticides containing this compound may result in disturbances in endocrine effects related to androgen action.

The examples we provide include both known and new protein associations with a given chemical, and many of the novel associations are supported by the literature. We compared our approach with STRING (version STRING 11 [232]) a high-confidence protein-protein association network, to see if the findings generated by the current approach are also found by other existing methods. The STRING network includes direct (physical) and indirect (functional) associations derived from diverse sources as genomic context, high throughput experiments, co-expression and literature. As a test example, we used the 15 proteins associated with DEHP in the CTD to query the P-PAN by a neighbor protein procedure. The same 15 proteins were also used to query the STRING network. Subsequently we compared the predicted molecular targets between the two networks (P-PAN and STRING). In the resulting STRING network none of the GABA receptors were found (see Figure S3). The STRING network showed a clear tendency to associate phthalates with kinases and nuclear receptors. This example demonstrated that our approach was complementary to other association approaches. This highlights the value of integrating various sources of data to understand potential toxic effects on human health caused by chemical exposure.

## Discussion

We propose an approach different from existing computational chemical biology networks, which primarily integrate drugs information, to identify new molecular targets for chemicals and to link them to diseases. In our approach we have integrated toxicogenomics data for drugs and environmental compounds. The ability to make new findings using a different network is illustrated by a comparison with a similar method, showing the capacity of our P-PAN to identify novel chemical-protein associations. Using phthalate as an example, our model suggests potential associations between DEHP and GABA receptors, which have not been predicted previously.

An extension of this network by integrating more data, for example other chemical-protein associations or dose levels for which a compound may affect human health, would be beneficial to the proposed approach. Paracelsus (1493–1541) is often cited for his quote, “all things are poisons and nothing is without poison, only the dose permits something not to be poisonous”. This emphasize that the dose of a chemical is an issue to consider in the deregulation of systems biology. Nevertheless, a global mapping could allow a better understanding of adverse effects of drugs and toxic effects of environmental compounds. This could be used as a new approach for risk assessment and regulatory decision-making for human health.

Among the examples presented, some predictions are supported by literature for other organisms. Regarding toxicogenomics, the available human data are generally sparse compared to rodents. Data on toxicity --- adverse effects of chemicals on humans --- can be acquired through epidemiologic studies and from occupational, accident-related exposures as intentional human testing of environmental compounds remains limited. However, differences exist between model animal and human responses to chemicals, including differences in the type of adverse effects experienced and the dosages at which they occur. The differences may reflect variations in the underlying biochemical mechanisms, in metabolism, or in the distribution of the chemicals. As an example, bisphenol A (BPA) does not affect proteins in a similar way across species

<i>Chemical</i>	<i>Known protein</i>	<i>Cpscore*</i>	<i>Novel protein</i>	<i>Cpscore*</i>	<i>Literature</i>
DEHP	CDO1	13.23	<b>GABA<math>\beta</math>1</b>	5.46	Yes
	PPARA	9.48	<b>POMC</b>	5.44	Yes
	SUOX	4.35	<b>CYP3A11</b>	5.40	Yes
	(15 proteins)		<b>GABA<math>\beta</math>2</b>	4.32	Yes
			<b>GABA<math>\gamma</math>2</b>	4.32	Yes
			<b>GABA<math>\alpha</math>1</b>	4.26	Yes
TCDD	HSPA9B	82.69	<b>PRKCE</b>	10.17	Yes
	SLC2A4	82.69	<b>POMC</b>	8.97	Yes
	TRIP11	82.69	<b>CPT1A</b>	6.96	Yes
	TSP1	82.69	<b>HSD11B1</b>	6.39	Yes
	EPHX2	75.77	<b>MVP</b>	6.77	No
	MT2A	10.85	<b>APOB</b>	5.61	Yes
PA	(90 proteins)				
			<b>CHST1</b>	5.19	No
			<b>CHST4</b>	5.19	No
			<b>CST</b>	3.19	Yes
			<b>ABCG5</b>	2.61	No
	CYP4X1	5.67	<b>C3</b>	2.80	Yes
	PPARA	2.53	<b>ADRA2A</b>	1.34	Yes
	CES1	1.45	<b>CYB5A</b>	1.21	No
	SULT2A1	0.87	<b>ADRA1A</b>	1.08	Yes
	CYP1A1	0.37	<b>CRHR2</b>	1.04	No
	(5 proteins)		<b>CYP2A13</b>	0.93	No
			<b>ALDH3</b>	0.91	Yes
Permethrin	AR	4.67	<b>CYP2B1</b>	4.43	Yes
	WNT10B	4.12	<b>SHBG</b>	3.51	Yes
	PGR	3.75	<b>CYP2B6</b>	2.89	No
	ESR1	3.31	<b>NR1I3</b>	2.64	Yes
	TFF1	3.15			
	NR1I2	2.94			
	(17 proteins)				

**Table 4.2:** Predicting novel molecular targets for chemicals. \*Proteins known to be associated to a compound were extracted from the CTD. In brackets is the total number of known proteins used to query the P-PAN. To find novel protein targets (in bold) associated to a chemical, a neighbor proteins procedure was used which scored the association between proteins and chemicals (cpscore). Among the novel predicted proteins (thus not input data), some are supported by literature, highlighting the usefulness of the P-PAN to identify new chemical-protein associations.

(Figure 4.5). In the human systems studied to date, BPA does not affect the proto-oncogene *c-FOS* (FOS) and the mitogen-activated protein kinase 8 (MAKP8) but seems to modify their expression in rodent species. BPA binds and modifies the activity of the estrogen receptor alpha (ESR1) in a very conservative way across organisms [181]. BPA has an ability to function as an estrogen like receptor (ER) agonist, and thus has the potential to disrupt normal endocrine signaling through regulation of ER target genes e.g. androgen receptors, estrogen receptor, progesterone receptors. There is a need to integrate data with cross-species extrapolation in order to have a more accurate understanding of the human risk from chemical exposure.

The major limitation of our integrative systems biology approach is that the molecular target predictions are limited to the 3,528 proteins present in our P-PAN, which represent only 15% of the estimated human proteome [233]. Hence, the current lack of high quality data is the limiting factor in approaches such as the one described here. Today high throughput methodologies result in available large scale data in both chemical biology and systems biology, but these data are discipline specific [234]. There is an evident need for the development of databases [235] to integrate disparate datasets such as toxicogenomics data in order progress in systems biology research. In addition, the results of the disease-compound association analysis will improve in the future as newer, more complete and curated data will become available.

## Material and Methods

### Data set

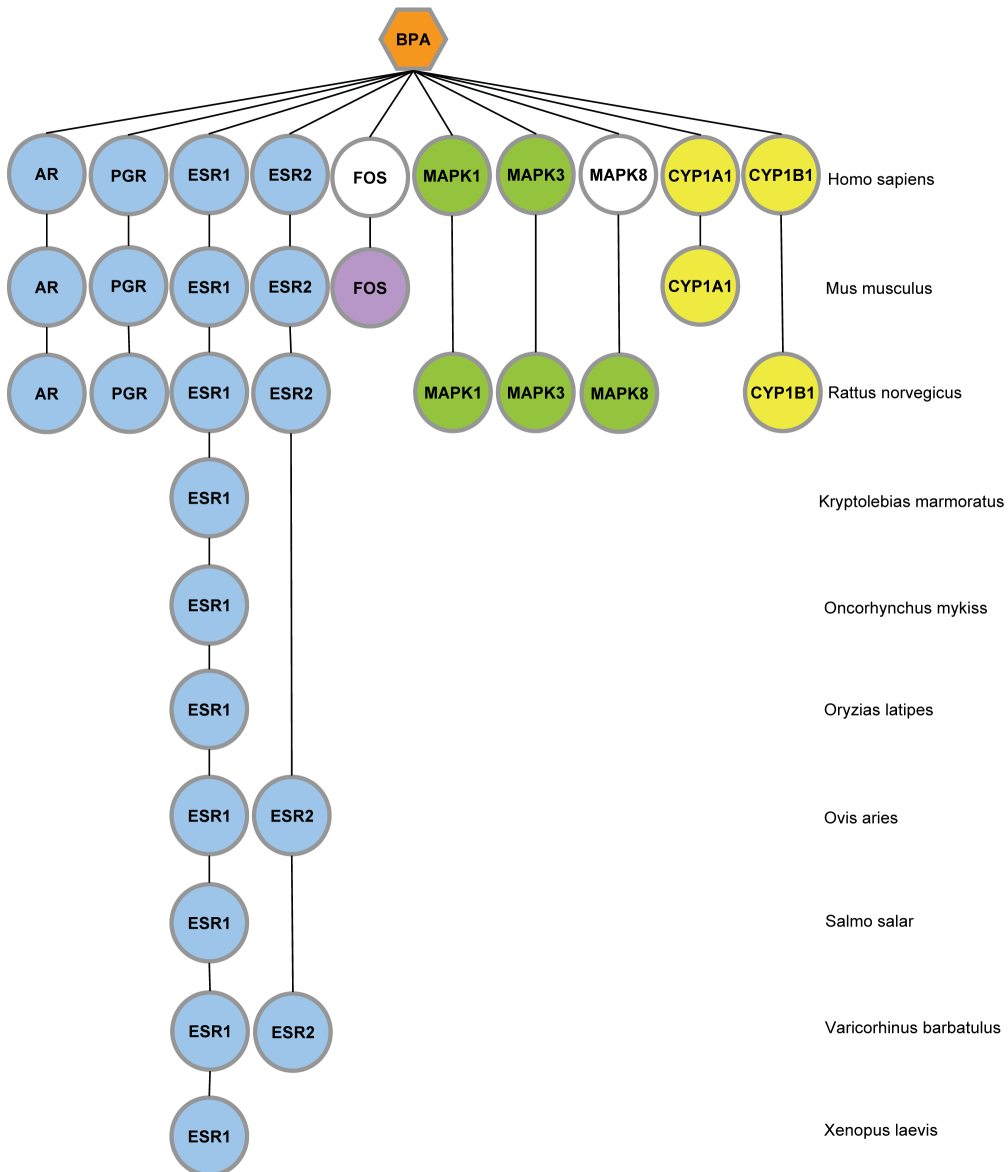
We downloaded the publicly available Comparative Toxicogenomics Database (CTD) as of June 26, 2008 [181]. The CTD contains curated information combining drug and environmental chemical data associated with proteins. We selected 42,194 associations between 2,490 unique compounds and 6,060 molecular targets known to be involved in human disease. Different associations are presented in the CTD such as ‘chemical x results in increased expression of protein z’ or ‘compound x binds to protein z’. Gene expression data are essentially present in the CTD such as a chemical can increase, decrease or affect a gene expression. However, only few binding data are present in CTD and therefore integrated in our network: 3189 in total among the 42,194 associations. Scripts were used to remove associations with negation such as ‘chemical x does not affect protein z’.

### Quality of chemical and protein annotations

To verify the uniqueness of chemicals, chemical names extracted from the CTD were checked using PubChem<sup>4</sup> to avoid synonymous names for the same compound. The few chemical names not retrieved via the database were manually verified. To determine overlaps with protein-protein interaction databases and facilitate further data integration, the CTD protein names were mapped to the corresponding Ensembl IDs [236] as of June 26, 2008. Only 1.5% of the 42,194 chemical-protein associations could not be clearly identified.

---

<sup>4</sup><http://pubchem.ncbi.nlm.nih.gov/>, as of June 26, 2008.



**Figure 4.5:** Cross-species comparative toxicogenomics for bisphenol A (BPA). Molecular targets are represented as nodes, and colored by gene family. Nodes presence represent available information extracted from the CTD and node absence are the unknown information. Colored nodes defined that BPA affect the protein, while nodes are not colored when BPA does not affect the protein. This figure highlights similarities and differences existing between animal model and human responses to chemical exposure.

### Structure-target relationship

To investigate chemical space of drugs and environmental compounds, 50 two-dimensional properties were calculated for each structure extracted from PubChem. To visualize them, principal component analysis (PCA) was performed. All necessary data were calculated using the MOE software<sup>5</sup>

### Generating a high confidence human Protein-Protein Associations Network

Relevant human chemical-protein associations collected from the CTD were used to create a P-PAN. The maximum number of molecular targets assigned to one compound 'tert-Butylhydroperoxide' was 1,189 and the maximum of chemicals assigned to one protein, the cytochrome P450 3A4 (CYP3A4), was 276. The P-PAN was generated by instantiating a node for each protein, and linking by an edge any protein-protein pair where at least one overlapping chemical was identified. Scripts were used to convert the protein-protein associations into a non-redundant list of associations. If proteins A and B are associated, the network may have two associations, A-B and B-A. Only one of these associations was retained in the P-PAN. We assigned two reliability scores to each protein-protein association: a score based on hypergeometric calculation and a weighted score. The weighted score was calculated as the sum of weights for overlapping compounds, where weights were inversely proportional to the number of assigned proteins. The resulting P-PAN is a complex structure containing a total of 2.44 million unique associations between 6,060 human proteins.

### Validating the protein-protein association score

The reliability of the weighted score was confirmed by fitting a calibration curve of different scores against Lage's PPIs [47] (version 2.9) and Vidal's PPIs [48]. Only 35,000 high confidence experimental interactions were extracted from Lage's PPI, which contains interactions present in the largest databases (Reactome, KEGG, ...) and data inferred from model organisms. Vidal's PPIs are based on an internal consistent single data source defined using yeast two-hybrid system and contains 3111 interactions. The overlaps of our P-PAN scores and Lage/Vidal PPIs are shown in Figure S2. The benchmark revealed that the weighted score is superior to a score calculated as the negative logarithm of p-values from a test in hypergeometric distribution and a simple overlap count. To estimate the robustness of the model, four thresholds selected from the 'weighted score' curves (5%, 8%, 12.5% and 17%) of the complete P-PAN were used to perform prediction for DEHP. At 5%, 73,000 associations between 2105 proteins were extracted. The number of proteins is relatively stable at 8% and 12.5%. However, the number of associations increased significantly from 200,080 to 306,000 including lower score associations in the output file of prediction. The threshold of 17% corresponds to 415,000 associations between 3894 proteins. All thresholds showed a good prediction with the GABA receptors for DEHP. As the 12% threshold already added some more noise in the prediction, we decided to not include more proteins, in order to keep the most significant associations. We then considered a threshold of 8%, represented by the vertical line in Figure S2, which captured a good overlap

<sup>5</sup>Chemical Computing Group version 2007.09.



between our P-PAN and the PPI networks. This selection represents 200,080 associations of the complete P-PAN.

Among the 200,000 high confidence associations selected, 3,528 proteins were identified, and these were significantly enriched among the high scoring protein-protein associations as shown in Figure S2 (861 Lage's PPI interactions corresponding to 24.4% were found among the top 5% of the high scoring protein-protein associations). By comparison, only 1,852 of the high confident interactions from Lage were identified in a random P-PAN created by node permutation, and no enrichment was seen for the random network. As example, the selection of high confidence associations allowed to conserve only 803 proteins from the 1189 proteins assigned to the 'tert-Butylhydroperoxide'.

### P-PAN clustering

A high confidence sub-network of 200,000 protein-protein associations was selected which contained 3,528 proteins. This sub-network was highly interconnected, with the majority of proteins belonging to a single large cluster. In order to increase the resolution and facilitate biological interpretation, two clustering methods were applied to the sub-network, MCODE [199] and MCL [200]. We used the default settings for MCODE (fluff option set to 0.1, mode score cutoff set to 0.2, degree cutoff set to 2), and obtained 35 clusters. One major drawback of this algorithm is that not all the proteins in the network were clustered. We used the MCL algorithm with scheme and granularity parameters set to 7 for highest performance and granularity. With the MCL approach we identified a total of 58 clusters as strongly interconnected, with a minimum size of 5 proteins. These clusters were linked together into a new network consisting of a scored cluster-cluster association network. The association score between each cluster pair was calculated from the mean of the P-PAN between each pair of clusters. Each cluster was investigated for functional analysis based on the three Gene Ontology categories (a) molecular function, (b) biological processes, and (c) cellular components as of January 2009. To reduce the noise and improve the quality of the functional annotation, we only used the functional annotation if it was experimentally supported or had traceable references. The following GO evidence codes were allowed: IMP (Inferred from Mutant Phenotype), IGI (Interfered from Genetic Interaction), IPI (Inferred from Physical Interactions) and IDA (Inferred from Direct Assay) and TAS (Traceable Author Statement). At time of use the molecular function category contained 5,981 proteins, the biological processes category 5,196 proteins, and the cellular components 5,151 proteins. We compared human proteins present in GO categories with proteins extracted from the CTD; 14.3% of the CTD proteins could not be annotated for the molecular function, 16.6% for biological processes and 14.9 % for cellular components.

To identify chemicals associated with disease, protein-specific information such as involvement in disease was integrated in each cluster. The Online Mendelian Inheritance in Man database (OMIM)<sup>6</sup> and the GeneCards database<sup>7</sup> were considered as sources of protein-disease connections. Various clusters were investigated. For example, cluster 1 contained 462 proteins. Using GeneCards, 269 proteins were retrieved with disease annotations. Amongst these 269

<sup>6</sup><http://www.ncbi.nlm.nih.gov/omim/>, July, 2009.

<sup>7</sup>February, 2008.

proteins, 128 were associated to breast cancer (with give a p-value of  $9.67\text{e-}18$  for breast cancer to cluster 1). Using OMIM, only 90 proteins among the 462 were retrieved with disease annotations. Looking at the cluster enrichment with OMIM, we obtained at the top a non significant p-value of 0.0048 (corresponding to two proteins for paget disease of bone). As another example, we analyzed the second cluster. Cluster 2 contained 433 proteins. 281 proteins were annotated to diseases in Genecards, for only 78 proteins in OMIM. Additionally, cluster 2 has a significant p-value of  $2.26\text{e-}12$  using GeneCards information for necrosis. According to these results we decided to use GeneCards as a source of protein-disease relationships. To avoid too many false positive from Genecards, we set a significance cut-off value of the GeneCards-AKS2 score based on a comparison with OMIM. This was done by overlapping common protein-disease associations from Genecards against OMIM (see Figure S4). The protein-disease connections were kept with a minimum AKS2 score of 60 and p-values were calculated for each disease present in clusters. Then, chemical information from the CTD was integrated with each cluster and p-values were assigned to each chemical. All p-values obtained were calculated using hypergeometric testing, and were corrected for multiple testing with Bonferroni correction. The significance cutoff for the corrected p-values was set to 0.05.

### **Neighbor protein procedure**

To predict molecular targets for a chemical, a network-neighbor's pull down was done in a three steps procedure: (1) Selection of the input protein(s): Extraction of the protein(s) known to be associated with the selected chemical from the CTD. (2) Identification of network(s) surrounding the input proteins by a neighbor proteins procedure. In this procedure, our P-PAN was queried for the input proteins, and associations between these were added. Next, the first order interactors of all the input proteins were queried and added. For each neighbor, a score was calculated taking into account the topology of the surrounding network, based on the ratio between total associations and associations with input proteins. Molecular targets with a score higher than the threshold (0.1) were kept in the final sub-network(s). This node inclusion parameter is in the conservative end of the optimal range for protein-protein interaction networks [47]. As a final step all proteins in the complex were checked for associations among them and the missing one were added. (3) Establishment of a confidence score for the surrounding network (cscore) and of a score for each protein (cpscore): Each of the pulled down complexes was tested for enrichment on our input set by comparing them against 1.0e4 random complexes for the protein-protein association set to establish a cscore for each sub-network and a cpscore for each connected proteins. The cpscore was used to rank proteins to select potential molecular targets for chemicals. An illustration of cpscore is available on Table S2 for approved drugs.

### **Postscript**

All the CTD human protein-chemical associations were extracted from the CTD on June 26, 2008. Subsequent updates of CTD, as of June 25, 2009, did not change the overall trends or conclusions of the present study.

## **Acknowledgements**

The authors would like to thank Daniel Edsgård for his technical help and Ramneek Gupta for critical reading of the manuscript.

## 4.3 Paper VI

# ChemProt: A Disease Chemical Biology Database

Olivier Taboureau<sup>1+\*</sup>, Sonny Kim Nielsen<sup>1+</sup>, Karine Audouze<sup>1</sup>, Nils Weinhold<sup>1</sup>, Daniel Edsgård<sup>1</sup>, Francisco S. Roque<sup>1</sup>, Irene Kouskoumvekaki<sup>1</sup>, Alina Bora<sup>2</sup>, Ramona Curpan<sup>2</sup>, Thomas Skøt Jensen<sup>1</sup>, Søren Brunak<sup>1</sup> and Tudor Oprea<sup>1,3\*</sup>

<sup>1</sup>Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark.

<sup>2</sup>Institute of Chemistry, Romanian Academy, Department of Computational Chemistry, Timisoara, Romania

<sup>3</sup>Division of Biocomputing School of Medicine, University of New Mexico, Albuquerque, NM, USA

<sup>+</sup> OT and SKN contributed equally to this work.

\*To whom correspondence should be addressed. Tel: +45 4525 2489; Fax: +45 4593 1585; Email: otab@cbs.dtu.dk (OT) or tuop@cbs.dtu.dk (TIO)

## Abstract

Systems pharmacology is an emergent area that studies drug action across multiple scales of complexity, from molecular and cellular to tissue and organism levels. There is a critical need to develop network-based approaches to integrate the growing body of chemical biology knowledge with network biology, and to understand the relationship between drug action and genetic susceptibility to disease. Here, we report ChemProt, a Disease Chemical Biology database, which is based on a compilation of multiple chemical-protein annotation resources, as well as disease-associated protein - protein interactions (PPI). We assembled more than 700,000 unique chemicals with biological annotation for 30,578 proteins. We gathered over two million chemical-protein interactions, which were integrated in a quality scored human PPI network of 428,429 interactions. The PPI network layer allows for studying disease and tissue specificity through each protein complex. ChemProt can assist in the *in silico* evaluation of environmental chemicals, natural products and approved drugs, as well as the selection of new compounds based on their activity profile against most known biological targets, including those related to adverse drug events. Results from the disease chemical biology database associate citalopram, an antidepressant, with osteogenesis imperfect and leukemia, and bisphenol A, an endocrine disruptor, with certain types of cancer, respectively. The server can be accessed at <http://www.cbs.dtu.dk/services/ChemProt-1.0/>.

## Introduction

The old drug design paradigm, i.e., drugs interact selectively with one or two targets (genes), resulting in treatment and prevention of disease, is now challenged by several studies that show most drugs interacting with multiple targets ('polypharmacology') [183, 187]. For example, celecoxib, often considered a selective cyclooxygenase-2 non-steroidal anti-inflammatory drug (NSAID), has been documented to be active on at least two additional targets, namely carbonic anhydrase II and 5-lipoxygenase [237]. Rosiglitazone, which has been used for the treatment of type II diabetes mellitus, not only stimulates the peroxisome proliferator activated receptor gamma, but also blocks interferon gamma-induced chemokine expression in Graves disease or ophthalmopathy [238]. Polypharmacology is not always beneficial, as it often causes side-effects: Cisapride, which acts as a serotonergic 5-HT<sub>4</sub> receptor agonist, as well as astemizole, which blocks histamine H<sub>1</sub> receptors (H<sub>1</sub>R), have both been withdrawn from all markets due to the risk of fatal cardiac arrhythmia associated with their blockade of the hERG potassium ion channel, an 'anti-target' associated to QT prolongation and 'torsades de pointes' [239]. However, 'target' and 'anti-targets' are dynamic attributes, as exemplified by the case of H<sub>1</sub>R antagonists and their (in)ability to achieve clinically significant levels in the brain, influenced by the ATP-binding cassette transporter ABCB1 (also known as P-glycoprotein), which effluxes some of these drugs from the brain [240]. Acquiring knowledge of the complete pharmacology profile has inspired new strategies to predict and to characterize drug-target associations in order to improve the success rates of current drug discovery paradigms, i.e. increase the efficacy and reduce toxicity and adverse effects [187].

As large-scale chemical bioactivity databases are being assembled, the polypharmacology and promiscuity (i.e., low affinity bioactivity across multiple gene families) of chemicals are profiled like pieces of a puzzle [241]. These studies are often focused on specific protein families, such

as G-protein coupled receptors [242], nuclear receptors [243] and kinases [244], but global pharmacology profiles of chemicals are considered as well [183, 187]. Recent chemoinformatics advances support the development of polypharmacology data mining, e.g., via iPHACE, an integrative web-based tool that enables pharmacological space navigation for small molecule drugs [245]. Biological information can also be retrieved for a large set of chemical compounds through PubChem [246], ChEBI and ChEMBL [247].

Two conceptual developments support polypharmacology: systems pharmacology, aimed at drug actions in the context of regulatory networks [248] and systems chemical biology [249], which introduces chemical awareness in systems biology. Since proteins rarely operate in isolation inside and outside cells, but rather function in highly interconnected cellular pathways, interactome networks have been developed by data integration. *Yildirim et al.* [188] combined FDA-approved drugs with a human protein-protein interaction (PPI) network (human interactome) in order to analyze the interrelationships between drug targets and disease-gene products. A similar work has been based on PubChem bioassay as source of polypharmacology [250]. The use of side-effect similarity has been proposed on the assumption that drugs with similar side-effects are likely to interact with similar target proteins [251]. Recent advances include a protein-protein association network based on the chemical toxicology of environmental chemicals [252] and a human disease network linking disorders and disease genes to various known phenotypes [46].

Our goal in the present work was to develop a disease chemical biology server, called ChemProt, based on the integration of chemical-protein annotation resources that are now accessible from large repositories, and curated disease-linked protein-protein interaction (PPI) data [47]. ChemProt is designed to assist the elucidation of drug actions in the context of cellular and disease networks. Further to that, it allows the identification of additional genes that may play major roles in modulating chemical response i.e. to drugs, environmental chemicals and natural products, thus leading to new options in drug discovery and environmental chemical evaluation. Lastly, the ChemProt server could contribute to drug repurposing as well as to the investigation of chemicals related to anti-targets and adverse drug events.

## Implementation

### Data sources

We first gathered chemical-protein interaction data from different open source databases i.e. ChEMBL (version chembl\_05) [247], BindingDB [253], PDSP Ki Database [254], DrugBank (version 2.5) [255], PharmGKB [256] and two commercial databases, WOMBAT (version 2009) and WOMBAT-PK (version 2008) [241]. Active compounds from the PubChem bioassay (2010) have been collected as well [246]. We considered only active compounds from ‘confirmatory’ assays in order to capture high confidence chemical-protein annotations from PubChem. Drug-target information was collected from DrugBank and PharmGKB. In addition, we integrated chemical-protein associations from CTD (version 2009) [181] and STITCH (version STITCH 2.0) [257]. These last two databases consider the effect or modulation (positive or negative) of a chemical on proteins, other than that defined as binding activity. Examples include gene expression or pathway data, where the deregulation of a gene by a chemical may be

not due to a physical interaction between the two entities but a response at a cellular level. Duplicate chemicals were found by using InChI keys and were merged into a single ChemProt ID. Overall, the final database contains 700,000 distinct molecules annotated for 30,578 proteins.

### Descriptors and similarity measurement

The chemical structure of the molecules was encoded using two rather different types of fingerprints. The 166 MACCS keys, encode the presence or absence of predefined substructural or functional groups [258]. On the other hand, a more complex 3-point pharmacophore fingerprint (GpiDAPH<sub>3</sub>) is based on an expansion of the PATTY pharmacophore feature recognition scheme of a 2D structure [259]. This scheme assigns one or more pharmacophore feature types to all atoms in a molecule using a predefined list of SMART queries. The list of pharmacophore feature types comprises: hydrogen-bond donor (D), hydrogen-bond acceptor (A), polar (P) and hydrophobic (H). In addition, an extra label (p or pi) is added to each feature if the originating atom or group is sp<sup>2</sup>-hybridized or planar for other reasons. The GpiDAPH<sub>3</sub> pharmacophore feature scheme is expressed in 2D as triplet feature combinations with a graph based inter-atom distance binning scheme. Both fingerprints are implemented in the Molecular Operating Environment (MOE - version 2008.10) [260]. The similarity between two molecules is measured using the Tanimoto coefficient (Tc), a method of choice for the computation of fingerprint-based similarity [261]. The Tanimoto coefficient is defined as the number of bits in common divided by the total number of used bits in both molecules. For any pair of chemicals, Tc assumes values between 0 and 1. A high Tanimoto coefficient represents high similarity.

### Protein-protein interaction network (PPI)

The human interactome used is an in-house protein-protein interaction network inferred from experiments in both humans and model organisms [47]. Using an elaborate scoring scheme, all interactions have been validated against a gold standard [48]. The current interactome contains 428,429 unique protein-proteins interactions derived from source databases such as BIND [90], GRID [91], MINT [262], dip\_full [263], HPRD [264], intact [265], mppi [266], MPact [267], Reactome [268] and KEGG [98]. Data are transferred between organisms by using the Inparanoid orthology database [160]. In total the human interactome comprises 22,997 genes.

Human disease genes and complexes. Based on a previous study [65], disease-associated protein complexes was associated to the chemical-protein annotation. By mining OMIM [63] and GeneCards [113], two data resources for genes association to diseases, we collected a list of 2,227 unique disease-related proteins and mapped the complexes of genes to disease. Similarly, complexes of genes were mapped to Gene Ontology (GO) terms [269] and tissues by using the expression data from 73 non-disease tissues from the Novartis Research Foundation Gene Expression Database (GNF) [115] and Human Protein Atlas [270]. Users of ChemProt can thus retrieve gene complexes that are related to a query chemical and visualize the annotations of each complex.

## Applications

### Query submission

There are several query options for the user to access the data in the ChemProt database. We have integrated a list of common names of chemicals that can be used as queries. A list of chemicals can also be retrieved by protein name using the 'search by Target' field. Another option is to perform a search by entering a chemical structure. The user can type a SMILES string or draw the structure in the JME Molecular Editor applet that is part of ChemProt's web interface. A search by chemical structure will provide an extended list of similar chemicals that are stored in the database. The user can choose the metric to be used for similarity search: MACCS for searching by substructure or functional group or Gpi-DAPH<sub>3</sub> for pharmacophore-based searches. All source databases of ChemProt (or a combination thereof) can be searched separately. The number of queries to WOMBAT and WOMBAT-PK has been restricted to 3 per user per day, since both databases are commercial.

### Annotation and prediction of small molecule bioactivity

A query will return several types of information. The query itself can be depicted by dragging the mouse to the position 'Results for this compound', where a 2D structure of the query is shown. The results are displayed in tabular format when multiple chemical-protein pairs are retrieved. The table is divided into two main categories; 'Annotated Compound', which includes information related to the query, and 'Similar Compound(s)', which contains chemicals that bear similarity to the input structure according to the chosen metric and cut-off. In each category, the results are grouped by species, including 'Human', 'Rat', 'Mouse', 'Bovine', 'Pig' and 'other species'.

Moving the mouse pointer to the 'compound ID', the 2D structure of the molecule is shown together with some computed physico-chemical properties such as Molecular Weight, LogP, number of hydrogen bond donors, number of hydrogen bond acceptors, number of rigid bonds and number of rings. Thanks to the Marvin java applet from Chemaxon for the depiction of the molecule [271].

The 'Type' column presents information on bioactivity, which can either be a biological end-point measurement (e.g., IC<sub>50</sub>, Ki), or a score between 1 and 1000, as defined by the STITCH system (the closer to 1000, the higher is the confidence of a true chemical-protein association). CTD, DrugBank, PharmGKB and PubChem do not have explicit bioactivity values, so this field is left blank for these databases. The 'Type' field is directly linked to the 'Value' field, which contains the numeric value for the specified bioactivity, and the 'Pharmacological effect' that describes the relation (substrate, inhibitor, agonist, antagonist) between the chemical and the protein.

The 'Target name' is the common name of the targeted protein. It is linked to the 'UniProt ID' field and 'Ensembl ID' where the user can get more information on the sequence and function of the protein. The 'TC' field contains the Tanimoto Coefficient, which is a value that assesses the chemical structure similarity of the query with other chemicals in ChemProt. The TC value can vary between 0 and 1. As previously reported (50), a TC of at least 0.85 can be used as a reasonable cut-off to indicate that two chemicals may share similar bioactivity. We



implement the same threshold of 0.85, when MACCS fingerprints are used as similarity metric, and compounds exceeding this threshold are listed under the field ‘Similar Compounds’. A threshold of 0.60 is used as default with GpiDAPH3 fingerprints, since the set of similar compounds captured using the threshold of 0.85 was too restrictive.

Finally, if a protein that is found to be deregulated by a chemical has been identified to be involved in protein-protein interaction networks in human, this protein is linked to the ‘Disease complexes’ server, where diverse biological and disease information is provided, as explained in the next section.

**1 SUBMISSION**

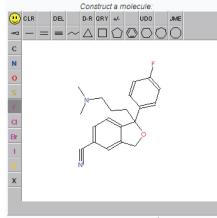
Type a compound name:

Choose fingerprint: ☐ MACC ☐ PH4

Search by Target

Type a target name:

Paste or import molecule in **SMI** format:



**2 Annotated Compound**

Human

Compound ID	Type	Value	Target Name	Species	Pharmacol. effect	UniProt ID	Ensembl ID	TC	Database	Diseases complexes
114367	Score	421	dopamine receptor D4	Homo sapiens	NULL		ENSG00000069696	1	Stitch	Diseases
114367	Score	408	Cytochrome P450, family 2 subfamily D, polypeptide 7	Homo sapiens	NULL		ENSP00000346625	1	Stitch	NA
114367	NA	NA	NA	Homo sapiens	NA		ENSG0000100197	1	CTD	Diseases

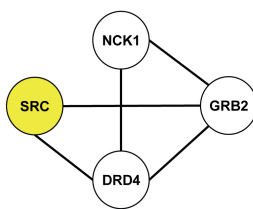
**3 Complex 145 from DRD4 (ENSG00000069696)**

Size: 4  
BioAlma Terms: 33

BioAlma

GO biological process

HPA



OMIM

GO cellular component

mRNA expression

**Figure 4.6:** Chemical-protein annotation and disease associations retrieved from ChemProt for the compound citalopram. 1) The compound can be queried using different formats (name, SMILES, and structure). 2) A query results in a table showing protein annotations and bioactivity predictions for the compound. 3) Finally, a protein-protein interaction network (protein-complex) for a target protein can be depicted and disease associations (OMIM and BioAlma) and other biological components (GO terms, HPA and mRNA expression) are displayed.

### From chemical-protein interactions to complex protein-disease associations

The unique feature of ChemProt is that it offers the user the possibility to get information at a cellular level, by linking chemically-induced biological perturbations to specific tissues and phenotypes.

Proteins that are both affected by a chemical and participate in one or more protein complexes are highlighted in the results table of the ChemProt server. By clicking on the protein, the user is redirected to the ‘Disease complexes’ server and has to choose which complex to visualize. On the ‘Disease complexes’ server, size and illustrations of the protein network are

provided. Additionally, enrichment analysis results of the proteins in the complex are shown, with respect to disease association (OMIM, BioAlma), Gene Ontology (GO) terms (biological process, cellular component) and tissue specificity (Human Protein Atlas, GNF). The table presenting the OMIM enrichment results is interactively linked with an illustration of the protein complex where proteins associated with the selected disease are colored yellow.

To get further information on chemicals and genes, we recommend the free service ‘Reflect’, developed by *Pafilis et al.* [272]. Reflect tags gene, protein and small molecule names in text and offers the opportunity to quickly view additional information on the ChemProt results, including synonyms, protein sequences, domains, three-dimensional structures and subcellular location.

## Examples

With the integration of several databases, ChemProt not only provides pharmacological information, but also includes biological data associated to environmental chemicals and natural products. As seen in the examples below, ChemProt can be queried for drugs as well as environmental chemicals. A search for citalopram, an antidepressant, illustrates the complementarity of the integrated databases within ChemProt (Figure 4.6 on the preceding page). Marketed as a selective serotonin reuptake inhibitor (SSRI) (DrugBank), this drug displays bioactivity on seven human proteins (ChEMBL). Via ChemProt, four other proteins (DRD<sub>3</sub>, 5HT<sub>1B</sub>, 5HT<sub>3</sub>, ADRA<sub>2A</sub>) are retrieved from the Ki database. Additional information on drug-target associations is provided by STITCH and CTD. From the first annotation to the D<sub>4</sub> dopamine receptor (DRD<sub>4</sub>), the disease term (under Disease Complexes) is highlighted, indicating that protein-protein interaction information for this protein is available. Using the link to the Disease Complexes server, one finds that DRD<sub>4</sub> interacts with 3 proteins (SRC, GRB2 and NCK1). According to OMIM, this protein network is associated to osteogenesis imperfecta and leukemia and, according to BioAlma, to several psychotic disorders. Gene Ontology (GO) enrichment indicates significant association of the protein complex to signal complex formation and vesicle membrane. Furthermore, tissue annotation suggests that this complex is mainly expressed in follicle and non-follicle cells (HPA) and dendritic cells (GNF). Although it might be surprising to see a connection between antidepressant and leukemia, it has been shown recently that antidepressants such as chlomipramine and fluoxetine reduce the growth of B-cell malignancies in leukemia [273].

The second query, ‘bisphenol A’ (BPA), is an environmental pollutant used as plasticizer [274]. BPA has biological activity on the estrogen receptor alpha (ESR1), the androgen receptor (AR) and the estrogen related receptor gamma (ERR3). However, several other proteins are retrieved from CTD and STITCH based on association data with this chemical. Looking at ESR1 in the Disease Complexes server, a complex of 17 proteins is depicted (complex 265) with significant associations to Li-FRAUMENI syndrome, breast cancer and neoplasms. Enrichment analysis indicates that the complex is found in the nucleus (GO Cellular component), involved in the regulation of metabolic processes and transcriptionally regulated by the RNA polymerase II promoter (GO Biological process). Furthermore, data from immunohistochemistry studies suggest that the complex is mainly located in the endometrium and the cerebral cortex (HPA). The Disease Chemical Biology network for BPA indicates that, under certain conditions, this chemical may be associated with certain types of cancers.

We have illustrated that ChemProt integrates molecular, cellular and phenotypic data associated to small molecules, which can lead to novel links and suggest new avenues for research. We envisage that the ChemProt server will find applications within a variety of chemogenomics, polypharmacology and systems chemical biology studies. ChemProt will be updated once a year with new compounds, new interactions and more sophisticated descriptors.

## **Funding**

This work has been supported by EU (DEER), the Innovative Medicines Initiative Joint Undertaking (eTOX), the Danish Research Council for Technology and Production Sciences, the Lundbeck foundation and the Villum Rasmussen Foundation. Sunset Molecular Discovery LLC ([www.sunsetmolecular.com](http://www.sunsetmolecular.com)) contributed with the WOMBAT databases.



## Epilogue

Looking back and examining all my work conducted as part of this PhD, I can't help but experiencing a feeling of achievement. Data integration in biology is far from being a trivial task. Putting together the different pieces of the puzzle that is human life is a painstaking endeavor that can only be achieved by interconnecting the different fields of science.

The work presented here focused on the *systems biology*, and is the result of close collaboration between several researchers. The holistic approach to disease systems biology enabled us to perform analyses that would be otherwise impossible.

The main goal for the thesis was to close the gap between genotype and phenotype. Taking disease descriptions and symptoms from medical databases, we combined the existing knowledge on the molecular level, to enhance our understanding of the disease etiology.

The methods used for extracting phenotypic descriptions from medical text, presented in Chapter 2, are generally applied to any EHR system. Future work could be the extension of the dictionary used for the extraction, by including terms from more terminologies (e.g., Snomed-CT). The patient stratification procedure shown in Paper I could be integrated with the patients' genetic material, in order to further distinguish patterns in the data.

The disease gene discovery procedures, pictured in Chapter 3, have been successfully used to discover novel proteins related to a disease. By combining different data types we achieve tissue and time resolution for elucidating the hidden causes of diseases. These approaches are extremely flexible and have been used in most of the analyses in this thesis.

Finally, Chapter 4, discusses approaches for discovering external causes of diseases, by combining small molecule information with genetic data.

Overall, the different Chapters of this thesis combine and make use of the available data in biology, achieving a much greater level of resolution than if focusing on them individually. With this I describe a generic framework for exploring the systems biology driving human disease, and paves way for discovering disease causality.



# Bibliography

- [1] Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881--8 (2008). URL <http://www.sciencemag.org/cgi/content/abstract/322/5903/881>.
- [2] Schadt, E. E. Molecular networks as sensors and drivers of common human diseases. *Nature* **461**, 218--23 (2009).
- [3] URL [http://www.nlm.nih.gov/databases/databases\\_medline.html](http://www.nlm.nih.gov/databases/databases_medline.html).
- [4] URL [http://www.nlm.nih.gov/bsd/medline\\_growth.html](http://www.nlm.nih.gov/bsd/medline_growth.html).
- [5] Iso/dtr 20514, health informatics - electronic health record - definition, scope, and context (2004).
- [6] Bleich, H. L. *et al.* Clinical computing in a teaching hospital. *N Engl J Med* **312**, 756--64 (1985).
- [7] Lium, J.-T., Tjora, A. & Faxvaag, A. No paper, but the same routines: a qualitative exploration of experiences in two norwegian hospitals deprived of the paper based medical record. *BMC Med Inform Decis Mak* **8**, 2 (2008). URL <http://www.biomedcentral.com/1472-6947/8/2>.
- [8] Ludwick, D. A. & Doucette, J. Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. *Int J Med Inform* **78**, 22--31 (2009). URL [http://www.ijmijournal.com/article/S1386-5056\(08\)00092-0/abstract](http://www.ijmijournal.com/article/S1386-5056(08)00092-0/abstract).
- [9] Garde, S., Knap, P., Hovenga, E. & Heard, S. Towards semantic interoperability for electronic health records. *Methods Inf Med* **46**, 332--43 (2007).
- [10] Kalra, D. Electronic health record standards. *Yearb Med Inform* 136--44 (2006).
- [11] Blobel, B. G. M. E., Engel, K. & Pharow, P. Semantic interoperability--hl7 version 3 compared to advanced architecture standards. *Methods Inf Med* **45**, 343--53 (2006).
- [12] Bernstein, K., Bruun-Rasmussen, M., Vingtoft, S., Andersen, S. & Nøhr, C. Modelling and implementing electronic health records in denmark. *Int J Med Inform* **74**, 213--220 (2005).
- [13] Häyrynen, K., Saranto, K. & Nykänen, P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform* **77**, 291--304 (2008). URL [http://www.ijmijournal.com/article/S1386-5056\(07\)00168-2/abstract](http://www.ijmijournal.com/article/S1386-5056(07)00168-2/abstract).
- [14] Bischof, F. & Pfeiffer, E. F. Experiences with a computerized diabetes information system. *Horm Metab Res Suppl* **26**, 146--8 (1992).
- [15] Safran, C. *et al.* Toward a national framework for the secondary use of health data: an american medical informatics association white paper. *J Am Med Inform Assoc* **14**, 1--9 (2007).

- [16] Prokosch, H. U. & Ganslandt, T. Perspectives for medical informatics. reusing the electronic medical record for clinical research. *Methods Inf Med* **48**, 38--44 (2009). URL <http://www.schattauer.de/en/magazine/subject-areas/journals-a-z/methods/contents/archive/issue/661/manuscript/10821.html>.
- [17] Ebidia, A., Mulder, C., Tripp, B. & Morgan, M. W. Getting data out of the electronic patient record: critical steps in building a data warehouse for decision support. *Proc AMIA Symp* 745--9 (1999).
- [18] Murphy, S. N., Mendis, M. E., Berkowitz, D. A., Kohane, I. & Chueh, H. C. Integration of clinical and genetic data in the i2b2 architecture. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium* 1040 (2006).
- [19] Loukides, G., Gkoulalas-Divanis, A. & Malin, B. Anonymization of electronic medical records for validating genome-wide association studies. *Proc Natl Acad Sci USA* **107**, 7898--903 (2010). URL <http://www.pnas.org/content/107/17/7898.long>.
- [20] Iacono, L. L. Multi-centric universal pseudonymisation for secondary use of the ehr. *Stud Health Technol Inform* **126**, 239--47 (2007).
- [21] Liu, H., Lussier, Y. A. & Friedman, C. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *J Biomed Inform* **34**, 249--61 (2001).
- [22] URL <http://www.google.com>.
- [23] URL <http://www.ncbi.nlm.nih.gov/pubmed/>.
- [24] Swanson, D. R. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* **30**, 7--18 (1986).
- [25] Cohen, A. M. & Hersh, W. R. A survey of current work in biomedical text mining. *Brief Bioinformatics* **6**, 57--71 (2005).
- [26] Jensen, L. J., Saric, J. & Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews genetics* **7**, 119--29 (2006). URL <http://www.nature.com/nrg/journal/v7/n2/abs/nrg1768.html>.
- [27] Ananiadou, S., Kell, D. B. & ichi Tsujii, J. Text mining and its potential applications in systems biology. *Trends Biotechnol* **24**, 571--9 (2006). URL [http://www.sciencedirect.com/science?\\_ob=ArticleURL&udi=B6TCW-4M3J0HP-2&user=10&coverDate=12%252F31%252F2006&rdoc=1&fmt=high&orig=search&sort=d&docanchor=&view=c&acct=C000050221&version=1&urlVersion=0&userid=10&md5=d56ffa72b432c3d07f7dd4b9541141c2](http://www.sciencedirect.com/science?_ob=ArticleURL&udi=B6TCW-4M3J0HP-2&user=10&coverDate=12%252F31%252F2006&rdoc=1&fmt=high&orig=search&sort=d&docanchor=&view=c&acct=C000050221&version=1&urlVersion=0&userid=10&md5=d56ffa72b432c3d07f7dd4b9541141c2).
- [28] Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C. & Hurdle, J. F. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 128--44 (2008). URL [http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list\\_uids=18660887&dopt=abstractplus](http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=18660887&dopt=abstractplus).
- [29] DesRoches, C. M. *et al.* Electronic health records in ambulatory care--a national survey of physicians. *N Engl J Med* **359**, 50--60 (2008).
- [30] Greenhalgh, T. *et al.* Adoption and non-adoption of a shared electronic summary



- record in england: a mixed-method case study. *BMJ* **340**, c3111 (2010).
- [31] Hoffman, S. Electronic health records and research: privacy versus scientific priorities. *Am J Bioeth* **10**, 19--20 (2010).
- [32] Kush, R. D. What the patient should order. *Sci Transl Med* **1**, 3cm3 (2009).
- [33] Scheuermann, R. H. & Milgrom, H. Personalized care, comparative effectiveness research and the electronic health record. *Curr Opin Allergy Clin Immunol* **10**, 168--70 (2010).
- [34] Tremblay, M. C., Berndt, D. J., Luther, S. L., Foulis, P. R. & French, D. D. Identifying fall-related injuries: Text mining the electronic medical record. *Inf Technol Manag* **10**, 253--265 (2009).
- [35] Walsh, S. H. The clinician's perspective on electronic health records and how they can affect patient care. *BMJ* **328**, 1184--7 (2004). URL <http://www.bmj.com/content/328/7449/1184.long>.
- [36] Park, J., Lee, D.-S., Christakis, N. A. & Barabási, A.-L. The impact of cellular networks on disease comorbidity. *Mol Syst Biol* **5**, 262 (2009). URL <http://www.nature.com/msb/journal/v5/n1/full/msb200916.html>.
- [37] Prather, J. C. *et al.* Medical data mining: knowledge discovery in a clinical data warehouse. *Proc AMIA Annu Fall Symp* 101--5 (1997). URL [http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list\\_uids=9357597&dopt=abstractplus](http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=9357597&dopt=abstractplus).
- [38] Embi, P. J., Jain, A., Clark, J. & Harris, C. M. Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. *AMIA Annual Symposium proceedings / AMIA Symposium* 231--5 (2005).
- [39] Liao, K. P. *et al.* Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res (Hoboken)* **62**, 1120--7 (2010).
- [40] Long, W. Extracting diagnoses from discharge summaries. *AMIA Annual Symposium proceedings / AMIA Symposium* 470--4 (2005).
- [41] Meystre, S. & Haug, P. J. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *J Biomed Inform* **39**, 589--99 (2006).
- [42] Bitterman, N., Lerner, E. & Bitterman, H. Evaluation of data display for patient-oriented electronic record of anticoagulant therapy. *Telemed J E Health* **16**, 799--806 (2010).
- [43] Patrick, J. & Li, M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc* **17**, 524--7 (2010).
- [44] Spasic, I., Sarafraz, F., Keane, J. A. & Nenadic, G. Medication information extraction with linguistic pattern matching and semantic rules. *J Am Med Inform Assoc* **17**, 532--5 (2010).
- [45] Aronson, A. R. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proc AMIA Symp* 17--21 (2001).
- [46] Goh, K.-I. *et al.* The human disease network. *Proc Natl Acad Sci USA* **104**, 8685--90 (2007). URL <http://www.pnas.org/content/104/21/8685.long>.

- [47] Lage, K., Karlberg, E., Størling, Z. & Ólason, P. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology* (2007). URL <http://www.nature.com/nbt/journal/v25/n3/abs/nbt1295.html>.
- [48] Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173--8 (2005). URL <http://www.nature.com/nature/journal/v437/n7062/full/nature04209.html>.
- [49] Paller, A. S. *et al.* Compound heterozygous mutations in the hairless gene in atrichia with papular lesions. *J Invest Dermatol* **121**, 430--2 (2003).
- [50] Colson, N. J., Lea, R. A., Quinlan, S., MacMillan, J. & Griffiths, L. R. The estrogen receptor 1 g594a polymorphism is associated with migraine susceptibility in two independent case/control groups. *Neurogenetics* **5**, 129--33 (2004).
- [51] Ikeda, A., Shibasaki, H., Shiozaki, A. & Kimura, J. Alopecia with carbamazepine in two patients with focal seizures. *J Neurol Neurosurg Psychiatr* **63**, 549--50 (1997).
- [52] Mercke, Y., Sheng, H., Khan, T. & Lippmann, S. Hair loss in psychopharmacology. *Ann Clin Psychiatry* **12**, 35--42 (2000). URL [http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list\\_uids=10798824&dopt=abstractplus](http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=10798824&dopt=abstractplus).
- [53] Muzina, D. J., El-Sayegh, S. & Calabrese, J. R. Antiepileptic drugs in psychiatry-focus on randomized controlled trial. *Epilepsy Res* **50**, 195--202 (2002).
- [54] Krasowska, D., Szymanek, M., Schwartz, R. A. & Myśliński, W. Cutaneous effects of the most commonly used antidepressant medication, the selective serotonin reuptake inhibitors. *J Am Acad Dermatol* **56**, 848--53 (2007). URL [http://www.eblue.org/article/S0190-9622\(06\)02850-7/abstract](http://www.eblue.org/article/S0190-9622(06)02850-7/abstract).
- [55] Whitty, C. W., Hockaday, J. M. & Whitty, M. M. The effect of oral contraceptives on migraine. *Lancet* **1**, 856--9 (1966).
- [56] Eaton, W. *et al.* Coeliac disease and schizophrenia: population based case control study with linkage of danish national registers. *BMJ* **328**, 438--9 (2004).
- [57] Bushara, K. O. Neurologic presentation of celiac disease. *Gastroenterology* **128**, S92--7 (2005).
- [58] Fessatou, S., Kostaki, M. & Karpathios, T. Coeliac disease and alopecia areata in childhood. *J Paediatr Child Health* **39**, 152--4 (2003).
- [59] Ma, H. *et al.* Use of four biomarkers to evaluate the risk of breast cancer subtypes in the women's contraceptive and reproductive experiences study. *Cancer Res* **70**, 575--87 (2010).
- [60] Plovnick, R. M. The progression of electronic health records and implications for psychiatry. *Am J Psychiatry* **167**, 498--500 (2010).
- [61] Hettne, K. M., van Mulligen, E. M., Schuemie, M. J., Schijvenaars, B. J. & Kors, J. A. Rewriting and suppressing umls terms for improved biomedical term identification. *J Biomed Semantics* **1**, 5 (2010).
- [62] Hidalgo, C. A., Blumm, N., Barabási, A.-L. & Christakis, N. A. A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol* **5**, e1000353 (2009).

- [63] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. On-line mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514--7 (2005).
- [64] Lage, K. *et al.* Dissecting spatio-temporal protein networks driving human heart development and related disorders. *Mol Syst Biol* **6**, 381 (2010). URL <http://www.nature.com/msb/journal/v6/n1/full/msb201036.html>.
- [65] Lage, K. *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc Natl Acad Sci USA* **105**, 20870--5 (2008). URL <http://www.pnas.org/content/105/52/20870>.
- [66] de Lusignan, S. & van Weel, C. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Fam Pract* **23**, --263 (2006). URL <http://dx.doi.org/10.1093/fampra/cmi106>.
- [67] Keim, D. Visual analytics: Scope and challenges. *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, Springer 76--90 (2008).
- [68] *LifeLines: Visualizing Personal Histories*. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.707>.
- [69] *Aligning temporal data by sentinel events: discovering patterns in electronic health records*. URL <http://dx.doi.org/10.1145/1357054.1357129>.
- [70] Wang, T. *et al.* Temporal summaries: Supporting temporal categorical searching, aggregation and comparison. *IEEE JVCG* **15**, -- 1056 (2009).
- [71] *Multi-modal presentation of medical histories*.
- [72] *KNAVE II: the definition and implementation of an intelligent tool for visualization and exploration of time-oriented clinical data*.
- [73] Bui, A. A. T., Aberle, D. R. & Kangarloo, H. Timeline: visualizing integrated patient records. *IEEE Trans Inf Technol Biomed* **11**, 462--73 (2007).
- [74] Kosara, R. & Miksch, S. Metaphors of movement: a visualization and user interface for time-oriented, skeletal plans. *Artif Intell Med* **22**, 111--31 (2001). URL [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=11348843](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11348843).
- [75] Shahar, Y., Miksch, S. & Johnson, P. An intention-based language for representing clinical guidelines. *Proc AMIA Annu Fall Symp* 592--6 (1996). URL [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=8947735](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8947735).
- [76] Chuah, M. Glyphs for software visualization. *5th International Workshop on Program Comprehension (IWPC '97) Proceedings* 183--191 (1997).
- [77] Aigner, W. & Miksch, S. Carevis: integrated visualization of computerized protocols and temporal patient data. *Artif Intell Med* **37**, 18 (2006).
- [78] *CareView: analyzing nursing narratives for temporal trends*.
- [79] Portet, F. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence* 789--816 (2009).
- [80] *Visualization of patient data at different temporal granularities on mobile devices*.

- [81] Bertini, E. & Lalanne, D. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration* 12--20 (2009).
- [82] Bork, P. *et al.* Predicting function: from genes to genomes and back. *J Mol Biol* **283**, 707--25 (1998).
- [83] Wu, X., Jiang, R., Zhang, M. Q. & Li, S. Network-based global inference of human disease genes. *Mol Syst Biol* **4**, 189 (2008).
- [84] Xu, J. & Li, Y. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics* **22**, 2800--5 (2006).
- [85] Stumpf, M. P. H. *et al.* Estimating the size of the human interactome. *Proc Natl Acad Sci USA* **105**, 6959--64 (2008). URL <http://www.pnas.org/content/105/19/6959.long>.
- [86] Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat Methods* **6**, 83--90 (2009). URL <http://www.nature.com/nmeth/journal/v6/n1/abs/nmeth.1280.html>.
- [87] Cusick, M. E. *et al.* Literature-curated protein interaction datasets. *Nat Methods* **6**, 39--46 (2009).
- [88] Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* (2008).
- [89] Ceol, A. *et al.* Mint, the molecular interaction database: 2009 update. *Nucleic Acids Res* **38**, D532--9 (2010).
- [90] Bader, G. D., Betel, D. & Hogue, C. W. V. Bind: the biomolecular interaction network database. *Nucleic Acids Res* **31**, 248--50 (2003).
- [91] Breitkreutz, B.-J. *et al.* The biogrid interaction database: 2008 update. *Nucleic Acids Res* **36**, D637--40 (2008).
- [92] Aranda, B. *et al.* The intact molecular interaction database in 2010. *Nucleic Acids Res* **38**, D525--31 (2010).
- [93] Xenarios, I. *et al.* Dip: the database of interacting proteins. *Nucleic Acids Res* **28**, 289--91 (2000).
- [94] Prasad, T. S. K. *et al.* Human protein reference database--2009 update. *Nucleic Acids Res* **37**, D767--72 (2009).
- [95] von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399--403 (2002).
- [96] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38**, D355--60 (2010).
- [97] Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res* **34**, D354--7 (2006).
- [98] Kanehisa, M. & Goto, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27--30 (2000).
- [99] Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**, 1041--52 (2001).
- [100] de Lichtenberg, U., Jensen, L. J., Brunak, S. & Bork, P. Dynamic complex formation during the yeast cell cycle. *Science* **307**, 724--7 (2005).

- [101] Matthews, L. *et al.* Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* **37**, D619--22 (2009).
- [102] Chao, E. C. & Lipkin, S. M. Molecular models for the tissue specificity of dna mismatch repair-deficient carcinogenesis. *Nucleic Acids Res* **34**, 840--52 (2006).
- [103] Vogelstein, B., Lane, D. & Levine, A. J. Surfing the p53 network. *Nature* **408**, 307--10 (2000).
- [104] Winter, E. E., Goodstadt, L. & Ponting, C. P. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res* **14**, 54--61 (2004).
- [105] Beyer, K. *et al.* Identification and characterization of a new alpha-synuclein isoform and its role in lewy body diseases. *Neurogenetics* (2007).
- [106] Kim, K. Y., Kee, M. K., Chong, S. A. & Nam, M. J. Galanin is up-regulated in colon adenocarcinoma. *Cancer Epidemiol Biomarkers Prev* **16**, 2373--2378 (2007).
- [107] Butland, G. *et al.* Interaction network containing conserved and essential protein complexes in escherichia coli. *Nature* **433**, 531--7 (2005).
- [108] Gavin, A. C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631--6 (2006).
- [109] Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141--7 (2002).
- [110] Ho, Y. *et al.* Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180--3 (2002).
- [111] Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature* **440**, 637--43 (2006).
- [112] van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G. & Leunissen, J. A. M. A text-mining analysis of the human phenome. *Eur J Hum Genet* **14**, 535--42 (2006). URL <http://www.nature.com/ejhg/journal/v14/n5/abs/5201585a.html>.
- [113] Safran, M. *et al.* Genecards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics* **18**, 1542--3 (2002).
- [114] Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957--68 (2005).
- [115] Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062--7 (2004).
- [116] Korbelt, J. O. *et al.* Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol* **3**, e134 (2005).
- [117] Eklund, A. C. & Szallasi, Z. Correction of technical bias in clinical microarray data improves concordance with known biological information. *Genome Biol* **9**, R26 (2008).
- [118] Polanco, J. C. & Koopman, P. Sry and the hesitant beginnings of male development. *Dev Biol* **302**, 13--24 (2007).
- [119] Patel, M. *et al.* Primate dax1, sry, and sox9: evolutionary stratification of sex-determination pathway. *Am J Hum Genet* **68**, 275--80 (2001).
- [120] Park, S. Y., Tong, M. & Jameson, J. L. Distinct roles for steroidogenic factor 1 and desert hedgehog pathways in fetal and adult leydig cell development. *Endocrinology* **148**, 3704--10 (2007).

- [121] Parker, K. L. The roles of steroidogenic factor 1 in endocrine development and function. *Mol Cell Endocrinol* **140**, 59--63 (1998).
- [122] Swain, A., Narvaez, V., Burgoyne, P., Camerino, G. & Lovell-Badge, R. Dax1 antagonizes sry action in mammalian sex determination. *Nature* **391**, 761--7 (1998).
- [123] Pop, R., Zaragoza, M. V., Gaudette, M., Dohrmann, U. & Scherer, G. A homozygous nonsense mutation in sox9 in the dominant disorder campomelic dysplasia: a case of mitotic gene conversion. *Hum Genet* **117**, 43--53 (2005).
- [124] MacLaughlin, D. T. & Donahoe, P. K. Sex determination and differentiation. *N Engl J Med* **350**, 367--78 (2004).
- [125] Shen, W. H., Moore, C. C., Ikeda, Y., Parker, K. L. & Ingraham, H. A. Nuclear receptor steroidogenic factor 1 regulates the mullerian inhibiting substance gene: a link to the sex determination cascade. *Cell* **77**, 651--61 (1994).
- [126] Meeks, J. J., Weiss, J. & Jameson, J. L. Dax1 is required for testis determination. *Nat Genet* **34**, 32--3 (2003).
- [127] Notarnicola, C., Malki, S., Berta, P., Poulat, F. & Boizet-Bonhoure, B. Transient expression of sox9 protein during follicular development in the adult mouse ovary. *Gene Expr Patterns* **6**, 695--702 (2006).
- [128] Bouma, G. J., Washburn, L. L., Albrecht, K. H. & Eicher, E. M. Correct dosage of fog2 and gata4 transcription factors is critical for fetal testis development in mice. *Proc Natl Acad Sci U S A* (2007).
- [129] Biason-Lauber, A. & Schoenle, E. J. Apparently normal ovarian differentiation in a prepubertal girl with transcriptionally inactive steroidogenic factor 1 (nr5a1/sf-1) and adrenocortical insufficiency. *Am J Hum Genet* **67**, 1563--8 (2000).
- [130] Schepers, G., Wilson, M., Wilhelm, D. & Koopman, P. Sox8 is expressed during testis differentiation in mice and synergizes with sf1 to activate the amh promoter in vitro. *J Biol Chem* **278**, 28101--8 (2003).
- [131] Ewing, R. M. *et al.* Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* **3**, 89 (2007).
- [132] Griffin, T. J. *et al.* Complementary profiling of gene expression at the transcriptome and proteome levels in *saccharomyces cerevisiae*. *Mol Cell Proteomics* **1**, 323--33 (2002).
- [133] Roch, K. G. L. *et al.* Global analysis of transcript and protein levels across the plasmodium falciparum life cycle. *Genome Res* **14**, 2308--18 (2004).
- [134] Mootha, V. K. *et al.* Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **115**, 629--40 (2003).
- [135] Kislinger, T. *et al.* Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* **125**, 173--86 (2006).
- [136] David, S. S., O'Shea, V. L. & Kundu, S. Base-excision repair of oxidative dna damage. *Nature* **447**, 941--50 (2007).
- [137] Falck, J., Mailand, N., Syljuasen, R. G., Bartek, J. & Lukas, J. The atm-chk2-cdc25a checkpoint pathway guards against radioreistant dna synthesis. *Nature* **410**, 842--7 (2001).
- [138] Petrocca, F. *et al.* Alterations of the tumor suppressor gene arlt1 in ovarian cancer. *Cancer Res* **66**, 10287--91 (2006).

- [139] Singh, S. K. *et al.* Identification of human brain tumour initiating cells. *Nature* **432**, 396--401 (2004).
- [140] Huminiecki, L., Lloyd, A. T. & Wolfe, K. H. Congruence of tissue expression profiles from gene expression atlas, sagemap and tissueinfo databases. *BMC Genomics* **4**, 31 (2003).
- [141] Jackson, D. A., Somers, K. M. & Harvey, H. H. Similarity measures: Measures of co-occurrence and association or simply measures of co-occurrence? *The American Naturalist* **133**, 436--453 (1989).
- [142] Ochiai, A. Zoogeographic studies on the soleoid fishes found in japan and its neighbouring regions. *Bull. Jpn. Soc. Sci. Fish.* **22**, 526--530 (1957).
- [143] Udoh, E. & Rhoades, J. Mining documents in a small enterprise using wordstat. 490--494 (2006).
- [144] Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science (New York, NY)* **320**, 539--543 (2008).
- [145] Bruneau, B. The developmental genetics of congenital heart disease. *Nature* **451**, 943--948 (2008). URL <http://dx.doi.org/10.1038/nature06801>. 10.1038/nature06801.
- [146] Chien, K., Domian, I. & Parker, K. Cardiogenesis and the complex biology of regenerative cardiovascular medicine. *Science (New York, NY)* **322**, 1494--1497 (2008).
- [147] Olson, E. A decade of discoveries in cardiac biology. *Nat Med* **10**, 467--474 (2004). URL <http://dx.doi.org/10.1038/nm0504-467>. 10.1038/nm0504-467.
- [148] Srivastava, D. Making or breaking the heart: from lineage determination to morphogenesis. *Cell* **126**, 1037--1048 (2006). URL <http://dx.doi.org/10.1016/j.cell.2006.09.003>. 10.1016/j.cell.2006.09.003.
- [149] Fishman, M. & Olson, E. Parsing the heart: genetic modules for organ assembly. *Cell* **91**, 153--156 (1997). URL [http://dx.doi.org/10.1016/S0092-8674\(00\)80397-9](http://dx.doi.org/10.1016/S0092-8674(00)80397-9). 10.1016/S0092-8674(00)80397-9.
- [150] Basson, C. *et al.* Mutations in human *tbx5* [lsqb]corrected[rsqb] cause limb and cardiac malformation in holt-oram syndrome. *Nat Genet* **15**, 30--35 (1997). URL <http://dx.doi.org/10.1038/ng0197-30>. 10.1038/ng0197-30.
- [151] Benson, D. *et al.* Mutations in the cardiac transcription factor *nkx2.5* affect diverse cardiac developmental pathways. *J Clin Invest* **104**, 1567--1573 (1999). URL <http://dx.doi.org/10.1172/JCI8154>. 10.1172/JCI8154.
- [152] Garg, V. *et al.* Gata4 mutations cause human congenital heart defects and reveal an interaction with *tbx5*. *Nature* **424**, 443--447 (2003). URL <http://dx.doi.org/10.1038/nature01827>. 10.1038/nature01827.
- [153] Gaudet, J., Muttumu, S., Horner, M. & Mango, S. Whole-genome analysis of temporal gene expression during foregut development. *PLoS Biol* **2**, e352 (2004). URL <http://dx.doi.org/10.1371/journal.pbio.0020352>. 10.1371/journal.pbio.0020352.
- [154] Gaudet, J. & Mango, S. Regulation of organogenesis by the *caenorhabditis elegans* *foxa* protein *pha-4*. *Science (New York, NY)* **295**, 821--825 (2002).

- [155] Kornberg, T. & Tabata, T. Segmentation of the drosophila embryo. *Curr Opin Genet Dev* **3**, 585--594 (1993). URL [http://dx.doi.org/10.1016/0959-437X\(93\)90094-6](http://dx.doi.org/10.1016/0959-437X(93)90094-6).
- [156] Pourquie, O. The segmentation clock: converting embryonic time into spatial pattern. *Science (New York, NY)* **301**, 328--330 (2003).
- [157] Schott, J. *et al.* Congenital heart disease caused by mutations in the transcription factor nkx2-5. *Science (New York, NY)* **281**, 108--111 (1998).
- [158] Weatherbee, S., Halder, G., Kim, J., Hudson, A. & Carroll, S. Ultrabithorax regulates genes at several levels of the wing-patterning hierarchy to shape the development of the drosophila haltere. *Genes Dev* **12**, 1474--1482 (1998). URL <http://dx.doi.org/10.1101/gad.12.10.1474>.
- [159] Bult, C., Eppig, J., Kadin, J., Richardson, J. & Blake, J. The mouse genome database (mgd): mouse biology and model systems. *Nucleic Acids Res* **36**, D724--D728 (2008). URL <http://dx.doi.org/10.1093/nar/gkm961>.
- [160] O'Brien, K., Remm, M. & Sonnhammer, E. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33**, D476 (2002). URL <http://dx.doi.org/10.1093/nar/gki107>.
- [161] Chen, F., Mackey, A., Vermunt, J. & Roos, D. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE* **2**, e383 (2007). URL <http://dx.doi.org/10.1371/journal.pone.0000383>.
- [162] Hulsen, T., Huynen, M., de Vlieg, J. & Groenen, P. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* **7**, R31 (2006). URL <http://dx.doi.org/10.1186/gb-2006-7-4-r31>.
- [163] Maduro, M. & Rothman, J. Making worm guts: the gene regulatory network of the caenorhabditis elegans endoderm. *Dev Biol* **246**, 68--85 (2002). URL <http://dx.doi.org/10.1006/dbio.2002.0655>.
- [164] Grego-Bessa, J. *et al.* Notch signaling is essential for ventricular chamber development. *Dev Cell* **12**, 415--429 (2007). URL <http://dx.doi.org/10.1016/j.devcel.2006.12.011>.
- [165] Wagner, G., Pavlicev, M. & Cheverud, J. The road to modularity. *Nat Rev Genet* **8**, 921--931 (2007). URL <http://dx.doi.org/10.1038/nrg2267>.
- [166] Garg, V. *et al.* Mutations in notch1 cause aortic valve disease. *Nature* **437**, 270--274 (2005). URL <http://dx.doi.org/10.1038/nature03940>.
- [167] Moorman, A. & Christoffels, V. Cardiac chamber formation: development, genes, and evolution. *Physiol Rev* **83**, 1267 (2003).
- [168] Olson, E. Gene regulatory networks in the evolution and development of the heart. *Science* **313**, 1922--1927 (2006). URL <http://dx.doi.org/10.1126/science.1132292>.



- [169] Kashtan, N. & Alon, U. Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci USA* **102**, 13773–13778 (2005). URL <http://dx.doi.org/10.1073/pnas.0503610102>. 10.1073/pnas.0503610102.
- [170] Lipson, H., Pollack, J. & Suh, N. On the origin of modular variation. *Evolution* **56**, 1549–1556 (2002).
- [171] Satou, Y. & Satoh, N. Gene regulatory networks for the development and evolution of the chordate heart. *Genes Dev* **20**, 2634–2638 (2006). URL <http://dx.doi.org/10.1101/gad.1485706>. 10.1101/gad.1485706.
- [172] Kim, S. *et al.* A gene expression map for caenorhabditis elegans. *Science (New York, NY)* **293**, 2087–2092 (2001).
- [173] Skeath, J. & Thor, S. Genetic control of drosophila nerve cord development. *Curr Opin Neurobiol* **13**, 8–15 (2003). URL [http://dx.doi.org/10.1016/S0959-4388\(03\)00007-2](http://dx.doi.org/10.1016/S0959-4388(03)00007-2). 10.1016/S0959-4388(03)00007-2.
- [174] Bergstrom, D. *et al.* Promoter-specific regulation of myod binding and signal transduction cooperate to pattern gene expression. *Mol Cell* **9**, 587–600 (2002). URL [http://dx.doi.org/10.1016/S1097-2765\(02\)00481-1](http://dx.doi.org/10.1016/S1097-2765(02)00481-1). 10.1016/S1097-2765(02)00481-1.
- [175] Bergholdt, R. *et al.* Integrative analysis for finding genes and networks involved in diabetes and other complex diseases. *Genome Biol* **8**, R253 (2007). URL <http://dx.doi.org/10.1186/gb-2007-8-11-r253>. 10.1186/gb-2007-8-11-r253.
- [176] D'Hertog, W. *et al.* Proteomics analysis of cytokine-induced dysfunction and death in insulin-producing ins-1e cells: new insights into the pathways involved. *Mol Cell Proteomics* **6**, 2180–2199 (2007). URL <http://dx.doi.org/10.1074/mcp.M700085-MCP200>. 10.1074/mcp.M700085-MCP200.
- [177] Lang, J. *et al.* A comparison of rna amplification techniques at sub-nanogram input concentration. *BMC Genomics* **10**, 326 (2009). URL <http://dx.doi.org/10.1186/1471-2164-10-326>. 10.1186/1471-2164-10-326.
- [178] Kim, J. *et al.* Polony multiplex analysis of gene expression (pmage) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1484 (2007).
- [179] Edwards, T. M. & Myers, J. P. Environmental exposures and gene regulation in disease etiology. *Cien Saude Colet* **13**, 269–81 (2008). URL [http://www.scielo.br/scielo.php?script=sci\\_arttext&pid=S1413-81232008000100030&lng=en&nrm=iso&tlng=en](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232008000100030&lng=en&nrm=iso&tlng=en).
- [180] Olden, K. & Wilson, S. Environmental health and genomics: visions and implications. *Nat Rev Genet* **1**, 149–53 (2000). URL <http://www.nature.com/doifinder/10.1038/35038586>.
- [181] Davis, A. P. *et al.* Comparative toxicogenomics database: a knowledge-base and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res* **37**, D786–92 (2009). URL [http://nar.oxfordjournals.org/cgi/content/full/37/suppl\\_1/D786?view=long&pmid=18782832](http://nar.oxfordjournals.org/cgi/content/full/37/suppl_1/D786?view=long&pmid=18782832).

- [182] Cooper, R. S. Gene-environment interactions and the etiology of common complex disease. *Ann Intern Med* **139**, 437--40 (2003).
- [183] Keiser, M. J. *et al.* Predicting new molecular targets for known drugs. *Nature* **462**, 175--81 (2009). URL <http://www.nature.com/nature/journal/v462/n7270/full/nature08506.html>.
- [184] Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263--6 (2008). URL <http://www.sciencemag.org/cgi/content/full/321/5886/263>.
- [185] Phillips, D. & Arlt, V. Genotoxicity: damage to dna and its consequences. *EXS* **99** (2009).
- [186] Hopkins, A. Network pharmacology. *Nat Biotechnol* **25** (2007).
- [187] Paolini, G., Shapland, R., van Hoorn, W., Mason, J. & Hopkins, A. Global mapping of pharmacological space. *Nat Biotechnol* **24** (2006).
- [188] Yildirim, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L. & Vidal, M. Drug-target network. *Nat Biotechnol* **25**, 1119--26 (2007). URL <http://www.nature.com/nbt/journal/v25/n10/abs/nbt1338.html>.
- [189] Keith, C., Borisy, A. & Stockwell, B. Multi-component therapeutics for networked systems. *Nat Rev Drug Discovery* **4** (2005).
- [190] Keiser, M., Roth, B., Armbruster, B., Ernsberger, P. & Irwin, J. Relating protein pharmacology by their ligand chemistry. *Nat Biotechnol* **25** (2007).
- [191] Morphy, R. & Rankovic, Z. Fragments network biology and designing multiple ligands. *Drug Discov Today* **12** (2007).
- [192] Lamb, J., Crawford, E., Peck, D., Modell, J. & Blat, I. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313** (2006).
- [193] Williams-Devane, C., Wolf, M. & Richard, A. Toward a public toxicogenomics capability for supporting predictive toxicology: survey of current resources and chemical indexing of experiments in geo and arrayexpress. *Toxicol Sci* **109** (2009).
- [194] Yang, L., Milutinovic, P., Brosnan, R., Eger, E. & Sonner, J. The plasticizers di(2-ethylhexyl)phthalate modulates gamma-aminobutyric acid type a and glycine receptor function. *Anesth Analg* **105** (2007).
- [195] Wishart, D., Knox, C., Guo, A., Cheng, D. & Shrivastava, S. Drugbank: a knowledge-base for drugs, drug actions and drug targets. *Nucleic Acids Res. Database issue* (2008).
- [196] Lipinski, C., Lombardo, F., Dominy, B. & Feeney, P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46** (2001).
- [197] Veber, D., Johnson, S., Chen, g. H., Smith, B. & Ward, K. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **45** (2002).
- [198] Ashburner, M. Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nature Genet.* **25** (2000).
- [199] Bader, G. & Hogue, C. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4** (2003).

- [200] van Dongen, S. A cluster algorithm for graphs technical report ins-00010. *version 1006 National Research Institute for Mathematics and Computer Science in the Netherlands Amsterdam*. (2000). URL <http://www.micans.org/mcl/>.
- [201] Brohé, S. & van Helden, J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7** (2006).
- [202] Vlasblom, J. & Wodak, S. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics* **10** (2009).
- [203] Rebhan, M., Chalifa-Caspi, V., Prilusky, J. & Lancet, D. Genecards: integrating information about genes proteins and diseases. *Trends in Genetics* **13** (1997).
- [204] Salehi, F., Turner, M., Phillips, K., Wigle, D. & Krewski, D. Review of the etiology of breast cancer with special attention to organochlorines as potential endocrine disruptors. *J. Toxicol. Environ. Health B Crit. Rev.* **11** (2008).
- [205] Geier, D. S. & Geier, M. A review of thimerosal (merthiolate) and its ethylmercury breakdown product: specific historical considerations regarding safety and effectiveness. *J. Toxicol. Env. Health* **10** (2007).
- [206] Westphal, G., Asgari, S., Schulz, T., Bünge, J. & Müller, M. Thimerosal induces micronuclei in the cytochalasin b block micronucleus test with human lymphocytes. *Arch. Toxicol.* **77** (2003).
- [207] Kuper, C., Stierum, R., Boorsma, A., Schijf, M. & Prinsen, M. The contact allergen dinitrochlorobenzene (dncb) and respiratory allergy in the th2-prone brown norway rat. *Toxicology* **246** (2008).
- [208] Sani, H., Rahmat, A., Ismail, M., Rosli, R. & Endrini, S. Potential anticancer effect of red spinach (*amaranthus gangeticus*) extract. *Asia Pac. J. Clin. Nutr.* **13** (2004).
- [209] Block, G., Patterson, B. & Subar, A. Fruit vegetables and cancer prevention: a review of the epidemiological evidence. *Nutr. Cancer* **18** (1992).
- [210] Krinsky, N. & Johnson, E. Carotenoid actions and their relation to health and disease. *Mol. Aspects Med.* **26** (2005).
- [211] Peters, U. Serum lycopene other carotenoids and prostate cancer risk: a nested case-control study in the prostate lung colorectal and ovarian cancer screening trial. *Cancer Epidemiol. Biomarkers Prev.* **16** (1997).
- [212] Toniolo, P., van Kappel, A., Akhmedkhanov, A., Ferrari, P. & Kato, I. Serum carotenoids and breast cancer. *Am. J. Epidemiol.* **153** (2001).
- [213] Martinasso, G., Maggiora, M., Trombetta, A., Angela, C. & Muzio, G. Effects of di(2-ethylhexyl) phthalate a widely used peroxisome proliferator and plasticizers on cell growth in the human keratinocyte cell line nctc 2544. *J. Toxicol. Env. Health* **69** (2006).
- [214] Latini, G. Potential hazards of exposure to di-2-ethylhexyl phthalate in babies: a review. *Biol. Neonate* **78** (2000).
- [215] Kim, H., Ishizuka, M., Kazusaka, A. & Fujita, S. Alterations of activities of cytosolic phospholipase a2 and arachidonic acid metabolizing enzymes in di-(2-ethylhexyl) phthalate induced testicular atrophy. *J. Vet. Med. Sci.* **66** (2004).
- [216] Turan, N., Waring, R. & Ramsden, D. The effect of plasticisers on "sulphate supply" enzymes. *Mol. Cell. Endocrinol.* **244** (2005).

- [217] Horiuchi, Y., Nakayama, J., Ishiguro, H., Ohtsuki, T. & Detera-Wadleigh, S. Possible association between a haplotype of the gaba-a receptor alpha 1 subunit gene (gabrar1) and mood disorders. *Biol. Psychiatry* **55** (2004).
- [218] Moon, B. A single administration of 2 and3 and7 and8-tetrachlorodibenzo-p-dioxin that produces reduced food and water intake induces long-lasting expression of corticotropin-releasing factor arginine vasopressin and proopiomelanocortin in rat brain. *Toxicol. Appl. Pharmacol.* **233** (2008).
- [219] Legare, M., Hanneman, W., Barhoumi, R., Burghardt, R. & Tiffany-Castoglioni, E. 2 and3 and7 and8-tetrachlorodibenzo-p-dioxin alters hippocampal astroglia-neuronal gap junctional communication. *Neurotoxicology* **21** (2000).
- [220] Nayyar, T., Zawia, N. & Hood, D. Transplacental effects of 2 and3 and7 and8-tetrachlorodibenzo-p-dioxin on the temporal modulation of sp1 dna binding in the developing cerebral cortex and cerebellum. *Exp. Toxicol. Pathol.* **53** (2002).
- [221] Kakeyama, M., Sone, H., Miyabara, Y. & Tohyama, C. Perinatal exposure to 2 and3 and7 and8-tetrachlorodibenzo-p-dioxin alters activity-dependent expression of bdnf mrna in the neocortex and male rat sexual behavior in adulthood. *Neurotoxicology* **24** (2003).
- [222] Kim, S. & Yang, J. Neurotoxic effects of 2 and3 and7 and8-tetrachlorodibenzo-p-dioxin in cerebellar granule cells. *Exp. Mol. Med.* **37** (2005).
- [223] Boverhof, D., Burgoon, L., Tashiro, C., Sharratt, B. & Chittim, B. Comparative toxicogenomics analysis of the hepatotoxic effects of tcdd in sprague dawley rats and c57bl/6 mice. *Toxicol. Sci.* **94** (2006).
- [224] Fletcher, N., Wahlström, D., Lundberg, R., Nilsson, C. & Nilsson, K. 2 and3 and7 and8-tetrachlorodibenzo-p-dioxin (tcdd) alters the mrna expression of critical genes associated with cholesterol metabolism bile acid biosynthesis and bile transport in rat liver: a microarray study. *Toxicol. Appl. Pharmacol.* **207** (2005).
- [225] Volz, D., Bencic, D., Hinton, D., Law, J. & Kullman, S. 2 and3 and7 and8-tetrachlorodibenzo-p-dioxin (tcdd) induces organ- specific differential gene expression in male japanese medaka (*oryzias latipes*). *Toxicol. Sci.* **85** (2005).
- [226] Lalwani, N., Reddy, M., Qureshi, S. & Reddy, J. Development of hepatocellular carcinomas and increased peroxisomal fatty acid beta-oxidation in rats fed [4-chloro-6-(23-xylidino)-2-pyrimidinylthio] acetic acid (wy-14643) in the semipurified diet. *Carcinogenesis* **2** (1981).
- [227] Suga, T. Hepatocarcinogenesis by peroxisome proliferators. *J. Toxicol. Sci.* **29** (2004).
- [228] Amacher, D., Adler, R., Herath, A. & Townsend, R. Use of proteomic methods to identify serum biomarkers associated with rat liver toxicity or hypertrophy. *Clin. Chem.* **51** (2005).
- [229] Bauer, D., Wolfram, N., Kahl, G. & Hirsch-Ernst, K. Transcriptional regulation of cyp2b1 induction in primary rat hepatocyte cultures: repression by epidermal growth chemical is mediated via a distal enhancer region. *Mol. Pharmacol.* **65** (2004).
- [230] Heder, A., Hirsch-Ernst, K., Bauer, D., Kahl, G. & Desel, H. Induction of cytochrome p450 2b1 by pyrethroids in primary rat hepatocyte cultures. *Biochem. Pharmacol.* **62** (2001).

- [231] Eil, C. & Nisula, B. The binding properties of pyrethroids to human skin fibroblast androgen receptors and to sex hormone binding globulin. *J. Steroid Biochem.* **35** (1990).
- [232] von Mering, C., Jensen, L., Kuhn, M., Chafon, S. & Doerks, T. String 7-- recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35** (2007).
- [233] Stein, L. D. Human genome: End of the beginning. *Nature* **431** (2004).
- [234] Oprea, T. & Tropsha, A. Target chemical and bioactivity databases -integration is key. *Drug. Discov. today technol.* **3** (2006).
- [235] Mestres, J., Gregori-Puigjané, E., Valverde, S. & Solé, R. Data completeness - the achilles heel of drug-target networks. *Nat. Biotechnol.* **26** (2008).
- [236] Birney, E., Andrews, T., Bevan, P., Caccamo, M. & Chen, Y. An overview of ensemble. *Genome Res.* **145** (2004).
- [237] Mestres, J., Gregori-Puigjane, E., Valverde, S. & Sole, R. The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol. Biosyst.* **5** (2009).
- [238] Antonelli, A. *et al.* Cytokines (interferon-gamma and tumor necrosis factor-alpha)-induced nuclear factor-kappaB activation and chemokine (c-x-c motif) ligand 10 release in graves disease and ophthalmopathy are modulated by pioglitazone. *Metabolism* (2010).
- [239] Vaz, R. & Klabunde, T. Antitargets: Prediction and prevention of drug side effects. In *Methods and Principles in Medicinal Chemistry* **Wiley-VCH** (2008).
- [240] Broccatelli, F., Carosati, E., Cruciani, G. & Oprea, T. Transporter-mediated efflux influences CNS side effects: Abcb1, from anti-target to target. *Mol. Inf.* **29** (2010).
- [241] Olah, M. *et al.* Wombat and wombat-pk: Bioactive databases for lead and drug discovery. In *Chemical Biology: From small molecules to systems biology and drug design* **Wiley-VCH** (2007).
- [242] Weill, N. & Rognan, D. Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G-protein coupled receptors and their ligands. *J. Chem. Inf. Model.* **49** (2009).
- [243] Mestres, J., Martin-Couce, L., Gregori-Puigjane, E., Cases, M. & Boyer, S. Ligand-based approach to in silico pharmacology: nuclear receptor profiling. *J. Chem. Inf. Model.* **46** (2006).
- [244] Knight, Z., Lin, H. & Shokat, K. Targeting the cancer kinome through polypharmacology. *Nat. Rev. Cancer* **10** (2010).
- [245] Garcia-Serna, R., Ursu, O., Oprea, T. & Mestres, J. iphace: integrative navigation in pharmacological space. *Bioinformatics* **26** (2010).
- [246] Wheeler, D. *et al.* Databases resources of the national center for biotechnology information. *Nucleic Acids Res.* **35** (2007).
- [247] de Matos, P. *et al.* Chemical entities of biological interest: an update. *Nucleic Acids Res.* **38** (2010).
- [248] Berger, S. & Iyengar, R. Network analyses in systems pharmacology. *Bioinformatics* **25** (2009).
- [249] Oprea, T., Tropsha, A., Faulon, J. & Rintoul, M. Systems chemical biology. *Nat. Chem. Biol.* **3** (2007).

- [250] Chen, B., Wild, D. & Guha, R. Pubchem as a source of polypharmacology. *J. Chem. Inf. Model.* **49** (2009).
- [251] Kuhn, M., Campillos, M., Letunic, I., Jensen, L. & Bork, P. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* **6** (2010).
- [252] Audouze, K. *et al.* Deciphering diseases and biological targets for environmental chemicals using toxicogenomics networks. *PLoS Comput Biol* **6**, e1000788 (2010). URL <http://www.ploscompbiol.org/article/info%253Adoi%252F10.1371%252Fjournal.pcbi.1000788>.
- [253] Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. Binding db: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **35** (2007).
- [254] Roth, B., Lopez, E., Beischel, S., Weskaemper, R. & Evans, J. Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for cns drug discovery. *Pharmacol* (2004).
- [255] Wishart, D. *et al.* Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34** (2006).
- [256] Hewett, M. *et al.* Pharmgkb: The pharmacogenetics knowledge base. *Nucleic Acids Res.* **30** (2002).
- [257] Kuhn, M. *et al.* Stitch 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.* **38** (2010).
- [258] Durant, J., Leland, B., Henry, D. & Nourse, J. Reoptimization of mdl keys for use in drug discovery. *J. Chem. Inf Comput. Sci.* **42** (2002).
- [259] Bush, B. L. & Sheridan, R. P. Patty: a programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **33** (1993).
- [260] Chemical Computing Group, C., Montreal. Moe (version 2007.09). URL [www.chemcomp.com](http://www.chemcomp.com).
- [261] Willet, P. Similarity-based virtual screening using 2d fingerprints. *Drug Discov* 1046--1053 (2006).
- [262] Zanzoni, A. *et al.* Mint: a molecular interaction database. *FEBS Lett.* **513** (2002).
- [263] Salwinski, L. *et al.* The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32** (2004).
- [264] Mishra, G. *et al.* Human protein reference database - 2006 update. *Nucleic Acids Res.* **34** (2006).
- [265] Hermjakob, H. *et al.* Intact: an open source molecular interaction database. *Nucleic Acids Res.* **32** (2004).
- [266] Pagel, P. *et al.* The mips mammalian protein-protein interaction database. *Bioinformatics* **21** (2005).
- [267] Guldener, U. *et al.* Mpaact: the mips protein interaction resource on yeast. *Nucleic Acids Res.* **34** (2006).
- [268] Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33** (2005).
- [269] Camon, E. *et al.* The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Res* **32 Database issue**, D262--6 (2004).
- [270] Ponten, F., K., J. & Uhlen, M. The human protein atlas - a tool for pathology. *J. Pathol.* **216** (2008).

- [271] Chemaxon. Marvin, version 5.3. URL <http://www.chemaxon.com>.
- [272] Pafilis, E. *et al.* Reflect: augmented browsing for the life scientist. *Nat. Biotechnol.* **27** (2009).
- [273] Chamba, A. *et al.* Slc6a4 expression and anti-proliferative responses to serotonin transporter ligands chlomipramine and fluoxetine in primary b-cell malignancies. *Leuk Res* **34**, 1103--6 (2010).
- [274] Halden, R. Plastics and health risks. *Annu. Rev. Public Health.* **31** (2010).





# Appendices



# Supplementary Information to Paper I

## Supplementary Text

### Dictionary generation

The dictionary is based on the Danish translation of the WHO International Classification of Diseases (ICD10), downloaded from the Danish national board of health 2.Nov 2009. ICD10 is divided into 22 chapters, and has a hierarchical structure with increased specification of terms in each lower level. Each term is uniquely matched to code of between 3 and 5 characters.

The core of the dictionary consists of all ICD10 terms in their original form in UTF8 format and uppcased. This is a total of 22261 unique terms matched 1:1 with an ICD10 code. In addition to this, a number of permutations of the core terms are created, with each created term variant pointing to the same ICD10 code as the term it was derived from. These permutations are:

1. Comma permutation. Many terms contain a comma, and in a number of these the term structure is such, that the term maintains its clinical meaning by swapping the right and left side of the comma, or by keeping only what is on the left side. Example: A060 AMØBEDYSENTERI, AKUT -> A060 AKUT AMØBEDYSENTERI and A060 AMØBEDYSENTERI
2. Abbreviations. Terms containing a number of standard abbreviations are added in a version where the abbreviated word is written in full. Example: B029 HERPES ZOSTER U KOMPLIKATION-> B029 HERPES ZOSTER UDEN KOMPLIKATION
3. Parenthesis. Some terms contain a parenthesis, which typically contain some further specification of the term. A variant with the parenthesis deleted is added to the dictionary. Example: A00 KOLERA FORÅRSAGET AF VIBRIO CHOLERAЕ (KLASSISK KOLERA) -> A00 KOLERA FORÅRSAGET AF VIBRIO CHOLERAЕ
4. Typical expressions. Seven typical expressions used as specifiers, indicators of causative agent or to express lack of were identified in many terms. Again for our

purposes this is typically not relevant, so a variant of these terms, with the expression and whatever follows, removed, is added to the dictionary. Example: B059 MÆSLINGER UDEN SPECIFIKATION -> B059 MÆSLINGER

5. ICD10 codes as terms. The ICD10 codes themselves are added as terms for their own code.

Finally all special characters are removed from the terms. Variants are created in an iterative way, in the order indicated above, such that type 3 permutation are also performed on the variations already created by type 1 and type 2 permutations. Some of the permutations result in addition of nonsense terms to the dictionary, but the very fact that they are nonsense terms makes them harmless. A more serious problem is created variants that are sensible terms, but that have lost the actual clinical meaning of the original term. Example: F019 DEMENS, VASKULÆR UDEN SPECIFIKATION -> F019 VASKULÆR UDEN SPECIFIKATION DEMENS -> F019 VASKULÆR. The first permutation is a comma permutation that leads to a nonsense term, which in the next permutation produces a term with single word vaskulær (vascular) pointing to F019. This is a simple word with no diagnostic or symptomatic meaning by itself, and can definitely not be said to be a synonym of F019. This type of variant is typically very short, so all dictionary terms of one and two words have been manually curated for this type of terms. They are not actually deleted from the dictionary but are added to a blacklist

The objective of dictionary construction is to get as many terms (sensible) as possible for each ICD10 codes. Essentially they are synonyms of the original term. So the goal is to change the 1:1 relationship of ICD10 code to Term into a 1:many relationship. The permutation procedure inevitably results in many situations where identical terms are created, that point to different ICD10 codes. In order to fix this, the complete dictionary is processed back into a 1:many structure by enforcing the following rules when two or more instances of the same term point to different ICD10 codes:

1. If the term is an original part of ICD10, then only the version pointing to the original ICD10 code is kept in the dictionary. Example:

DA37 KIGHOSTE

DA370 KIGHOSTE FORÅRSAGET AF BORDETELLA PERTUSSIS -> DA370 KIGHOSTE

Two codes with the same term, but since DA37 KIGHOSTE is an original, DA370 KIGHOSTE is deleted

1. If the identical terms are all variations resulting from permutations, then it is checked if the ICD10 codes become identical at a higher level in the hierarchy, going to a maximum of level 3. If this is the case, one copy of the term is kept, and set to point to this ICD10 code. These contributions to the dictionary are counted as mixed types. If no common ICD10 stem is found, all the variations are removed. Example:

---

A395 HJERTESYGDOM FORÅRSAGET AF MENINGOKOKKER -> A395 HJERTESYGDOM

I518 HJERTESYGDOM ANDRE DÅRLIGT DEFINEREDE -> I518 HJERTESYGDOM

I519 HJERTESYGDOM UDEN SPECIFIKATION -> I519 HJERTESYGDOM

Permutations create 3 (actually more) identical terms HJERTESYGDOM pointing to 3 different codes. Ignoring the first term, the last two share a 3 characters stem A51 that would have resulted in I51 HJERTESYGDOM to be added to the dictionary. However since this stem is not shared with the top candidate A395, which points too a whole different chapter, the term HJERTESYGDOM is not added to the dictionary at all.

1. (Actually the original ICD10 classification does in fact contain a few cases of the same term pointing to two different codes. In these instances, only the term pointing to the lowest level in the classification is kept.)

As a final addition to the dictionary, a number of terms have been added manually. These are obvious variants of frequent diseases missing from the dictionary that have been discovered during the process of working with the data, or variations with clinical sense that point to a wrong ICD10 code. This accounts for approximately 50 dictionary entries. The final dictionary consists of 53452 terms.

### **Extracting ICD10 codes from patient records**

The corpus was parsed in units of the individual text entry. Hits consist in a 1:1 match between candidate strings in the corpus and the complete dictionary. After tokenizing the text, a stepping algorithm is used to move through the tokens, creating candidate strings from each token, by joining up to the 10 following tokens.

Candidates that match a dictionary term, and do not match any blacklist term, are then checked for negations or family mention in the preceding tokens of the current sentence. We consider the negations 'aldrig', 'ingen', 'intet', 'uden' ('Never', 'None', 'No', 'Without') and disregard candidates with any of these within the preceding sentence. We similarly look for any mention of family members and relatives in the preceding sentence. Words like 'kone', 'søn', 'kæreste', ('Wife', 'Son', 'Girlfriend') will also disqualify a candidate term, since very often the hit will pertain to the family member and not the patient.

Using this dictionary approach, we get hits covering a range of scenarios ranging from very specific phenotypes like 'PARANOID SKIZOFRENI' and 'SOCIAL FOBI' (paranoid schizophrenia and social fobia), to more general ones such as 'HOVEDPINE' (headache).

The majority of the hits fall in the general category. This category is generally characterized by terms consisting of only one word, typically describing a disease or symptom in its most generic form. Since created term variants point to the same ICD10 code as the term it is derived from, the permutation process will create some terms that point to an ICD10 code with an original meaning that is more specific than the variant term can support.

That means a very specific code can be assigned to a hit with only a generic description of disease or symptom.

To deal with this, while also creating a more homogenous data material, all mined codes were converted to level 3 in the ICD10 classification. All subterms are treated as synonyms for the upper term. In some cases a very specific code could have been mined from a correct context, thus resulting in lost information, but in the more frequent case of a generic context, it increases the precision of the mined term. Furthermore, level 3 also represents a good general distinction of different diseases.

We text mined a total of 218963 hits, where a candidate in the corpus matched a dictionary entry, and was not disqualified by the blacklist, negations or family mention. Negations and family mention disqualified around 10% of candidates.

The hits cover a total of 1229 different dictionary terms pointing to 1064 different ICD10 codes. Rounding up to level 3 reduced the number ICD10 codes from 1064 to 677. Out of the 5543 patients in the Sct. Hans database, terms were mined for 3259 patients. The remaining patients either had no text entries, or the text entries were too few and/or short to generate any hits.

In addition to the mined ICD10 codes, assigned codes were also extracted from structured fields in the EPR system. 31734 assigned codes were found for 2803 patients. Adding assigned and mined data, we find ICD10 codes for a total of 3290 patients. Of these, 66 patients have contributions only from assigned codes, 487 have contributions only from mining, and the remaining 2737 have contributions from both. Counting each code once per patient and distinguishing between assigned and mined codes, we found the following number of codes (Averages are all based on 3290 patients):

	Number of ICD10 codes retrieved	Average retrieved codes per patient
Assigned only:	4974	1.51
Mined only:	31662	9.62
Assigned-mined overlap	3798	1.15
Total	40434	12.28

**Table 1:** ICD10 code contribution from physician assignment and text mining.

The percentage of assigned codes that are also recovered by mining the text is 43%. For each patient a vector is created with a unique list of all codes, assigned or mined, associated with this patient. As seen from the table, mining adds almost 10 additional terms to the 2.6 terms coming from assigned codes.

### Validation of textmining

The precision of our textmining was investigated by manually checking all 2724 mining hits for 48 patients. The validation set covered 214 full level ICD10 codes, corresponding

---

to I51 level 3 codes. A hit was considered a correctly mined association when it was possible from the immediate record context to see directly (or deduce with good certainty) a clinical link between the term and the patient. We defined precision in two ways: Incidence precision of all curated hits, and association precision where an ICD10 code is considered correctly associated with a patient if it has at least one correct incidence. In both cases we also considered how precision was distributed among the different chapters. The low precision in chapter 1 (Infectious and parasitic diseases) is largely caused by a number of false associations of the term 'AIDS', which mostly comes from somatic delusions. In chapter 19 (Injury, poisoning and certain other consequences of external causes) which includes the term 'bivirkninger' (sideeffects), a standard expression stating that the patient has been informed of possible sideeffects, cause many false associations, resulting in the low precision. We found the total incidence precision to be 87.78% and the association precision to be 84.03%. The 333 false associations were further subdivided into subcategories with this distribution: Negations:105, Wrong individual:17, Delusion:9, Putative:40, Polysemi: 10, Patient information: 92, Other:60.

Negations cover negation constructs not caught by our negation detection. Eg. 'Investigated for dyslexia, but nothing was found'. Wrong individual covers cases where the association is really to another person. Eg. 'An acquaintance of his recently died from a heartattack' Delusions cover cases where patients are delusional about a disorder. Eg. 'pt. is paranoid about contacting AIDS' Putative covers cases where the clinical link is vague, speculative or awaiting confirmation. Eg. 'pt should be examined by an ophthalmologist on suspicion of glaucoma'. Polysemic covers cases where a term is used in a nonclinical context. In the sample the ambiguity came from a geographical location that is also a clinical term. Patient information covers information delivered to patient where the clinical association is eg. a future issue. Eg. 'pt. has been informed of the possible sideeffects of the new drug'. Other covers any other false association.

## Supplementary Table 1 to Paper I

**Table S1:** The final list of 802 candidate ICD10 pairs, out of the full list of possible pairs, resulting from ranking based on p-values and filtering based on Co-occurrence score followed by a False Discovery Rate cutoff of 1% - corrected p-value.

<i>ICD10 A</i>	<i>ICD10 B</i>	<i>#pt.'s A</i>	<i>#pt.'s B</i>	<i>#pt.'s A∩B</i>	<i>Co-occurrence score</i>	<i>p-value</i>	<i>corrected p-value</i>
F11	K77	350	293	140	2.13189778	1.10E-67	1.32E-63
F11	F14	350	201	116	2.386034582	1.03E-66	1.06E-62
F06	F07	222	183	92	2.800571645	2.67E-65	2.42E-61
F19	K77	441	293	149	1.897025061	5.58E-61	4.09E-57
F11	F19	350	441	163	1.775149744	2.59E-60	1.78E-56
J00	R05	578	852	314	1.063838281	1.69E-59	1.06E-55
F07	R41	183	254	91	2.604387085	1.15E-57	6.49E-54
F06	R41	222	254	98	2.448316922	2.33E-56	1.20E-52
B18	K77	58	293	55	3.183172232	2.94E-56	1.48E-52
L29	L30	618	398	203	1.429040079	6.78E-56	3.34E-52
R05	R53	852	564	299	1.02858564	7.67E-53	3.34E-49
J10	R53	756	564	278	1.095110225	8.84E-53	3.78E-49
J00	J10	578	756	280	1.070308495	3.52E-51	1.43E-47
F40	F41	394	610	190	1.366968428	6.52E-48	2.51E-44
R05	R12	852	517	271	1.011869091	1.10E-45	4.03E-42
H53	R07	453	493	176	1.361568121	9.91E-43	3.12E-39
F11	F13	350	291	114	1.847409805	1.36E-42	4.24E-39
J00	R53	578	564	220	1.142810304	7.53E-42	2.22E-38
F11	F12	350	665	176	1.302805581	3.50E-41	9.80E-38
F14	F15	201	123	57	2.76804399	4.29E-39	1.16E-35
F14	F19	201	441	99	1.839464245	1.83E-37	4.67E-34
B17	K77	77	293	51	2.726379118	1.25E-36	3.08E-33
F12	F14	665	201	119	1.527420052	2.77E-36	6.68E-33
J42	J45	129	297	65	2.383866043	5.17E-36	1.24E-32
R12	R53	517	564	193	1.114026045	2.70E-34	6.00E-31
J45	R05	297	852	164	1.082528759	1.11E-29	2.19E-26
F18	R41	76	254	43	2.679651402	1.82E-29	3.54E-26
F10	G62	1149	98	87	1.320882032	2.81E-29	5.39E-26
F12	F15	665	123	81	1.664806678	3.28E-29	6.21E-26
B18	F19	58	441	45	2.390250285	4.25E-29	7.96E-26
F01	F06	67	222	38	2.820479766	2.09E-28	3.80E-25
J42	R05	129	852	92	1.434541737	3.12E-28	5.62E-25
F19	Z04	441	469	145	1.19284933	1.55E-27	2.71E-24
F14	K77	201	293	70	1.909386388	9.68E-27	1.65E-23
M11	M19	109	165	40	2.664552362	5.40E-26	8.81E-23
E10	E16	52	32	18	3.657421229	5.76E-26	9.33E-23
K25	R12	283	517	114	1.338601626	6.02E-26	9.68E-23
B18	F11	58	350	39	2.479912165	2.28E-25	3.49E-22
R07	R30	493	115	64	1.83392552	4.94E-25	7.41E-22
E10	E14	52	113	25	3.222235016	3.63E-24	5.12E-21
R05	R50	852	131	88	1.349560915	4.73E-24	6.62E-21
E11	E14	94	113	31	2.919829651	6.97E-24	9.58E-21
F01	F07	67	183	32	2.803546253	1.23E-23	1.66E-20
F22	Z04	561	469	161	1.00049273	2.42E-23	3.14E-20
B34	R05	83	852	64	1.530885061	6.22E-23	7.83E-20
L30	R46	398	624	153	1.009645201	1.21E-22	1.50E-19
F06	I69	222	39	26	2.894277848	3.25E-22	3.95E-19
F20	G24	1414	105	92	1.011598621	4.72E-22	5.66E-19
K77	R12	293	517	110	1.238512087	5.59E-22	6.60E-19
E66	R53	388	564	139	1.05216212	1.23E-21	1.42E-18
E66	R12	388	517	131	1.090862799	2.14E-21	2.44E-18
J00	L30	578	398	143	1.021759852	2.68E-21	3.04E-18
F12	K77	665	293	126	1.07643233	7.59E-21	8.44E-18
L30	R60	398	178	67	1.593486974	1.08E-19	1.13E-16
B34	R53	83	564	50	1.743716636	4.33E-19	4.42E-16
F01	R41	67	254	32	2.418505443	4.67E-19	4.74E-16
L29	R60	618	178	84	1.303550872	4.85E-19	4.89E-16
L30	R12	398	517	128	1.0215692	9.26E-19	9.13E-16
J45	R53	297	564	111	1.109295235	1.02E-18	1.00E-15
J10	R50	756	131	76	1.307845169	1.14E-18	1.11E-15
K77	R53	293	564	109	1.102482958	3.41E-18	3.25E-15
L29	L50	618	65	44	1.768324411	3.47E-18	3.29E-15



<i>ICD10 A</i>	<i>ICD10 B</i>	<i>#pt.'s A</i>	<i>#pt.'s B</i>	<i>#pt.'s A&amp;B</i>	<i>Co-occurrence score</i>	<i>p-value</i>	<i>corrected p-value</i>
K77	R46	293	624	116	1.048347519	4.06E-18	3.83E-15
B15	K77	45	293	27	2.483235946	5.55E-18	5.14E-15
I64	I69	56	39	15	3.265492149	7.25E-18	6.68E-15
Ro5	R60	852	178	98	1.071822503	1.61E-17	1.43E-14
M19	M54	165	654	80	1.260924609	1.85E-17	1.62E-14
G93	R41	45	254	25	2.538821536	3.93E-17	3.38E-14
L29	R26	618	288	112	1.036238341	3.93E-17	3.38E-14
M53	M54	23	654	23	2.106757799	5.32E-17	4.49E-14
F22	K77	561	293	106	1.070134032	7.74E-17	6.43E-14
G24	R44	105	901	69	1.23420733	9.02E-17	7.39E-14
F13	F60	291	635	114	1.008413355	9.45E-17	7.68E-14
N30	Ro5	136	852	80	1.161157154	1.06E-16	8.58E-14
F13	F41	291	610	111	1.027191638	1.07E-16	8.61E-14
F15	F19	123	441	53	1.626657291	1.11E-16	8.96E-14
J45	L29	297	618	113	1.005348945	1.93E-16	1.52E-13
B17	F19	77	441	40	1.856587263	2.77E-16	2.17E-13
E66	L30	388	398	101	1.089346005	4.28E-16	3.29E-13
B34	J00	83	578	47	1.623175937	4.79E-16	3.67E-13
G43	G44	210	196	46	1.798564922	4.85E-16	3.70E-13
Ro5	R39	852	180	96	1.026599467	6.10E-16	4.62E-13
G24	G47	105	990	71	1.143313271	8.20E-16	6.12E-13
F13	K77	291	293	69	1.378901323	8.58E-16	6.38E-13
R26	R53	288	564	103	1.045905976	8.85E-16	6.54E-13
F13	F14	291	201	55	1.576351188	1.28E-15	9.43E-13
E66	R60	388	178	60	1.47182405	1.45E-15	1.06E-12
K77	Z04	293	469	92	1.120696103	1.90E-15	1.36E-12
I69	R41	39	254	22	2.519620762	2.19E-15	1.55E-12
M40	M54	60	654	40	1.665230639	2.29E-15	1.62E-12
N30	R53	136	564	62	1.373547619	2.55E-15	1.79E-12
N30	R30	136	115	28	2.333466073	2.67E-15	1.87E-12
B17	F11	77	350	35	1.969626351	2.79E-15	1.94E-12
B17	B18	77	58	17	2.932699782	2.82E-15	1.96E-12
F11	F15	350	123	45	1.707463401	5.11E-15	3.50E-12
G43	M54	210	654	89	1.074180099	9.59E-15	6.33E-12
F91	Z04	112	469	49	1.559285246	1.10E-14	7.23E-12
Fo6	I64	222	56	24	2.387230924	1.51E-14	9.82E-12
L30	R23	398	81	37	1.815130526	2.52E-14	1.60E-11
F10	H55	1149	80	61	1.09924269	2.54E-14	1.61E-11
R53	R60	564	178	72	1.2118905	2.79E-14	1.76E-11
R33	R39	57	180	22	2.481428594	3.22E-14	2.02E-11
B17	R53	77	564	42	1.59844573	4.58E-14	2.83E-11
K25	R10	283	70	29	2.095157233	5.24E-14	3.22E-11
Fo1	I64	67	56	15	2.902102357	7.14E-14	4.36E-11
L30	L40	398	100	41	1.681123813	8.88E-14	5.38E-11
Ro5	R13	852	113	66	1.146595508	9.94E-14	5.92E-11
F32	F33	219	293	55	1.449546252	1.01E-13	5.99E-11
R12	R26	517	288	93	1.022984598	1.24E-13	7.29E-11
B17	R44	77	901	52	1.262779597	1.27E-13	7.43E-11
F13	F33	291	293	65	1.294012425	1.63E-13	9.51E-11
R12	R60	517	178	67	1.230903921	1.75E-13	1.02E-10
Fo1	I61	67	128	20	2.541643224	1.79E-13	1.04E-10
L30	Ro6	398	368	92	1.030795667	2.00E-13	1.16E-10
R30	Z60	115	160	28	2.137110431	2.14E-13	1.23E-10
I61	R41	128	254	37	1.80404682	2.32E-13	1.33E-10
J10	R60	756	178	84	1.02043934	2.32E-13	1.33E-10
M41	M54	37	654	28	1.795338543	2.78E-13	1.58E-10
G24	Z04	105	469	45	1.526442546	3.63E-13	2.05E-10
K25	Ro7	283	493	88	1.035877922	4.05E-13	2.28E-10
G43	J00	210	578	79	1.078045145	6.27E-13	3.48E-10
F22	G24	561	105	49	1.403217075	7.17E-13	3.95E-10
M17	M19	17	165	12	2.810901081	8.52E-13	4.68E-10
F14	Z04	201	469	67	1.19734736	1.04E-12	5.70E-10
K77	R60	293	178	47	1.510090646	1.17E-12	6.37E-10
J18	Ro5	26	852	24	1.692803593	1.21E-12	6.57E-10
N30	R31	136	55	18	2.537068753	1.59E-12	8.55E-10
N30	R32	136	79	21	2.36666484	1.86E-12	9.99E-10
M54	R60	654	178	75	1.06271165	2.25E-12	1.20E-09
L29	R23	618	81	43	1.440156922	2.29E-12	1.21E-09
B15	R46	45	624	30	1.70097027	2.31E-12	1.22E-09
D50	E61	24	41	9	2.944428771	3.26E-12	1.71E-09
Fo6	I61	222	128	33	1.81886645	3.62E-12	1.89E-09
F91	R26	112	288	35	1.736397268	3.87E-12	2.02E-09

<i>ICD10 A</i>	<i>ICD10 B</i>	<i>#pt.'s A</i>	<i>#pt.'s B</i>	<i>#pt.'s A∩B</i>	<i>Co-occurrence score</i>	<i>p-value</i>	<i>corrected p-value</i>
B17	Z04	77	469	36	1.627307379	4.14E-12	2.15E-09
F07	T90	183	9	9	2.736381034	4.22E-12	2.17E-09
J45	J98	297	34	19	2.297147128	4.39E-12	2.25E-09
F01	I69	67	39	12	2.85707896	4.67E-12	2.39E-09
R05	R30	852	115	64	1.078392351	5.04E-12	2.56E-09
B24	Z04	120	469	47	1.406536031	5.59E-12	2.84E-09
F06	G40	222	337	57	1.288744048	6.39E-12	3.22E-09
L30	R26	398	288	75	1.084423887	6.57E-12	3.30E-09
R06	R26	368	288	71	1.11620622	8.84E-12	4.38E-09
E14	E16	113	32	13	2.737592172	9.30E-12	4.60E-09
L30	N30	398	136	46	1.429245262	9.52E-12	4.69E-09
F07	F62	183	41	17	2.455988561	1.05E-11	5.17E-09
G24	R46	105	624	50	1.285966587	1.12E-11	5.46E-09
L29	N30	618	136	59	1.17644059	1.26E-11	6.13E-09
F07	G40	183	337	50	1.36901099	1.28E-11	6.21E-09
F13	F19	291	441	80	1.017691709	1.52E-11	7.30E-09
G24	R53	105	564	47	1.337034987	1.72E-11	8.27E-09
B15	F19	45	441	25	1.886522108	1.93E-11	9.20E-09
B34	J10	83	756	47	1.257825562	2.12E-11	1.01E-08
B24	R44	120	901	67	1.005815503	2.13E-11	1.01E-08
J10	N30	756	136	66	1.054814346	2.23E-11	1.06E-08
R44	T73	901	113	64	1.024795358	2.64E-11	1.24E-08
B95	L29	62	618	35	1.509292896	2.72E-11	1.27E-08
F04	R41	22	254	14	2.474743471	3.46E-11	1.61E-08
E11	I10	94	282	30	1.775140297	4.88E-11	2.23E-08
B18	F14	58	201	20	2.208524443	4.91E-11	2.24E-08
E51	F10	26	1149	25	1.366981183	5.55E-11	2.52E-08
E14	R73	113	52	15	2.521795298	6.65E-11	3.00E-08
J00	R50	578	131	54	1.195520461	7.13E-11	3.21E-08
M54	R50	654	131	58	1.125580882	7.99E-11	3.58E-08
J10	T73	756	113	57	1.104913642	8.11E-11	3.63E-08
F15	R46	123	624	54	1.176761986	8.44E-11	3.76E-08
R05	T73	852	113	61	1.034702628	9.26E-11	4.12E-08
R50	R53	131	564	53	1.202932107	9.54E-11	4.23E-08
B24	R46	120	624	53	1.184431953	9.69E-11	4.29E-08
F50	Z50	183	119	27	1.877725065	1.00E-10	4.43E-08
B90	R05	62	852	40	1.265350737	1.24E-10	5.46E-08
J00	J42	578	129	53	1.190312889	1.26E-10	5.54E-08
E51	H55	26	80	10	2.752597114	1.58E-10	6.85E-08
J42	R53	129	564	52	1.197207579	1.71E-10	7.35E-08
B15	B17	45	77	12	2.662571533	1.74E-10	7.47E-08
K76	K77	25	293	15	2.310055118	1.80E-10	7.70E-08
A46	R60	49	178	17	2.301608111	1.95E-10	8.29E-08
B34	R50	83	131	19	2.215960695	2.05E-10	8.67E-08
L29	L40	618	100	46	1.248312427	2.43E-10	1.02E-07
G44	J00	196	578	70	1.002682957	2.49E-10	1.04E-07
F06	F18	222	76	23	1.969483234	2.71E-10	1.13E-07
B34	L30	83	398	32	1.579630527	3.76E-10	1.56E-07
F04	F18	22	76	9	2.729093944	4.52E-10	1.87E-07
L29	R30	618	115	50	1.174058061	4.70E-10	1.93E-07
I10	I61	282	128	34	1.5477596	6.39E-10	2.60E-07
F07	I69	183	39	15	2.335835356	6.65E-10	2.70E-07
J00	T14	578	53	29	1.540743796	7.58E-10	3.08E-07
R11	R26	79	288	26	1.770206741	7.76E-10	3.14E-07
L89	R05	36	852	27	1.43959299	8.04E-10	3.25E-07
J98	R05	34	852	26	1.461389997	8.12E-10	3.27E-07
B17	B24	77	120	17	2.240698076	9.02E-10	3.62E-07
J00	R13	578	113	47	1.202829294	9.25E-10	3.70E-07
E11	R73	94	52	13	2.493694443	9.29E-10	3.71E-07
B34	M54	83	654	41	1.263109581	9.65E-10	3.85E-07
J10	R13	756	113	55	1.054287568	1.01E-09	4.01E-07
L29	R50	618	131	54	1.102876765	1.02E-09	4.04E-07
B17	R46	77	624	38	1.321534615	1.03E-09	4.07E-07
B23	B24	12	120	8	2.646172402	1.09E-09	4.30E-07
R05	R10	852	70	42	1.168676301	1.18E-09	4.60E-07
J42	L30	129	398	41	1.338730673	1.25E-09	4.84E-07
J45	R50	297	131	35	1.488944056	1.26E-09	4.88E-07
F07	G91	183	16	10	2.541068586	1.33E-09	5.17E-07
R53	T73	564	113	46	1.206113602	1.42E-09	5.47E-07
M54	N30	654	136	57	1.048841818	1.57E-09	6.04E-07
N30	R39	136	180	27	1.729987234	1.70E-09	6.53E-07
E11	E78	94	89	16	2.262549548	1.76E-09	6.73E-07

<i>ICD10 A</i>	<i>ICD10 B</i>	<i>#pt.'s A</i>	<i>#pt.'s B</i>	<i>#pt.'s A∩B</i>	<i>Co-occurrence score</i>	<i>p-value</i>	<i>corrected p-value</i>
G24	L29	105	618	46	1.181399741	1.80E-09	6.85E-07
B17	K76	77	25	9	2.657348426	1.98E-09	7.50E-07
F07	F18	183	76	20	2.006236112	2.05E-09	7.74E-07
L30	R50	398	131	41	1.31786196	2.10E-09	7.92E-07
R26	R60	288	178	41	1.34079193	2.19E-09	8.23E-07
F15	K77	123	293	33	1.508028826	2.43E-09	9.09E-07
R11	R53	79	564	36	1.347214537	2.49E-09	9.26E-07
B24	R12	120	517	45	1.211975805	2.63E-09	9.75E-07
L29	T73	618	113	48	1.140524343	2.78E-09	1.02E-06
R50	R60	131	178	26	1.739186918	2.78E-09	1.02E-06
B18	F12	58	665	32	1.374981297	3.18E-09	1.16E-06
B17	B34	77	83	14	2.349822101	3.25E-09	1.18E-06
L30	R39	398	180	50	1.16304138	3.32E-09	1.20E-06
L30	T73	398	113	37	1.373139515	3.37E-09	1.22E-06
B24	F22	120	561	47	1.161249473	3.53E-09	1.27E-06
R39	T73	180	113	24	1.79939605	3.63E-09	1.30E-06
F06	I10	222	282	46	1.230601235	3.99E-09	1.43E-06
K25	N30	283	136	34	1.462699077	4.02E-09	1.43E-06
B86	L29	29	618	20	1.703596258	4.16E-09	1.48E-06
I48	I49	19	13	5	2.480523866	4.54E-09	1.61E-06
R46	R50	624	131	53	1.063007187	4.72E-09	1.66E-06
L21	L30	18	398	13	2.139459347	4.80E-09	1.69E-06
L89	R33	36	57	9	2.622635703	4.81E-09	1.69E-06
J10	Z25	756	18	16	1.726769827	5.07E-09	1.78E-06
B24	K77	120	293	32	1.49757001	5.12E-09	1.79E-06
K77	R17	293	26	14	2.177663479	5.30E-09	1.85E-06
J42	L29	129	618	52	1.07076004	5.44E-09	1.89E-06
B95	R46	62	624	32	1.370920176	5.44E-09	1.89E-06
R11	R50	79	131	17	2.118346619	5.56E-09	1.93E-06
K25	R50	283	131	33	1.470588923	5.65E-09	1.95E-06
J98	R53	34	564	21	1.687847828	5.94E-09	2.04E-06
M54	T73	654	113	49	1.091564341	6.08E-09	2.08E-06
B15	R44	45	901	31	1.264076241	6.34E-09	2.16E-06
F07	G93	183	45	15	2.191392738	6.78E-09	2.30E-06
I50	R18	19	29	6	2.583960542	6.80E-09	2.30E-06
B17	R12	77	517	33	1.375967935	8.09E-09	2.70E-06
F98	Z04	72	469	30	1.460570777	8.46E-09	2.82E-06
G24	R25	105	37	11	2.460071253	8.84E-09	2.93E-06
F98	R53	72	564	33	1.349467118	9.88E-09	3.26E-06
B17	R60	77	178	19	1.952892335	1.07E-08	3.52E-06
J44	J45	35	297	16	2.031026896	1.10E-08	3.59E-06
F03	R41	35	254	15	2.111645356	1.10E-08	3.60E-06
J00	T73	578	113	45	1.141428749	1.12E-08	3.64E-06
F15	Z04	123	469	42	1.214159071	1.14E-08	3.70E-06
J45	K77	297	293	56	1.054147783	1.15E-08	3.75E-06
E66	N30	388	136	40	1.26679122	1.17E-08	3.80E-06
M10	M14	19	6	4	2.272784642	1.18E-08	3.83E-06
F06	G93	222	45	16	2.074367184	1.21E-08	3.91E-06
A46	B95	49	62	10	2.515769603	1.24E-08	3.99E-06
B18	Z04	58	469	26	1.54261621	1.37E-08	4.37E-06
B24	R53	120	564	46	1.123539034	1.40E-08	4.46E-06
F84	J00	92	578	39	1.220705148	1.42E-08	4.50E-06
I10	R41	282	254	49	1.134703293	1.44E-08	4.58E-06
F07	I61	183	128	25	1.679003194	1.48E-08	4.67E-06
B17	F22	77	561	34	1.30861518	1.65E-08	5.20E-06
B17	M54	77	654	37	1.220562614	1.71E-08	5.36E-06
E16	R73	32	52	8	2.579418717	1.76E-08	5.47E-06
F62	R41	41	254	16	2.029025274	1.79E-08	5.55E-06
H10	L29	47	618	26	1.457905764	1.90E-08	5.89E-06
B34	R12	83	517	34	1.317518428	1.93E-08	5.97E-06
F98	R12	72	517	31	1.377738958	2.05E-08	6.31E-06
R06	R39	368	180	46	1.153112751	2.06E-08	6.34E-06
J42	R06	129	368	37	1.300338091	2.13E-08	6.54E-06
R06	T73	368	113	34	1.359562717	2.20E-08	6.74E-06
R26	R33	288	57	20	1.80984195	2.24E-08	6.87E-06
F06	F62	222	41	15	2.08675044	2.26E-08	6.90E-06
B24	R50	120	131	20	1.86171838	2.45E-08	7.48E-06
B34	J45	83	297	25	1.614215552	2.51E-08	7.62E-06
G47	J98	990	34	26	1.265472627	2.73E-08	8.27E-06
B17	R50	77	131	16	2.063867725	2.78E-08	8.39E-06
E88	G47	50	990	34	1.12517784	2.84E-08	8.57E-06
K59	R46	31	624	20	1.609985344	2.86E-08	8.62E-06

<i>ICD10 A</i>	<i>ICD10 B</i>	<i>#pt.'s A</i>	<i>#pt.'s B</i>	<i>#pt.'s A∩B</i>	<i>Co-occurrence score</i>	<i>p-value</i>	<i>corrected p-value</i>
E51	G93	26	45	7	2.561043874	2.99E-08	8.98E-06
N30	R60	136	178	25	1.637272516	3.01E-08	9.03E-06
J00	M40	578	60	29	1.378190086	3.21E-08	9.60E-06
G91	R41	16	254	10	2.298990026	3.35E-08	1.00E-05
A49	R05	40	852	27	1.301633847	3.37E-08	1.00E-05
A41	R33	32	57	8	2.533560425	3.74E-08	1.11E-05
J00	N30	578	136	50	1.03475661	3.82E-08	1.13E-05
B17	R10	77	70	12	2.300832259	4.05E-08	1.19E-05
R12	R13	517	113	41	1.162949239	4.23E-08	1.24E-05
L20	L30	10	398	9	2.178060315	4.57E-08	1.34E-05
M54	R02	654	17	14	1.776179998	5.00E-08	1.45E-05
G24	K77	105	293	28	1.486273852	5.00E-08	1.45E-05
J45	N30	297	136	33	1.356583418	5.30E-08	1.53E-05
F51	J00	85	578	36	1.215495501	5.37E-08	1.55E-05
R13	R53	113	564	43	1.110956369	5.50E-08	1.58E-05
K25	R26	283	288	52	1.04011776	5.83E-08	1.66E-05
F19	K76	441	25	15	1.878631819	5.87E-08	1.67E-05
A41	R05	32	852	23	1.369760726	5.87E-08	1.67E-05
B95	L30	62	398	24	1.55634176	5.97E-08	1.70E-05
I46	R53	121	564	45	1.08109233	5.98E-08	1.70E-05
R12	R50	517	131	45	1.091557028	6.13E-08	1.74E-05
J40	J45	15	297	10	2.224253964	6.21E-08	1.75E-05
G62	R27	98	47	11	2.321928095	6.67E-08	1.88E-05
B24	L30	120	398	36	1.253701894	7.44E-08	2.08E-05
F98	R11	72	79	12	2.252133206	7.58E-08	2.12E-05
I64	R41	56	254	18	1.835578386	7.72E-08	2.16E-05
F98	R06	72	368	25	1.521964812	7.92E-08	2.21E-05
B95	R60	62	178	16	1.964986483	8.46E-08	2.36E-05
B90	L29	62	618	30	1.293564205	8.85E-08	2.45E-05
F98	R44	72	901	41	1.019509244	9.22E-08	2.55E-05
T73	Z04	113	469	38	1.188759951	9.44E-08	2.61E-05
L29	T00	618	12	11	1.882702432	9.52E-08	2.63E-05
F98	L29	72	618	33	1.227034321	1.04E-07	2.85E-05
B18	R44	58	901	35	1.092349522	1.14E-07	3.11E-05
I25	R07	16	493	12	1.935937127	1.16E-07	3.17E-05
J10	R61	756	43	26	1.311168003	1.21E-07	3.29E-05
R26	R50	288	131	31	1.35990234	1.25E-07	3.38E-05
R06	R60	368	178	44	1.105729842	1.26E-07	3.38E-05
E51	R41	26	254	12	2.111973403	1.26E-07	3.40E-05
A41	R26	32	288	14	1.980429662	1.26E-07	3.40E-05
F84	R46	92	624	39	1.116438604	1.29E-07	3.46E-05
B23	R11	12	79	6	2.442058918	1.31E-07	3.50E-05
B15	F22	45	561	23	1.468389427	1.38E-07	3.69E-05
E66	R50	388	131	37	1.207978479	1.40E-07	3.73E-05
F22	F98	561	72	31	1.269120577	1.48E-07	3.92E-05
L29	M40	618	60	29	1.289766498	1.51E-07	3.99E-05
J18	J98	26	34	6	2.464011906	1.52E-07	4.03E-05
M11	R50	109	131	18	1.831054925	1.54E-07	4.06E-05
I46	R46	121	624	47	1.003036214	1.64E-07	4.31E-05
E78	I10	89	282	24	1.534734467	1.84E-07	4.79E-05
R07	Z60	493	160	49	1.001403912	1.96E-07	5.09E-05
G40	I46	337	121	32	1.300854924	1.99E-07	5.16E-05
G44	M40	196	60	16	1.893858843	2.00E-07	5.18E-05
B17	L29	77	618	34	1.178457259	2.04E-07	5.26E-05
F19	R60	441	178	49	1.008126494	2.09E-07	5.38E-05
B18	F13	58	291	19	1.706019561	2.14E-07	5.50E-05
J00	L40	578	100	39	1.107151344	2.16E-07	5.53E-05
A52	F02	3	21	3	1.972635101	2.24E-07	5.72E-05
R26	R41	288	254	47	1.046756465	2.28E-07	5.82E-05
K50	K63	4	14	3	1.975650139	2.45E-07	6.23E-05
B23	F11	12	350	9	2.13504996	2.61E-07	6.59E-05
F22	R50	561	131	46	1.009998991	2.65E-07	6.69E-05
F15	F22	123	561	44	1.034156625	2.66E-07	6.72E-05
B17	J10	77	756	38	1.060928381	2.73E-07	6.88E-05
F05	R26	89	288	24	1.507848359	2.75E-07	6.91E-05
M06	M19	21	165	9	2.284059909	2.83E-07	7.09E-05
L29	R13	618	113	44	1.017667595	2.87E-07	7.17E-05
T14	T63	53	30	7	2.431206436	2.87E-07	7.18E-05
I44	I45	38	14	5	2.368722306	2.91E-07	7.27E-05
E66	T73	388	113	33	1.24685421	2.94E-07	7.33E-05
A49	R53	40	564	21	1.485426827	2.99E-07	7.44E-05
R10	R53	70	564	30	1.253756592	3.02E-07	7.51E-05

<i>ICD10 A</i>	<i>ICD10 B</i>	<i>#pt.'s A</i>	<i>#pt.'s B</i>	<i>#pt.'s A&amp;B</i>	<i>Co-occurrence score</i>	<i>p-value</i>	<i>corrected p-value</i>
M11	R26	109	288	27	1.409327315	3.21E-07	7.93E-05
G44	R06	196	368	46	1.035839448	3.35E-07	8.27E-05
E11	R60	94	178	19	1.716501492	3.38E-07	8.34E-05
R60	T73	178	113	21	1.628835984	3.52E-07	8.66E-05
A41	B37	32	17	5	2.364199748	3.58E-07	8.79E-05
K51	K52	12	5	3	1.973926488	3.69E-07	9.03E-05
F98	R26	72	288	21	1.590994628	3.76E-07	9.17E-05
B34	Z04	83	469	30	1.272531737	3.83E-07	9.33E-05
A49	B34	40	83	9	2.315365407	3.86E-07	9.39E-05
B17	L30	77	398	26	1.388230465	4.12E-07	1.00E-04
R33	R53	57	564	26	1.325749806	4.32E-07	0.000104783
F15	H53	123	453	38	1.120626698	4.49E-07	0.000108649
E66	E78	388	89	28	1.33491483	4.50E-07	0.000108788
B34	R06	83	368	26	1.392573242	4.53E-07	0.000109354
F60	F61	635	40	22	1.399173477	4.62E-07	0.000111225
L29	T14	618	53	26	1.301287867	4.66E-07	0.000112062
L21	L40	18	100	7	2.370421928	4.70E-07	0.000112799
G24	R12	105	517	37	1.118644497	4.71E-07	0.000112869
K73	K77	6	293	6	2.189730596	4.76E-07	0.000113887
B17	J45	77	297	22	1.532414054	5.06E-07	0.000120761
G43	M40	210	60	16	1.815503206	5.23E-07	0.000124377
B34	L29	83	618	35	1.117606369	5.31E-07	0.000126046
F91	F98	112	72	13	2.020313764	5.38E-07	0.000127747
L30	R13	398	113	33	1.212674843	5.40E-07	0.000127965
J40	J42	15	129	7	2.332656547	6.00E-07	0.000141398
H26	H40	38	16	5	2.340315994	6.24E-07	0.000146694
E11	I11	94	4	4	2.165808893	6.26E-07	0.000146932
B34	R46	83	624	35	1.104503547	6.77E-07	0.000158547
B18	B24	58	120	12	2.060975297	6.88E-07	0.000160767
F65	R46	30	624	18	1.505927857	7.08E-07	0.000164745
L30	M11	398	109	32	1.217996317	7.10E-07	0.000165024
J18	J45	26	297	12	1.957522692	7.16E-07	0.000165997
F60	F90	635	22	15	1.608727025	7.32E-07	0.000169464
B15	B18	45	58	8	2.327297631	7.38E-07	0.000170895
R23	Z04	81	469	29	1.257642063	7.60E-07	0.000175567
L30	L50	398	65	23	1.43713127	7.63E-07	0.000175964
B95	R53	62	564	27	1.267752955	8.16E-07	0.000187453
B18	F15	58	123	12	2.036690218	9.02E-07	0.000205245
R05	R09	852	33	22	1.268681238	9.19E-07	0.000208804
I61	I64	128	56	12	2.031992232	9.32E-07	0.000211543
G24	L30	105	398	31	1.223671974	9.32E-07	0.000211428
A46	E66	49	388	19	1.56091449	9.74E-07	0.000220524
N30	R06	136	368	35	1.150920761	9.76E-07	0.000220708
I46	L30	121	398	34	1.162327511	9.81E-07	0.000221631
E10	R73	52	52	8	2.304493502	9.89E-07	0.00022311
F84	R53	92	564	35	1.101991325	1.01E-06	0.000227331
F02	R41	21	254	10	2.069162025	1.04E-06	0.000234245
I46	R02	121	17	7	2.299357907	1.10E-06	0.00024558
R23	R53	81	564	32	1.148535754	1.12E-06	0.000249909
K59	L29	31	618	18	1.477500088	1.20E-06	0.000266455
F07	I64	183	56	14	1.866035463	1.22E-06	0.000270409
F91	G40	112	337	29	1.266230289	1.25E-06	0.000274473
R12	T73	517	113	38	1.056034035	1.27E-06	0.000279652
B34	E66	83	388	26	1.323471823	1.29E-06	0.000284297
R05	R31	852	55	31	1.069909788	1.30E-06	0.000285802
J45	T73	297	113	27	1.321810642	1.31E-06	0.000286645
B17	R17	77	26	7	2.314274522	1.34E-06	0.000294406
F51	M54	85	654	36	1.047835183	1.37E-06	0.000299049
B95	Z04	62	469	24	1.345447454	1.40E-06	0.000303978
R53	T14	564	53	24	1.309614833	1.43E-06	0.000310913
E86	N30	16	136	7	2.267602124	1.48E-06	0.000321351
R11	Z04	79	469	28	1.241893637	1.50E-06	0.000324602
F98	L30	72	398	24	1.364380378	1.52E-06	0.000329024
K25	R11	283	79	21	1.496801028	1.53E-06	0.000330306
E11	E66	94	388	28	1.262750159	1.57E-06	0.00033658
I61	R07	128	493	40	1.022658622	1.57E-06	0.000336969
I46	R09	121	33	9	2.175482834	1.67E-06	0.000357963
G21	R53	50	564	23	1.326228232	1.70E-06	0.000361717
J45	R10	297	70	20	1.52064152	1.71E-06	0.000363332
R05	T14	852	53	30	1.073978247	1.72E-06	0.000365176
K76	R18	25	29	5	2.297650097	1.74E-06	0.000370256
F98	R46	72	624	31	1.126587677	1.74E-06	0.000370083

<i>ICD10 A</i>	<i>ICD10 B</i>	<i>#pt.'s A</i>	<i>#pt.'s B</i>	<i>#pt.'s A∩B</i>	<i>Co-occurrence score</i>	<i>p-value</i>	<i>corrected p-value</i>
B95	H10	62	47	8	2.254813899	1.77E-06	0.000374793
I50	R05	19	852	15	1.43431394	1.87E-06	0.000394367
R26	R55	288	20	10	1.99960141	1.90E-06	0.00039904
E11	E16	94	32	8	2.233118828	1.97E-06	0.000412689
I46	Z04	121	469	37	1.058187055	2.05E-06	0.000428159
G43	L65	210	38	12	1.924111692	2.07E-06	0.000431026
R12	R30	517	115	38	1.032061209	2.08E-06	0.000433583
F07	F91	183	112	20	1.538364232	2.09E-06	0.000433899
E78	I25	89	16	6	2.288490747	2.13E-06	0.000442905
N30	R33	136	57	12	1.953597701	2.20E-06	0.000455737
M54	R33	654	57	27	1.183172232	2.22E-06	0.000459086
R35	T73	8	113	5	2.234723223	2.26E-06	0.00046606
F21	F50	280	183	35	1.119034335	2.28E-06	0.000469816
R39	R60	180	178	26	1.330153242	2.29E-06	0.000470577
A49	J10	40	756	23	1.235669507	2.30E-06	0.000473015
N30	T73	136	113	17	1.666290142	2.37E-06	0.000486889
R10	R50	70	131	13	1.886210343	2.38E-06	0.000487437
A49	J00	40	578	20	1.38739261	2.39E-06	0.000488255
B18	F22	58	561	25	1.255511696	2.39E-06	0.000488818
G24	R39	105	180	19	1.568177916	2.42E-06	0.000492227
J42	K25	129	283	28	1.261480803	2.42E-06	0.000491813
B95	J10	62	756	31	1.069564618	2.47E-06	0.000502333
I46	R07	121	493	38	1.027515766	2.54E-06	0.00051461
K59	R10	31	70	7	2.269186633	2.55E-06	0.000517355
R18	R53	58	564	25	1.248522153	2.65E-06	0.000535469
F07	I10	183	282	35	1.10938346	2.70E-06	0.000544093
G47	L70	990	27	20	1.202552935	2.83E-06	0.00056716
F50	N91	183	6	5	2.169486559	2.90E-06	0.000580164
G93	R26	45	288	15	1.695719771	2.91E-06	0.000582859
B17	F51	77	85	11	2.005125032	2.96E-06	0.000591988
A46	R05	49	852	28	1.082997721	2.97E-06	0.000592406
R10	T73	70	113	12	1.933100475	2.98E-06	0.000595127
K25	M11	283	109	25	1.325262929	3.00E-06	0.000598541
G43	M11	210	109	21	1.46712601	3.04E-06	0.000605295
B90	R53	62	564	26	1.215285535	3.13E-06	0.000621783
F22	F23	561	66	27	1.192161904	3.22E-06	0.000638994
R11	R12	79	517	29	1.161192265	3.33E-06	0.00065874
A41	N30	32	136	9	2.106065421	3.38E-06	0.000668375
J42	R60	129	178	21	1.46316377	3.44E-06	0.000678
M54	R23	654	81	34	1.03230384	3.46E-06	0.000680925
R02	R26	17	288	9	2.006857012	3.46E-06	0.000681774
R23	R46	81	624	33	1.055104705	3.59E-06	0.000703613
J10	M40	756	60	30	1.067915994	3.61E-06	0.000706796
N30	R10	136	70	13	1.846243935	3.65E-06	0.000713569
R05	R33	852	57	31	1.021704207	3.66E-06	0.000716208
E61	K25	41	283	14	1.728415689	3.70E-06	0.000722415
F25	F31	145	172	22	1.422492316	3.70E-06	0.000722343
I46	L89	121	36	9	2.105310479	3.72E-06	0.000725415
E16	F06	32	222	11	1.925371026	3.77E-06	0.000733211
F13	F32	291	219	40	1.009141334	3.82E-06	0.000743154
R05	R73	852	52	29	1.052270375	3.86E-06	0.000750391
B15	L29	45	618	22	1.282806865	3.92E-06	0.000759224
L89	R07	36	493	17	1.493086931	3.99E-06	0.00077172
J44	R05	35	852	22	1.192454434	4.03E-06	0.000778404
F62	I64	41	56	7	2.23627201	4.05E-06	0.000782214
F21	Z50	280	119	26	1.278809218	4.07E-06	0.000784375
F13	G44	291	196	37	1.051307077	4.10E-06	0.000789728
B24	R60	120	178	20	1.486889265	4.12E-06	0.00079187
A41	R39	32	180	10	1.99960141	4.22E-06	0.000810695
I84	L29	39	618	20	1.334722305	4.26E-06	0.000817056
G24	I46	105	121	15	1.718538497	4.34E-06	0.000832085
K59	L30	31	398	14	1.658916924	4.36E-06	0.00083465
G43	K25	210	283	38	1.032636148	4.37E-06	0.000835609
N30	R50	136	131	18	1.566433817	4.50E-06	0.000859215
K70	R41	15	254	8	2.060193561	4.53E-06	0.000862781
G43	Y06	210	62	15	1.690402707	4.60E-06	0.000875356
L29	R31	618	55	25	1.198197347	4.60E-06	0.000874822
F11	R60	350	178	39	1.004611709	4.89E-06	0.000930504
L89	R26	36	288	13	1.753768172	4.95E-06	0.000940301
B15	R05	45	852	26	1.093423433	5.10E-06	0.000966273
F98	G40	72	337	21	1.393329338	5.15E-06	0.000973489
K25	K26	283	14	8	2.02963366	5.17E-06	0.000975968

<i>ICD10 A</i>	<i>ICD10 B</i>	<i>#pt.'s A</i>	<i>#pt.'s B</i>	<i>#pt.'s A∩B</i>	<i>Co-occurrence score</i>	<i>p-value</i>	<i>corrected p-value</i>
M19	R60	165	178	24	1.332490888	5.18E-06	0.00097663
L21	R06	18	368	10	1.868051944	5.22E-06	0.000984391
B17	R26	77	288	20	1.439904542	5.35E-06	0.001007551
E87	T73	40	113	9	2.07469313	5.41E-06	0.001016458
I85	K74	3	5	2	1.578399813	5.54E-06	0.001040412
H53	R30	453	115	34	1.055947204	5.63E-06	0.001055585
F04	F06	22	222	9	2.008973435	5.67E-06	0.001061555
I10	R00	282	27	11	1.856264523	5.68E-06	0.001062633
J18	J44	26	35	5	2.232660757	5.68E-06	0.00106192
A46	L29	49	618	23	1.233863504	5.69E-06	0.001062341
J10	T30	756	39	22	1.207169684	5.92E-06	0.001102927
F41	F42	610	37	19	1.347365309	5.93E-06	0.001103164
B34	F22	83	561	31	1.078479165	5.96E-06	0.001106819
K25	R13	283	113	25	1.278198536	6.02E-06	0.001116979
K25	T73	283	113	25	1.278198536	6.02E-06	0.001116067
F18	G93	76	45	8	2.141699819	6.05E-06	0.001120239
B95	R06	62	368	20	1.404095489	6.10E-06	0.001127562
E78	R07	89	493	30	1.112577957	6.15E-06	0.001136108
A49	L30	40	398	16	1.541764811	6.23E-06	0.001147802
L65	M54	38	654	20	1.29575205	6.31E-06	0.001159329
E66	G21	388	50	18	1.462030394	6.61E-06	0.00121308
G44	J45	196	297	37	1.023453676	6.72E-06	0.001230391
A46	F19	49	441	19	1.401999782	6.99E-06	0.001278627
A49	H53	40	453	17	1.467799695	7.17E-06	0.001309704
E11	R07	94	493	31	1.084888898	7.20E-06	0.001313065
F06	R40	222	76	17	1.554445734	7.21E-06	0.001313621
I84	R44	39	901	24	1.097820244	7.36E-06	0.001337977
B24	F65	120	30	8	2.103508602	7.41E-06	0.001343649
R13	R39	113	180	19	1.477467955	7.53E-06	0.001364871
K73	K76	6	25	3	1.935679019	7.64E-06	0.001382003
L29	R10	618	70	29	1.084268917	7.84E-06	0.001413822
E66	Z72	388	46	17	1.48624561	7.96E-06	0.001434246
R26	R32	288	79	20	1.407636662	8.18E-06	0.001468283
A36	B24	5	120	4	2.080245524	8.19E-06	0.001468547
I20	R07	10	493	8	1.848874192	8.22E-06	0.001472265
B95	R12	62	517	24	1.218550355	8.26E-06	0.001475674
J45	Z25	297	18	9	1.929652435	8.27E-06	0.001477318
R10	R12	70	517	26	1.169925001	8.44E-06	0.001504
F42	J10	37	756	21	1.211181029	8.51E-06	0.00151346
B24	F19	120	441	34	1.034615691	8.63E-06	0.001530113
G21	R60	50	178	12	1.810901081	8.66E-06	0.001534398
M19	R26	165	288	32	1.095441144	8.70E-06	0.001541157
J10	L50	756	65	31	1.005766943	8.73E-06	0.001543879
F41	F45	610	26	15	1.458815097	9.11E-06	0.001603461
F10	T50	1149	11	11	1.30946629	9.14E-06	0.001607582
F51	K77	85	293	21	1.360151765	9.64E-06	0.001689328
J00	R61	578	43	20	1.295649524	9.79E-06	0.00171365
R30	R39	115	180	19	1.455654426	9.83E-06	0.001718835
I39	K77	11	293	7	2.014765357	1.00E-05	0.001747795
N30	R26	136	288	28	1.16810407	1.02E-05	0.001774478
M11	M17	109	17	6	2.162832352	1.06E-05	0.001841654
L29	L55	618	11	9	1.705447403	1.06E-05	0.001840459
J42	R26	129	288	27	1.187660066	1.07E-05	0.00185771
A41	R31	32	55	6	2.189159118	1.07E-05	0.001856318
L30	R18	398	29	13	1.634801262	1.07E-05	0.001856656
I10	I69	282	39	13	1.688710426	1.08E-05	0.001871903
A46	L89	49	36	6	2.188016842	1.11E-05	0.00190952
R05	T30	852	39	23	1.112514235	1.13E-05	0.001950425
I61	R60	128	178	20	1.405864989	1.14E-05	0.001961775
K25	R60	283	178	33	1.059667736	1.17E-05	0.002004278
J10	R33	756	57	28	1.040575454	1.18E-05	0.002029546
M40	R60	60	178	13	1.72118239	1.21E-05	0.002073963
B18	M54	58	654	26	1.107632488	1.21E-05	0.002076523
G93	I10	45	282	14	1.626782676	1.22E-05	0.002092104
F98	G24	72	105	11	1.863426947	1.23E-05	0.002097738
A41	H40	32	16	4	2.113257147	1.24E-05	0.002113087
R18	R60	29	178	9	1.960722912	1.27E-05	0.002162592
B90	R26	62	288	17	1.485699703	1.27E-05	0.002163779
E16	R05	32	852	20	1.177115648	1.27E-05	0.002163432
F04	F07	22	183	8	2.01695751	1.28E-05	0.002180453
I95	R61	42	43	6	2.176077228	1.29E-05	0.002181059
F25	R06	145	368	34	1.02336655	1.32E-05	0.00223676

<i>ICD10 A</i>	<i>ICD10 B</i>	<i>#pt.'s A</i>	<i>#pt.'s B</i>	<i>#pt.'s A∩B</i>	<i>Co-occurrence score</i>	<i>p-value</i>	<i>corrected p-value</i>
I47	Ro5	23	852	16	1.289156999	1.33E-05	0.002241216
K25	R33	283	57	16	1.526004842	1.33E-05	0.002244874
E87	R00	40	27	5	2.175416805	1.36E-05	0.002300508
B95	E66	62	388	20	1.337147092	1.38E-05	0.002320815
A46	J45	49	297	15	1.560801289	1.41E-05	0.002363517
E66	I46	388	121	31	1.067380458	1.42E-05	0.002390224
G47	R09	990	33	22	1.073328424	1.43E-05	0.002394321
R30	R31	115	55	10	1.912232345	1.43E-05	0.002394895
M11	M13	109	11	5	2.136655988	1.44E-05	0.002405253
L29	R11	618	79	31	1.014543864	1.44E-05	0.002411
H40	J10	16	756	12	1.474980994	1.44E-05	0.002410861
B95	F65	62	30	6	2.160870074	1.46E-05	0.002442425
M11	T73	109	113	14	1.660856838	1.47E-05	0.002449105
E88	H53	50	453	19	1.342909104	1.47E-05	0.002457165
K25	R39	283	180	33	1.04453116	1.49E-05	0.002488287
K02	Z04	26	469	13	1.572736198	1.51E-05	0.002505422
J42	N30	129	136	17	1.507144463	1.51E-05	0.002504887
I67	I69	5	39	3	1.916928928	1.52E-05	0.00251888
J42	M19	129	165	19	1.420896167	1.52E-05	0.002528318
L53	Ro5	13	852	11	1.458463569	1.53E-05	0.002536412
G47	K59	990	31	21	1.090905255	1.55E-05	0.002558254
F05	I46	89	121	13	1.71202043	1.55E-05	0.002565852
B15	R53	45	564	20	1.268935007	1.56E-05	0.002567777
K59	M54	31	654	17	1.329500032	1.56E-05	0.002568163
E66	T30	388	39	15	1.514729792	1.57E-05	0.002585875
E87	I21	40	5	3	1.914860548	1.64E-05	0.002686436
I10	R73	282	52	15	1.551854189	1.66E-05	0.002722758
F01	F18	67	76	9	1.972721157	1.68E-05	0.002747081
A41	F22	32	561	16	1.396702724	1.70E-05	0.0027649
G43	I46	210	121	21	1.334540371	1.70E-05	0.002767937
B17	I46	77	121	12	1.762374199	1.71E-05	0.002787358
M54	R25	654	37	19	1.259285642	1.72E-05	0.002798292
E53	G44	16	196	7	2.034166603	1.73E-05	0.002808566
A69	E61	5	41	3	1.912795128	1.77E-05	0.00286201
E10	F07	52	183	12	1.739779292	1.78E-05	0.00287644
E66	I61	388	128	32	1.035813943	1.78E-05	0.002878245
B95	K59	62	31	6	2.143605422	1.79E-05	0.002886621
H10	Ro7	47	493	19	1.31422	1.80E-05	0.002898927
B34	B95	83	62	9	1.963456584	1.82E-05	0.002940506
A36	J20	5	5	2	1.574041214	1.84E-05	0.0029718
R12	R23	517	81	28	1.078871391	1.85E-05	0.002981024
K77	R18	293	29	11	1.743925416	1.86E-05	0.002993095
G24	N30	105	136	15	1.583045298	1.87E-05	0.002995646
A49	R26	40	288	13	1.636942771	1.87E-05	0.002994161
E87	R26	40	288	13	1.636942771	1.87E-05	0.002992045
B17	Ro6	77	368	22	1.258610347	1.89E-05	0.00303323
K29	R12	38	517	17	1.368470681	1.92E-05	0.003065263
B90	J45	62	297	17	1.448123538	1.92E-05	0.00307161
N13	R33	4	57	3	1.903332101	1.95E-05	0.003115819
N47	R33	4	57	3	1.903332101	1.95E-05	0.003113625
F70	H50	27	43	5	2.148920584	1.96E-05	0.003129867
E87	R18	40	29	5	2.149244749	1.97E-05	0.003141915
M40	R53	60	564	24	1.147430364	1.98E-05	0.003155696
E51	F18	26	76	6	2.128734985	1.99E-05	0.003159039
B90	Z04	62	469	22	1.22515322	2.01E-05	0.003192033
B90	L30	62	398	20	1.304802993	2.03E-05	0.003214972
B18	B22	58	4	3	1.901692675	2.06E-05	0.003257466
F13	F61	291	40	13	1.625300246	2.09E-05	0.003307716
G81	I69	15	39	4	2.085819368	2.10E-05	0.003316886
E66	R32	388	79	23	1.218050413	2.10E-05	0.003319267
E66	T71	388	16	9	1.792391923	2.16E-05	0.00340778
J10	L89	756	36	20	1.179383888	2.16E-05	0.003406633
F50	T14	183	53	12	1.719308843	2.19E-05	0.003443955
L89	R31	36	55	6	2.127639544	2.19E-05	0.003442946
H10	H53	47	453	18	1.346543394	2.19E-05	0.003442916
G91	G94	16	2	2	1.570998011	2.22E-05	0.003483608
B23	K77	12	293	7	1.951280433	2.22E-05	0.003481993
R13	T73	113	113	14	1.619668026	2.24E-05	0.003503793
A49	K77	40	293	13	1.617590434	2.25E-05	0.00352414
J44	J98	35	34	5	2.139551352	2.26E-05	0.003534841
R00	R46	27	624	15	1.386239079	2.27E-05	0.003544515
F19	F51	441	85	26	1.123362114	2.37E-05	0.003691778



<i>ICD10 A</i>	<i>ICD10 B</i>	<i>#pt.'s A</i>	<i>#pt.'s B</i>	<i>#pt.'s A∩B</i>	<i>Co-occurrence score</i>	<i>p-value</i>	<i>corrected p-value</i>
L30	T14	398	53	18	1.358152195	2.41E-05	0.003744379
J44	L30	35	398	14	1.518964942	2.41E-05	0.003745273
F32	G44	219	196	29	1.094720119	2.43E-05	0.003784524
M54	R10	654	70	29	1.008208813	2.48E-05	0.003846435
F19	G24	441	105	30	1.040161119	2.49E-05	0.003858285
Bo6	B26	3	10	2	1.571866843	2.49E-05	0.003861888
J10	K59	756	31	18	1.225843073	2.50E-05	0.003871224
J10	T14	756	53	26	1.034748781	2.51E-05	0.003884528
A41	J00	32	578	16	1.360220994	2.51E-05	0.003882889
Fo6	I63	222	7	5	2.02685122	2.51E-05	0.003887754
L25	L30	5	398	5	1.902512155	2.53E-05	0.003914913
G62	H18	98	3	3	1.876516947	2.56E-05	0.003951978
L89	R02	36	17	4	2.075801901	2.60E-05	0.004001797
B68	L40	7	100	4	2.043626932	2.62E-05	0.004037089
E66	I20	388	10	7	1.876114465	2.63E-05	0.004048884
Ro6	R50	368	131	31	1.031643085	2.67E-05	0.004102346
F25	R33	145	57	11	1.77260473	2.68E-05	0.004110139
To0	T14	12	53	4	2.066955506	2.70E-05	0.0041443
Fo3	I46	35	121	8	1.976321004	2.71E-05	0.00415723
R26	R27	288	47	14	1.552357835	2.72E-05	0.004164579
J44	M54	35	654	18	1.255621905	2.81E-05	0.004291046
I10	R60	282	178	32	1.021392294	2.86E-05	0.004360934
Fo5	I61	89	128	13	1.649465904	2.87E-05	0.004368489
J00	L73	578	6	6	1.768846156	2.88E-05	0.004380514
H40	M54	16	654	11	1.521270739	2.91E-05	0.004427464
A46	F91	49	112	9	1.906123409	2.92E-05	0.004433628
F60	Yo6	635	62	26	1.058163026	2.97E-05	0.004502221
R40	T90	76	9	4	2.049423806	3.04E-05	0.004598107
F33	G62	293	98	22	1.241469214	3.04E-05	0.004599238
K59	Z04	31	469	14	1.468824298	3.06E-05	0.004613407
E87	L30	40	398	15	1.45430197	3.06E-05	0.004620099
F22	R23	561	81	29	1.018210257	3.06E-05	0.004617263
Fo5	R41	89	254	19	1.345358313	3.07E-05	0.004628929
Bo1	Bo5	8	4	2	1.570998011	3.10E-05	0.004658711
J10	J98	756	34	19	1.182333203	3.12E-05	0.004689699
B90	J98	62	34	6	2.093017529	3.12E-05	0.00469108
B17	R31	77	55	8	1.976321004	3.13E-05	0.004692086
F22	F65	561	30	15	1.387529184	3.13E-05	0.004702303
E78	G43	89	210	17	1.429893104	3.14E-05	0.004709467
K59	Ro5	31	852	19	1.147527505	3.15E-05	0.004712383
A49	F22	40	561	18	1.280635545	3.15E-05	0.004711729
J00	R11	578	79	29	1.01168228	3.19E-05	0.004765451
L30	R32	398	79	23	1.184856501	3.20E-05	0.004773811
I95	R26	42	288	13	1.581896198	3.37E-05	0.005004299
B95	N90	62	11	4	2.050150056	3.42E-05	0.005073378
F15	R23	123	81	12	1.690280237	3.44E-05	0.005096263
K76	R17	25	26	4	2.061820049	3.47E-05	0.005132695
E61	Ro5	41	852	23	1.046718719	3.49E-05	0.005162128
F19	T40	441	7	6	1.852564625	3.49E-05	0.005162959
L30	L89	398	36	14	1.485999926	3.51E-05	0.00518573
B18	M41	58	37	6	2.08289705	3.51E-05	0.005185503
R33	R60	57	178	12	1.670495507	3.61E-05	0.005322238
B20	B23	3	12	2	1.569261916	3.65E-05	0.00537522
G44	N30	196	136	21	1.273227798	3.69E-05	0.005422846
F21	Yo6	280	62	16	1.437480549	3.69E-05	0.005421409
B95	R10	62	70	8	1.956329528	3.78E-05	0.005548992
G62	R41	98	254	20	1.293702912	3.79E-05	0.005557124
X60	X78	5	7	2	1.569695744	3.87E-05	0.00565485
B18	B23	58	12	4	2.045073969	3.88E-05	0.005662543
B18	K72	58	12	4	2.045073969	3.88E-05	0.005658899
J10	R02	756	17	12	1.405779867	3.88E-05	0.005662826
K77	R11	293	79	19	1.315529111	3.90E-05	0.00568281
F14	F16	201	8	5	2.010857311	3.90E-05	0.005685049
K56	R11	28	79	6	2.065486366	3.91E-05	0.005698202
E11	R35	94	8	4	2.024946357	4.01E-05	0.005821889
L70	M54	27	654	15	1.329346979	4.08E-05	0.005917267
E66	R02	388	17	9	1.734628772	4.12E-05	0.005951591
G40	Z50	337	119	27	1.08605208	4.12E-05	0.005953898
Mo6	M11	21	109	6	2.045435955	4.17E-05	0.006017718
G24	H53	105	453	30	1.0039666174	4.22E-05	0.006090264
I26	I80	3	13	2	1.567961214	4.32E-05	0.006210194
L30	R10	398	70	21	1.216358944	4.33E-05	0.006223417

<i>ICD10 A</i>	<i>ICD10 B</i>	<i>#pt.'s A</i>	<i>#pt.'s B</i>	<i>#pt.'s A∩B</i>	<i>Co-occurrence score</i>	<i>p-value</i>	<i>corrected p-value</i>
J42	J44	129	35	8	1.923613953	4.33E-05	0.006225268
E05	G44	18	196	7	1.948738984	4.34E-05	0.006237364
L30	M40	398	60	19	1.276073018	4.38E-05	0.006286067
R31	R53	55	564	22	1.141092319	4.41E-05	0.006325676
B34	R02	83	17	5	2.070082404	4.42E-05	0.006326808
F91	I46	112	121	14	1.550986615	4.42E-05	0.00633379
F50	O00	183	5	4	1.967909878	4.44E-05	0.006349173
R23	R26	81	288	19	1.305685409	4.45E-05	0.006360884
M40	M53	60	23	5	2.079627535	4.46E-05	0.006370616
A41	J10	32	756	18	1.185600003	4.50E-05	0.006419999
E56	G62	73	98	10	1.792916745	4.54E-05	0.006466484
N30	R13	136	113	15	1.496365141	4.56E-05	0.006501821
K25	K29	283	38	12	1.606645306	4.57E-05	0.00651302
A46	R26	49	288	14	1.503796961	4.58E-05	0.006510849
E16	R26	32	288	11	1.658501567	4.63E-05	0.006586591
G43	J42	210	129	21	1.252469238	4.64E-05	0.006596563
B95	F98	62	72	8	1.93307185	4.66E-05	0.006616328
M53	M70	23	2	2	1.564930796	4.68E-05	0.006636803
J45	R11	297	79	19	1.288386904	4.72E-05	0.006687935
G47	N90	990	11	10	1.351733575	4.72E-05	0.006691753
B34	G43	83	210	16	1.432598327	4.82E-05	0.006816008
I95	R11	42	79	7	1.993873897	4.83E-05	0.006827435
E66	G24	388	105	27	1.065027567	4.83E-05	0.006831663
G44	K29	196	38	10	1.752865798	4.88E-05	0.006885998
L65	M40	38	60	6	2.047765178	5.01E-05	0.007057193
E78	G44	89	196	16	1.431623863	5.03E-05	0.007090805
L30	R25	398	37	14	1.453771352	5.04E-05	0.007090627
F07	S06	183	33	9	1.818293236	5.05E-05	0.007105574
B95	L97	62	12	4	2.027804595	5.06E-05	0.007107741
B23	R05	12	852	10	1.421136347	5.06E-05	0.007105216
L30	R02	398	17	9	1.710031029	5.06E-05	0.007103564
F25	F30	145	6	4	1.98343215	5.07E-05	0.007108861
E66	R73	388	52	17	1.335512549	5.08E-05	0.007118533
L29	L65	618	38	18	1.223254311	5.19E-05	0.00723728
I10	I15	282	4	4	1.89662226	5.29E-05	0.00736067
F10	K70	1149	15	13	1.166132189	5.32E-05	0.007382742
A46	J10	49	756	24	1.028019191	5.33E-05	0.007396096
F07	R40	183	76	14	1.520809285	5.39E-05	0.007451958
I95	R60	42	178	10	1.749108777	5.41E-05	0.007472435
R05	Z30	852	14	11	1.375343222	5.46E-05	0.007537128
K59	R53	31	564	15	1.341380458	5.48E-05	0.007546299
B24	K25	120	283	24	1.142775253	5.56E-05	0.007648345
J10	L53	756	13	10	1.464043328	5.60E-05	0.007697899
G81	R41	15	254	7	1.890268559	5.66E-05	0.007768942
F22	K29	561	38	17	1.266957082	5.75E-05	0.007884875
I60	I61	3	128	3	1.840735958	5.76E-05	0.007889321
R02	R12	17	517	10	1.583090086	5.77E-05	0.007900731
N91	O92	6	7	2	1.566661684	5.80E-05	0.007936718
B18	K76	58	25	5	2.058163026	5.80E-05	0.007935428
E61	R33	41	57	6	2.033076543	5.82E-05	0.007960915
I61	K25	128	283	25	1.11423531	5.87E-05	0.008021346
B24	T30	120	39	8	1.893432861	5.91E-05	0.0080547
B15	R06	45	368	15	1.407020482	5.92E-05	0.008069911
K02	R15	26	2	2	1.562338351	6.01E-05	0.008177556
L89	R50	36	131	8	1.88693097	6.02E-05	0.008189892
R11	R13	79	113	11	1.692231929	6.08E-05	0.008256989
B34	L03	83	4	3	1.861301038	6.08E-05	0.008258728
R05	R25	852	37	21	1.055923524	6.09E-05	0.00825492
F23	Z04	66	469	22	1.143870214	6.10E-05	0.008262658
J18	L29	26	618	14	1.350120179	6.11E-05	0.008267767
H40	R02	16	17	3	1.885400068	6.16E-05	0.008336055
B22	K77	4	293	4	1.882325165	6.17E-05	0.008349531
R11	R60	79	178	14	1.507948127	6.19E-05	0.008371541
B86	L30	29	398	12	1.527886058	6.27E-05	0.008463544
G40	M40	337	60	17	1.332809952	6.30E-05	0.008490023
E78	I69	89	39	7	1.960850937	6.36E-05	0.008570591
E87	I50	40	19	4	2.022093771	6.37E-05	0.008580381
R17	R18	26	29	4	2.024232682	6.37E-05	0.008575533
R05	R19	852	25	16	1.185550595	6.41E-05	0.008617671
R09	R26	33	288	11	1.625654601	6.41E-05	0.008623607
F70	J45	27	297	10	1.678119743	6.48E-05	0.008709995
K02	R33	26	57	5	2.048456042	6.51E-05	0.008741732

<i>ICD10 A</i>	<i>ICD10 B</i>	<i>#pt.'s A</i>	<i>#pt.'s B</i>	<i>#pt.'s A&amp;B</i>	<i>Co-occurrence score</i>	<i>p-value</i>	<i>corrected p-value</i>
F51	G43	85	210	16	1.403646954	6.53E-05	0.008769729
B23	F12	12	665	9	1.545600068	6.55E-05	0.008786024
H10	L30	47	398	16	1.346381139	6.64E-05	0.008896511
E11	T73	94	113	12	1.620269369	6.67E-05	0.00893651
J42	M11	129	109	14	1.508031272	6.69E-05	0.00894801
A46	Z04	49	469	18	1.250615885	6.81E-05	0.00909684
R17	R46	26	624	14	1.338540551	6.83E-05	0.009113459
R10	R60	70	178	13	1.548162583	6.89E-05	0.00918046
R11	R39	79	180	14	1.494871001	7.01E-05	0.009325566
J45	M11	297	109	23	1.146693921	7.09E-05	0.009424146
K56	R26	28	288	10	1.672390461	7.13E-05	0.009468646
A46	R12	49	517	19	1.200912694	7.13E-05	0.009468183
G24	R26	105	288	22	1.174268962	7.17E-05	0.009499916
B17	R11	77	79	9	1.811504796	7.20E-05	0.009534535
R12	R33	517	57	21	1.14369979	7.24E-05	0.009584135
R12	T14	517	53	20	1.170661258	7.26E-05	0.009602679
B22	B23	4	12	2	1.56406613	7.29E-05	0.009634522
R60	R73	178	52	11	1.653894522	7.37E-05	0.009730949
M15	M17	3	17	2	1.562770102	7.52E-05	0.009919199
A41	B17	32	77	6	2.000877285	7.54E-05	0.009944004

## Supplementary Table 2 to Paper I

**Table S2:** ICD10 pairs curated by a medical doctor. The pairs were assigned one of three categories: 1. similar diagnosis, 2. consequences or side effects, 3. surprising correlations.

ICD10 A	Danish description A	ICD10 B	Danish description B	Category
B34	Virussygdom Uden Specifikation Lokalisation	J10	Influenza forårsaget Af Anden Identifieret Type Infl.Virus	1
J42	Kronisk Bronkit Uden Specifikation	R06	Abnorm Vejtrækning	1
F20	Skizofreni	G24	Dystoni	2
G24	Dystoni	R44	Sympt Og Abnorme Fund Vedrørende Opfattelsesevne, Andre	2
F22	Paranoide Psykoser	G24	Dystoni	2
F06	Psykiske Lidelser Hjerneorganisk Betinget, Andre	G40	Epilepsi	2
F07	Personligheds- Og Adfærdsforstyrrelser Hjerneorg Betinget	G40	Epilepsi	2
J10	Influenza forårsaget Af Anden Identifieret Type Infl.Virus	R13	Synkebesvær	2
B24	Hiv Sygdom Og Aids Uden Specifikation	K77	Leverlidelse Ved Sygdom Klassificeret Andetsteds	2
G47	Søvnforstyrrelser	J98	åndedrætssygdom, Anden	2
B24	Hiv Sygdom Og Aids Uden Specifikation	L30	Dermatit, Andre Former	2
J10	Influenza forårsaget Af Anden Identifieret Type Infl.Virus	R61	øget Svedsekretion	2
B17	Leverbetændelse, Anden Akut Viral	J10	Influenza forårsaget Af Anden Identifieret Type Infl.Virus	2
B17	Leverbetændelse, Anden Akut Viral	L30	Dermatit, Andre Former	2
B34	Virussygdom Uden Specifikation Lokalisation	R06	Abnorm Vejtrækning	2
A49	Bakteriel Infektion U Angivelse Af Lokalisation	J10	Influenza forårsaget Af Anden Identifieret Type Infl.Virus	2
B95	Strepto- Og Stafylokokker Som årsag Til Sygd I Andre Kap	J10	Influenza forårsaget Af Anden Identifieret Type Infl.Virus	2
B95	Strepto- Og Stafylokokker Som årsag Til Sygd I Andre Kap	R06	Abnorm Vejtrækning	2
E88	Stofskifteforstyrrelser, Andre	H53	Synsforstyrrelser	2
B90	Tuberkulose, Følgende Efter	J45	Astma	2
E66	Fedme, Overvægt	R32	Urininkontinens Uden Specifikation	2
F32	Depressiv Enkeltepisode	G44	Hovedpine Syndromer, Andre	2
A41	Anden Blodforgiftning	J10	Influenza forårsaget Af Anden Identifieret Type Infl.Virus	2
E66	Fedme, Overvægt	R73	Forhøjet Blodsukker	2
A46	Rosen	J10	Influenza forårsaget Af Anden Identifieret Type Infl.Virus	2
J10	Influenza forårsaget Af Anden Identifieret Type Infl.Virus	L53	Erytematøse Tilstande, Andre	2
G24	Dystoni	R26	Gangbesvær Og Mobilitetsforstyrrelser	2
F22	Paranoide Psykoser	K77	Leverlidelse Ved Sygdom Klassificeret Andetsteds	3
N30	Blærebetændelse	R05	Hoste	3
E66	Fedme, Overvægt	L30	Dermatit, Andre Former	3
G43	Migræne	M54	Rygmerter	3
B17	Leverbetændelse, Anden Akut Viral	R44	Sympt Og Abnorme Fund Vedrørende Opfattelsesevne, Andre	3
L30	Dermatit, Andre Former	R06	Abnorm Vejtrækning	3
L30	Dermatit, Andre Former	R26	Gangbesvær Og Mobilitetsforstyrrelser	3
L30	Dermatit, Andre Former	N30	Blærebetændelse	3
B24	Hiv Sygdom Og Aids Uden Specifikation	R44	Sympt Og Abnorme Fund Vedrørende Opfattelsesevne, Andre	3
J10	Influenza forårsaget Af Anden Identifieret Type Infl.Virus	N30	Blærebetændelse	3
B34	Virussygdom Uden Specifikation Lokalisation	M54	Rygmerter	3
J42	Kronisk Bronkit Uden Specifikation	L30	Dermatit, Andre Former	3
M54	Rygmerter	N30	Blærebetændelse	3
B24	Hiv Sygdom Og Aids Uden Specifikation	F22	Paranoide Psykoser	3
F06	Psykiske Lidelser Hjerneorganisk Betinget, Andre	I10	Blodtryksforhøjelse Af Ukendt årsag	3

<i>ICD10 A</i>	<i>Danish description A</i>	<i>ICD10 B</i>	<i>Danish description B</i>	<i>Category</i>
K25	Mavesår	N30	Blærebetændelse	3
B15	Akut Leverbetændelse a	R44	Sympt Og Abnorme Fund Vedrørende Opfattelsesevne, Andre	3
J45	Astma	K77	Leverlidelse Ved Sygdom Klassificeret An- detsteds	3
E66	Fedme, Overvægt	N30	Blærebetændelse	3
I10	Blodtryksforhøjelse Af Ukendt årsag	R41	Sympt Og Abnorme Fund Vedrørende Opfattelsesevne, Andre	3
B17	Leverbetændelse, Anden Akut Viral	F22	Paranoide Psykoser	3
B17	Leverbetændelse, Anden Akut Viral	M54	Rygsmerte	3
E88	Stofskiftesyndromer, Andre	G47	Søvnsforstyrrelser	3
G24	Dystoni	K77	Leverlidelse Ved Sygdom Klassificeret An- detsteds	3
J45	Astma	N30	Blærebetændelse	3
K25	Mavesår	R26	Gangbesværet Og Mobilitetsforstyrrelser	3
B18	Kronisk Viral Leverbetændelse	R44	Sympt Og Abnorme Fund Vedrørende Opfattelsesevne, Andre	3
B15	Akut Leverbetændelse a	F22	Paranoide Psykoser	3
G40	Epilepsi	I46	Hjertestop	3
G44	Hovedpine Syndromer, Andre	R06	Abnorm Vejrtrekning	3
L30	Dermatit, Andre Former	R13	Synkebesværet	3
G24	Dystoni	L30	Dermatit, Andre Former	3
N30	Blærebetændelse	R06	Abnorm Vejrtrekning	3
I46	Hjertestop	L30	Dermatit, Andre Former	3
B34	Virussygdom Uden Specifikation Lokali- sation	E66	Fedme, Overvægt	3
M54	Rygsmerte	R33	Urineretention	3
B18	Kronisk Viral Leverbetændelse	F22	Paranoide Psykoser	3
J42	Kronisk Bronkit Uden Specifikation	K25	Mavesår	3
F07	Personligheds- Og Adfærdssyndromer Hjerneorg Betinget	I10	Blodtryksforhøjelse Af Ukendt årsag	3
G47	Søvnsforstyrrelser	L70	Akne	3
J10	Influenza forårsaget Af Anden Identifi- ceret Type Infl.Virus	M40	Krum Ryg	3
F13	Sedativa-Hypnotikabetingede Psykiske Forstyrrelser	G44	Hovedpine Syndromer, Andre	3
G43	Migræne	K25	Mavesår	3
B17	Leverbetændelse, Anden Akut Viral	R26	Gangbesværet Og Mobilitetsforstyrrelser	3
B34	Virussygdom Uden Specifikation Lokali- sation	F22	Paranoide Psykoser	3
K25	Mavesår	R13	Synkebesværet	3
L65	Hårtab Urdannelse, Andre Former	M54	Rygsmerte	3
E66	Fedme, Overvægt	G21	Sekundær Rystelammelse	3
G44	Hovedpine Syndromer, Andre	J45	Astma	3
a49	Bakteriel Infektion U Angivelse Af Lokali- sation	H53	Synsforstyrrelser	3
I84	Hæmoroider	R44	Sympt Og Abnorme Fund Vedrørende Opfattelsesevne, Andre	3
F42	Obsessiv-Kompulsiv Tilstand	J10	Influenza forårsaget Af Anden Identifi- ceret Type Infl.Virus	3
M19	Slidigt, Andre Former	R26	Gangbesværet Og Mobilitetsforstyrrelser	3
J10	Influenza forårsaget Af Anden Identifi- ceret Type Infl.Virus	L50	Nældede	3
N30	Blærebetændelse	R26	Gangbesværet Og Mobilitetsforstyrrelser	3
J42	Kronisk Bronkit Uden Specifikation	R26	Gangbesværet Og Mobilitetsforstyrrelser	3
J10	Influenza forårsaget Af Anden Identifi- ceret Type Infl.Virus	R33	Urineretention	3
B18	Kronisk Viral Leverbetændelse	M54	Rygsmerte	3
B90	Tuberkulose, Følgende Efter	R26	Gangbesværet Og Mobilitetsforstyrrelser	3
F25	Psykozer, Skizo-Affektive	R06	Abnorm Vejrtrekning	3
B95	Strepto- Og Stafylokokker Som årsag Til Sygdom Andre Kap	E66	Fedme, Overvægt	3
E66	Fedme, Overvægt	I46	Hjertestop	3
G47	Søvnsforstyrrelser	R09	Symptomer Og Fund Fra Kredsløbs- Og åndedrætsorganer, Andre	3
H40	Grøn Stær	J10	Influenza forårsaget Af Anden Identifi- ceret Type Infl.Virus	3
J42	Kronisk Bronkit Uden Specifikation	N30	Blærebetændelse	3
J42	Kronisk Bronkit Uden Specifikation	M19	Slidigt, Andre Former	3
G47	Søvnsforstyrrelser	K59	Forstyrrelser I Tarmfunktionen, Andre	3
K59	Forstyrrelser I Tarmfunktionen, Andre	M54	Rygsmerte	3

<i>ICD10 A</i>	<i>Danish description A</i>	<i>ICD10 B</i>	<i>Danish description B</i>	<i>Category</i>
a41	Anden Blodforgiftning	F22	Paranoide Psykoser	3
G43	MigræNe	I46	Hjertestop	3
M54	Rygsmarter	R25	Abnorme Ufrivillige Bevægelses	3
E66	Fedme, Overvægt	I61	Hjerneblødning	3
B17	Leverbetændelse, Anden Akut Viral	Ro6	Abnorm Vejrtrekning	3
B90	Tuberkulose, Følgende Efter	L30	Dermatit, Andre Former	3
J10	Influenza forårsaget Af Anden Identificeret Type Infl.Virus	L89	Liggesår	3
H10	Betændelse I øjets Bindehinde	H53	Synsforstyrrelser	3
J10	Influenza forårsaget Af Anden Identificeret Type Infl.Virus	K59	Forstyrrelser I Tarmfunktionen, Andre	3
J44	Kronisk Obstruktiv Lungesygdom, Anden	M54	Rygsmarter	3
F33	Periodisk Depression	G62	Polyneuropatier, Andre	3
E78	Forstyrrelser I Lipoproteinstofskiftet Og Andre LipidæMier	G43	MigræNe	3
a49	Bakteriel Infektion U Angivelse Af Lokalisation	F22	Paranoide Psykoser	3
L30	Dermatit, Andre Former	R32	Urininkontinens Uden Specifikation	3
G44	Hovedpine Syndromer, Andre	N30	Blærebetændelse	3
G62	Polyneuropatier, Andre	R41	Sympt Og Abnorme Fund Vedr Erkendelsesevne, Andre	3
L70	Akne	M54	Rygsmarter	3
G24	Dystoni	H53	Synsforstyrrelser	3
L30	Dermatit, Andre Former	M40	Krum Ryg	3
N30	Blærebetændelse	R13	Synkebesvær	3
a46	Rosen	R26	Gangbesvær Og Mobilitetsforstyrrelser	3
G43	MigræNe	J42	Kronisk Bronkit Uden Specifikation	3
G47	Sævnforstyrrelser	N90	Andre Ikke Inflammatoriske Sygdomme I Kvindens Ydre Kønslidelser	3
B34	Virussygdom Uden Specifikation Lokalisation	G43	MigræNe	3
E66	Fedme, Overvægt	G24	Dystoni	3
E78	Forstyrrelser I Lipoproteinstofskiftet Og Andre LipidæMier	G44	Hovedpine Syndromer, Andre	3
L30	Dermatit, Andre Former	R25	Abnorme Ufrivillige Bevægelses	3
B24	Hiv Sygdom Og Aids Uden Specifikation	K25	Mavesår	3
F22	Paranoide Psykoser	K29	Mavekatar	3
I61	Hjerneblødning	K25	Mavesår	3
B15	Akut Leverbetændelse a	Ro6	Abnorm Vejrtrekning	3
G40	Epilepsi	M40	Krum Ryg	3
H10	Betændelse I øjets Bindehinde	L30	Dermatit, Andre Former	3

# Supplementary Table 3 to Paper I

Cluster no.	Most distinguishing code	pt.'s in cluster	mined	assigned
27	K77: Liver Disor.	53	53	0
66	F14: Psyk. Disor. Due to Cocaine	47	1	46
71	F15: Psyk. Disor. Due to Other Stimulants	33	0	33
38	F11: Psyk. Disor. Due to Opioids	28	0	28
45	F19: Psyk. Disor. Due to Multiple Drug Use	39	0	39
8	Z04: Examination and Observation	138	4	134
1	F20: Schizophrenia	197	5	192
10	F22: Persistent Delusional Disor.	45	29	16
6	J45: Asthma	40	33	7
2	F21: Schizotypal Disor.	120	13	107
101	F50: Eating Disorders	54	25	29
15	F25: Schizoaffective Disor.	75	11	64
25	G40: Epilepsy	65	31	34
22	F43: Stress Reaction & Adjustment Disor.	38	6	32
77	G43: Migraine	47	44	3
53	F31: Bipolar Affective Disor.	74	5	69
14	L30: Other Dermatitis	27	26	1
4	L40: Psoriasis	31	19	12
48	F60: Specific Personality Disor.	74	0	74
65	F33: Recurrent Depressive Disor.	57	3	54
103	F40: Phobic Anxiety Disorders	64	15	49
54	F10: Psyk. Disor. Due to Alcohol	43	0	30
11	F07: Psyk. Disor. Due to Brain Dis/Damage/Dysfunction	24	0	24
78	F06: Other Psyk. Disor. Due to Brain Dis/Damage/Dysfunction	31	1	30
72	I10: Hypertension	27	11	16
26	E10: Insulin-Dependent Diabetes Mellitus	26	7	19
Total		1497	342	1142

**Table S3:** The table shows how the members of the 26 clusters in figure 3 are associated with the ICD10 code that is most distiguishing for that cluster. Mined contains those patients where the association comes only from mining, and assigned contains those patients where association comes from assignment only or from both assignment and mining. Cluster 54 contains 13 patients that are in fact not associated to F10 at all.