Environmental and Life Style Factors in Relation to Male Reproductive Disorders

Krysiak-Baltyn, Konrad; Brunak, Søren; Jensen, Thomas Skøt

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Environmental and Life Style Factors in Relation to Male Reproductive Disorders

Konrad Krysiak-Baltyn

January, 2011

Preface

In 2007, the Center for Disease Systems Biology (CDSB) was founded by the Villum Kann Rasmussen foundation, involving two research groups; the Center for Biological Sequence Analysis at the Technical University of Denmark, a center specializing in bioinformatics and systems biology, and the Department for Growth and Reproduction, Rigshospitalet, University of Copenhagen, a world known group studying infertility and pubertal disorders including testicular cancer. The ambition was to apply Systems Biology for solving a difficult biological problem; finding the cause for the rapid rise of male reproductive disorders in the western world, particularly Denmark.

The work within my PhD is part of the larger group effort at the CDSB, and has been carried out under the main supervision of Professor Søren Brunak and Thomas Skøt Jensen at the Center for Biological Sequence Analysis, DTU, as well as in close collaboration with Professor Niels Erik Skakkebœk and Katharina Maria Main at the Department of Growth and Reproduction, Rigshospitalet.

Contents

Abstract

During the past four decades, the incidence rates of testicular cancer and other male reproductive disorders have been increasing at a rapid rate, predominantly in developed and industrialized countries. This increase is considered too great to be explained by genetic factors alone, and thus environmental factors have strongly been suspected to play a major role. There is a large amount of clinical research which has tried to pinpoint the mechanism of action for this trend.

Although the exact mechanism of action has not been elucidated, a number of genetic factors as well as environmental chemicals have been found, mostly in animal studies, to act as risk factors for the disorders. The common consensus today is that there exists a common causal mechanism for a number of different male reproductive disorders which takes place before birth, during fetal development, and is termed Testicular Dysgenesis Syndrome (TDS). TDS occurs when certain critical developmental events are disturbed, and has a profound effect that propagates into adulthood, which may lead to lower sperm concentration, cryptorchidism and testicular cancer.

The work within this PhD thesis has primarily focused on the environmental aspects of TDS, generating further support for the hypothesis that environmental factors may play a critical role in the observed trends.

This thesis is divided into four parts. In the first part I introduce male reproductive disorders and the current state of affairs. In the second part, I focus on studies of environmental chemicals and their possible impact on reproductive health. In the third part, I discuss network biology as a powerful tool for the study of gene-gene and chemical-gene interactions. In the fourth part, I discuss association mining of clinical data as a means to find interesting and unexpected associations between life style factors and disease. The thesis ends with concluding remarks.

Dansk resumé

I løbet af de seneste fire årtier er forekomsten af testikelkræft og andre mandlige reproduktionsforstyrrelser steget kraftigt. Denne stigning ses overvejende i udviklede og industrialiserede lande og betragtes for stor til at kunne forklares af genetiske faktorer alene. Miljømæssige faktorer er derfor mistænkt for at spille en vigtig rolle for denne trend, og klinisk forskning forsøger ihærdigt at identificere virkningsmekanismerne bag. Disse er endnu ukendt, men en række genetiske faktorer og miljømæssige kemikalier er identificeret til at være risikofaktorer for mandlige reproduktionsforstyrrelser - de fleste fundet ved dyreforsøg.

En general hypotese for årsagen til en række mandlige reproduktionsforstyrrelse kaldes testikulært dysgenese syndrom (TDS), hvor forstyrrelserne finder sted allerede under fosterudviklingen. Forstyrrelserne har en dybtgående indvirkning, og konsekvenserne forplanter sig ind i voksenlivet I form af dårlig sædkvalitet, kryptorkisme og testikelkræft.

Arbejdet i denne ph.d.-afhandling har primært fokuseret på de miljømæssige aspekter af TDS. Det er forsøgt yderligere at underbygge hypotesen om, at miljøfaktorer kan spille en afgørende rolle i den øgede hyppighed af mandlige reproduktionsforstyrrelser.

Denne afhandling er inddelt i fire dele. I den første del giver jeg en introduktion til mandlige reproduktionsforstyrrelser og den aktuelle situation på området. I den anden del fokuserer jeg på undersøgelser af kemikalier i miljøet og deres mulige indvirkning på reproduktiv sundhed. I tredje del diskuterer jeg netværksbiologi som et stærkt værktøj til undersøgelse af gen-gen og kemikalie-gen interaktioner. I den fjerde del diskuterer jeg "association mining" af kliniske data som et middel til at finde interessante og uventede sammenhænge mellem livsstilsfaktorer og sygdom.

Acknowledgements

The three years of my PhD have been a true learning experience, both professionally and personally. It has helped me build courage to go into areas I am not familiar with, and face obstacles and failures with an attitude of acceptance. Many things I have tried did not work out as I hoped, and that is just part of life experience. However, some things did work out successfully, for which I am grateful.

The project I have been involved in has included many people from different scientific disciplines. It has all been a great team-effort, and could not have been accomplished by any single person. I would like to thank the following people for their contribution and dedication:

- My supervisors, Professor Søren Brunak, who has created a great work environment at CBS, and supported me during times when things were not working out.

- My second supervisor, Thomas Skøt Jensen, for his dedication and his enthusiasm in attending many of my project meetings.

- My collaborators at the Dept. of Growth and Reproduction, Rigshsopitalet, Professor Niels Erik Skakkebæk and Katharina Maria Main, as well as Professor Jorma Toppari. Your enthusiasm and dedication has been truly inspirational.

- Olivier Taboureau, for his supervision early in my PhD, as well as Karine Audouze, with whom I worked in different projects.

- My colleague and office mate Daniel Edsgärd, who often shared his knowledge and ideas. I especially appreciate his late friday jokes, which had a level of quality strongly affected by a long week of hard work. A thanks to my colleagues Nils Weinhold and Thomas Stranzl, who enjoy singing as much as I do. Also a thanks to all other colleagues and friends at CBS who made my time here very pleasant.

- A special thanks goes to my family and close friends: your support during times of stress was amazing.

Publications

Manuscript I

Krysiak-Baltyn K, Toppari J, Skakkebaek NE, Jensen TS, Virtanen HE, Schramm KW, Shen H, Vartiainen T, Kiviranta H, Taboureau O, Brunak S, Main KM. 2010. **Country-specific chemical signatures of persistent environmental compounds in breast milk.** International journal of andrology 33: 270-278.

Manuscript II

Krysiak-Baltyn K, Toppari J, Skakkebaek NE, Jensen TS, Virtanen HE, Schramm KW, Shen H, Vartiainen T, Kiviranta H, Taboureau O, Brunak S, Main KM. **Association between chemical pattern in breast milk and congenital cryptorchidism of newborn boys.** (Manuscript in preparation.)

Manuscript III

Audouze K, Juncker AS, Roque F, Krysiak-Baltyn K, Weinhold N, Taboureau O, Jensen TS, Brunak S. 2010. **Deciphering diseases and biological targets for environmental chemicals using toxicogenomics networks.** PLoS computational biology 6: e1000788.

Manuscript IV

Krysiak-Baltyn K, Nordahl Pedersen T, Audouze K, Brunak S. **Compass: a hybrid method for clinical and biobank data mining.** (Manuscript in preparation.)

Part I

Introduction

Chapter 1

Systems Biology

From the year 2000 and onward, the term "Systems Biology" was steadily gaining momentum into becoming the new buzz word in the life sciences. It was a term signifying the importance to treat living organisms, their internal systems as well as external environment, in their entirety as opposed to the reductionist view which had been more prevalent earlier. The reductionist approach would typically concern itself with studying one small biological subsystem at a time, such as the possible biochemical function of one enzyme. Although this type of approach is useful, and often necessary, it became a common consensus that this was not enough to understand the behavior of complex biological organisms.

On a biochemical level, biological organisms consist of vast and complex networks (divided into regulatory and metabolic pathways) consisting of hundreds of thousands of biochemical compounds, known and unknown, which interact in specific ways at different time points and different locations, be it intracellular or extracellular. It is known that relatively simple networks may (sometimes unexpectedly so) give rise to emergent properties, e.g. exhibiting periodic oscillations over time or spontaneously giving rise to spatial patterns of different kinds. A well known biological example of periodic oscillation is the circadian rhythm (which governs our daily sleep pattern). Spatial patterns are well known to occur during fetal develop-

ment, as the embryo forms from a single zygote which gradually grows into a multi cellular organism, exhibiting body segmentations in the early stages and gradually growing more complex. It has been shown that other types of patterns, such as the color patterns on some animal furs [1], can be roughly reproduced by computationally simulating relatively simple biochemical reactions. Thus, there is strong indication that many morphological features of biological organisms may spontaneously arise from the mathematical properties of biochemical networks; a phenomenon which cannot be observed from a strictly reductionist framework.

Additional layers of complexity are added when considering other forces of nature in the behavior of biological systems such as that of electromagnetic fields. This area has been of particular interest during recent years due to the heavily discussed association between cancer and mobile phone use. However, the study of the effects of external electromagnetic fields on biological systems dates back far earlier into the 1950's, particularly thanks to the fascinating research done by Robert O. Becker [2], who is known for studying the effects of electromagnetic fields on bone regeneration and other aspects of wound healing.

It is thus evident that knowing the basic building blocks is not enough to understand the behavior of biological systems, but a holistic approach is indeed needed; an idea which was well summarized by Aristotle roughly 2300 years ago saying: "the whole is greater than the sum of its parts".

It is likely that the holistic ideology of Systems Biology came about as a result of the increased use of new high throughput techniques in the biosciences, which were able to measure the activities and levels of thousands of genes or proteins simultaneously. In particular, the Human Genome Project may be considered a catalyst for this trend, as it provided scientists with easy access to the human genome, along with new possibilities to unlock its secrets. In general, Systems Biology is nowadays a scientific field which involves often heavy, computational analyses of biological data from many levels, including the chemical environment, genome, proteome and metabolome.

Chapter 2

Emergence of Male Reproductive Disorders

There are a number of disorders that concern the male reproductive organs which affect the sexual function. Within the framework of my PhD, my work has primarily concerned itself with four disorders; Cryptorchidism, Hypospadias, Oligospermia and Testicular Cancer. This chapter briefly describes the clinical manifestation and prevalence of these disorders, and explains the concept which ties them together into a common syndrome; Testicular Dysgenesis Syndrome.

Cryptorchidism

Cryptorchidism, also called *Undescended Testis*, is a congenital malformation which is characterized by the absence of one or both testes from the scrotum. The prevalence of this disorder varies considerably between geographical regions. In most countries, it has been reported to occur in roughly 3-5% newborn boys born on term, although it may be as high as 9% in Denmark. It is considered to be the most common congenital malformation which affects male reproductive organs [3].

During fetal development, the gonad (testis) is first formed in the abdomen and subsequently migrates down into the scrotum. This migrations

is proposed to occur in two phases; the transabdominal phase, and the inguinoscrotal phase. In the transabdominal phase, the developing gonad travels through the abdomen to the pelvic cavity as a result of the swelling of the gubernaculum. It is believed that this phase is largely androgen-independent, and influenced by hormones such as *Anti-Mullerian Hormone* (AMH) and *Insulin-like 3* (INSL3). In the second phase, the gonad enters the inguinal canal where it migrates into the scrotum, to a large part assisted by contractions of the cremaster muscle, and is thought to be androgen dependent [4].

Cryptorchidism may manifest in various forms of severity, ranging from mild forms where a testis is located in close proximity to the scrotum, to severe forms where the testis may be completely absent or located further up the abdomen. There are also cases of so called *reascensus testis*, where the testis is located in the scrotum at birth but later reascends into the abdomen. Boys born prematurely are more likely to have cryptorchidism at birth.

Hypospadias

Hypospadias is a penile malformation where the urethral opening (urinary meatus) is not located on the tip of the glans, but instead located anywhere along the midline on the underside of the penis. Hypospadias is categorized in the three *degrees* of severity. In the first degree, the urethral opening is located on the glans, and covers about 75% of cases [5]. Second degree cases have the urethral opening on the underside of the shaft, and the third degree, which is the most severe form, is characterized by a urethral opening that is located on the perineum.

The incidence rate of hypospadias varies between countries, but has been reported to be occuring in 0.3-0.8% of newborn boys [6], and is less common than cryptorchidism.

Oligospermia

Oligospermia is characterized by semen with a low concentration of sperm. According to a recent study by the World Health Organization (WHO), sperm concentration among fertile men (Time to Pregnancy < 12 months) has a median of 73 million/mL, with the first and third quartile having 41 and 116 million/mL, respectively [7]. According to the same study, the criteria for setting the diagnosis of oligospermia is a concentration of less than 15 million/mL which corresponds to the fifth centile of the distribution of healthy fertile men. However, it has been shown that fertillity starts to be reduced already at 40 million/mL, and reduces linearly below this point while concentrations above this value do not exhibit any concentration dependent increase [8]. Besides concentration, other factors relating to sperm also play a role in the reproductive health such as sperm morphology and motility.

Testicular Cancer

In many western countries, testicular cancer is the most common cancer among men aged between 20 and 35 years, with a substantially lower risk of acquiring this disease before puberty and after the age of 40. The peak incidence at a relatively young age makes this cancer quite unusual, as it does not follow the expected pattern of risk increase with age (presumably due to accumulated genetic damage) as is common with other types of cancer. The vast majority of testicular cancers (more than 95%) are germ cell tumors, which can be divided into two main classes; seminomas and non-seminomas. These two classes exhibit clinical differences, with seminomas being less aggressive and associated with a higher 5-year survival rate among cases. Non-seminomas are characterized by a more aggressive growth and undifferentiated cellular histology, and may be composed of teratomas, choriocarcinomas, poly-embryomas and yolk-sac tumors.

It is believed that the vast majority of testicular tumors originate from carcinoma in situ (CIS) cells which are likely gonocotyes or primordial germ cells that have retained their stem cell-like qualities, without fully differenti-

ating and adopting their proper cellular role. This view is supported by the fact that CIS and testicular tumors express a number of biomarkers which are characteristic of stem cells or certain cells in early fetal development, such as the expression of c-KIT [9], OCT-4 [10, 11, 12] and placental-like alkaline phosphatase (PLAP) [13]:

- The c-KIT gene encodes a cell membrane cytokine receptor which binds to a cytokine called stem cell factor (which can exist both as a soluble protein and transmembrane protein) and plays an important role in hematopoesis as well as early germ cell survival. It has been shown to be crucial for germ cell survival in early development [14, 15].

- The OCT-4 encodes a transcription factor which is crucial for cells to retain the pluripotency. Too low, and also too high, levels of this factor promotes cellular differentiation [16].

- PLAP is a membrane bound glycosylated enzyme and is one of the most commonly used markers for CIS. Although it was identified as a marker for primordial germ cells in mice relatively early [17] its function is still largely unknown.

The stem cell-like qualities of germ cells are also further demonstrated by the very unique feature of testicular tumors of being able to mimic any other tissue in the body [18], which is a sign of pluripotency.

On a global level, the rate of incidence of testicular cancer varies greatly between countries, with Scandinavian men in Denmark and Norway exhibiting rates as high as roughly 9.5 cases per 100,000 man years, while the incidence in Setif, Algeria is as low as 0.2 cases per 100,000 man years (see Figure 2.2) [19]. Interestingly, different ethnicities within the same geographic region exhibit significant differences, with caucasian men having a 4-fold higher rate of incidence than black men in USA [20]. This observation does suggest there is a genetic component which strongly affects the predisposition for this disease.

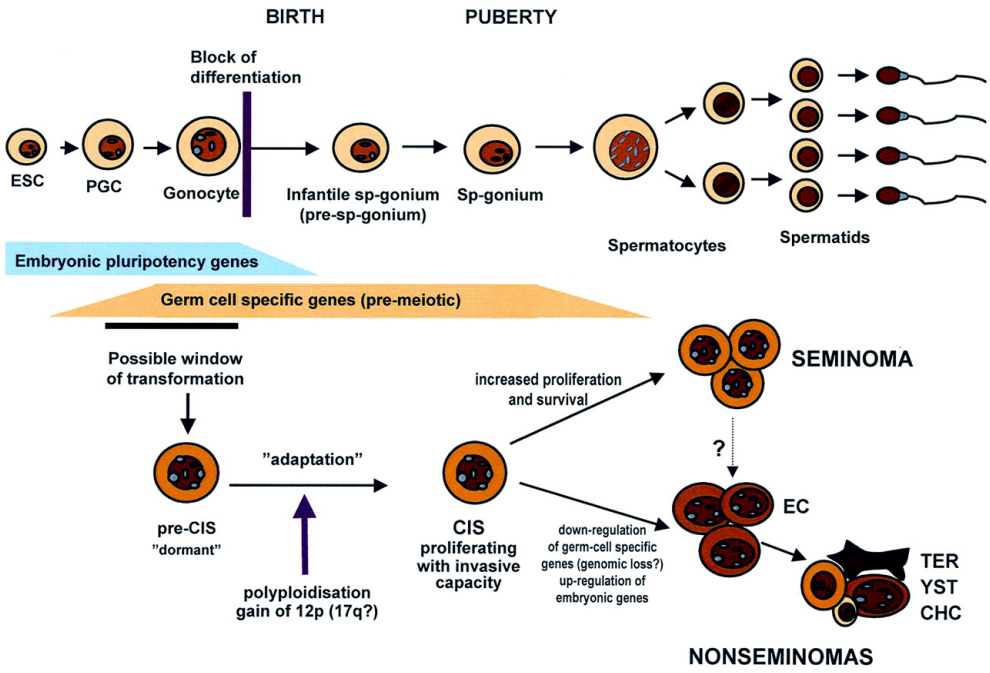

Figure 2.1: Illustrating a model for the pathogenesis of CIS in the testis. Abbreviations: EC, embryonal carcinoma; TER, teratoma; YST, yolk sac tumour; CHC, choriocarcinoma. Figure reproduced with permission from Ewa Reypert-De Meyts [18].

Numerous studies have reported a steady increase in testicular cancer from 1940's and onward in different geographic regions [20, 21, 22]. Using available cancer regisitries, Adami and collegues estimated the average annual increase in age-standardized incidence to be 2-5% (depending on geographic region), with Denmark having the highest rates of incidence

throughout the time period covered in their study. In 1940, the estimated incidence in Denmark was 2.55, and by the end of 1988 it was estimated to be 7.70; corresponding to a 300% increase. Such a rapid increase can hardly be explained by genetic factors alone, but strongly suggests that environmental factors, introduced into our modernized western society, are involved.

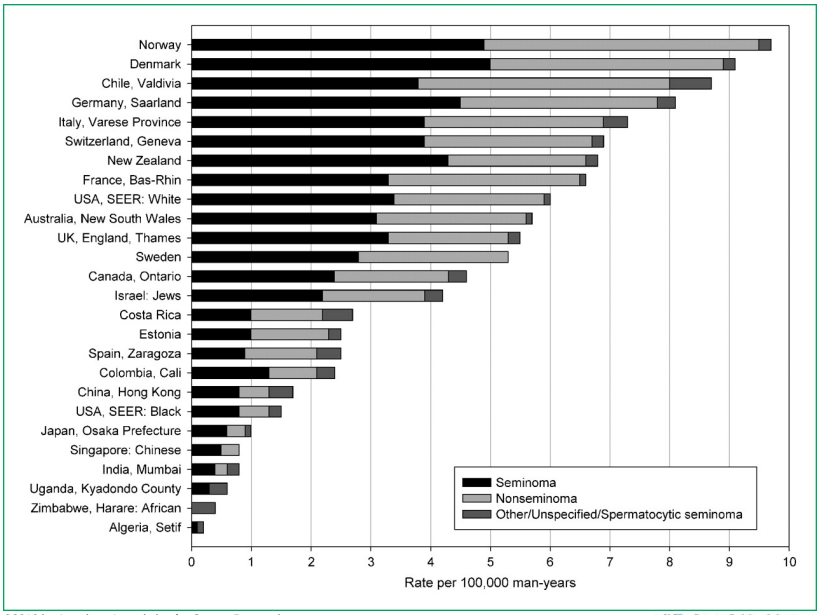

Figure 2.2: Incidence rates of testicular cancer (per 100,000 man-years) age-standardized to the world population (1998-2002). Incidences are compared between between different geographical regions. Figure reproduced with permission by Katherine McGlynn [19]

2.1 Testicular Dysgenesis Syndrome

In the clinical sciences, the collection of data on cancers in general, including testicular cancer, has traditionally been comprehensive. However, data on the incidence of other reproductive disorders, such as congenital cryptorchidism (absence of one or both testes from the scrotum), hypospadias (abnormal placement of urethral opening) and bad sperm quality, have generally been more sparse and less reliable than data on testis cancer, particularly during the 20th century. For this reason, comprehensive studies on yearly changes in the incidence rates of these disorders over longer periods of time are lacking, although analysis of available data does suggest increased incidence rates for these disorders as well [23, 24, 25]. Interestingly, evidence has suggested that the different male reproductive disorders and testicular cancer may be interrelated [26], with men born with cryptorchidism, and other reproductive disorders, having a significantly higher risk for developing testicular cancer later in life [27, 18]. Moreover, there is evidence that testicular cancer and reproductive disorders are geographically linked [28, 5].

The apparent association between the various reproductive disorders and testicular cancer gives a strong indication that the incidence of cryptorchidism, hypospadias and lower sperm counts have, like testicular cancer, also been increasing during the past 4-5 decades [29]. Moreover, the observed association has also given rise to the hypothesis that testicular cancer and other male reproductive disorders in fact share a common etiology, termed Testicular Dysgenesis Syndrome (TDS) [30]. As some of the reproductive disorders are manifest already at birth, it also indicates that the underlying cause for TDS is to be found in utero, during fetal development. Many animal studies have successfully demonstrated that male reproductive disorders can be induced through exposure of chemicals during fetal development, but that the timing of the exposure may be crucial for an effect to be observed [31].

Biology of TDS

The development of sperm cells is a process involving many stages of cell differentiation (top of Figure 2.1). Starting during fetal development, embryonic stem cells (ESC) differentiate into primordial germ cells (PGC). These cells subsequently become enclosed into the inner walls of seminiferous tubules in the testis as gonocytes. Inside these tubules the gonocytes continue to develop from pre-spermatogonium into immature spermatids over several stages. Importantly, at the stage of primary spermatocyte, the cells undergo meiosis I and meiosis II to produce haploid spermatids. During this whole process, the cells continuously stay in close contact with Sertoli cells which make up the epithelium of the seminiferous tubules and are thought to provide metabolic support to the germ cells. At the end of the process, the immature spermatids detach from the sertoli cells and travel along the seminiferous tubules into the epididimus. Here, the final stages of maturation occur which involves the growth of a tail, the formation of an axoneme with an accumulation of mitochondria and tight packaging of the DNA. The formation of a mature spermatid, from a pre-spermatogonium has been estimated to last about 64 days [32]. As the gonocytes retain their ability to divide indefinitely, the production of sperm cells in males is a life-long continuous process.

The underlying reason for the perturbation in normal germ cell development in TDS is unknown, but is hypothesized to be mediated by a dysfunction in both Sertoli cells and Leydig cells (Figure 2.3), which may in turn explain the observed associations between the different types of male reproductive disorders. Sertoli cells are activated by follicle stimulating hormone (FSH), which causes them to produce androgen binding protein (ABP), which raises the testosterone concentration in the seminiferous tubules and in turn stimulates spermatogenesis. The presence of FSH is therefore critical for the initiation of spermatogenesis. As the main function of Sertoli cells is to nurture the developing sperm cells, a dysfunction in these cells may lead to abnormal germ cell differentiation, which can cause low levels of sperm production or testicular cancer. Leydig cells are located between the seminiferous tubules and are known to produce

androgens (such as testosterone) and insulin-like factor 3 (INSL3) when stimulated by luteinizing hormone (LH). INSL3 stimulates the development of the gubernaculum, which is a structure that guides the testis down the inguinal canal during fetal development in the early phases. In the later phases, testosterone stimulates the passage of the testis from the inguinal canal down into the scrotum. A dysfunction in leydig cells may therefore cause abnormal testicular descent, leading to congenital cryptorchidism. The production of testosterone in adequate amounts during fetal development is necessary in order for skin fibroblasts to migrate and enclose the urethral groove. Thus, impaired Leydig cell function may increase the risk of hypospadias via androgen insufficiency.

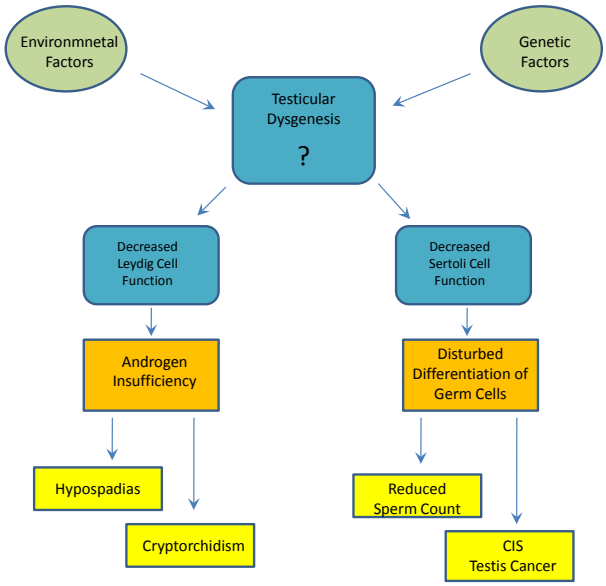

Figure 2.3: Flow chart illustrating events leading to male reproductive disorders.

The hormone producing cells in the gonads are part of a larger hormonal system called the hypothalamic-pituitary-gonadal axis. As the name implies, this system involves the hypothalamus, the pituitary gland and the gonads, and the various hormones secreted by these glands. The hypothalamus produces a hormone called gonadotropin releasing hormone (GnRH), which has the effect of stimulating the pituitary gland to produce LH and FSH. FSH stimulates the Sertoli cells in the gonads to produce inhibin, while LH stimulates the leydig cells in the gonads to produce testosterone. Both inhibin and testosterone inhibit the production of GnRH by the hypothalamus, thus creating a feedback loop within the system.

Part II

The Hunt for Environmental Factors

Chapter 3

Background: Environmental Chemicals

Envrionmental chemicals are present everywhere and affect us all the time. From the things we use in everyday life to the food we eat, we are constantly exposed to low levels of harmful chemicals; the plastic wrappers around our sandwhiches contain phthalates, the chair, sofa and computer contains flame retardants to prevent fire while the fish cought in the Baltic contains high levels of fat soluble PCBs. As humans are continuously exposed to a vast number of chemical pollutants produced by the industry, it is possible that a large number of chemicals are simultaneously responsible for the negative trends in reproductive health. Even at low levels, mixtures of chemicals can have more adverse effects than individual chemicals alone. Animal studies have demonstrated synergistic effects for mixtures of chemicals, inducing stronger and sometimes new symptoms that are not present in the case of exposure to any individual chemical [33, 34, 35]. Thus the effect of a cocktail of chemicals cannot be extrapolated from toxicological studies on individual chemicals alone. In order to assess the correlation between the chemical exposure and a given outcome, analyzing one individual chemical at a time to a disease may not be optimal. Instead, an approach which takes into account a large number of chemicals simulta-

neously is needed, and is common and well researched within the area of chemometrics.

3.1 Chemometrics

Chemometrics is a scientific area which concerns itself with applying data-driven approaches to extract relevant information from chemical systems and measurement data. The term was originally coined in 1974 by Svante Wold, the son of the late Herman Wold who is credited for having introduced Partial Least Squares (PLS) [36] regression (one of the most widely used methods in chemometrics). As chemometrics is heavily dominated by multivariate analyses on potentially large data sets, the increased use of computers in scientific research in the 70's was crucial for the development of the field at that time.

Since its introduction, chemometrics has evolved and is nowadays applied in a number of different areas. One very important area is classification and pattern recognition. This area of chemometrics is concerned with predicting or estimating certain features of interest, such as analyte concentration, based on a set of measured descriptors, e.g. readings from NIR spectra. A real-world example is the non-invasive estimation of crude lipid content in the muscle of rainbow trout from readings of short-wavelength near-infrared (SW-NIR) spectroscopy [37]. Fish are irradiated with light from the near-infrared region of the electromagnetic spectrum, while the back-scattered light is detected and measured to generate a spectrum. Such a spectrum amounts to creation of a data set containing a large number of descriptors (often in the thousands). By measuring the crude lipid content with a traditional chemical approach, such as with acid hydrolysis, a mathematical model can be trained to correlate the NIR spectra to the lipid content. This model can subsequently be used to estimate the lipid content of new fish samples without resorting to the traditional chemical analysis, thus making the process much faster and cheaper. Examples of other types of descriptors, besides NIR, is Nuclear Magnetic Resonance (NMR) spectra, Raman spectra, liquid chromatography (LC) and mass spectrometry (MS).

In particular, the combined use of LC-MS has the advantage of enabling the analyst to both quantify the levels and possibly identify the atomic composition and structure of the analyte.

3.2 Geographical Differences in Enviornmental Factors

As the increase in TDS-related phenotypes is likely to be caused by environmental factors, a natural approach to explore the validity of this hypothesis is to compare the levels of environmental chemicals in different geographical regions. When I started my phd, Niels Erik Skakkebæk and his colleagues had been working along this line for many years already, mainly making comparisons between Denmark and Finland. While Denmark seems to have one of the highest incidences of male reproductive disorders in the world, Finland on the contrary has a very low incidence. To look for relevant environmental differences between these two nations was therefore a reasonable approach. One major challenge with such an endeavor is the sheer number of environmental chemicals that exist. Measuring them all would be impossible, both practically as well as financially. Thus a narrowing down of the number of chemicals to measure is crucial. In the case of my project, a decision had been made to focus on endocrine disrupting chemicals (EDCs), i.e. chemicals which are known to affect the internal hormone levels. The chemicals which were measured included polychlorinated biphenyls (PCBs), polybrominated diphenyl ethers (PBDEs), dioxins, pesticides and phthalates. Furthermore, as TDS was believed to originate before birth, the chemicals were measured in human breast milk. Due to the high fat content of breast milk, this matrix is considered a good proxy for estimating the exposure levels of fat soluble persistent organic pollutants (POPs) to a developing fetus.

I analyzed this data using traditional statistical techniques as well as multivariate approaches commonly used in chemometrics. The results of the analysis yielded very strong support for the hypothesis that there existed a difference in levels of environmental chemicals between countries

[see Manuscript I]. As the chemical profiles between the two countries were distinct, and the vast majority of chemicals tended to be higher in Danish samples than in Finnish, our publication yielded a lot of attention in the scientific media.

3.3 Contacting the Environment Agencies

The distinct country difference came as a surprise to me and my colleagues. I was curious about what was causing the environmental difference we observed. For a period of time, I decided to do detective work to find out; something which was more difficult than I anticipated. I contacted both the Danish and the Finnish Ministries of the Environment, as well as visiting the European Environment Agency (EEA), which has its office in Copenhagen, Denmark. I was keen on being able to compare the release of chemicals into air and water between Denmark and Finland. The data I obtained by the ministries contained estimates of the release of various types of chemicals into the air, water or sediment. However, it was not well suited to be used for comparing the two nations, as some estimates were too imprecise to be useful and some measurements were lacking. I was also unable to obtain data on the rate of industrial production of the chemicals of interest, as the industry rarely reveals all chemicals in their products (according to my contact at the EEA). Thus, in my endeavor as a detective I was unable to confirm our observed country difference in other sources of data.

3.4 Enviornmental Factors related to Cryptorchidism

Having observed the distinct chemical profiles between two Nordic countries, I wanted to know if a more direct connection could be made between chemicals and male reproductive disorders [see Manuscript II]. The data from the Finnish and Danish cohort included both healthy children and

children born with congenital cryptorchidism. My analysis on the association between endocrine disrupting chemicals and the outcome of cryptorchidism indicates that there is a stronger correlation in Danish samples than in Finnish samples (i.e. a clearer association of the chemical profile to the outcome of cryptorchidism could be observed in the Danish samples). This observation fits well with the hypothesis that the adverse trend in Denmark is tied to environmental chemicals, whereas in Finland it is expected that the genetic component plays a bigger role. To my surprise, our analysis also indicated that PCBs tended to be higher among controls in the Danish samples, indicating a possible protective effect against cryptorchidism. The finding of a protective effect by PCBs may seem counter-intuitive due to their well known toxicity, but this observation is supported by previous studies [38, 39].

Chapter 4

Manuscript I: Country-Specific Chemical Signatures of Persistent Environmental Compounds in Breast Milk

international journal of andrology ISSN 0105-6263

ORIGINAL ARTICLE

Country-specific chemical signatures of persistent environmental compounds in breast milk

K. Krysiak-Baltyn,*† J. Toppari,‡ N. E. Skakkebaek,† T. S. Jensen,* H. E. Virtanen,‡ K.-W. Schramm,§¶ H. Shen,§ T. Vartiainen,** H. Kiviranta,** O. Taboureau,* S. Brunak* and K. M. Main†

*Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark, †University Department of Growth and Reproduction, Rigshospitalet, Copenhagen, Denmark, ‡Department of Physiology and Paediatrics, University of Turku, Turku, Finland, §Helmholtz Zentrum München, Institute of Ecological Chemistry, Neuherberg, Germany, ¶Department für Biowissenschaftliche Grundlagen, Technische Universität München, Weihenstephaner, Freising, Germany, and **Department of Environmental Health, National Institute for Health & Welfare, Kuopio, Finland

Keywords:
breast milk, chemical signature, endocrine disrupting chemicals, geographical differences, semen quality, testicular dysgenesis syndrome

Correspondence:
Niels E. Skakkebaek, University Department of Growth and Reproduction, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark. E-mail: nes@rh.dk

Received 6 July 2009; revised 18 August 2009; accepted 19 August 2009

doi:10.1111/j.1365-2605.2009.00996.x

Summary

Recent reports have confirmed a worldwide increasing trend of testicular cancer incidence, and a conspicuously high prevalence of this disease and other male reproductive disorders, including cryptorchidism and hypospadias, in Denmark. In contrast, Finland, a similarly industrialized Nordic country, exhibits much lower incidences of these disorders. The reasons behind the observed trends are unexplained, but environmental endocrine disrupting chemicals (EDCs) that affect foetal testis development are probably involved. Levels of persistent chemicals in breast milk can be considered a proxy for exposure of the foetus to such agents. Therefore, we undertook a comprehensive ecological study of 121 EDCs, including the persistent compounds dioxins, polychlorinated biphenyls (PCBs), pesticides and flame retardants, and non-persistent phthalates, in 68 breast milk samples from Denmark and Finland to compare exposure of mothers to this environmental mixture of EDCs. Using sophisticated, bioinformatic tools in our analysis, we reveal, for the first time, distinct country-specific chemical signatures of EDCs with Danes having generally higher exposure than Finns to persistent bioaccumulative chemicals, whereas there was no country-specific pattern with regard to the non-persistent phthalates. Importantly, EDC levels, including some dioxins, PCBs and some pesticides (hexachlorobenzene and dieldrin) were significantly higher in Denmark than in Finland. As these classes of EDCs have been implicated in testicular cancer or in adversely affecting development of the foetal testis in humans and animals, our findings reinforce the view that environmental exposure to EDCs may explain some of the temporal and between-country differences in incidence of male reproductive disorders.

Introduction

A considerable increase in testicular cancer incidence among young men during the last century has taken place worldwide and large scale geographical differences in the incidence of this disease exist (Bray *et al.*, 2006). There is a remarkable three to fourfold higher incidence of testicular cancer in Denmark in comparison with the nearby country Finland. We previously tested the hypothesis that testicular cancer incidence may be a 'whistleblower' for occurrence of other reproductive health problems in a population by carrying out large, coordinated, prospective studies of cohorts of newborn boys and their mothers in Denmark and Finland. These showed that the incidence of cryptorchidism and hypospadias was also three to fourfold higher in Denmark than in Finland (Boisen *et al.*, 2004, 2005). Even among healthy newborn boys, there were significant differences

as Finnish boys had larger testes than Danish and higher levels of inhibin B, a marker of testicular Sertoli cells (Main *et al.*, 2006b). Prospective studies of the general adult populations have also revealed higher sperm counts in Finnish than in Danish men (Jørgensen *et al.*, 2002).

A crucial question is to what extent these conspicuous differences in occurrence of reproductive problems between two Nordic countries are because of environmental factors. Studies of immigrants' testicular cancer risk have shown that second generation immigrants have similar risk of cancer to that of the host country population (Hemminki & Li, 2002; Myrup *et al.*, 2008). This, together with the increasing trends of male reproductive health problems, strongly suggests that environmental rather than genetic factors play a major role. As humans have been widely exposed over the same time period to man-made persistent EDCs, their aetiological involvement has been suspected. In reality, humans are exposed not to single EDCs, but to complex mixtures and the latest evidence from animal studies shows that such mixtures can have profound effects on male reproductive development at concentrations at which the individual EDCs have no effect (Christiansen *et al.*, 2008; Kortenkamp, 2008; Rider *et al.*, 2009).

Therefore, we have undertaken an ecological study to examine whether exposures to EDC pollutants is higher in Denmark than in Finland. We measured 121 chemicals (listed in Tables S1 and S2) in 68 breast milk samples from 36 Danish and 32 Finnish women who gave birth to healthy boys. Chemicals studied included flame retardants, such as polybrominated diphenyl ethers (PBDE) and biphenyls (PBB), organochlorine pesticides (OC), polychlorinated dibenzo-*p*-dioxins (PCDD/F dioxins) and biphenyls (PCB), and phthalates, all known for their potential as endocrine disruptors. Breast milk was chosen because concentrations of pollutants in milk fat are considered to represent human exposures (Smith, 1999; Wang & Needham, 2007).

Materials and methods

The data set for this analysis was obtained from a joint prospective bi-national study of pregnant women and their offspring between 1997 and 2001. This study aimed at assessing the current prevalence of congenital cryptorchidism and hypospadias in Denmark and Finland as well as identifying environmental and lifestyle factors possibly associated with testis development and function. Questionnaires and breast milk samples were obtained. The design of the study was previously described, as well as details on breast milk sample collection and selection for chemical analysis; some of the data were included in other investigations (Main *et al.*, 2006a; Shen *et al.*, 2006, 2007).

The original data set consisted of 130 breast milk samples from mothers of newborn children. Sixty-eight of the newborns (36 Danish and 32 Finnish) were healthy and without signs of reproductive malformations and 62 were born cryptorchid. As breast milk from women who delivered a boy with cryptorchidism may be a major confounder in an analysis of the general exposure levels to EDCs in a population, we only included breast milk of the 68 mothers who gave birth to healthy boys. A total of 121 chemicals were analysed; however, 12 chemicals with non-detectable levels in all samples were excluded from the final statistical analysis (Tables S1 and S2).

During all chemical analyses, the laboratories and technicians were blinded for country of origin. All laboratories participated in external quality control programmes. Pesticides including enantiomeric compounds and polybrominated biphenyls were analysed at the Institute of Ecological Chemistry, Neuherberg, Germany (Damgaard *et al.*, 2006) and polybrominated diphenyl ethers, dioxins, PCB's and furans at the Department of Environmental Health, National Public Health Institute, Kuopio, Finland (Main *et al.*, 2007). All phthalate analyses were performed at chemical laboratory at the Department of Growth and Reproduction, Copenhagen, Denmark (Mortensen *et al.*, 2005; Main *et al.*, 2006a).

The Danish mothers were slightly younger than the Finnish and more of them participated with their first child in the study. Moreover, Danish samples were collected on average 1.8 years later than Finnish samples. These potential confounders, and others, were adjusted for in the analysis.

To assess the extent of differences in exposure to individual chemicals between Denmark and Finland, chemical concentrations in breast milk samples were analysed using linear multiple regression. The *p*-values were corrected for multiple testing by the method of Bonferroni. Potential confounders, known to affect the level of chemicals in breast milk samples, were added as covariates in the analysis, including maternal age, maternal body mass index (BMI), year of milk sampling, maternal smoking (yes/no), maternal diabetes (yes/no) and parity.

We investigated the differences in combined chemical exposures between the two countries using machine-learning classifiers, which simultaneously take all chemical concentrations into account. These classifiers can detect any combination or pattern of chemicals which discriminates between Danish and Finnish samples. Such patterns may describe for example if the sum of two or more chemicals must be above a certain level, or the level of one chemical is high whilst the level of another chemical is low. Three Machine Learning Classifiers were applied

for comparison, two of which were linear methods [Partial Least Squares (PLS) (Wold, 1966) and Sequential Minimal Optimization (SMO) with 1st order polynomial kernel], and one was a non-linear method (Multilayer Perceptron with one hidden layer of 5 nodes).

For the machine learning classifiers, analysis was performed both with and without adjusting for confounders. Confounders were adjusted for by interpolating the data using regression coefficients. In addition, confounders were adjusted for in the analysis with PLS by adding the confounding factors, along with country, as response-variables.

Two different software programs were used; Simca-P 10.5 (by Umetrics Inc., Umeå, Sweden) was used for performing PLS, and Weka 3.5.3 (Witten & Frank, 2005) was used for performing analysis with the other classifiers. One fourth of the samples (17 samples) in the original data were randomly removed and used as a test set for external validation. A balanced number of samples from each country were included in the test set.

Non-detectable sample measurements were treated in three separate ways: they were either set to 0, to half the Limit of Quantification (LOQ), or to LOQ. All analyses presented were repeated for each case. In addition to the measured values of the chemicals, we analysed sums and toxic equivalencies (TEQs) of PCBs, PBDEs and PBBs. Finally, as phthalates differ in their chemical properties and exposure routes compared with all the other compounds in our data set, they were analysed together in a separate model.

Results

After correcting for multiple testing, six chemicals exhibited significant differences between the two countries and all were higher in Danes than in Finns (Table 1, Fig. 1). Without statistical correction for multiple testing, higher concentrations in Danish samples were observed for the vast majority (54 out of 58) of chemicals that exhibited a significant between-country difference (Tables 2A and 2B). Chemicals which did not differ significantly are listed in Table 2C (resulting *p*-values for all chemicals, including TEQs, can be found in Tables S3 and S4).

Analyses with the machine-learning classifiers showed that the chemical exposures in the two countries were so distinct that perfect, or near perfect, separation of samples with respect to country of origin was possible (Fig. 2). Obviously, not all chemicals contribute equally to this difference. To examine the importance of each chemical, we examined the weights of each chemical in the models and performed feature selection by training the different machine-learning methods using only a subset of chemicals (Tables S5–S8 for performance of the machine learning analysis for various models). Each of the methods achieved perfect separation of Danish and Finnish samples using a slightly different set of chemicals (Tables S9–S11). Several of the chemicals were used by two machine-learning methods, and two chemicals, 1,2,3,4,7,8-HCDD and 1,2,3,6,7,8-HCDD, were consistently selected by all three methods. Indeed, the combination of these two chemicals alone perfectly separated the Danish and Finnish samples (Fig. 3). Moreover, the clear between-country difference was robust and did not disappear if either of these two chemicals was left out of the analyses. These results did not change when the chemical levels below the limit of quantification were assigned to 0, one half of LOQ or LOQ.

As phthalates differ from the persistent compounds in their chemical properties, exposure routes and persistence, they were also analysed in a separate model. The phthalate levels alone did not exhibit any strong separation between the two nations.

Discussion

Our comprehensive analysis of more than one-hundred environmental chemicals in contemporary breast milk from Finland and Denmark revealed conspicuous

Table 1 Chemicals with significantly higher concentrations in Danish than in Finnish breast milk samples in a linear multiple regression analysis after correction for multiple testing. Percentiles show unadjusted concentrations

	Percentile, Denmark			Percentile, Finland				
Chemical	25th	50th	75th	25th	50th	75th	*p*-value	Higher in
1,2,3,4,7,8-HCDD	2.39e-3	3.32e-3	4.76e-3	0.78e-3	1.06e-3	1.28e-3	2.18e-4	Denmark
PCB 209	0.088	0.11	0.16	0.045	0.061	0.092	3.27e-5	Denmark
PCB 156	4.21	5.66	8.62	2.83	3.59	4.83	1.07e-2	Denmark
PCB 157	0.70	0.86	1.27	0.44	0.60	0.80	2.25e-2	Denmark
Dieldrin	3.06	4.66	5.98	1.86	2.21	3.10	2.30e-4	Denmark
Hexachlorobenzene	8.80	11.78	14.16	6.87	7.60	8.55	1.32e-4	Denmark

Levels below LOQ were assigned the value 0. Data are given as mean ± SD. LOQ, Limit of Quantification.

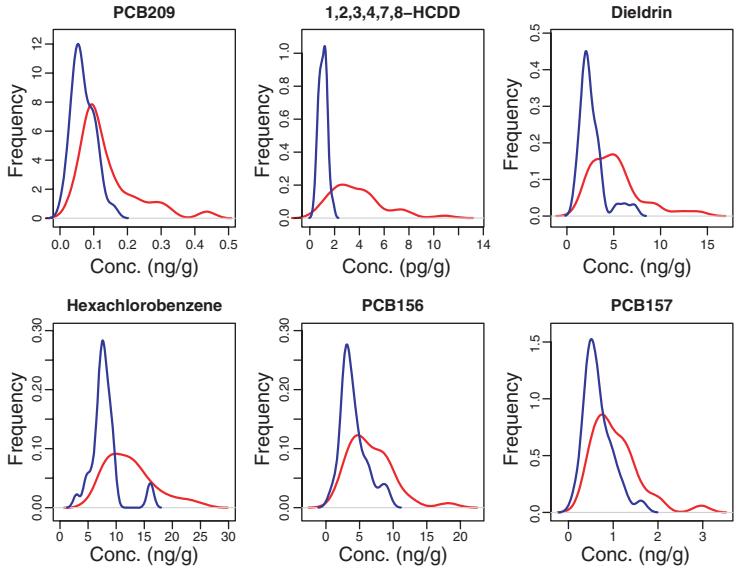

Figure 1 Plots show the distribution of concentrations of chemicals which exhibited significant differences between Denmark (red) and Finland (blue) in a linear multiple regression. Chemicals were measured in breast milk samples from Danish and Finnish mothers of healthy boys.

Table 2A Chemicals that had significantly higher concentrations in Danish than in Finnish breast milk samples in a linear multiple regression analysis *before* correction for multiple testing ($p < 0.05$). Chemicals that still differed significantly after correction are shown in bold (see supplementary Tables S3 and S4 for details)

Class of chemicals	Congeners
PBDE	BDE-153
PBB	2-BB, 4-BB, 22′-BB, 344′-BB, 33′44′-BB, 22′45′6-BB, 33′44′5-BB, 33′44′55′-BB
Organochlorine pesticides	Hexachlorocyclohexanes [(+)-α-HCH, (–)-α-HCH, α-HCH (sum of enantiomers), β-HCH, ε-HCH]
	o,p′-DDT, 1,1,1-trichloro-2-(2-chlorophenyl)-2-(4-chlorophenyl)ethane
	p,p′-DDT, 1,1,1-trichloro-2,2-bis(4-chlorophenyl)ethane
	o,p′-DDE, 1,1-dichloro-2-(2-chlorophenyl)-2-(4-chlorophenyl)ethene
	p,p′-DDE, 1,1-dichloro-2,2-bis(4-chlorophenyl)ethene
	Dieldrin, **Hexachlorobenzene (HCB)**, Aldrin, Trans-chlordane, Heptachlor, (+)-Oxychlordane (OXC), (–)-Oxychlordane, (OXC), Oxychlordane (Sum of enantiomers), (+)-Cis-heptachlor epoxide (HE), (–)-Cis-heptachlor epoxide (HE), Trans-heptachlor epoxide, Endosulfan-I, Pentachlorobenzene
PCB	PCB-28/31, -49, -60, -66, -74, -77, -81, -99, -110, -114, -128, -138, -153, **-156**, **-157**, -167, -169, -170, -180, -183, -187, -189, -194, -206, **-209**, Sum of PCBs, WHO-TEQ
PCDD/PCDF's	2378-TCDD, 12378-PD, 123478-HF, 123678-HF, **123478-HD**, 123789-HD

PBDE, polybrominated diphenyl ethers; PBB, polybrominated biphenyls; PCDD/F dioxins, polychlorinated dibenzo-*p*-dioxins; PCB, polychlorinated biphenyls.

differences, particularly with regard to concentrations of persistent organic pollutants (POPs). In fact, an analysis of only two dioxins could totally separate the Danish breast milk from the Finnish breast milk. Another important finding was that the levels of chemicals were generally higher in the Danish samples, where the concentration range of POPs was also much broader and included some quite high values. Thus, taken together, the exposure levels of the examined chemicals and their mixture pattern seemed quite different in Denmark and Finland.

Three classes of chemicals were represented by the compounds that were found in significantly higher concentrations and with broader distribution spectrums in Danish samples; PCBs, organochlorine pesticides and PCDDs. Several of these have been implicated in impairment of foetal testis development or testis cancer

Table 2B Chemicals that had significantly higher concentrations in Finnish than in Danish breast milk samples in a linear multiple regression analysis *before* correction for multiple testing ($p < 0.05$). No chemical was significantly higher in Finland after correction for multiple testing (see supplementary Tables S3 and S4 for details)

Class of chemicals	Congeners
Phthalate Monoesters	Mono-butylphthalate, mono-benzylphthalate
Organochlorine pesticides	Methoxychlor
PCB	PCB-51

PCB, polychlorinated biphenyls.

(Toppari *et al.*, 1996; Hardell *et al.*, 2003; Main *et al.*, 2006a; Fowler *et al.*, 2007; Andersen *et al.*, 2008; Damgaard *et al.*, 2008; McGlynn *et al.*, 2008).

An important question is whether the apparent different exposure levels can explain the marked differences in male reproductive health problems between the two countries. Most previous studies examining links between exposures and reproductive health problems, including our own previous investigations, have focused on possible effects of single agents or a group of related agents at a time (Main *et al.*, 2007; Kortenkamp, 2008). However, it seems important that evaluations of effects of chemicals on human health should include as many as possible of the agents constituting the 'total pollution cocktail' to estimate the combined effect (Christiansen *et al.*, 2008). Recent animal studies have, in fact, shown that combined exposures to multiple chemicals had significant adverse effects, although previous dose–response studies had shown no effects when the chemicals were administered one at the time at low concentrations (Christiansen *et al.*, 2008; Rider *et al.*, 2008). In the present study, we therefore made use of advanced bioinformatics software programs to extract the total information of all analysed chemicals in all breast milk samples.

Why more POPs in Danish milk?

We were unable to find data which could explain the generally higher levels of EDCs in Danish samples. The major source of human exposure to POPs is from fatty foods (Wang & Needham, 2007). According to the National Danish Implementation Plan of the Stockholm Convention, Miljøstyrelsen (Danish Ministry of the Environment) (2006), the levels of POPs in Danish foodstuff should not raise cause for concern. Furthermore, regulation of the most significant chemicals in Table 1 show no striking differences between Finland and Denmark. Differences in regulatory practices may therefore not account for the specific chemical signatures of the two populations.

Polychlorinated biphenyls have been produced since 1929 and used in many applications such as in paints, plastizisers and dielectric fluids in capacitors and transformers. The sale of PCBs and PCB containing products was banned in Denmark in 1986. In Finland, the manufacture and use of PCB containing products was banned in the early 1990's [Finnish Environment Institute, 2006; Miljøstyrelsen (Danish Ministry of the Environment), 2006].

Polychlorinated dibenzo-*p*-dioxins are unintentionally produced as byproducts in many industrial processes (e.g. paper bleaching), traffic and waste combustion. In Denmark, the total emission of chlorinated dioxins into air (including PCDD and PCDF) in 2000–2002 was estimated to be 11–162 g I-TEQ/year (Hansen & Hansen, 2003). In Finland, the corresponding number was

Table 2C Chemicals that did not differ significantly ($p > 0.05$) between Danish and Finnish breast milk samples in a linear multiple regression analysis of their concentrations (see supplementary Tables S3 and S4 for details)

Class of chemicals	Congeners
PBDE	BDE-28, -47, -66, -75, -77, -85, -99, -100, -119, -138, -154, -183, Sum of PBDEs
PBB	2-BB, 4-BB, 22′-BB, 24′5-BB, 344′-BB, 22′55′-BB, 22′45′-BB, 33′55′-BB, 22′45′6-BB, 22′455′-BB, 33′44′5-BB, 22′44′66′-BB, 22′44′55′-BB, 33′44′55′-BB, Pentabromobenzene (PeBB), Hexabromobenzene (HeBB)
Phthalate monoesters	Mono-methylphthalate, mono-ethylphthalate, mono-2-ethylhexylphthalate, mono-isononylphthalate
Organochlorine pesticides	Hexachlorocyclohexanes (γ-HCH, δ-HCH, ε-HCH) o,p′-DDD, 1,1-dichloro-2-(2-chlorophenyl)-2-(4-chlorophenyl)ethane p,p′-DDD, 1,1-dichloro-2,2-bis(4-chlorophenyl)ethane Octachlorostyrene (OCS), Pentachloroanisole (PCA), Aldrin, Cis-chlordane, Heptachlor, Trans-heptachlor epoxide, Mirex, Endosulfan-II
PCB	PCB-18, -33, -47, -52, -101, -105, -118, -122, -123, -126, -141
PCDD/PCDF's	2378-TCDF, 12378-PF, 23478-PF, 1234678-F, 1234789-F, 234678-HF, 123789-HF, 123678-HD, 1234678-D, OCDF, OCDD, WHO-TEQ, Sum PCDD/F

PBDE, polybrominated diphenyl ethers; PBB, polybrominated biphenyls; PCDD/F dioxins, polychlorinated dibenzo-*p*-dioxins; PCB, polychlorinated biphenyls.

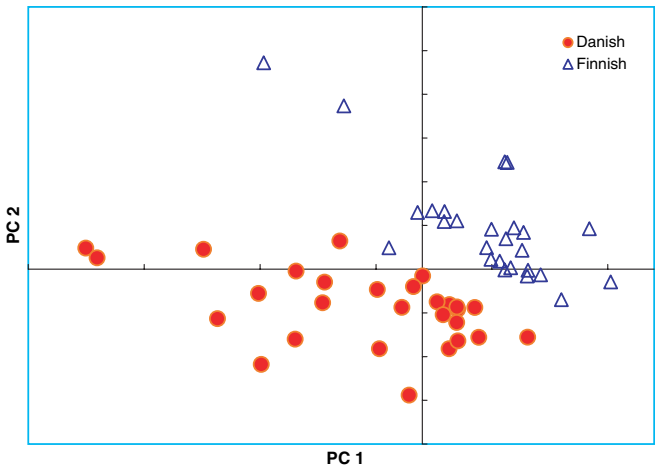

Figure 2 Scatter plot generated from the Partial Least Squares (PLS) model. Each point represents one milk sample, blue: Finnish, red: Danish. The location of each point is a reflection of the chemical concentration profile in the breast milk. The y- and x-axis are the first and second principal components, respectively, which are linear combinations of the concentration of the chemicals. The top 10 most important chemicals in each of the two principal components are listed as follows: PCI: 1,2,3,4,7,8-HCDD, PCB 209, PCB 156, PCB 189, PCB 170, PCB 157, PCB 194, PCB 180, o.p'-DDE, PCB 81. PC2: 1,2,3,6,7,8-HCDD, 1,2,3,4,6,7,8-HepCDD, Mirex, 1,2,3,4,6,7,8-HepCDF, OCDD, PeBB, BDe 154,1,2,3,4,7,8-HCDD, PCB 49, Octachlorostyrene.

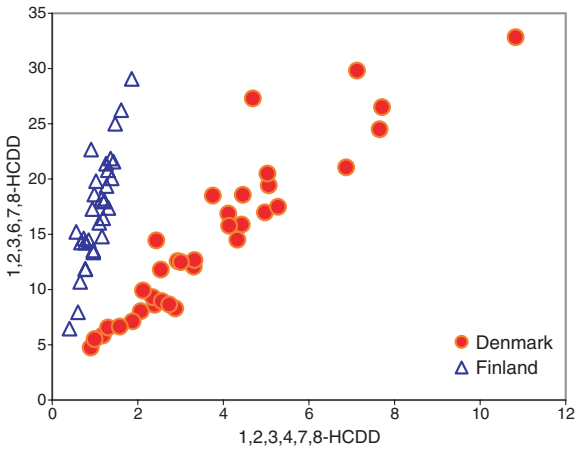

Figure 3 Two-dimensional scatter plot shows the concentration of the two chemicals 1,2,3,4,7,8-HCDD (x-axis) and 1,2,3,6,7,8-HCDD (y-axis) in each breast milk sample. The Danish (red) and Finnish (blue) samples are completely separated into two distinct groups. In each country, the two chemicals show clear linear correlations but with different slopes.

estimated to be 32.24 g I-TEQ/year in 2002 (Finnish Ministry of the Environment, 2006). As the Danish estimate is imprecise, we are unable to assess in which country the emission is higher. In our dataset, seven different PCDD congeners were present including 2,3,7,8-TCDD, which has the highest TEQ factor of all PCDDs and therefore considered the most toxic PCDD. Among the PCDDs measured, only 1,2,3,4,7,8-HCDD differed significantly between the nations and the levels of this compound was somewhat higher than the similar 2,3,7,8-TCDD. However, different PCDD-congeners vary in their biological effects (Niittynen *et al.*, 2007) and therefore the TEQ factors for the compounds do not necessarily reflect their endocrine disrupting potential and their effect on male reproductive health.

Hexachlorobenzene and dieldrin are both organochlorine pesticides that were introduced at about the same time in the 1940s. In Denmark and Finland, this hexachlorobenzene was banned from use as a pesticide in 1993 and 1996 respectively and was totally banned in 2003 and 2002, indicating that the regulation has been similar in both countries [Finnish Environment Institute, 2006; Miljøstyrelsen (Danish Ministry of the Environment), 2006]. Dieldrin, which recently has been shown to be toxic to foetal Leydig cells at low concentrations (Fowler *et al.*, 2007) has been used in Denmark in small amounts between 1956 and 1988 [Miljøstyrelsen (Danish Ministry of the Environment), 2006]. In Finland, dieldrin was stopped being used as a pesticide in 1970, but it was still manufactured for treatment of plywood for export

until 2002. An extensive study of possibly contaminated sites in Finland in the early 1990s indicated that levels of dieldrin were very low (Finnish Ministry of the Environment, 2006).

Limitations and strengths

The participating lactating women had a narrow age distribution and were mainly from higher social classes in both countries. Therefore, they may not represent the populations in general. However, as most of the EDCs that we have measured are widely distributed in the environment and/or tend to accumulate in fat and in the food chain, it is likely that our findings in general are applicable to the wider population. Danish samples were on average collected 1.8 years later than Finnish samples, which is a confounder as environmental concentrations of the measured POPs have generally been declining during the past decade (Zietz *et al.*, 2008). However, as we detected higher levels of POPs in Danish samples, the true differences in exposures between Danes and Finns are likely to be even greater than we observed.

In addition, although our study included more than a hundred reproductive toxicants, it should be remembered that current environmental exposures involve many thousand chemicals which were not included in our study, but could still be part of the problem. These include perfluorinated compounds and several non-persistent chemicals for example currently used pesticides and industrial chemicals like bisphenol A, several phthalates and phthalate metabolites not included in our study, certain sun blockers, phytoestrogens and mycotoxins. Furthermore, we know little about the genetic variations in the metabolism of, and susceptibility to, these drugs. Therefore, although our study was extensive, future studies relating chemical exposures to diseases should aim at including an even larger list of these ubiquitous chemicals along with genotype data. Thus, exposure to the chemicals we analysed here may not alone explain the difference in incidence of male reproductive problems between the two nations.

Implications

Persistent chemicals obviously give rise to exposure of newborn babies through breastfeeding after birth. However, the levels of chemicals in breast milk can also be considered a proxy for exposure of the foetus during pregnancy by transfer across the placenta, as these persistent chemicals with very long metabolic half lives in the body show strong correlations between levels in breast milk and concentrations in maternal and foetal serum (Wang & Needham, 2007). Although little is known about the possible reproductive effects in the foetus of most of the measured chemicals, a number of them have already been implicated in such effects in animal and human studies. Additionally, more and more data suggest that the foetal testis is inherently more susceptible to endocrine-disrupting effects than the adult gonad (Andersen *et al.*, 2008; Welsh *et al.*, 2008). Our findings of distinctly different chemical exposure patterns and significantly higher levels of persistent compounds in samples from Denmark than from Finland could therefore play a causal role in the (yet unexplained) higher Danish incidence of male reproductive disorders.

Our data reinforce current thinking of the need to minimize human exposure to EDCs on precautionary grounds. In this regard, it should be noted that some such compounds (e.g. dioxins and flame retardants) are still being unintentionally or intentionally produced and released. Although human exposure to most POPs has decreased, this may not yet be discernible in incidence rates of testis cancer or other male reproductive disorders. For reasons of the persistent nature of these compounds, exposure will be passed on to the next one or even two generations and there is a long latency between perinatal exposure and (adult) manifestations of many reproductive disorders.

Conclusion

This comprehensive study on endocrine disrupting chemicals in Danish and Finnish breast milk samples revealed conspicuous differences: specific chemical signatures were found in the two countries. In addition, the levels of persistent compounds were significantly higher in samples from Denmark, where higher incidences of testis cancer, cryptorchidism, hypospadias and poor semen quality are present. Our findings are important, as these compounds are known for their endocrine disrupting abilities. Furthermore, animal studies, as well as recent human studies, have shown associations between some of the same agents and male reproductive problems.

Funding

This work was made possible by grants from Villum Kann Rasmussen Foundation, The Danish Medical Research Council, The Novo Nordisk Foundation, The Svend Andersen's Foundation, The program commission on Nanoscience, Biotechnology and IT (NABITT), EU (EDEN, DEER, EXPORED), Academy of Finland, Sigrid Jusélius Foundation, Pediatric Research Foundation, Turku University Hospital. The study sponsors played no role in the planning and execution of the research work. All researchers were independent from the study sponsors.

Competing interest declaration

None declared.

Ethical approval

The study was approved by the ethical committees in Finland and Denmark, which included the Joint Commission on Ethics of the Turku University and the Turku University Central Hospital, Turku, Finland; the Ethical Committees of Copenhagen and Frederiksberg County, Copenhagen, Denmark; and the Danish Data Protection Agency, Copenhagen, Denmark. The study complied with the Helsinki II declaration (World Medical Association 2004), after informed oral and written consent of the parents.

References

Andersen, H. R., Schmidt, I. M., Grandjean, P., Jensen, T. K., Budtz-Jorgensen, E., Kjaerstad, M. B., Baelum, J., Nielsen, J. B., Skakkebæk, N. E. & Main, K. M. (2008) Impaired reproductive development in sons of women occupationally exposed to pesticides during pregnancy. *Environmental Health Perspectives* 116, 566–572.

Boisen, K. A., Kaleva, M., Main, K. M., Virtanen, H. E., Haavisto, A.-M., Schmidt, I. M. *et al.* (2004) Difference in prevalence of congenital cryptorchidism in infants between two Nordic countries. *Lancet* 363, 1264–1269.

Boisen, K. A., Chellakooty, M., Schmidt, I. M., Kai, C. M., Damgaard, I. N., Suomi, A. M., Toppari, J., Skakkebæk, N. E. & Main, K. M. (2005) Hypospadias in a cohort of 1072 Danish newborn boys: prevalence and relationship to placental weight, anthropometrical measurements at birth, and reproductive hormone levels at 3 months of age. *Journal of Clinical Endocrinology and Metabolism* 90, 4041–4046.

Bray, F., Ferlay, J., Devesa, S. S., McGlynn, K. A. & Møller, H. (2006) Interpreting the international trends in testicular seminoma and nonseminoma incidence. *Nature Clinical Practice. Urology* 3, 532–543.

Christiansen, S., Scholze, M., Axelstad, M., Boberg, J., Kortenkamp, A. & Hass, U. (2008) Combined exposure to anti-androgens causes markedly increased frequencies of hypospadias in the rat. *International Journal of Andrology* 31, 241–248.

Damgaard, I. N., Skakkebæk, N. E., Toppari, J., Virtanen, H. E., Shen, H., Schramm, K. W., Petersen, J. H., Jensen, T. K., The Nordic Cryptorchidism Study Group & Main, K. M. (2006) Persistent pesticides in human breast milk and cryptorchidism. *Environmental Health Perspectives* 114, 1133–1138.

Damgaard, I. N., Jensen, T. K., Petersen, J. H., Skakkebæk, N. E., Toppari, J. & Main, K. M. (2008) Risk factors for congenital cryptorchidism in a prospective birth cohort study. *PLoS ONE* 3, e3051.

Finnish Environment Institute. (2006) References to the restrictions of intentionally produced POPs. http://www.miljo.fi/download.asp?contentid=51365&lan=EN (accessed 07 September 2009).

Finnish Ministry of the Environment. (2006) National implementation plan for the stockholm convention on persistent organic pollutants. http://www.ymparisto.fi/download.asp?contentid=51364&lan=EN (accessed 7 September 2009).

Fowler, P. A., Abramovich, D. R., Haites, N. E., Cash, P., Groome, N. P., Al Qahtani, A., Murray, T. J. & Lea, R. G. (2007) Human fetal testis Leydig cell disruption by exposure to the pesticide dieldrin at low concentrations. *Human Reproduction* 22, 2919–2927.

Hansen, E. & Hansen, C. (2003) Substance Flow Analysis for Dioxin 2002. Environmental Project no. 811. The Danish Environmental Protection Agency. http://www2.mst.dk/Udgiv/publications/2003/87-7972-675-5/pdf/87-7972-676-3.pdf (accessed 07 September 2009).

Hardell, L., van Bavel, B., Lindstrom, G., Carlberg, M., Dreifaldt, A. C., Wijkstrom, H., Starkhammar, H., Eriksson, M., Hallquist, A. & Kolmert, T. (2003) Increased concentrations of polychlorinated biphenyls, hexachlorobenzene, and chlordanes in mothers of men with testicular cancer. *Environmental Health Perspectives* 111, 930–934.

Hemminki, K. & Li, X. (2002) Cancer risks in Nordic immigrants and their offspring in Sweden. *European Journal of Cancer* 38, 2428–2434.

Jørgensen, N., Carlsen, E., Nermoen, I., Punab, M., Suominen, J., Andersen, A.-G. *et al.* (2002) East-West gradient in semen quality in the Nordic-Baltic area: a study of men from the general population in Denmark, Norway, Estonia and Finland. *Human Reproduction* 17, 2199–2208.

Kortenkamp, A. (2008) Low dose mixture effects of endocrine disrupters: implications for risk assessment and epidemiology. *International Journal of Andrology* 31, 233–240.

Main, K. M., Mortensen, G. K., Kaleva, M., Boisen, K., Damgaard, I., Chellakooty, M. *et al.* (2006a) Human breast milk contamination with phthalates and alterations of endogenous reproductive hormones in three months old infants. *Environmental Health Perspectives* 114, 270–276.

Main, K. M., Toppari, J., Suomi, A.-M., Kaleva, M., Chellakooty, M., Schmidt, I. M. *et al.* (2006b) Larger testes and higher inhibin B levels in Finnish than in Danish newborn boys. *Journal of Clinical Endocrinology and Metabolism* 91, 2732–2737.

Main, K. M., Kiviranta, H., Virtanen, H. E., Sundqvist, E., Tuomisto, J. T., Tuomisto, J., Vartiainen, T., Skakkebæk, N. E. & Toppari, J. (2007) Flame retardants in placenta and breast milk and cryptorchidism in newborn boys. *Environmental Health Perspectives* 115, 1519–1526.

McGlynn, K. A., Quraishi, S. M., Graubard, B. I., Weber, J. P., Rubertone, M. V. & Erickson, R. L. (2008) Persistent organochlorine pesticides and risk of testicular germ cell tumors. *Journal of the National Cancer Institute* 100, 663–671.

Miljøstyrelsen (Danish Ministry of the Environment). (2006) National Implementation Plan for the Stockholm Convention on Persistent Organic Pollutants. http://www2.mst.dk/Udgiv/publications/2006/87-7052-067-4/pdf/87-7052-068-2.pdf (accessed 07 September 2009).

Mortensen, G. K., Main, K. M., Andersson, A.-M., Leffers, H. & Skakkebæk, N. E. (2005) Determination of phthalate monoesters in human breast milk, consumer milk and infant formula by tandem mass spectrometry(LC/MC/MS). *Analytical Bioanalytical Chemistry* 382, 1084–1092.

Myrup, C., Westergaard, T., Schnack, T., Oudin, A., Ritz, C., Wohlfahrt, J. & Melbye, M. (2008) Testicular cancer risk in first- and second-generation immigrants to Denmark. *Journal of the National Cancer Institute* 2008 Jan.2;100(1): 41.-7.Epub.2007.Dec.25 . 100, 41–47.

Niittynen, M., Simanainen, U., Syrjala, P., Pohjanvirta, R., Viluksela, M., Tuomisto, J. & Tuomisto, J. T. (2007) Differences in acute toxicity syndromes of 2,3,7,8-tetrachlorodibenzo-p-dioxin and 1,2,3,4,7,8-hexachlorodibenzo-p-dioxin in rats. *Toxicology* 235, 39–51.

Rider, C. V., Furr, J., Wilson, V. S. & Gray, L. E. Jr. (2008) A mixture of seven antiandrogens induces reproductive malformations in rats. *International Journal of Andrology* 31, 249–262.

Rider, C. V., Wilson, V. S., Howdeshell, K. L., Hotchkiss, A. K., Furr, J. R., Lambright, C. R. & Gray, L. E. Jr. (2009) Cumulative effects of in utero administration of mixtures of antiandrogens on male rat reproductive development. *Toxicologic Pathology* 37, 100–113.

Shen, H., Virtanen, H. E., Main, K. M., Kaleva, M., Andersson, A. M., Skakkebæk, N. E., Toppari, J. & Schramm, K. W. (2006) Enantiomeric ratios as an indicator of exposure processes for persistent pollutants in human placentas. *Chemosphere* 62, 390–395.

Shen, H., Main, K. M., Virtanen, H. E., Damgaard, I. N., Haavisto, A.-M., Kaleva, M. *et al.* (2007) From mother to child: investigation of prenatal and postnatal exposure to persistent bioaccumulationg toxicants using breast milk and placenta biomonitoring. *Chemosphere* 67, 256–262.

Smith, D. (1999) Worldwide trends in DDT levels in human breast milk. *International Journal of Epidemiology* 28, 179–188.

Toppari, J., Larsen, J. C., Christiansen, P., Giwercman, A., Grandjean, P., Guillette, L. J. Jr *et al.* (1996) Male reproductive health and environmental xenoestrogens. *Environmental Health Perspectives* 104, 741–803.

Wang, R. Y. & Needham, L. L. (2007) Environmental chemicals: from the environment to food, to breast milk, to the infant. *Journal of toxicology and environmental health. Part B, Critical reviews* 10, 597–609.

Welsh, M., Saunders, P. T., Fisken, M., Scott, H. M., Hutchison, G. R., Smith, L. B. & Sharpe, R. M. (2008) Identification in rats of a programming window for reproductive tract masculinization, disruption of which leads to hypospadias and cryptorchidism. *Journal of Clinical Investigation* 118, 1479–1490.

Witten, I. H. & Frank, E. (2005) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco.

Wold, H. (1966) Multivariate Analysis. Estimation of Principal Components and Related Models by Iterative Least squares. Academic Press, New York.

Zietz, B. P., Hoopmann, M., Funcke, M., Huppmann, R., Suchenwirth, R. & Gierden, E. (2008) Long-term biomonitoring of polychlorinated biphenyls and organochlorine pesticides in human milk from mothers living in northern Germany. *International Journal of Hygiene and Environmental Health* 211, 624–638.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 All chemicals which were measured in mother's breast milk.

Table S2 Chemicals excluded from analysis, due to low or non-detectable levels.

Tables S3 (a–c) Results from investigating the country difference of each chemical individually, using multiple regression analysis. *p*-values have been corrected for multiple testing.

Tables S4 (a–c) Results for investigating the country difference of each chemical individually, using multiple regression analysis.

Tables S5 (a–c) Matthews Correlation Coefficients calculated from the performance of the machine learning classifiers on raw data.

Tables S6 (a–c) Matthews Correlation Coefficients calculated from the performance of the machine learning classifiers on interpolated data using regression coefficients.

Tables S7 (a–c) Matthews Correlation Coefficients calculated from the performance of the PLS classifier with country and all confounders as response variables.

Table S8 Performance when only including phthalates in the Machine Learning analyses.

Tables S9 (a–c) Chemicals selected as most important by the machine learning classifiers applied on raw data.

Tables S10 (a–c) Chemicals selected as most important by the machine learning classifiers applied on interpolated data using regression coefficients.

Tables S11 (a–c) Chemicals selected as most important by the PLS with classifier with country and confounders included as response variables.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Chapter 5

Manuscript II: Association Between Chemical Pattern in Breast Milk and Congenital Cryptorchidism of Newborn Boys.

Association between chemical pattern in breast milk and congenital cryptorchidism of newborn boys.

K. Krysiak-Baltyn[1,2], J. Toppari[3], N. E. Skakkebaek[2], T. S. Jensen[1], H. E. Virtanen[3], K.-W. Schramm[4,5], H. Shen[4], T. Vartiainen[6], H. Kiviranta[6], O. Taboureau[1], K. Audouze[1], S. Brunak[1] and K. M. Main[2]

1. Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark.
2. University Department of Growth and Reproduction, Rigshospitalet, Copenhagen, Denmark.
3. Department of Physiology and Paediatrics, University of Turku, Turku, Finland.
4. Helmholtz Zentrum München, Institute of Ecological Chemistry, Neuherberg, Germany.
5. Department für Biowissenschaftliche Grundlagen, Technische Universität München, Weihenstephaner, Freising, Germany.
6. Department of Environmental Health, National Institute for Health & Welfare, Kuopio, Finland.

ABSTRACT

During the past four decades, there has been an increase in the incidence rate of male reproductive disorders in the west, with some countries having notably higher rates than others, such as Denmark, whereas Finland has much lower levels. The observed trend is strongly suspected to be due to environmental factors, as the increase has been too rapid to be explained by genetic factors alone. Moreover, it has further been hypothesized that the environmental factors primarily exert there negative effect on the testis at specific time windows during fetal development.

In order to study a possible correlation between chemical exposure and the outcome of cryptorchidism, the most common congenital malformation of male genitalia, we undertook an ecological study of 121 endocrine disrupting chemicals (EDCs), measured in 130 breast milk samples from Danish and Finnish mothers of newborn boys. Half of the newborns were healthy controls, while the other half were boys born with cryptorchidism.

We demonstrate that Danish samples exhibit a stronger correlation between chemical patterns in breast milk with the outcome of cryptorchidism than Finnish samples. As Finland has much lower incidence rates of male reproductive disorders, and therefore likely not affected by environmental factors to the same extent as Denmark, our results lend further support to

the current belief that environmental factors may be behind the observed trends in Denmark and other industrialized countries.

INTRODUCTION

Cryptorchidism, occurring in 2-9% of newborns, is one of the most common malformations among boys and is associated with decreased semen quality and a higher risk of testis cancer. Some cases are familiar and mutations leading to cryptorchidism have been described, such as loss of function of INSL3. However, in most cases no obvious genetic aberration can be revealed. Recent findings of increased frequency in certain geographical areas over a relatively short period of time suggest that environmental factors are important, as genetic factors alone cannot explain the sudden and rapid increase. In particular, Denmark and Norway both exhibit the highest incidence rates of testicular cancer in the world, whereas Sweden and Finland have markedly lower incidence rates. In relation to these temporal and geographical trends, there has been much speculation that exposures to endocrine disrupting chemicals during fetal development may play a major role, as animal studies have conclusively demonstrated a link between the outcome of cryptorchidism in male offspring and the mother's exposure to such chemicals.

However, in reality, people are not exposed to only a few chemicals, but rather to a large amount of different environmental chemicals that are present in our environment as a consequence of the modernized lifestyle in the west. It is therefore unlikely that any single chemical can be strongly correlated to phenotypic outcomes in the general population (with the exception of special cases such as environmental disasters), but rather the combined exposure may be more informative. Investigation of the simultaneous exposures of hundreds of chemicals in minute amounts is extremely complex as, in theory, they may have additive, synergistic or antagonistic actions and effects may not be linear. Exposure levels are also difficult and often expensive to establish.

In order to investigate the possible correlation between fetal exposure of environmental chemicals with the outcome of congenital cryptorchidism, we performed an ecological study of 130 Danish and Finnish mothers and their newborn boys. In order to estimate the chemical exposure burden of the fetus during pregnancy, breast milk samples were obtained from the mothers, and analyzed for more than hundred persistent organic pollutants. These chemicals were chosen as they are known to have endocrine disrupting effects, and are thus linked to

reproductive disorders. The presence or absence of cryptorchidism in the newborns had been established by careful clinical examination.

To assess any possible correlation between exposure patterns from a large number of chemicals with cryptorchidism, we performed computational analysis with a number of different machine learning classifiers. These classifiers are able to simultaneously take into account all chemicals in order to find optimal exposure patterns that are able to separate the case group from the control group.

MATERIALS AND METHODS

The data was acquired from a previous study on pregnant women and their newborn boys from two Nordic countries; Denmark and Finland. Breast milk samples, clinical examinations and questionnaire data was collected between 1997-2001, with the purpose of examining any possible influence of environmental and lifestyle factors on the prevalence of cryptorchidism and hypospadias.

The original data set consisted of 130 breast milk samples from mothers of newborn children from Denmark (65 sampels) and Finland (65 samples). 68 of the newborns (36 Danish and 32 Finnish) were healthy and without signs of reproductive malformations and 62 were born cryptorchid. Presence or absence of cryptorchidism was established by careful clinical examination [1]. The details of the design of the study and analytical methods of chemical measurements have been described previously [2,3,4].

Measurements from 121 chemicals were present in the data, but 15 were removed due to low or non-detectable levels in all samples (all chemicals are listed in supplementary tables S1 and S2).

Three versions of the data set were created, in each case treating non-detectable levels differently. The non-detecable levels were either set to 0, to one half of *Limit of Detection* (LOD) or to LOD. The same analyses were performed for each of the three data sets.

To asses the association between individual chemical levels and the outcome of cryptorchidism, we analyzed the breast milk samples with logistic regression on log-transformed values. The p-values were corrected for multiple testing by the method of Holm [5]. Potential confounders, known to be risk factors for cryptorchidism, were added as covariates in the analysis, including maternal age, maternal body mass index (BMI), parity, weight for gestational age and gestational age at birth.

To investigate the possible influence of combined chemical exposures on the outcome of cryptorchidism, we analyzed the data using machine learning classifiers. These classifiers can find patterns of chemical levels which may correlate with any given phenotype. Three machine learning classifiers were applied, the linear PLS (Partial Least Squares) and SVM (Support Vector Machine) and a non linear feed forward neural network with one hidden layer of 2-5 nodes. The analysis was performed in the R-statistical language (version 2.12.0), using package "plspm" to perform PLS, package "svmpath" for SVM, and package "nnet" for the neural network classifier.

The performance of the classifiers, on correctly predicting the case/control status of the samples, was estimated using 10-fold cross validation. The data was first divided into 10 stratified folds, creating ten sets of training and validation samples. The analysis was then performed in ten iterations, once for each fold. Within the training samples, 10-fold cross validation was performed in order to estimate the optimal number of components (in the case of PLS) or the optimal number of hidden nodes (in the case of neural network). The obtained model was subsequently used to predict the case/control status on the validation samples. Four measures of performance were obtained from the classifiers: Matthew Correlation Coefficient (MCC) [6], Accuracy (fraction of correctly predicted samples), Sensitivity (fraction of correctly predicted cases) and Specificity (fraction of correctly predicted controls). To asses the statistical significance of the obtained MCC, we applied Fisher's Exact Test on the confusion matrixes. The statistical significance of the remaining three performance measures were assessed using binomial tests.

In the analysis with PLS,. In addition, analyses without

Analysis with PLS was done both with and without adjusting for confounders. Adjustment for confounders was done by adding the confounding factors, along with case/control status, as response-variables. For the other two classifiers, no adjustment was done.

RESULTS

The results differed somewhat for the three ways that non-detectable levels were handled. After correction for multiple testing of the logistic regression analyses, one three and five chemicals were significantly different between cases and controls in the Danish cohort, respectively (Table 1 and supplementary Table S3a). Before correction for multiple testing, 25, 30 and 30 chemicals were significant, for each data set respectively (supplementary table S3b). These chemicals also included the sum and toxic equivalence of dioxins. The

statistically most significant chemicals mainly included BDEs, PCBs and dioxins. Interestingly, among all the nominally significant chemicals (p-value less than 0.05), all PCBs tended to be higher in controls (healthy), whereas BDEs and dioxins tended to be higher in cases.

In the Finnish cohort, after correction for multiple testing, no chemicals exhibited a statistically significant difference between cases and controls. Before multiple testing correction, five chemicals were significant in all data sets, and included mainly PCBs that tended to be higher in cases than controls (supplementary table S4).

Table 1)

	Percentile				
Chemical	**25th**	**50th**	**75th**	**P-val**	**Higher in**
BDE85	5.00E-4	4.70E-3	2.09E-2	7.98E-3	Cases
PCB18	2.10E-3	1.36E-2	5.75E-2	1.01E-2	Controls
PCB33	1.30E-3	4.40E-3	3.36E-2	1.62E-2	Controls

Chemicals exhibiting statistically significant differences between cases and controls in the Danish cohort, after multiple testing correction with Holm's method. Here results are shown from the data set with non-detectables set to LOD / 2.Units are in ng/g lipid.

Analysis by the machine learning classifiers indicated that the correlation between case/control status and chemical profile is stronger in Danish samples than in Finnish samples. This is illustrated by the score scatter plots from the PLS model, shown in figure 1. Table 2 lists four different measures of performance of the PLS classifier, with confounders included as response variables, in predicting the case/control status for each sample. It is noticeable that the p-values in the Finnish cohort are much less significant than in the Danish cohort. In the Danish cohort, all measures showed very significant p-values. The obtained sensitivity was somewhat worse than the specificity, which indicates that the classifier had greater difficulty in correctly predicting cases than controls.

The analyses of PLS without accounting for confounders showed the same pattern (supplementary Tables S6a and S6b).

The performance of the SVM classifier was comparable to that of PLS, but tended to have a more balanced success rate in predicting cases and controls, with sensitivity and specificity being more equal in magnitude (supplementary Tables S7a and S7b). The neural network classifier nnet was unable to find any patterns in both the Danish and the Finnish data set (supplementary Tables S8a and S8b). This is likely due to the fact that the neural network searches for non-linear patterns on data with a relatively small number of samples and large number of variables, which may make it more prone to overfitting.

From the PLS model, with confounders included as response, on Danish samples we extracted the coefficients in order to examine the variable importance. As in the case for linear regression, the PLS model indicated that the most important variables correlating with case/control status are PCBs, BDEs and dioxins, with PCBs tending to be higher in controls than cases (supplementary table S9).

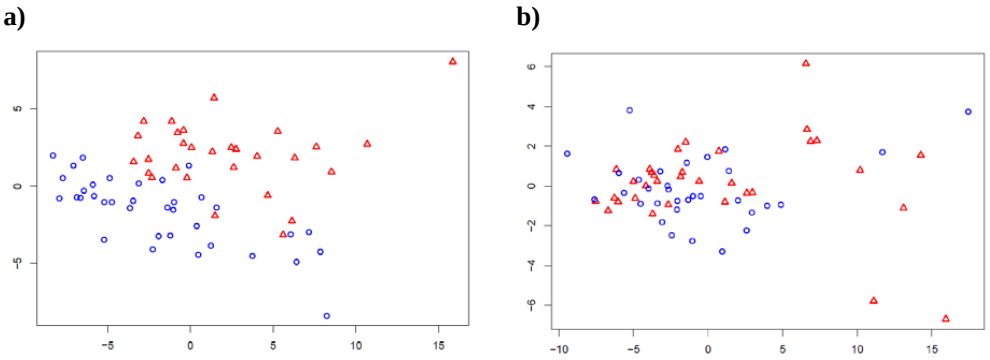

Figure 1) Scatter plots from the PLS models, illustrating the separation between cases and controls. Red triangles represent controls, blue circles represent cases. a) Danish samples. b) Finnish samples.

Table 2) Performance of PLS on Danish Samples

	Danish		Finnish	
	Value	**P-val**	**Value**	**P-val**
MMC	0.60	1.56E-6	0.17	1.26E-1
Accuracy	0.80	5.84E-7	0.58	1.07E-1
Sensitivity	0.69	3.07E-2	0.52	5.00E-1
Specificity	0.89	9.71E-7	0.66	5.51E-2

Performance of the PLS models with confounding factors added as response variables. Results are shown from data set with non-detectables set to LOD / 2. The four different performance measures given are; the Matthews Correlation Coefficient (MCC), Accuracy (fraction of correctly predicted samples), Sensitivity (TP / (TP+FN)), and Specificity (TN / (TN+FP)). P-values were obtained by applying Fisher's Exact Test for the MCC, while binomial tests were performed to assess significance for Accuracy, Sensitivity and Specificity.

DISCUSSION

Our results indicate that within Danish samples there seems to be a clearer correlation between chemical exposures and outcome of Cryptorchidism than in Finnish samples. This observation is what we would expect, based on current knowledge about the time trends and geographical differences in incidence rates of male reproductive disorders. The rise in reproductive disorders in the west during the past four decades has been too rapid to be explained by genetic factors alone, which suggests that environmental factors are involved. However, certain countries, such as Finland, have had a markedly lower incidence rate of these disorders, which may indicate that the environment in Finland is not yet as detrimental for male reproductive health as in e.g. Denmark.

Although the machine learning classifiers are not able to perfectly predict case/control status in the Danish samples, the correlation is nevertheless statistically significant. It is our belief that a major reason for the non-perfect performance in the Danish samples is likely due to the fact that a number of boys in the cohort acquired Cryptorchidism due to a strong genetic component. Thus the chemical profile of their mother's breast milk would not have any causal link to the disease. This, of course, introduces some noise into the data which will cause the machine learning classifiers to have more difficulty in correctly classifying the samples.

Interestingly, a number of PCBs seem to be correlated with healthy samples in the Danish cohort, indicating that they may have a protective effect. This observation may seem counter intuitive, as PCBs are known for their toxic effects. However, some previous studies have made similar observations. A similarly protective effect in relation to testis cancer, which is associated with cryptorchidism, was recently found in a study on a large number of military personel [7]. Moreover, a study in rats showed an increase in testis size and sperm production in PCB exposed cases, suggesting a masculinizing effect by PCBs [8]. However, not all studies indicate that PCBs have a strictly masculinizing effect. It has been shown that PCBs may have a femininizing effect in males and a masculinizing effect in females [9], while some studies show no clear effects at all [10].

To investigate the ´protective effect´ hypothesis at a molecular level, a systems biology approach was used. All six relevant PCB congeners, which were significant before multiple testing, were PCB 51, PCB 81, PCB 52, PCB 49, PCB 47 and PCB 77 were computationally screened against ChemProt [11]. ChemProt is a newly established chemical biology database allowing identification of chemical-gene/protein associations. We found gene and protein data for PCB 52 and PCB 77, although data for the other congeners was lacking. Using all species information, 12 proteins are know to be connected to PCB 52, only one for human, which is a phospholipase A2 (PLA2G4A). Regarding PCB 77, 31 proteins were extracted from ChemProt. Among them seven are linked to human. All PCB 52, PCB 77 and protein connections are shown on Figure 2. We went one step further and performed pathway analysis independently for both gene lists using KEGG pathway (version 2010)[12]. The results show two interesting pathways which are common to PCB 52 and PCB 77; the gonadotropin releasing hormone pathway (GnRH) and the arachidonic acid metabolism pathway. GnRH is known to be implicated in reproductive health, as this hormone affects the production of Luteinizing Hormone (LH) in the pituitary gland, which in turn stimulates production of testosterone in the leydig cells located in the testis.
The connection between the arachidonic acid pathway and reproductive health may be mediated through prostaglandins (one of the major downstream metabolites of arachidonic acid) as some studies have reported differences in prostaglandin levels in relation to fertility [13], including masculinization in animal models [14].

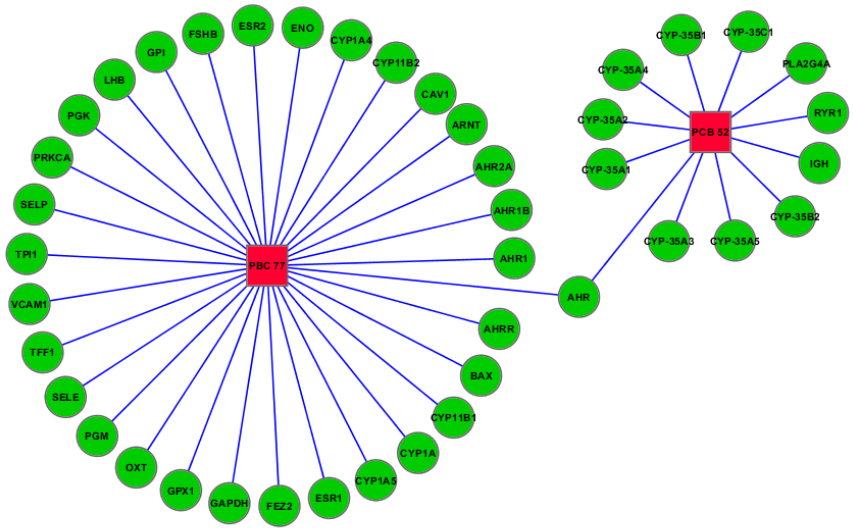

Figure 2: Chemical-protein association network. Proteins are represented as green circles, and PCB congeners as red squares. Edges represent at least one know association between a PCB and a protein, extracted from the ChemProt database. For example, PCB 77 is known to bind to the aryl hydrocarbon receptor (AHR) in human.

We did observe that in the Finnish cohort, among all chemicals significant before correction for multiple testing, PCBs were higher in case samples, which is contrary to what we found in the Danish cohort. We are not able to explain this discrepancy, although this may be a chance observation as the general chemical exposure patterns related to disease outcome is much weaker in the Finnish than in Danish samples. Another explanation may be that the effects of PCBs are context dependent, depending both on the genetic and environmental factors present at any given time. As the Finnish population has a different genetic and environmental background than the Danish population, we speculate that the effect of PCBs on the outcome of cryptorchidism may differ between the two nations.

.

References

1. Boisen KA, Kaleva M, Main KM, Virtanen HE, Haavisto AM, et al. (2004) Difference in prevalence of congenital cryptorchidism in infants between two Nordic countries. Lancet 363: 1264-1269.
2. Shen H, Main KM, Andersson A-M, Damgaard IN, Virtanen HE, et al. (2007) From mother to child: investigation of prenatal and postnatal exposure to persistent bioaccumulationg oxicants using breast milk and placenta biomonitoring. Chemosphere 67: 256–262.
3. Shen H, Virtanen HE, Main K, Kaleva M, Andersson AM, et al. (2006) Enantiomeric ratios as an indicator of exposure processes for persistent pollutants in human placentas. Chemosphere 62: 390-395.
4. Main K, Mortensen G, Kaleva M, Boisen K, Damgaard I, et al. (2006) Human breast milk contamination with phthalates and alterations of endogenous reproductive hormones in infants three months of age. Environmental health perspectives 114: 270-276.
5. Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6: 65-70.
6. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et biophysica acta 405: 442-451.
7. McGlynn K, Quraishi S, Graubard B, Weber J-P, Rubertone M, et al. (2009) Polychlorinated biphenyls and risk of testicular germ cell tumors. Cancer research 69: 1901-1909.
8. Cooke PS, Zhao YD, Hansen LG (1996) Neonatal polychlorinated biphenyl treatment increases adult testis size and sperm production in the rat. Toxicology and applied pharmacology 136: 112-117.
9. Vreugdenhil H, Slijper F, Mulder P, Weisglas-Kuperus N (2002) Effects of perinatal exposure to PCBs and dioxins on play behavior in Dutch children at school age. Environmental health perspectives 110.
10. Gellert R, Wilson C (1979) Reproductive function in rats exposed prenatally to pesticides and polychlorinated biphenyls (PCB). Environmental Research 18: 437-443.
11. Taboureau O, Nielsen S, Audouze K, Weinhold N, Edsgärd D, et al. (2011) ChemProt: a disease chemical biology database. Nucleic Acids Research 39: D367-D372.
12. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucl Acids Res 38: D355-D360.
13. Lewy R, Bills T, Dalton J, Smith J, Silver M (1979) 19-Hydroxy-prostaglandin e and infertility in human males1. Prostaglandines and Medicine 2: 367-371.
14. Wright C, McCarthy M (2009) Prostaglandin E2-induced masculinization of brain and behavior requires protein kinase A, AMPA/kainate, and metabotropic glutamate receptor signaling. The Journal of neuroscience : the official journal of the Society for Neuroscience 29: 13274-13282.

Part III

Linking the Environment and Genes

Chapter 6

Background: Network Biology

Much of systems biology today concerns itself with analysis and use of available data on protein-protein and chemical-protein interaction networks, often as a tool to find new candidate disease genes, find new potential drugs for the treatment of disease or to elucidate the function of genes with unknown function. As proteins rarely perform their function alone, but rather interact with other proteins in a larger system, interaction data is a powerful tool in the study of biological systems. Protein-protein interactions are separated into two major classes: stable and transient. Stable interactions involve proteins involved in the formation of stable "long-term" complexes such as that of hemoglobin which consists of four globular protein subunits. Transient interactions are "short-term" interactions, such as those observed in signalling cascades where the physical interactions are very brief.

6.1 Availability of Data

Today, a large amount of interaction data is available in the form of data bases (both free and commercial) whose purpose is to collect and curate data from various heterogeneous sources. Examples of such data bases are

DrugBank [40] (includes mostly data on drugs and their genetic interaction partners), Comparative Toxicogenomics Database [41] (contains text mined and manually curated gene-gene, gene-protein and gene-disease associations) and STITCH [42]. The data provided by these databases is primarily collected from text mining and from high throughput experimental methods. Text mining methods screen the body of scientific literature in order to automatically extract relevant information, which can subsequently be curated manually. High throughput experimental methods screen large numbers of proteins for interactions to other proteins simultaneously.

There exists two types of experimental high throughput techniques that are used to detect protein-protein interactions: yeast-two-hybrid screening and complex affinity purification. The yeast-two-hybrid system was invented by Fields and Song in 1989 [43], and has been shown to be more suitable for detecting transient than stable interactions. This system makes use of the functional domains of transcription factors in order to detect the interactions between two proteins, say protein A and B. Protein A is hybridized to the DNA-binding domain of a transcription factor, while protein B is hybridized to an activating domain. An interaction taking place between A and B will lead to assembly of a functional transcription factor, which in turn causes transcription of a reporter gene. In the complex affinity purification approach, a protein of interest (bait) is fused to an organic molecule, such as glutathione S-transferase (GST). The bait is then immobilized on a solid support which has a high affinity to the fused molecule. Any new protein (prey) interacting with the bait will bind and likewise be immobilized, while non-binding proteins will be washed away. The prey proteins can be seperated by 2d-gele electrophoresis and identified with mass-spectrometry [44]. This approach is generally better at detecting stable interactions. High throughput gene-chemical interaction data is typically generated using gene expression analysis of micro arrays. Such experiements may indicate if genes are up- or down regulated in the presence of various chemicals.

6.2 Dealing with Noise

A major problem with large scale protein-protein interaction data is that it contains a large number of false positives. Some strategies for dealing with high levels of noise exist and often exploit the fact that protein networks are highly modular, i.e. groups of proteins tend to work together [45, 46, 47]. Proteins within the network are expected to have high local cluster coefficients, and two proteins that interact should have many common interaction partners [48]. A scoring scheme assigning higher scores to interactions within modules can then be used to filter out false positives which are expected to occur more often in the less dense areas of the network between modules.

6.3 Increasing Coverage

The various databases use different approaches for collection and curation of interaction data, and as a conseuqnce do not have complete overlap between each other. In order to increase the coverage of interaction data, one can integrate several such databases into one. Such integration may also involve interaction data obtained from different species, which requires mapping of the homologous proteins between species. Integrating data from different species is a reasonable and sound strategy, considering the fact that some species are more common and easier to work with than others. Yeast and Rat are for example much easier to study than humans, and may thus provide a large amount of data where human data is unavailable.

The next chapter (Manuscript III) describes a nove approach to increase the coverage of protein-protein interaction data. The idea is that if two genes are known to be affected by the same (or similar) chemical compounds, there is a high probability that these genes interact in some way. Utilizing such protein-chemical interaction data, one can then generate protein-protein associations (PPAs). This approach may likely predict new and unknown candidate genes and chemicals involved in disease, as new and previously unknown protein-protein associations are formed.

Chapter 7

Manuscript III: Deciphering Diseases and Biological Targets for Environmental Chemicals using Toxicogenomics Networks

OPEN ACCESS Freely available online

Deciphering Diseases and Biological Targets for Environmental Chemicals using Toxicogenomics Networks

Karine Audouze, Agnieszka Sierakowska Juncker, Francisco J. S. S. A. Roque, Konrad Krysiak-Baltyn, Nils Weinhold, Olivier Taboureau, Thomas Skøt Jensen, Søren Brunak*

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

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 permethrine. 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.

Citation: Audouze K, Juncker AS, Roque FJSSA, Krysiak-Baltyn K, Weinhold N, et al. (2010) Deciphering Diseases and Biological Targets for Environmental Chemicals using Toxicogenomics Networks. PLoS Comput Biol 6(5): e1000788. doi:10.1371/journal.pcbi.1000788

Editor: Olaf G. Wiest, University of Notre Dame, United States of America

Received September 11, 2009; **Accepted** April 15, 2010; **Published** May 20, 2010

Funding: This work has been supported by BioSim (NoE), FP6, LSHB-CT-2004-005137, the Villum Kann Rasmussen Foundation and the Innovative Medicines Initiative Joint Undertaking (IMI-JU) for the eTOX project (115002). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: brunak@cbs.dtu.dk

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 [1–2]. 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 [3–5]. 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 [6]. 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 [7–10]. For example Keiser *et al.* [8] applied network analysis to drugs and their targets. The authors identified unexpected molecular targets such as muscarinic acetylcholine receptor M_3, alpha-2 adrenergic receptor and neurokinin NK2 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 [11–12]. Lamb et al. 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 [3]. 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

Author Summary

Exposure to environmental chemicals and drugs may have a negative effect on human health. An essential step towards understanding the effect of chemicals on human health is to identify all possible molecular targets of a given chemical. Recently, various network-oriented chemical pharmacology approaches have been published. However, these methods limit the protein prediction to already known molecular drug targets. New findings can for example be made by using high-confidence protein-protein association databases. Here, we describe a generic, computational systems biology model with the aim of understanding the underlying molecular mechanisms of chemicals and the biological pathways they perturb. We present a novel and complementary approach to existing models by integrating toxicogenomics data, chemical structures, protein-protein interaction data, disease information and functional annotation of proteins. The high confidence protein-protein association network proposed reveals unexpected connections between chemicals and diseases or human proteins. We provide literature support to demonstrate the validity of some predictions, and thereby illustrate the power of an approach that integrates toxicogenomics data with other data types.

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 [13], 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) [14], we constructed a human P-PAN. A workflow of the strategy is shown on Figure 1. 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, http://www.nlm.nih.gov/mesh/MBrowser.html) 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 [15]. 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 (PDPK1). 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 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 1). 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 [16] and Veber [17] rules. Again, overlaps were observed, indicating that environmental chemicals mimic drug properties (see Figure S1). 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 [18] and (2) PPIs based on an internal consistent single data source [19]. 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 ~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) [20]. For example cell proliferation (p-values of 3.22e-36 and 1.46e-27 for the large network and the sub-network, respectively) and protein binding (p-values of 1.2e-72 and 4.13e-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

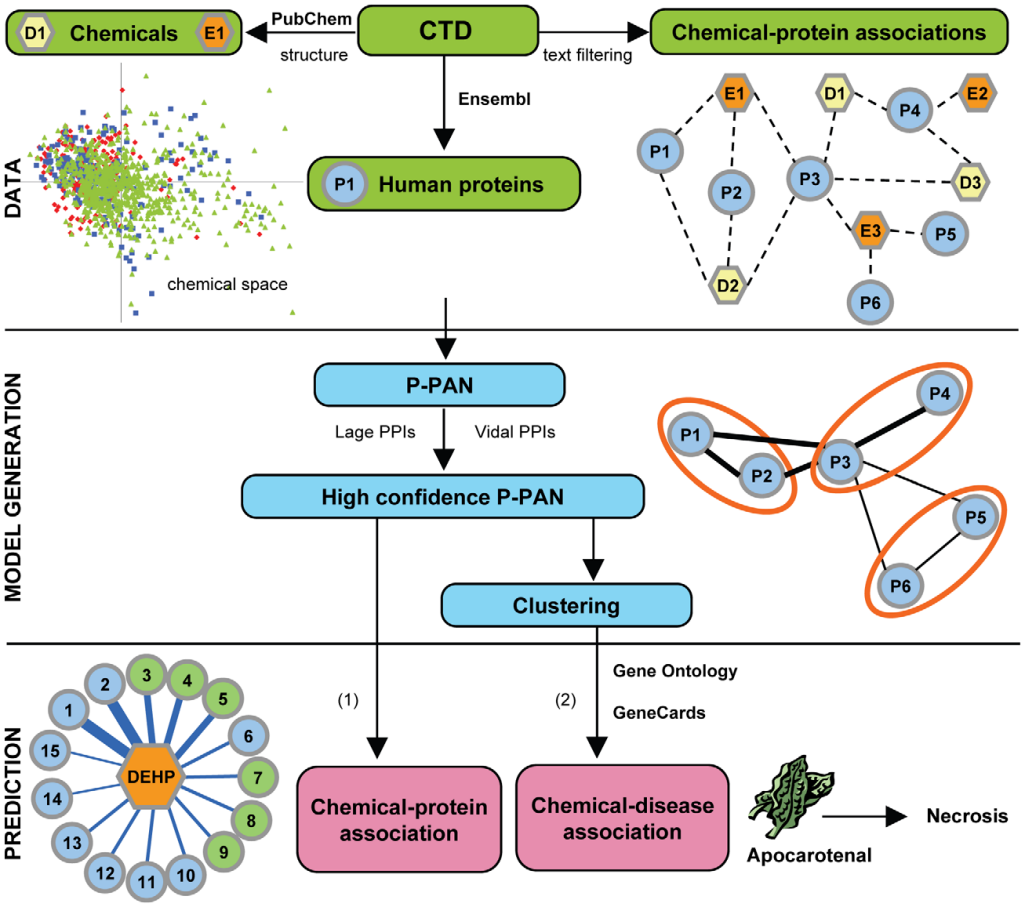

Figure 1. Workflow of the strategy for generating a human P-PAN and predicting novel associations. DATA: Extraction and filtering of human protein-chemical associations from CTD. The visualization of the chemical space by Principal Component Analysis projection confirms that drugs (D) and environmental 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.
doi:10.1371/journal.pcbi.1000788.g001

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 [21] that allows multiple clusters assignation for a protein, mimicking the reality as a protein can participate in several complexes simultaneously. On the other hand, the markov cluster algorithm (MCL) [22] which assign one protein to a unique cluster has been shown to be superior to other graph clustering methods in recent studies [23–24]. 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

biologically 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 6.60e-25, cluster 4) and glucuronosyl transferase activity (p-value of 2.34e-21, cluster 12). Regulation of apoptosis (p-value of 1.67e.17, cluster 2) and oxidation reduction (p-value of 6.67e-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 [25]. We identified several diseases associated with specific clusters. These included the two most common types of cancer, breast cancer (cluster 1, p-value of 9.67e-18) and lung cancer (cluster 12, p-value of 4.84e-12), as well as necrosis (cluster 2, p-value of 2.26e-12), ichthyosis (a skin disorder associated to cluster 4, p-value of 1.41e-5), retinoblastoma (cluster 7, p-value of 9.46e-8) and inflammation (cluster 8, p-value of 1.55e-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 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.

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 [26]. 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 [27] as well as possible gene-toxic effects to human lymphocytes [28]. 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 [29], 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 [30] 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 [31] due to the presence of carotenoids. Furthermore, carotenoids have been defined as chemopreventive agents [32]. Studies have established associations between carotene and beta-carotene with reduced risk of prostate cancer [33] or breast cancer [34]. 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 1) are discussed in the supplementary text (see Text S1).

Table 1. Mining the P-PAN for chemicals associated with breast cancer, lung cancer and necrosis, using a clustering procedure.

Cluster ID	Disease	Chemical name	p-Value
1 (462 proteins)	Breast cancer (128 proteins)	*estradiol*	7.68e-134
		bisphenol A	4.46e-92
		PCBs	1,15e-88
		genistein	2.20e-78
		fulvestrant	7.05e-63
12 (59 proteins)	Lung cancer (29 proteins)	**thimerosal**	1.57e-26
		(10 proteins)	
		DNCB	3.29e-22
		(12 proteins)	
		styrene	7.78e-06
2 (433 proteins)	Necrosis (122 proteins)	*arsenic disulfide*	4.76e-35
		apocarotenal	1.63e-29
		(8 proteins)	
		doxorubicin	2.66e-26

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.

doi:10.1371/journal.pcbi.1000788.t001

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 [8]. 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 2 (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.

Table 2. Predicting novel molecular targets for chemicals.

Chemical	Known protein	Cpscore*	Novel protein	Cpscore*	Literature
DEHP	CDO1	13.23	**GABAß1**	5.46	Yes
	PPARA	9.48	**POMC**	5.44	Yes
	SUOX	4.35	**CYP3A11**	5.40	Yes
	(15 proteins)		**GABAß2**	4.32	Yes
			GABAγ2	4.32	Yes
			GABAα1	4.26	Yes
TCDD	HSPA9B	82.69	**PRKCE**	10.17	Yes
	SLC2A4	82.69	**POMC**	8.97	Yes
	TRIP11	82.69	**CPT1A**	6.96	Yes
	TSP1	82.69	**HSD11B1**	6.39	Yes
	EPHX2	75.77	***MVP***	6.77	No
	MT2A	10.85	**APOB**	5.61	Yes
	(90 proteins)				
PA	CYP4X1	5.67	***CHST1***	5.19	No
	PPARA	2.53	***CHST4***	5.19	No
	CES1	1.45	**CST**	3.19	Yes
	SULT2A1	0.87	***ABCG5***	2.61	No
	CYP1A1	0.37	**C3**	2.80	Yes
			ADRA2A	1.34	Yes
			CYB5A	1.21	No
			ADRA1A	1.08	Yes
			CRHR2	1.04	No
			CYP2A13	0.93	No
			ALDH3	0.91	Yes
	(5 proteins)				
Permethrin	AR	4.67	**CYP2B1**	4.43	Yes
	WNT10B	4.12	**SHBG**	3.51	Yes
	PGR	3.75	***CYP2B6***	2.89	No
	ESR1	3.31	**NR1I3**	2.64	Yes
	TFF1	3.15			
	NR1I2	2.94			
	(17 proteins)				

*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.

doi:10.1371/journal.pcbi.1000788.t002

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) [35]. 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 [36]. Using our approach, several proteins were identified as being associated with DEHP (Table 2). Cysteine dioxygenase type I (CD01) and peroxisome proliferator-activated receptor alpha (PPARA), the two top scoring proteins, are already known in the CTD and from the literature [37–38] 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 [13]. 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 (GABAα1) [39], and DEHP is also associated with neurological defects [36]. In addition to GABA receptors, we identified several other candidates including proopiomelanocortin (POMC) and a cytochrome P450 (CYP3A11) (discussed in the Text S2). 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 [40]. TCDD is also known to be a neurotoxin leading to neurodevelopmental and neurobehavioral deficits [41–42], and accumulating in the brain as well as other organs [43]. We identified six proteins associated with TCDD that are not recorded in the CTD for human (Table 2). Among them five are supported by literature (see Text S2). This included protein kinase C elipson (PRKCE), known to be involved in brain tumors

[44], carnitine palmitoyltransferase I (CPT1A), 11β-hydroxysteroid dehydrogenase type 1 (HSD11B1) and apolipoprotein B (APOB) which are all linked to obesity [45–47]. 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 [48]. To date there is no evidence that PA promotes carcinogenesis in humans [49], 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 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 [50]. 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) [51–52] and sex hormone-binding globulin (SHBG) [53], which are proteins linked to the endocrine system. These findings suggest a mechanism by which chronic exposure 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 1) [54] 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 (Figure 2). 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 [14]. 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 [55]. 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 [56]. There is an evident need for the development of databases [57] 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.

Materials and Methods

Data set

We downloaded the publicly available Comparative Toxicogenomics Database (CTD) as of June 26, 2008 [14]. 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

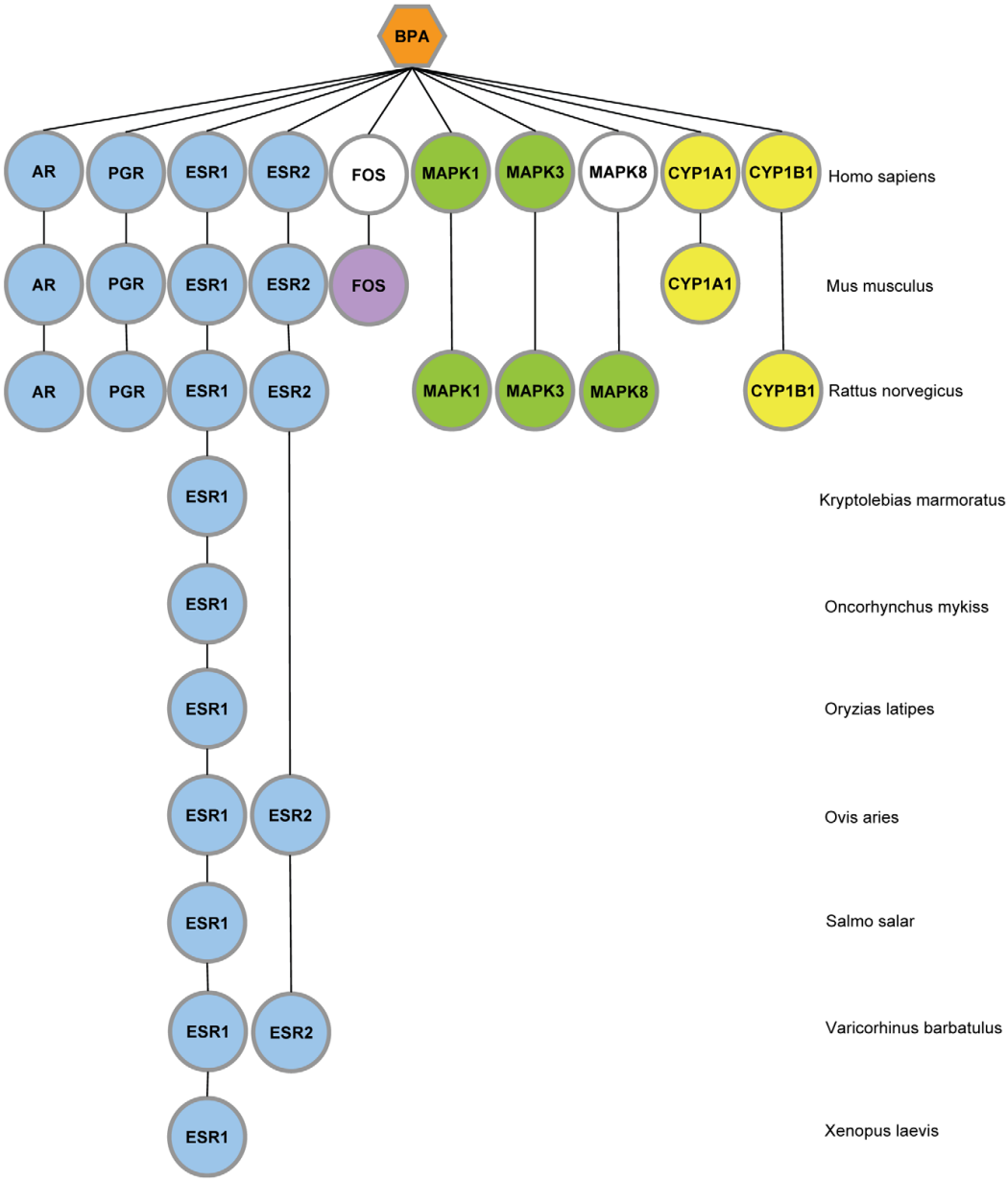

Figure 2. Cross-species comparative toxicogenomics for bisphenol A (BPA). Molecular targets are represented as nodes, and colored by gene family. Nodes presence represent available information extracted from the CTD and node absence are the unknown information. Colored nodes defined that BPA affect the 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.
doi:10.1371/journal.pcbi.1000788.g002

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 (http://

pubchem.ncbi.nlm.nih.gov/) as of June 26, 2008 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 [58] as of June 26, 2008. Only 1.5% of the 42,194 chemical-protein associations could not be clearly identified.

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 (Chemical Computing Group version 2007.09)

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[18] (version 2.9) and Vidal's PPIs[19]. 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 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 [21] and MCL [22]. 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) [59] (July, 2009) and the GeneCards database [25] (February, 2008) 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 proteins, 128 were associated to breast cancer (with give a p-value of 9.67e-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 2.26e-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 [60]. 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[18]. 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.

Supporting Information

Figure S1 Structure-target relationship: Oral bioavailability profiles.For drugs, permeability and absorption are properties considered to be important for effective delivery systems, and they receive special attention in pharmaceutical research. We chose to focus on the oral bioavailability properties based on standard Lipinski and Veber rules. It is important to keep in mind that the rules serve as guidelines only - some classes of chemicals, like antibiotics, do not respect the rules. The selected properties are the molecular weight, the octanol/water partition coefficient (an indication of the ability of a molecule to cross biological membranes), the number of hydrogen bond-donor, the number of hydrogen bond-acceptor and the number of rotatable bond. The distributions of the different molecular properties have partial overlaps indicating that small environmental molecules could mimic drug properties. As an example, the distribution of the molecular weight shows a similar profile for each of the three MeSH categories, with a light tendency for 'Toxic Actions' chemicals to have a smaller molecular weight (MW). The mean of MW for 'Toxic Actions' is 264 daltons whereas the mean of MW for 'Pharmaceutical Actions' chemicals is 386 daltons.
Found at: doi:10.1371/journal.pcbi.1000788.s001 (0.06 MB DOC)

Figure S2 Comparing overlaps between protein-protein associations and protein-protein interactions. To assess the reliability of our protein-protein association scores, we fitted a calibration curve of the different PPA scores against overlaps with two PPI databases: the Vidal's interactome and a highly confident set from Lage et al. Vidal's PPIs are based on an internal consistent single data source defined using yeast two-hybrid system. Lage's PPIs contain interactions present in the largest databases and data inferred from model organisms. All the interactions used from Lage et al for the calibration curve are experimental (extracted from Reactome, KEGG and experimental data from small scale experiments). In both comparison, the weighted score (wscore, in red) appears to be superior compared to the score derivates from a hypergeometric test (hscore, in green) and to the random scores. The vertical line represent the threshold selected, which correspond to 8% of the complete P-PAN i.e. 200,080 proteins.
Found at: doi:10.1371/journal.pcbi.1000788.s002 (0.07 MB DOC)

Figure S3 Molecular target predictions for DEHP: novelty of the P-PAN. The novelty of our P-PAN is supported by comparing the predicted proteins associated to DEHP using our approach and an existing method String [1]. Blue nodes are the 15 input proteins known to be associated to this chemical in CTD, green nodes are the predicted proteins from String. Purple nodes are the proteins predicted for DEHP using our P-PAN (dark purple are the proteins with a high confidence score). Green edges are the protein-protein interactions predicted from the String database and purple edges are the protein-protein associations suggested by P-PAN. In the String output network none of the GABA receptors were found, which were identified as potential molecular targets for DEHP using our P-PAN. Considering high confidence score for both methods (String score$>$0.98), no overlaps between predicted proteins were found. The interactions between predicted proteins were removed for more clarity.
Found at: doi:10.1371/journal.pcbi.1000788.s003 (0.26 MB DOC)

Figure S4 Distributions of the gene- disease scores from GeneCards-AKS2 and OMIN. To integrate disease information to the clusters, GeneCards was used as a source of disease-protein connections. In order to limit the use of false positives present in GeneCards, we mapped shared protein-disease association from OMIN and GeneCards. According to the overlap curves, we set a significant cut-off value of the GeneCards-AKS2 score (in red) of 60.
Found at: doi:10.1371/journal.pcbi.1000788.s004 (0.28 MB DOC)

Text S1 Mining the P-PAN for chemicals associated with diseases.
Found at: doi:10.1371/journal.pcbi.1000788.s005 (0.06 MB DOC)

Text S2 Molecular targets predictions for chemicals.

Found at: doi:10.1371/journal.pcbi.1000788.s006 (0.07 MB DOC)

Table S1 Example of molecular target predictions for chemicals. References: 1. Mahgoub AA, El-Medany AH (2001) Evaluation of chronic exposure of the male rat reproductive system to the insecticide methomyl. Pharmacol. Res. 44:73–80. 2. Bernard L, Martinat N, Lécureuil C, Crépieux P, Reiter E, Tilloy-Ellul A, Chevalier S, Guillou F (2007) Dichlorodiphenyltrichloroethane impairs follicle-stimulating hormone receptor-mediated signaling in rat Sertoli cells. Reprod. Toxicol. 23:158–164. 3. Saqib TA, Naqvi SN, Siddiqui PA, Azmi MA (2005) Detection of pesticide residues in muscles liver and fat of 3 species of Labeo found in Kalri and Haleji lakes. J. Environ. Biol. 26:433–438. 4. Flodström S, Hemming H, Warngard L, Ahlborg UG (1990) Promotion of altered hepatic foci development in rat liver cytochrome P450 enzyme induction and inhibition of cell-cell communication by DDT and some structurally related organohalogen pesticides. Carcinogenesis 11:1413–1417. 5. Sakai H, Iwata H, Kim EY, Tsydenova O, Miyazaki N, Petrov EA, Batoev VB, Tanabe S (2006) Constitutive androstane receptor (CAR) as a potential sensing biomarker of persistent organic pollutants (POPs) in aquatic mammal: molecular characterization expression level and ligand profiling in Baikal seal (Pusa sibirica). Toxicol. Sci. 94:57–70 6. Ding X, Staudinger JL (2005) Repression of PXR-mediated induction of hepatic CYP3A gene expression by protein kinase C. Biochem. Pharmacol. 69:867–873. 7. Matsuura I, Saitoh T, Tani E, Wako Y, Iwata H, Toyota N, Ishizuka Y, Namiki M, Hoshino N, Tsuchitani M, Ikeda Y (2005) Evaluation of a two-generation reproduction toxicity study adding endpoints to detect endocrine disrupting activity using lindane. J. Toxicol. Sci. Spec No 135–161.

Found at: doi:10.1371/journal.pcbi.1000788.s007 (0.04 MB DOC)

Table S2 Illustration of cpscore for approved drugs.

Found at: doi:10.1371/journal.pcbi.1000788.s008 (0.09 MB DOC)

Acknowledgments

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

Author Contributions

Conceived and designed the experiments: KA SB. Performed the experiments: ASJ FJSSAR KKB NW OT. Analyzed the data: KA. Wrote the paper: KA TSJ.

References

1. Edwards TM, Myers JP (2008) Environmental exposures and gene regulation in disease etiology. Cien saude Colet 13: 269–281.
2. Phillips DH, Arlt VM (2009) Genotoxicity: damage to DNA and its consequences. EXS 99: 87–110.
3. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. Nat Biotechnol 24: 805–815.
4. Hopkins AL (2007) Network pharmacology. Nat Biotechnol 25: 1110–1111.
5. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M (2007) Drug-target network. Nat Biotechno l 25: 1119–1126.
6. Keith CT, Borisy AA, Stockwell BR (2005) Multicomponent therapeutics for networked systems. Nat Rev Drug Discovery 4: 71–78.
7. Morphy R, Rankovic Z (2007) Fragments network biology and designing multiple ligands. Drug Discov Today 12: 156–160.
8. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, et al. (2007) Relating protein pharmacology by their ligand chemistry. Nat Biotechnol 25: 197–206.
9. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P (2008) Drug target identification using side-effect similarity. Science 321: 263–266.
10. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, et al. (2009) Predicting new molecular targets for known drugs. Nature 462: 175–81.
11. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313: 1929–35.
12. Williams-Devane CR, Wolf MA, Richard AM (2009) 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: 358–71.
13. Yang L, Milutinovic PS, Brosnan RJ, Eger EI, 2nd, Sonner JM (2007) The plasticizers di(2-ethylhexyl) phthalate modulates gamma-aminobutyric acid type A and glycine receptor function. Anesth Analg 105: 393–396.
14. Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, et al. (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. Nucleic Acids Res Database issue: D786–92.
15. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, et al. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res Database issue: D901–6.
16. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46: 3–26.
17. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, et al. (2002) Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 45: 2615–2623.
18. Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25: 309–316.
19. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437: 1173–8.
20. Ashburner M, et al. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet 25: 25–29.
21. Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4: 2.
22. van Dongen S (2000) A cluster algorithm for graphs Technical Report INS-R0010. version 1006 National Research Institute for Mathematics and Computer Science in the Netherlands Amsterdam Available from: http://www.micans.org/mcl/.
23. Brohée S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7: 488.
24. Vlasblom J, Wodak SJ (2009) Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. BMC Bioinformatics 10: 99.
25. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D (1997) GeneCards: integrating information about genes proteins and diseases. Trends in Genetics 13: 163.
26. Salehi F, Turner MC, Phillips KP, Wigle DT, Krewski D, et al. (2008) Review of the etiology of breast cancer with special attention to organochlorines as potential endocrine disruptors. J Toxicol Environ Health B Crit Rev 11: 276–300.
27. Geier DA, Sykes LK, Geier MR (2007) A review of thimerosal (merthiolate) and its ethylmercury breakdown product: specific historical considerations regarding safety and effectiveness. J Toxicol Env Health 10: 575–596.
28. Westphal GA, Asgari S, Schulz TG, Bünger J, Müller M, et al. (2003) Thimerosal induces micronuclei in the cytochalasin B block micronucleus test with human lymphocytes. Arch Toxicol 77: 50–55.
29. Kuper CF, Stierum RH, Boorsma A, Schijf MA, Prinsen M, et al. (2008) The contact allergen dinitrochlorobenzene (DNCB) and respiratory allergy in the Th2-prone Brown Norway rat. Toxicology 246: 213–221.
30. Sani HA, Rahmat A, Ismail M, Rosli R, Endrini S (2004) Potential anticancer effect of red spinach (Amaranthus gangeticus) extract. Asia Pac J Clin Nutr 13: 396–400.
31. Block G, Patterson B, Subar A (1992) Fruit vegetables and cancer prevention: a review of the epidemiological evidence. Nutr Cancer 18: 1–29.
32. Krinsky NI, Johnson EJ (2005) Carotenoid actions and their relation to health and disease. Mol Aspects Med 26: 459–516.
33. Peters U, et al. (1997) Serum lycopene other carotenoids and prostate cancer risk: a nested case-control study in the prostate lung colorectal and ovarian cancer scrrening trial. Cancer Epidemiol Biomarkers Prev 16: 109–126.
34. Toniolo P, Van Kappel AL, Akhmedkhanov A, Ferrari P, Kato I, et al. (2001) Serum carotenoids and breast cancer. Am J Epidemiol 153: 1142–1147.
35. Martinasso G, Maggiora M, Trombetta A, Angela CR, Muzio G (2006) Effetcs 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: 353–365.
36. Latini G (2000) Potential hazards of exposure to di-2-ethylhexyl phthalate in babies: a review. Biol Neonate 78: 268–276.
37. Turan N, Waring RH, Ramsden DB (2005) The effect of plasticisers on "sulphate supply" enzymes. Mol Cell Endocrinol 244: 15–19.
38. Kim HS, Ishizuka M, Kazusaka A, Fujita S (2004) 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: 1119–1124.

39. Horiuchi Y, Nakayama J, Ishiguro H, Ohtsuki T, Detera-Wadleigh SD, et al. (2004) Possible association between a haplotype of the GABA-A receptor alpha 1 subunit gene (GABRA1) and mood disorders. Biol Psychiatry 55: 40–45.
40. Moon BH, et al. (2008) A single administration of 2,3,7,8-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: 314–322.
41. Legare ME, Hanneman WH, Barhoumi R, Burghardt RC, Tiffany-Castoglioni E (2000) 2,3,7,8-tetrachlorodibenzo-p-dioxin alters hippocampal astroglia-neuronal gap junctional communication. Neurotoxicology 21: 1109–1116.
42. Nayyar T, Zawia NH, Hood DB (2002) Transplacental effects of 2,3,7,8-tetrachlorodibenzo-p-dioxin on the temporal modulation of Sp1 DNA binding in the developing cerebral cortex and cerebellum. Exp Toxicol Pathol 53: 461–468.
43. Kakeyama M, Sone H, Miyabara Y, Tohyama C (2003) Perinatal exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin alters activity-dependent expression of BDNF mRNA in the neocortex and male rat sexual behavior in adulthood. Neurotoxicology 24: 207–217.
44. Kim SY, Yang JH (2005) Neurotoxic effects of 2,3,7,8-tetrachlorodibenzo-p-dioxin in cerebellar granule cells. Exp Mol Med 37: 58–64.
45. Boverhof DR, Burgoon LD, Tashiro C, Sharratt B, Chittim B, et al. (2006) Comparative toxicogenomics analysis of the hepatotoxic effects of TCDD in Sprague Dawley rats and C57BL/6 mice. Toxicol Sci 94: 398–416.
46. Fletcher N, Wahlström D, Lundberg R, Nilsson CB, Nilsson KC, et al. (2005) 2,3,7,8-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: 1–24.
47. Volz DC, Bencic DC, Hinton DE, Law JM, Kullman SW (2005) 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD) induces organ- specific differential gene expression in male Japanese medaka (Oryzias latipes). Toxicol Sci 85: 572–84.
48. Lalwani ND, Reddy MK, Qureshi SA, Reddy JK (1981) 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: 645–650.
49. Suga T (2004) Hepatocarcinogenesis by peroxisome proliferators. J Toxicol Sci 29: 1–12.
50. Amacher DE, Adler R, Herath A, Townsend RR (2005) Use of proteomic methods to identify serum biomarkers associated with rat liver toxicity or hypertrophy. Clin Chem 51: 1796–1803.
51. Bauer D, Wolfram N, Kahl GF, Hirsh-Ernst KI (2004) Transcriptional regulation of CYP2B1 induction in primary rat hepatocyte cultures: repression by epidermal growth chemical is mediated via a distal enhancer region. Mol Pharmaco 65: 172–180.
52. Heder AF, Hirsch-Ernst KI, Bauer D, Kahl GF, Desel H (2001) Induction of cytochrome P450 2B1 by pyrethroids in primary rat hepatocyte cultures. Biochem Pharmacol 62: 71–79.
53. Eil C, Nisula BC (1990) The binding properties of pyrethroids to human skin fibroblast androgen receptors and to sex hormone binding globulin. J Steroid Biochem 35: 409–414.
54. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, et al. (2007) STRING 7– recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35: 358–362.
55. Stein LD (2004) Human Genome: End of the Beginning. Nature 431: 915–916.
56. Oprea TI, Tropsha A (2006) Target chemical and bioactivity databases -integration is key. Drug Discov today technol 3: 357–365.
57. Mestres J, Gregori-Puigjané E, Valverde S, Solé RV (2008) Data completeness- -the Achilles heel of drug-target networks. Nat Biotechnol 26: 983–984.
58. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, et al. (2004) An Overview of Ensembl. Genome Res 145: 925–928.
59. Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/.
60. Bonferroni CE (1935) Il calcolo delle assicurazioni su gruppi di teste In Studi in Onore del Professore Salvatore Ortu Carboni Rome: . pp Italy13–60.

Part IV

Life Style Factors: A Fishing Expedition Into the Unexpected

Chapter 8

Background: Data Mining

Within epidemiology and clinical sciences there is a large body of data which has been collected over periods of time spanning years, with the aim of testing hypothesis relating various clinical outcomes to life style factors, environmental exposures or bioligcal features. Although such data is likely generated with the goal of testing certain specific hypotheses, it may potentially contain patterns yet undiscovered, things not yet thought of, by a human mind. In order to extract new and unexpected information from large data sets, it is not feasible to manually test associations between the variables, in particular if interactive effects between variables are taken into account. The number of possible combinations of variables to test simply grows too large. Therefore automated methods are used to mine a given data set for associations. A scientific area which concerns itself specifically with such data mining is *Association Mining* (AM).

8.1 Association Mining

Association Mining (AM) is a type of unsupervised learning technique that has successfully been employed for market basket analysis. The term AM applies to a number of different algorithms and techniques which concern themselves with generating association rules. Association rules were first

popularized by Agrawal [33], usually taking the form "$A \rightarrow B$" (read "if A then B"). An example of an association rule generated from market basket data could thus be $\{carrots\} \rightarrow \{cabbage\}$, which would imply that customers who buy carrots also buy cabbage. The term on the left hand side of the implication arrow is called the *antecedent*, and the right hand side is called the *consequent*. Higher order associations, involving more than one variable on either side, may also be generated such as $\{carrots + beef\}$ $\rightarrow \{potatoes\}$, which is read as "if *carrots* and *beef* then *potatoes*". A famous example of an interesting association rule is $\{diapers\} \rightarrow \{beer\}$, indicating that many customers who buy diapers also buy beer, and has been attributed to young fathers going shopping.

One of the central challenges in AM is the sheer number of possible associations that can be generated. In a market basket data set, the number of association rules A_{tot} is given by the formula:

$$A_{tot} = \sum_{k=n}^{N} \binom{V}{k} \cdot \left(2^k - 2\right) \tag{8.1}$$

Where N is the maximum number of variables in each rule which are mined for, n is the minimum number of variables in each rule and V is the number of variables in the data set. Thus, for a reasonable small data set containing 200 variables, and limiting the number of variables in each rule to between two and six, the total number of association rules that can be generated becomes $A_{tot} = 5.19 \cdot 10^{12}$.

In many cases, mining through all possible associations is not feasible. To reduce the search space in a sensible way is therefore crucial. Perhaps the most well known strategy to limit the search space is with the *Apriori*-algorithm [49]. In order to explain this algorithm, a few definitions need to be introduced.

Item set, is simply a group of variables such as $\{carrots, cabbage\}$, or $\{smoking, drinking, dancing\}$.

Support is a measure of defined as the proportion of samples which contain an item set. For example, the support for $\{carrots, cabbage\}$ would be

calculated by counting how many transactions included both carrots and cabbage, divided by the total number of transactions in the data set.
Confidence is a measure that pertains to association rules, and is defined as the proportion of samples satisfying the antecedent that also satisfy the consequent, and is given by the formula:

$$Confidence = \frac{support(antecedent + consequent)}{support(antecedent)} \tag{8.2}$$

The Apriori-algorithm uses a minimum support constraing in order to limit the search space. First a minimum support cutoff C_{supp} is defined. The Method then generates all possible item sets containing a single variable and calculates the support for each of them. If a variable, say {carrots}, has a support lower than C_{supp}, all supersets (e.g. $\{carrots,potatoes\}$) will also have a support smaller than C_{supp}. This way, the Apriori-algorithm will filter out all itemsets that contain the variable *carrots*, , thereby pruning the search space of item sets. After filtering out all item sets that don't fulfill the minimum support criteria, association rules are generated and filtered out based on a minimum confidence constraint. As an example, for the item set $\{potatoes, cabbage\}$, there are two rules which can be generated: $\{potatoes\} \rightarrow \{cabbage\}$ and $\{cabbage\} \rightarrow \{potatoes\}$. These two rules, although being mirror opposites of each other, do not necessarily have the same confidence, and the rule with the highest confidence would usually be considered. Although the support and confidence measures are not directly related to the strength of a rule, i.e. its statistical significance, the approach is nevertheless useful as it will naturally filter out combinations of variables which have zero, or close to zero, support and thus would very likely generate rules with a high p-value.

Although AM has traditionally been used in analysis of market basket data, it can also be applied on other types of data, such as clinical data. In the context of mining clinical data, the goals may be slightly different than for mining market basket data:

- In order for a clinical association to be interesting it has to be statistically significant. In other words, there needs to be a correlation

between the antecedent and consequent that is significantly stronger than anticipated by chance.

- The interestingness of a clinical associationstrongly depends upon the type of variables involved in the antecedent and consequent. For example, in a data set containing information about smoking and drinking habits in relation to different diseases, the association rule $\{smoking\} \rightarrow \{drinking\}$ is most likely not interesting, whereas the rule $\{smoking\} \rightarrow \{cancer\}$ may be considered interesting.

These criteria do not necessarily apply to market basket data, as a store manager is likely more interested in knowing which goods are bought together, so they can conveniently be placed together, potentially stimulating sales. In this case, a high support, and confidence, may be enough to make a rule interesting.

8.2 Rule Filtering

In a typical Association Mining experiment, the data set analyzed may generate anywhere between tens of thousands to millions of rules. Reduction in the number of rules is therefore desirable. Perhaps the most common strategy for this task is the application of rule filters, which filter out rules that are known to be uninteresting due to some special qualities. A simple example is rule filtering based on prior knowledge. Let's consider a data set containing chemical measurements as well different clinical symptoms that are potentially associated. A clinician will be much more interested to find out which chemicals confer risk for aquiring some of the clinical symptoms, rather than identifying correlations between different chemicals. Therefire it would be reasonable to filter out all rules which only contain chemicals.

Another very useful rule filter is the application of a minimum improvement constraint, originally suggested by Bayardo et al [34]. Suppose we have a rule antecedent which includes a number of elements $A = A_1 + A_2 + ... + A_n$, and likewise a rule consequent $B = B_1 + B_2 + ... + B_m$. The improvement of the rule A $\rightarrow$ B is then defined as:

$$improvement(A \rightarrow B) = \\ confidence(A \rightarrow B) - max(confidence(Z \rightarrow B)) \tag{8.3}$$

where $Z \in A$, that is Z can be any subset of A. This rule filter discards uninteresting rules, consisting of more than two variables, which are either redundant or if they contain any element in the antecedent that is independent of the consequent, given the other elements in the antecedent. A redundant rule is a rule where elements in the antecedent are entailed by each other. An example of a redundant rule is $\{father + man\} \rightarrow \{married\}$, where father entails man (although not all men are fathers).

8.3 The Compass

Although traditional AM-approaches may be used for mining clinical data and succesfully generate association rules, I was looking for a different approach that could be more suitable in the context of mining clinical data. In particular, I was looking for a way to avoid binning of numerical variables before analysis (i.e. grouping a numerical variable into several discrete classes), and I also wanted an approach that could generate an output that could be manually screened by a clinician in order to find interesting associations.

As the interestingness of any newfound clinical association depends on the variables involved, it can only be assessed by a human scientist and is thus a subjective measure. However, as a typical clinical data set may generate anywhere between tens of thousands to millions of rules, it is also not feasible for a clinician to simply browse through such a huge list.

For a period of time, I had a constructive collaboration with a group of people at the Department of Informatics and Mathematical Modeling at the Technical University of Denmark, who were testing a new method suitable for my purposes. Our strategy involved the application of Non-negative Matrix Factorization (NMF) with consensus clustering of the output to

find clusters of variables which were associated. Unfortunately, after several months of hard work, we discovered that the method was, in fact, not working as expected. As a consequence, I decided to terminate the collaboration and find another strategy.

Seeing that a large amount of my own work and time was spent in vain, I felt somewhat discouraged to continue this particular project. However, I was later able to find a another strategy on my own which was suitable for mining clinical data [manuscript IV]. The method uses Self-Organizing Maps (SOM) to generate groups of highly associated variables, and then performs mining of Association Rules on these groups. The application of SOM is a means for handling numerical variables without binning a priori, and it naturally produces *Associative Variable Groups* (AVGs) i.e. groups of variables which are highly assocaited. The AVGs allows for the generation of a condensed output, such that a large number of rules can be manually screened for interestingness and unexpectedness.

Chapter 9

Manuscript IV: Compass: A Hybrid Method for Clinical and Biobank Data Mining

Compass: a hybrid method for clinical and biobank data mining

K. Krysiak-Baltyn[1], T. Nordahl Pedersen[1], K. Audouze[1], Niels Jørgensen[2], L. Ängquist[3] and S. Brunak[1].

1. Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark,
2. University Department of Growth and Reproduction, Rigshospitalet, Copenhagen, Denmark.
3. Institute of Preventative Medicine, Copenhagen University Hospital, Centre for Health and Society, Copenhagen, Denmark.

Abstract
We describe a new method for identification of confident associations within large clinical data sets. The method is a hybrid of two existing methods; Self-Organizing Maps and Association Mining. We utilize Self-Organizing Maps as the initial step to reduce the search space, and then apply Association Mining in order to find association rules. We demonstrate that this procedure has a number of advantages compared to traditional Association Mining; it allows for handling numerical variables without a priori binning and is able to generate variable groups which act as "hotspots" for statistically significant associations. We showcase the method on infertility-related data from Danish military conscripts. The clinical data we analyzed contained both categorical type questionnaire data and continuous variables generated from biological measurements, including missing values. From this data set, we successfully generated a number of interesting association rules, which relate an observation with a specific consequence and the p-value for that finding. Additionally, we demonstrate that the method can be used on non-clinical data containing chemical-disease associations in order to find associations between different phenotypes, such as prostate cancer and breast cancer.

1. Introduction
Due to the existence of a vast amount of data, such as biobank data, collected by a large number of scientific groups over the years, there is a growing interest in mining such data for the purpose of new knowledge discovery [Nat. Gen., Editorial. vol 42, p. 467]. Traditional hypotheses testing approaches are typically not ideal in more comprehensive data mining aiming for new and unexpected patterns due to the immensely large search space, particularly in high-volume data sets.

Methods for unsupervised data mining have commonly been employed in market basket analysis and fall under the category of Association Mining (AM). The main goal of AM in market basket analyses is to find interesting associations of the form "{chips} → {beer}" which would indicate that people who buy chips are likely to also buy beer. More complex rules involving more items may also be formed such as "{ham+cheese} → {milk+bread}, i.e. those who buy ham and cheese are likely simultaneously to also buy milk and bread. In these rules items appearing on the left side of the arrow are called *antecedent*, while items on the right side are called *consequent*. As market basket data sets tend to be large, and the number of possible combinations between items may be extremely big, the central effort in this type of mining has been to restrict the search space in a sensible way.

The concept of association rules was popularized by Agrawal et al [Agrawal R., Imielinski T., Swami A. Mining Association Rules Between Sets of Items in Large Databases, SIGMOD Conference 1993: 207-216] in 1993, although the concept may have already been created as far back as 1966 [1]. AM has traditionally consisted of two steps: first to find frequent itemsets, and

second to generate rules by calculating the *confidence*, which may give an estimate of the interestingness of the rule. Frequent itemsets are collections of items, such as "{bread,milk,butter}", which appear together (customers often buy these together) more often than a specified threshold called *support*. The confidence indicates the probability of the consequent given the antecedent in a given rule, say "{bread} → {milk,butter}", and is often used to restrict which rules are retained. There are other measures of interestingness besides confidence, such as *lift*, *leverage* and *conviction*, which we will not discuss here.

The various types of existing AM methods typically address relatively simple dichotomous data sets containing only 1s and 0s. Applying AM to mine other types of data, such as clinical data, has been done in several, previous studies [2], [Siri K et al. An Efficient Interestingness based Algorithm for Mining Association Rules in Medical Databases. K. Elleithy (ed.), Advances and Innovations in Systems, Computing Sciences and Software Engineering, 167–172]. These approaches utilized the support measure to control the size and shape of the search space of associations. However, restricting the search space by applying one strict minimum support threshold, thus generating frequent itemsets, as has traditionally been done in AM, is not necessarily an optimal strategy when analyzing clinical data.

The reason is primarily due to the nature of such data, where some features and associations, may only be present in a subset of samples e.g. certain types of tests that only apply to certain diseases. When generating frequent itemsets, if setting the support threshold too high, there is a risk that one will not find associations which contain these non-frequent, but interesting, features. However, if setting the threshold low enough to detect these associations, the combinatorial explosion of the number of rules generated may prevent detailed manual inspection.

In this paper we present *Compass*, a new hybrid approach using a combination of Self-Organizing Maps (SOM) and AM, which is suitable for mining unexpected patterns in clinical data and other similar types of data. In this approach, SOM is applied as a first step before the application of AM. The reason for applying SOM is that it acts as a navigating pointer to areas in the search space where one is more likely to find statistically significant association rules, hence acting as a compass to reduce the search space.

SOM, also known as Kohonen neural networks, is an unsupervised neural-network method which clusters the data into a set of interconnected nodes [3,4]. These nodes are typically arranged in two-dimensional maps with rectangular or hexagonal grids, but may also be arranged in multi-dimensional maps with other types of grids. The training process is started by randomly assigning to each node a model vector, which has the same dimensionality as the data samples. All samples are then iteratively assigned to the nodes whose' model vector is most similar (often measured with Euclidian distance), and gradually regressing the model vectors towards each sample that is assigned. After the training process is finished, the patterns of interest are extracted from the model vectors.

Since the time of the introduction of SOM in 1982, different variants of this method have found their way into many practical application in different scientific fields including biology, economics and physics [Pöllä M et al. Bibliography of Self-Organizing Map (SOM) Papers: 2002-2005 Addendum. 2009. TKK Reports in Information and Computer Science, Helsinki University of Technology, Report TKK-ICS-R24. http://www.cis.hut.fi/research/som-bibl/].

The idea of combining SOM with association rule mining is not new [Shangming Y, Yi Z (2004). Advances in Neural Networks. Berlin: Springer.; Tangsripairoj S (2004) A growing hierarchical self-organizing map with mining association rules for software repository organization and visualization: Oklahoma State University]. However, previous studies did not fully explore the

potential of this approach on complex medical data. In particular, they did not explore the ability of SOM to address two important challenges in the context of association mining; handling numerical variables and dealing with a large output of association rules to find the most interesting and unexpected ones.

Our method can handle numerical variables without resorting to a priori binning, i.e. grouping numerical variables into classes before the analysis. Such grouping is common when studying e.g. income, wherein people are assigned to a number of arbitrary and discrete classes depending on their level of income. While previous studies have explored novel ways to deal with numerical variables without binning, to our knowledge these approaches rely on setting a strict support cutoff to find frequent itemsets. [Calders T et al. Mining rank-correlated sets of numerical attributes; 2006; Philadelphia, PA]

Our approach makes it also much easier by manual inspection to identify new, interesting and unexpected patterns by introducing the concept of *Associative Variable Groups* (AVGs). AVGs are simply groups of variables which are much more likely to contain statistically significant association rules than any variables randomly chosen from the data, and which arise naturally using SOM as the initial step. Instead of generating a long list of possibly thousands of association rules as the output, our method generates a list of AVGs, which drastically reduces the size of the output and makes it much more comprehensible for subsequent manual inspection.

2. Data

Clinical data from military conscripts
For the analysis, we obtained two, independent medium sized clinical data sets from the Jørgensen group [5], containing questionnaire data as well as biological measurements, such as sperm concentration, from Danish military conscripts. We used one data set for training, and the other as a test set to validate associations that were found on the training data. The original training set included 1,444 samples, and 163 variables. The original test set included 2,532 samples and 377 variables, but only the 163 variables also present in the training set were included in the analysis. The raw data was subjected to an extensive cleaning process.

Variables removed
Variables with more than 90% missing values were removed. Moreover, variables that were deemed either irrelevant or highly unreliable by the clinicians who generated the data were also removed.

Variables modified
Variables representing times and dates were converted into scalars, reflecting time in years from a reference time point. Geographical data was provided in the form of zip codes, which was subsequently converted into latitudes and longitudes (http://geonames.org).

Variables added
New variables were added from performing operations on other existing variables e.g. the age of each military conscript was calculated by taking the difference between “date of questionnaire fill-in” and “birth date”.

Variable encoding
Numerical values were scaled to a mean value of zero and standard deviation of one, while categorical variables were converted into binary form with 1s and 0s. As the majority of the categorical variables contained more than one attribute value, each categorical variable had to be “expanded” such that each attribute value was represented by one new variable. As an example,

consider the variable “Smoking” with the possible attribute values “Yes” and “No”. From this variable, two new variables are generated; “Smoking-Yes” and “Smoking-No”. Thus, 67 categorical variables were converted into a total of 184 attribute variables. Categorical variables with non-informative attribute values, with less than 20 samples coded as 1’s, or more than 90% samples coded as 1’s, were removed. Moreover, we merged the binary categorical variables that were identical above a cutoff of 95% identity. In our case, we found for example that a low disposition in one testis strongly implied a low disposition in the other testis, thus making one of these variables redundant. Nearly identical variables like these are bound to end up together in many association rules, but do not confer any new or interesting information. The cleaned data set included 95 categorical variables encoded in binary form.

Some contradictions and errors in the data were discovered and removed. Examples included people who claim to be non-smokers but then subsequently specify that they smoke more than 0 cigarettes per day. The resulting cleaned data had 145 variables. No samples were removed from either the training or test sets.

Table 1. Data cleaning summary

Description of cleaning process	Count
Variables in raw training set	163
Variables in raw test set	377
Variables removed, missingness	15
Variables removed, unsuitable or irrelevant	40
Variables modified	13
Variables added	7
Attribute-value variables generated	184
Attribute-value variables removed	88
Variables in cleaned data	145

The table shows the number of variables removed, modified or added during the extensive cleaning process of the raw data.

Chemical-disease associations from CTD

We extracted chemical-disease annotations from the Comparative Toxicogenomics Database (CTD) [6]. The CTD contains direct and inferred chemical–disease associations. Direct chemical–disease associations are curated from the published literature. These associations are either demonstrated experimentally in model physiological systems or through epidemiological studies. Inferred relationships are established via CTD–curated chemical–gene interactions (e.g., chemical A is associated with disease B because chemical A has a curated interaction with gene C, and gene C has a direct relationship with disease B).

The data was downloaded from CTD on September 28 2010, and contained 424,266 chemical-disease relationships, consisting of 5,915 unique chemicals and 3,436 diseases (annotated by unique MeSH terms).

We excluded all inferred chemical-disease associations. Furthermore, we only kept diseases which were directly associated to at least 20 chemicals. We converted the data into a binary matrix, representing existing associations as 1, and non-reported associations as 0. The resulting cleaned data set contained 2,057 chemicals and 65 diseases.

3. Method

The Compass workflow is divided into four steps, where SOM and AM form the backbone. The main idea is that the SOM is employed as a first step to find “hotspots”, i.e. areas in the multi-dimensional search space where we are more likely to find strong associations. AM is then subsequently applied to extract the association rules from the hotspots.

The four steps of the Compass are the following: SOM, Associative Variable Group extraction, AM and post-processing. Theworkflow of the method is shown in Figure 1 (a more detailed description of each step in the method is included in the supplementary material):

Step1. SOM
SOM is performed on the training set. To perform SOM, we used a variant available in the R-package (ver. 2.11.0) "kohonen" [7]. It can handle missing values and has the built-in feature where the user may assign arbitrary weights to different variables in the data, thus controlling the degree of influence these variables have on the training process. This feature turns out to be extremely useful, as it affects what area of the search space is covered, thus granting the user some degree of control over which associations are found. We explored maps of size 3x3, 5x5 and 7x7 nodes.

The clusters generated from the SOM may not necessarily be optimal, i.e. some clusters may be sufficiently similar to be merged. For this reason, it is common to perform clustering on the model vectors. To perform the clustering we applied a neural gas algorithm available in the “cclust” package [Edvenia, 2009] in R.

Step 2. Associative Variable Group Extraction
From the clusters *Associative Variable Groups* (AVGs) were extracted, as well as the estimated numeric intervals of any corresponding variables that were involved in the cluster and its associations. This allowed for converting these variables into binary form, which was subsequently used to perform AM in step 4. Specifically, the interesting categorical variables were primarily those that had high model vector values, above a user specified cutoff, while the interesting numerical variables had high and low model vector values, below or above some specified cutoffs, respectively. Intermediate ranges for each numerical variable were deemed potentially interesting, if the standard deviation of all samples in a cluster was below a cutoff.

The boundaries for the numerical ranges were obtained from the 1^{st} and 3^{rd} quartiles of the sample values in each cluster.

Step 3. Association Mining (AM)
Each AVG was subjected to a thorough search for association rules, iterating through all combinations of the variables in the group. Within each combination of variables, all possible association rules were in turn generated and subjected to Fisher’s Exact Test [8] to obtain a p-value. The confidence measure was used to assess the direction of the rule. We limited the number of variables in each AVG to a maximum of 20, and limited the number of variables in each itemset to 5. This decreased the computational load, as well as filtered out more complex association rules with 6 or more variables.

To reduce the number of rules generated, we implemented two types of rule filters in the AM step, which can filter out certain types of rules based on prior knowledge; the *notopp* (not opposite) and *notdom* (not dominant) filter. The notopp filter removes all rules, where certain variables occur on opposite sides of the implication arrow. E.g. if we know that the sizes of the left and right testis correlate, we also know that whenever these two variables occur on the opposite sides, along with

any other variables on either side, we can be confident that the rule will be uninteresting. The notdom filter removes all rules where certain variables are dominant, i.e. rules which only contain a specified group of variables, but not others. In our data set, there were a number of variables related to smoking and drinking habits. As these are associated, they are likely to be grouped into the same AVGs. However, as our main interest is the effects of smoking or drinking on clinical outcomes, association rules which only contain drinking or smoking are not of primary interest.

Step 4. Post-processing
The AVGs obtained in step 3 and their corresponding association rules obtained in step 4 were processed into a human readable output. To reduce the output size, the AVGs were clustered into groups based on similarity, i.e. the proportion of variables in common. For each such AVG cluster, their corresponding AVGs and association rules were made easily accessible. As the output is structured around the AVGs, the amount of information a human scientist has to inspect is vastly reduced, compared to an exhaustive list of all association rules.

To further reduce the amount of information in the output, we opted to present the association rules as fuzzy rules, i.e. any numeric ranges in the association rules were coded as either 'H' (high), 'L' (low) or 'I' (intermediate). When presented with a list of many associations at once, it may be less important to specifically know exactly the size and endpoints of any intervals involved in each rule. However, if the association is deemed interesting, the relevant intervals can be easily extracted from the output.

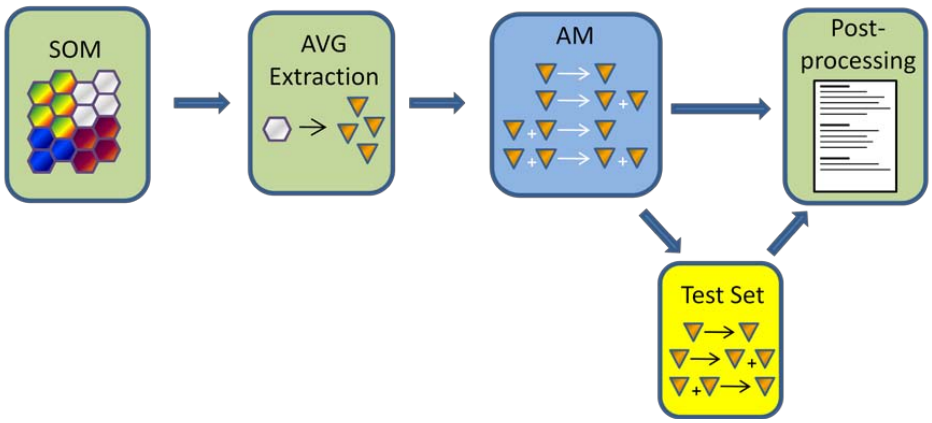

Fig 1) Workflow of the Compass method.
The four steps of the Compass. 1) SOM (Self Organising Map) produces model vectors representing each node (cluster). 2) AVG (Associative Variable Group) extraction produces groups of variables which are hotspots for associations. 3) AM (Association Mining) mines for association rules in the given AVGs, iterating through possible combinations of itemsets and rules. The rules found in the training set are then validated on the test set. 4) Post-processing clusters similar AVGs and produces a human readable output of the AVGs and their corresponding association rules.

Reliability of associations
In the type of unsupervised data mining we present here, there is an extreme risk of finding false positives due to the large number of associations tested. Strategies to deal with this problem have been discussed in literature, such as directly adjusting for multiple testing, validating the

associations on a test set [9], and data randomization [10]. In this study, we explored the use of the direct adjustment approach and validation on test set, in order to limit the first type of error. We briefly comment on these approaches in the discussion section.

In the analysis of the military conscript's data, we used the strategy of validating the newfound associations on a test set. From the results on the training set, we generated 81 hypotheses of interest that we subsequently tested in the test set. The p-values obtained from the test set were corrected for multiple testing using the Holms method [11].

In the analysis of the data from the Comparative Toxicogenomics Database, we applied the direct adjustment approach. To correct for multiple testing, Holms test was used, setting the number of hypotheses tested to the size of the total search space, which was equal to $2.57 \cdot 10^8$ associations.

4. Results

Performance
The primary objective of the Compass is to allow scientists to find new, unexpected and interesting associations in clinical data, and as such, its usefulness is ultimately measured in subjective terms. However, as a central feature of the Compass is its ability to generate associative variable groups (AVGs) which contain statistically significant association rules, a possible approach to assess the usefulness is to compare the statistical significance of the rules generated from the AVGs to the rules obtained from randomly generated variable groups (RVG).

RVGs are generated by randomly choosing variables from the data to form variable groups, and are then subjected to AM. Thus, in this case SOM is not used as an initial step. In the process of generating RVGs, care was taken to not include more than one categorical variable from the same attribute into each group, as this would artificially cause the RVGs to produce poorer associations. We also examined the performance of Compass after the non-missing values in each data variable were randomly shuffled. This may give an idea of the tendency of the method to find spurious associations given the margins and incompleteness of the data set at hand.

Figure 2 illustrates the performance of the Compass (green lines), the RVGs (blue lines), and on randomly shuffled data (red line). The fraction of significant associations is significantly higher for the Compass than RVGs, thus justifying the use of SOM as an initial step in the analysis. We noted that smaller map sizes in the SOM step give rise to stronger associations. We presume this is due to the fact that smaller maps may find associations that are represented by a larger number of samples in the data set. The baseline for the Compass performance on randomly shuffled data is shown in red. In our case the method was unable to find any spurious associations with a p-value of 1e-6 or lower.

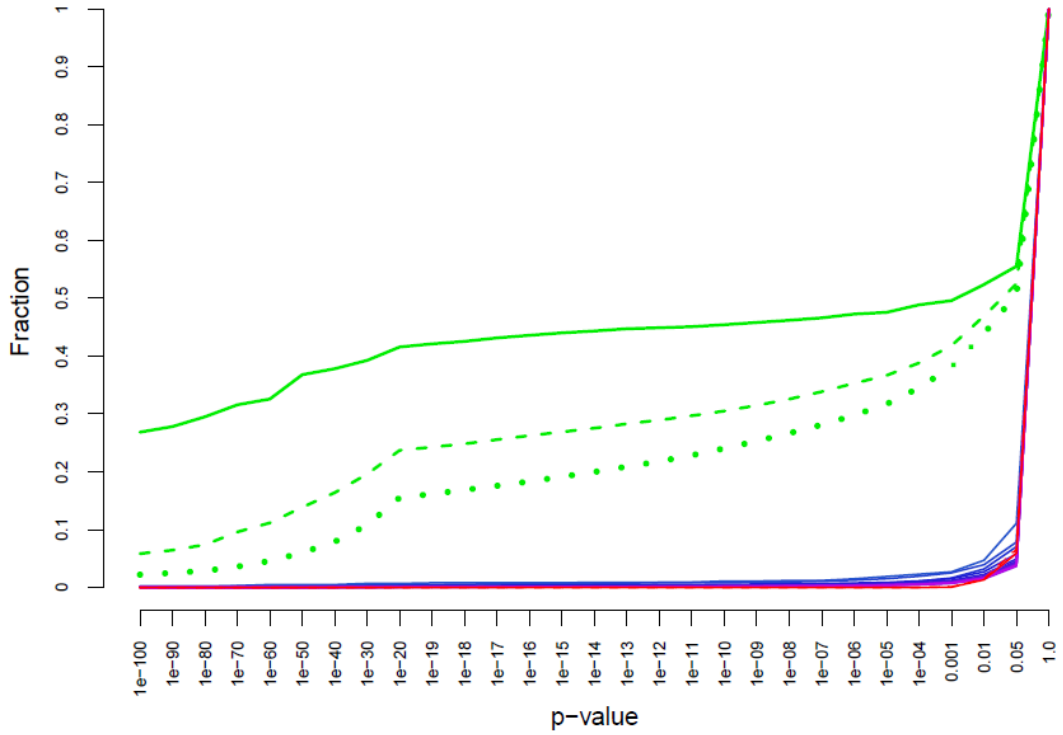

Fig 2. Cumulative distribution plot illustrating the performance of the Compass method. a) Each point on the graph shows the fraction of rules (y-axis) that have a p-value equal to or lower than the corresponding point on the x-axis. The green lines represent results from using Compass, with different map sizes in the AM-step; solid line=3x3, dashed line=5x5, dotted line=7x7. The blue lines represent analysis performed on randomly generated AVGs of various sizes, ranging from 2-15 variables each. The red line shows performance of Compass on randomly shuffled data.

By default, the SOM does not grant the user any control over the area in the search space where association rules are mined. However, there may be many circumstances where one would be interested in finding associations involving certain specific variables, in this case, such as e.g. birth weight. It is possible to obtain some control over this by employing a *weighted search*, by assigning higher weights to certain variables of choice during the SOM-step. Figure 3 illustrates the performance of six different analyses, each assigning a higher weight to one variable during the SOM. The performance differs considerably, depending on which variable is weighted higher. Assigning higher weights to certain variables affects which AVGs are created, and thus which associations are found. Our analysis showed that this approach allows for spreading out the search into a greater area of the multidimensional space, thus finding a larger number of diverse rules (data not shown).

However, assigning a higher weight to a certain variable does not guarantee that all AVGs, and their corresponding association rules, will contain that particular variable. This is true especially in cases with weakly associated variables, where the Compass method may still generate other highly associated variable groups that do not necessarily contain the variable of interest.

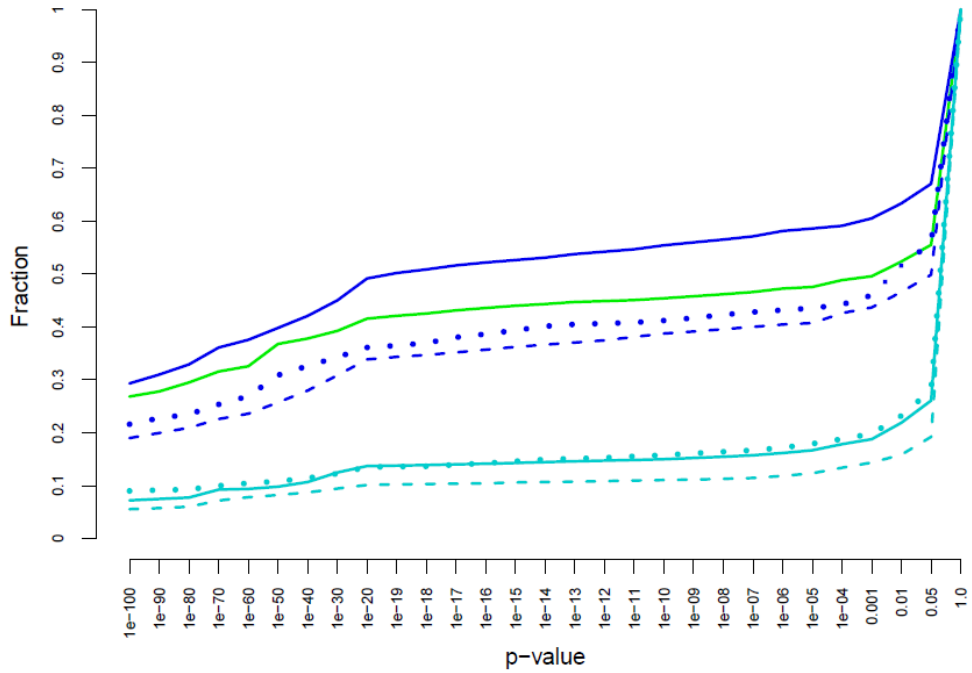

Fig 3. Variable dependent performance.
Cumulative distribution plot illustrating the performance when assigning higher weights to certain variables during the SOM. The weighted variables are: solid blue = "Low testis disposition", dashed blue = "testis size", dotted blue = "proportion of immotile sperm", solid cyan = "Absence of cicatrices", dashed cyan = "early post fetal illness", dotted cyan = "working posture". All analyses were done with SOM of size 3x3. The solid green line represents non-weighted search on 3x3 map, as shown in figure 2.

Clinical findings from the military conscripts data

The Compass method generated 111 AVG clusters and 373,745 fuzzy association rules from the training set with a p-value lower than 0.001. Examples of some of the AVGs found are shown in Table S1 in the supplementary material. By browsing the 111 AVG clusters generated from the analysis, we found 81 interesting associations, and subsequently evaluated them on the test set. The p-values obtained from the test set were corrected for multiple testing using Holms test, taking into account that 81 associations were found. Examples of these associations are listed in Table 2. Note, the confidence given in the table is defined as the conditional probability of the consequent given the antecedent. This definition is commonly employed in Association Mining, and differs from that used in traditional statistics.

Table 2. Examples of associations found in data from military conscripts.

Association Rule	P-value		Confidence	
	Training Set	Test Set	Training Set	Test Set
1. "Right testis volume high" → "Left testis volume high"	1.07E-169	8.10E-319	0.90	0.91 (597)
2. "Low level inhibinb" → "High level FSH"	1.34E-31	1.28E-80	0.69	0.58 (914)
3. "High level FSH" → "Low testis volume"	4.26E-12	5.22E-25	0.88	0.78 (826)
4. "Right testis consistency soft" → "Right testis volume low"	9.06E-07	1.85E-23	0.68	0.85 (62)
5. "Smoking many cigarettes" → "Less likely to go to school"	2.36E-21	3.24E-16	0.59	0.47 (265)
6. "Mother smoking during pregnancy" → "Low birth weight of child"	1.98E-10	8.28E-8	0.60	0.57 (441)
7. "High alcohol consumption" → "Increased levels of free androgen index"	8.78E-04	1.35E-4	0.58	0.49 (620)
8. "High testis volume" → "High inhibinb"	7.61E-25	7.42E-4	0.85	0.90
9. "Mother smoking during pregnancy" → "Conscript is less likely to eat organic food"	3.00E-05	8.69E-4	0.5	0.24
10. "Low testosterone level" → "High BMI"	4.75E-11	4.46E-3	0.45	0.39

The table shows ten selected examples of association rules discovered by applying the Compass method on the military conscripts data set. Each rule is divided into a left-hand side (antecedent) and a right-hand side (consequent) separated by an implication arrow. Each rule should be read "if lhs then rhs". P-values and confidences are provided for both training and test sets. The values in the parentheses represent the number of samples that fulfil the criteria of the rule in question. The p-values for the test set have been corrected for multiple testing using Holms test, taking into account that 81 associations found in the training were subsequently tested.

The statistically most significant association presented in Table 2 demonstrates a correlation in size between the right and left testis. However, it is important to point out that the associations with the lowest p-values are not necessarily the most interesting; they often tend to be trivial. In fact, in our study, we observed that many associations with higher p-values were actually more interesting from a biological and scientific point of view, as these patterns are usually less visible to scientists in the field and therefore more likely to be unexpected. Association 7 in Table 2, which indicates that there is a positive correlation between alcohol intake and free androgen index (levels of free unbound testosterone in blood), has a relatively high p-value. However, it is biologically interesting, and has been discussed in the literature previously [12].

Previous studies have indicated that people in their late teens who smoke perform more poorly in school than non-smokers [13,14]. We find in association 5 in Table 2, an indication that smokers are less likely to go to school, which may be correlated to poor performance. We find this to be true for 47% of smokers in our test set.

We found the unusual association 9 which indicates that if the mother smoked during pregnancy, the military conscript is less likely to eat organic food. The intuitive explanation for this association is that most of the conscripts we studied are in their late teens and early 20's and may still live at home. Therefore, they will eat what their parents bring home. A mother who smokes during pregnancy is less likely to be as health conscious as a non-smoking mother, and therefore is less likely to buy organic food for the household. To our knowledge, this association has not been published previously.

Disease associations found in Comparative Toxicogenomics Database

On this completely different data set the Compass generated a total of three association rules from the Comparative Toxicogenomics Database (CTD) data with a p-value lower than 0.001. To correct for multiple testing we applied the direct adjustment approach, using the Holms method [11] and taking into account the size of the search space. Table 3 lists the rules, along with their p-value and confidence. These rules associate diseases based on the chemical-disease associations present in the data. In other words, diseases or symptoms that share a common set of chemicals, larger than expected by chance, will form association rules together. As such, these rules may imply that two diseases share a common genetic mechanism, or that they may co-occur together in the general population more often than expected by chance.

However, care must be taken when interpreting these results, as they are based on chemical-disease associations reported in literature. As such, they may be heavily biased towards chemicals and diseases that have been studied extensively. Moreover, chemical-disease associations that were not present in CTD, and likely not reported extensively in the literature, were assumed to not exist.

Table 3. Associations found in CTD.

Association Rule	P-value		Confidence
	Unadjusted	Adjusted	
1. "Pain" → "Inflammation"	5.61e-18	1.44e-09	0.43 (20)
2. "Liver Neoplasms" → "Lung Neoplasms"	9.14e-16	2.35e-07	0.35 (38)
3. “Prostatic Neoplasms” → "Breast Neoplasms"	2.31e-15	5.94e-07	0.33 (22)

The table shows three rules discovered from the chemical-disease relations obtained from CTD. The rules in the table associate diseases together based on which chemicals are associated (or not associated) with each disease. Adjusted and unadjusted p-values are shown as well as confidences. The values in the parentheses represent the number of samples that fulfil the criteria of the rule in question (support).Note, the confidence is defined, as used in Association Mining, as the conditional probability of the consequent given the antecedent.

The statistically most significant association in Table 3 relates pain to inflammation, indicating that 43% of chemicals associated with pain are also associated with inflammation. These chemicals mostly include known anti-inflammatory painkillers available on the market, but also natural substances found in plants such as Capsaicins, Mangifer Indica extract or Desmodium Gangeticum extract. This association may reasonably be regarded as a trivial association, as pain is one of five cardinal signs in acute inflammation.

Association 2 links lung cancer with prostate cancer via chemicals such as epoxy compounds, polyvinyl chlorides, arsenite and butylated hydroxytoluene. We were unable to find any reports in the scientific literature linking these two cancers specifically, although it is known that metastatic cancers originating from other tissues, not only lung, may spread to the liver.

Association 3 in Table 3 links prostatic cancer with breast cancer; both being gender-specific for males and females, respectively. The chemicals in CTD linking these two diseases mainly consist of different estrogens and androgens, which are endogenous sex hormones. Familial co-occurrence between prostate cancer and breast cancer has been reported previously in a number of different studies [15,16].

5. Discussion

We have developed the Compass method, an unsupervised approach that can successfully mine data for interesting associations in clinical data with missing values and mixed data types. We have also demonstrated the use of this approach on non-clinical data containing chemical-disease associations based on text mining from CTD, where we successfully generated associations between different phenotypes. Our approach can find associations with very low p-values, but is also sufficiently sensitive to find interesting associations with relatively high p-values, such as the relation between alcohol consumption and levels of free androgen index, which we were able to confirm in the literature.

The Compass method is divided into 4 steps, where SOM and AM are the two initial components of the pipeline. SOM is applied as a first step in order to reduce the search space covered by AM in a later step. Several benefits of this approach has been discussed, including handling of numerical variables, the possibility for a weighted search (the ability to control which part of the search space is covered), and generation of AVGs which reduces the search space covered to find associations with low p-values and also allows for easier handling of a large number of associations in the result output.

Handling numerical variables
We are able to handle numerical variables without a priori binning, and in particular, without using the support measure to restrict the search space by applying SOM in the first step. The resulting clusters obtained from the SOM output can suggest approximate intervals in the numerical variables that are relevant to the associations found in that cluster. Although the method is able to find numeric intervals in any intermediate range that may be associated with a group of variables, the main driving force in the Compass pipeline are the values in the high or low end of the spectrum, as these are more likely to be greater distances apart from other non-similar samples during the learning process of the SOM. The values that are to be considered high or low are parameters specified by the user before the analysis. In our study, we considered values to be low or high if they were below or above the 1st and 3rd quartiles, respectively.

Weighted Search - increasing control and sensitivity
In the first step of the analysis, SOM is employed in order to point towards hotspots in the search space with strong associations. The locations of the hotspots that the SOM will point towards are normally outside the control of the user. However, by performing a weighted search, assigning higher weights to one or more variables in the SOM, the method may be coerced to find associations that include particular variables of interest. This grants the user some control over which associations are found, and may also increase sensitivity. The increase in sensitivity arises because variables that are involved in patterns represented by very few samples will have a higher chance of being found if they are weighted higher.

Our analysis indicated that iterating through all variables in a data set, assigning a higher weight to a single variable in each iteration, will essentially “force” the SOM to spread out over a bigger area in the search space, thus finding a greater diversity of associations than in the case where a non-weighted analysis is performed (data not shown).

Dealing with large number of associations
Typically, one will find thousands of associations, which fulfil certain criteria of reliability, such as low p-value and high replicability. However, the p-value and replicability of an association does not confer any information about its interestingness. In our study, we found many associations with extremely low p-values (some as low as 1e-100) which turned out to be either already known or trivial. One such example is the finding that the size of the left testis strongly correlates with the

size of the right testis (see Table 2). This association makes intuitive sense, but would be considered too trivial to be publishable on its own. The only true measure of interestingness for our purposes is the subjective opinion given by an expert in the field, who can decide whether an association is new and/or interesting.

However, it is nearly impossible for a human brain to browse through all newfound associations, and this procedure does not give a good overview of the structure of the data either. It is therefore important to reduce the amount of information inspected manually. For this purpose, we propose the approach where a number of AVGs are presented in the result output. Given that the AVGs are "hotspots" for statistically significant associations, experts in the field can with a glance easily identify if any potential association between the variables occurring together in a group would be interesting. Thus the amount of information that a human has to process is vastly reduced, as each AVG can, depending on its size, represent hundreds or even thousands of association rules.
While the approach with AVGs does considerably reduce the amount of information needed to manually process, it may in cases of larger data sets even still be too much. We have therefore also implemented the additional measures of rule filters and generation of fuzzy rules (described in the methods section) and merging nearly identical categorical variables (described in the data section).

Statistically sound associations
Another challenge with dealing with large number of associations is, of course, the extreme risk of finding false positives. We briefly discuss two approaches that deal with this challenge.

Direct adjustment
In the case of directly adjusting the p-values for multiple testing, the correction factor is dependent upon the size of the search space, i.e. the theoretical number of associations that can be tested in any given data. As the possible number of rules is large, this strategy has a tendency to filter out many true associations.

In simple binary data sets, calculating the size of the theoretical search space is usually straightforward. However, the presence of numerical variables may complicate the matter due to the fact that any given numeric interval may theoretically be involved in any association. As an example, consider two numeric variables N1 and N2 that are linearly correlated. These two variables could produce association rules like "N1 High → N2 High", and "N1 Low → N2 Low", as well as "N1 Intermediate → N2 Intermediate". Due to the fact that these two numerical variables are continuous, there can theoretically be an extremely large number of intermediate intervals in which N1 and N2 are associated. If all these intervals are to be taken into account, the theoretical search space may easily become astronomical, even for relatively small data sets. Only extremely significant associations would then pass as reliable.

Validation on test set
Validating associations from a training set on a test set is a common strategy to control for false discoveries. The correction factor for multiple testing is in this case the number of associations that were considered interesting in the training set, and subsequently validated on the test set, and is thus much smaller then the correction factor used in the direct adjustment approach.

In many cases, if no test set is available, it is common to divide a given data set into a training and test set. A disadvantage with this approach is that the splitting of the data into two smaller sets reduces the number of samples, and hence reduces the power to detect new associations. Another problem is that the splitting may not be optimal with regards to missingness and skewed distributions. By chance, variables may be unevenly divided such that more missing values appear in the test set than the train set (and vice versa). Thus, certain associations may fail to be validated, not because of being spurious, but rather due to the fact that the degree of missingness is too high

for the variables involved. These problems are more pronounced for higher order associations, but should be less problematic for data sets containing large numbers of samples. Moreover there may be a general issue of similarities between training and test sets, which may limit to what degree newfound associations can be generalized to e.g. other parts of the population if the data sets are too similar. This has not been discussed extensively in the biobank questionnaire analysis field but has often been discussed within molecular level bioinformatics, where the performance of prediction algorithms depends strongly on the similarity between training and test set examples (e.g. the sequence similarity between two proteins) [17,18].

In general, we would not recommend using the direct adjustment approach, as it is in many cases needlessly strict and sets the p-value cutoff extremely low due to the usually large search space. As a consequence, only associations with extremely low p-values will be retained, and potentially interesting associations with higher p-values will be overlooked. In special cases, it may however be used, as in our analysis of the data from CTD, which was a relatively small data set that generated a small number of very significant rules.

It is likely that other techniques than SOM may be used in a similar way to restrict the search space for association rules. We compared the performance of the Compass workflow, if using k-means clustering [19] instead of SOM as the first step of the procedure. K-means clustering is similar to a special case of SOM, where the degree of influence on neighbouring nodes is set to 0. We found that k-means performs worse than SOM (data not shown).

The interest for the type of unsupervised analysis presented here has been growing in the life sciences, particularly due to the large amounts of data available today. It is our opinion that the Compass method is well suited for data sets with a much higher number of samples than that used in our study (roughly 2,000 samples in the CTD data), as the analysis scales somewhat linearly in the SOM step. However, data sets with a much larger number of variables may create a challenge when using our approach. The computational time for the SOM step will increase linearly with the number of variables, but for the AM step the load may be considerably heavier due to the combinatorial explosion.

It is a future prospect of ours to investigate our method on other larger data sets, and how to improve its performance. We are currently investigating ideas involving genetic algorithms or a direct modification of the SOM algorithm itself.

References

1. Hájek P, Havel I, Chytil M (1966) The GUHA method of automatic hypotheses determination. Computing 1: 293-308.
2. Delgado M, Sanchez D, Martn-Bautista MJ, Vila MA (2001) Mining association rules with improved semantics in medical databases. Artificial Intelligence in Medicine 21: 241-245.
3. Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biological Cybernetics 43: 59-69.
4. Kohonen T (2001) Self-Organizing Maps. New York: Springer.
5. Jørgensen N, Carlsen E, Nermoen I, Punab M, Suominen J, et al. (2002) East-West gradient in semen quality in the Nordic-Baltic area: a study of men from the general population in Denmark, Norway, Estonia and Finland. Human reproduction (Oxford, England) 17: 2199-2208.
6. Davis A, King B, Mockus S, Murphy C, Saraceni-Richards C, et al. The Comparative Toxicogenomics Database: update 2011. Nucleic Acids Research.
7. Wehrens R, Lutgarde MCB (2007) Self- and Super-organizing Maps in R: The kohonen Package. Journal of Statistical Software 21.
8. Fisher RA (1922) On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society 85: 87-94.
9. Webb G (2007) Discovering Significant Patterns. Mach Learn 68: 1-33.
10. Gionis A, Mannila H, Mielikäinen T, Tsaparas P (2007) Assessing data mining results via swap randomization. ACM Trans Knowl Discov Data 1: 14.
11. Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6: 65-70.
12. Sarkola T, Eriksson P (2003) Testosterone increases in men after a low dose of alcohol. Alcoholism, clinical and experimental research 27: 682-685.
13. Hu T-w, Lin Z, Keeler TE (1998) Teenage smoking, attempts to quit, and school performance. American Journal of Public Health 88: 940–943.
14. Borland BL, Rudolph JP (1975) Relative effects of low socio-economic status, parental smoking and poor scholastic performance on smoking among high school students. Social Science & Medicine (1967) 9: 27-30.
15. Jennifer LB-D, Elizabeth AD, Rodney LD, Cathryn HB, James EM, et al. (2006) Association between family history of prostate and breast cancer among African-American men with prostate cancer. Urology 68: 1072-1076.
16. Lopez-Otin C, Diamandis E (1998) Breast and Prostate Cancer: An Analysis of Common Epidemiological, Genetic, and Biochemical Features. Endocr Rev 19: 365-396.
17. Frimurer T, Bywater R, Nærum L, Lauritsen L, Brunak S (2000) Improving the Odds in Discriminating “Drug-like” from “Non Drug-like” Compounds. Journal of Chemical Information and Computer Sciences 40: 1315-1324.
18. Nielsen H, Engelbrecht J, von Heijne G, Brunak S (1996) Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. Proteins 24: 165-177.
19. Lloyd S (1982) Least squares quantization in PCM. IEEE Transactions on Information Theory 28: 129-137.

Part V

Epilogue

Chapter 10

Concluding Remarks

10.1 Environmental Chemicals

As many of the chemicals I analyzed have been banned for a number of years, their levels have been decreasing over time. The results produced in my studies should therefore be regarded as a snapshot at the point in time when the chemicals were measured. As the data I used for my analyses was generated roughly 8-10 years ago, the difference between the chemical environment between Denmark and Finland (reported in Manuscript I) may not necessarily be the same today.

Eventhough many chemicals have been banned, the industry is often able to find substitues. These substitutes may be chemically and structurally similar and therefore likely to have similar negative health effects as the banned chemicals. Despite this, it will take time before the legislature catches up with the change in order to ban any new chemicals introduced into the market. Environmental chemicals with negative health effects will therefore still remain in society for a long time. I am very curious to know how the chemical exposure patterns look today, which chemicals dominate, and what their effects are. The type of comprehensive studies required for such a task would require meausirng over a hundred chemical compounds in a relatively large number of samples from different geographical regions,

which unfortunately is very expensive and time consuming with todays technology.

10.2 Mining for Associations in Clinical Data

During my project I was working with two clinical data sets; one set containing biological and questionnaire data from military conscripts, and the other containing longitudinal data, including biological and chemical measurements, from newborn boys. I was able to find the greatest number of interesting associations in the data on military conscripts. However, most of the associations were already known. Moreover, a number of potentially very interesting associations turned out to be artifacts of some systematic errors, such as examiner bias. With an approach that is able to mine clinical data for interesting rules, the philosophical question therefore remains how useful such a method is for finding new and unexpected associations. Many data sets are usually created for some specific purposes in mind and to test certain hypotheses. For that reason, these data sets may be most suitable for testing a limited number of pre-conceived hypotheses. However, new and unexpected associations would most likely fall outside such pre-conceived hypotheses, and as such are not optimally suitable to be tested and/or discovered in the given data. And even if they are found, they still need to be evaluated for interestingness and whether they are publishable. Nevertheless, I feel confident that with more data and given enough time, there will certainly come a chance to catch that big fish.

Bibliography

[1] Murray JD (1993) Mathematical Biology. Springer, second corrected edition edition.

[2] Becker R, Selden G (1998) The Body Electric: Electromagnetism and the Foundation of Life. Harper Paperbacks, first edition.

[3] Main KM, Skakkebæk NE, Virtanen HE, Toppari J (2010) Genital anomalies in boys and the environment. Best Practice & Research Clinical Endocrinology & Metabolism 24: 279–289.

[4] Hutsona JM, Hasthorpe S (2005) Testicular descent and cryptorchidism: the state of the art in 2004. Journal of Pediatric Surgery Lecture 40: 297–302.

[5] Boisen KA, Chellakooty M, Schmidt IM, Kai CM, Damgaard IN, et al. (2005) Hypospadias in a cohort of 1072 Danish newborn boys: prevalence and relationship to placental weight, anthropometrical measurements at birth, and reproductive hormone levels at three months of age. The Journal of clinical endocrinology and metabolism 90: 4041–4046.

[6] Paulozzi LJ (1999) International trends in rates of hypospadias and cryptorchidism. Environmental Health Perspectives 107: 297–302.

[7] Cooper TG, Noonan E, von Eckardstein S, Auger J, Baker GW, et al. (2010) World Health Organization reference values for human semen characteristics. Human reproduction update 16: 231–245.

[8] Bonde JP, Ernst E, Jensen TK, Hjollund NH, Kolstad H, et al. (1998) Relation between semen quality and fertility: a population-based study of 430 first-pregnancy planners. Lancet 352: 1172–1177.

[9] Strohmeyer T, Peter S, Hartmann M, Munemitsu S, Ackermann R, et al. (1991) Expression of the hst-1 and c-kit protooncogenes in human testicular germ cell tumors. Cancer research 51: 1811–1816.

[10] Palumbo C, van Roozendaal K, Gillis AJ, van Gurp RH, de Munnik H, et al. (2002) Expression of the PDGF alpha-receptor 1.5 kb transcript, OCT-4, and c-KIT in human normal and malignant tissues. Implications for the early diagnosis of testicular germ cell tumours and for our understanding of regulatory mechanisms. The Journal of pathology 196: 467–477.

[11] Gidekel S, Pizov G, Bergman Y, Pikarsky E (2003) Oct-3/4 is a dose-dependent oncogenic fate determinant. Cancer cell 4: 361–370.

[12] RajpertâĂŘDe Meyts E, Hanstein R, Jørgensen N, Græm N, Vogt PH, et al. (2004) Developmental expression of POU5F1 (OCTâĂŘ3/4) in normal and dysgenetic human gonads*. Human Reproduction 19: 1338–1344.

[13] Hustin J, Collette J, Franchimont P (1987) Immunohistochemical demonstration of placental alkaline phosphatase in various states of testicular development and in germ cell tumours. International Journal of Andrology 10: 29–35.

[14] Chabot B, Stephenson DA, Chapman VM, Besmer P, Bernstein A (1988) The proto-oncogene c-kit encoding a transmembrane tyrosine kinase receptor maps to the mouse W locus. Nature 335: 88–89.

[15] Huang E, Nocka K, Beier DR, Chu TY, Buck J, et al. (1990) The hematopoietic growth factor KL is encoded by the Sl locus and is the ligand of the c-kit receptor, the gene product of the W locus. Cell 63: 225–233.

[16] Niwa H, Miyazaki J, Smith AG (2000) Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nature genetics 24: 372–376.

[17] Jacobsen GK, Nørgaard-Pedersen B (1984) Placental alkaline phosphatase in testicular germ cell tumours and in carcinoma-in-situ of the testis. An immunohistochemical study. Acta pathologica, microbiologica, et immunologica Scandinavica Section A, Pathology 92: 323–329.

[18] Rajpert-De Meyts E (2006) Developmental model for the pathogenesis of testicular carcinoma in situ: genetic and environmental aspects. Human reproduction update 12: 303–323.

[19] Chia VM, Quraishi SM, Devesa SS, Purdue MP, Cook MB, et al. (2010) International trends in the incidence of testicular cancer, 1973-2002. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 19: 1151–1159.

[20] Huyghe E, Matsuda T, Thonneau P (2003) Increasing incidence of testicular cancer worldwide: a review. The Journal of urology 170: 5–11.

[21] Møller H, Jørgensen N, Forman D (1995) Trends in incidence of testicular cancer in boys and adolescent men. Int J Cancer 61: 761–764.

[22] Adami HO, Bergström R, Möhner M, Zatoôski W, Storm H, et al. (1994) Testicular cancer in nine northern european countries. Int J Cancer 59: 33–38.

[23] Andersen AG, Jensen TK, Carlsen E, Jørgensen N, Andersson AM, et al. (2000) High frequency of sub-optimal semen quality in an unselected population of young men. Human Reproduction 15: 366–372.

[24] Swan SH, Elkin EP, Fenster L (1997) Have sperm densities declined? A reanalysis of global trend data. Environmental health perspectives 105: 1228–1232.

[25] (1992) Cryptorchidism: a prospective study of 7500 consecutive male births, 1984-8. John Radcliffe Hospital Cryptorchidism Study Group. Archives of Disease in Childhood 67: 892–899.

[26] Skakkebaek NE, Rajpert-De Meyts E, Jørgensen N, Carlsen E, Petersen PM, et al. (1998) Germ cell cancer and disorders of spermatogenesis: an environmental connection? APMIS : acta pathologica, microbiologica, et immunologica Scandinavica 106.

[27] Møller H, Prener A, Skakkebaek NE (1996) Testicular cancer, cryptorchidism, inguinal hernia, testicular atrophy, and genital malformations: case-control studies in Denmark. Cancer causes & control : CCC 7: 264–274.

[28] Boisen KA, Kaleva M, Main KM, Virtanen HE, Haavisto AMM, et al. (2004) Difference in prevalence of congenital cryptorchidism in infants between two Nordic countries. Lancet 363: 1264–1269.

[29] Skakkebaek NE, Rajpert-De Meyts E, Jørgensen N, Main KM, Leffers H, et al. (2007) Testicular cancer trends as 'whistle blowers' of testicular developmental problems in populations. International journal of andrology 30: 198–205.

[30] SkakkebÃÂęk NE, Rajpert-De Meyts E, Main KM (2001) Testicular dysgenesis syndrome: an increasingly common developmental disorder with environmental aspects: Opinion. Human Reproduction 16: 972–978.

[31] Welsh M, Saunders PTK, Fisken M, Scott HM, Hutchison GR, et al. (2008) Identification in rats of a programming window for reproductive tract masculinization, disruption of which leads to hypospadias and cryptorchidism. Journal of Clinical Investigation 118: 1479–1490.

[32] Heller CG, Clermont Y (1963) Spermatogenesis in Man: An Estimate of Its Duration. Science 140: 184–186.

[33] Rider CV, Furr J, Wilson VS, Gray LE (2008) A mixture of seven antiandrogens induces reproductive malformations in rats. International Journal of Andrology 31: 249–262.

[34] Christiansen S, Scholze M, Axelstad M, Boberg J, Kortenkamp A, et al. (2008) Combined exposure to anti-androgens causes markedly increased frequencies of hypospadias in the rat. International Journal of Andrology 31: 241–248.

[35] Kortenkamp A (2008) Low dose mixture effects of endocrine disrupters: implications for risk assessment and epidemiology. International journal of andrology 31: 233–240.

[36] Wold H (1966) Estimation of Principal Components and Related Models by Iterative Least squares, New York: Academic Press. pp. 391–420.

[37] Lee MH, Cavinato AG, Mayes DM, Rasco BA (1992) Noninvasive short-wavelength near-infrared spectroscopic method to estimate the crude lipid content in the muscle of intact rainbow trout. Journal of Agricultural and Food Chemistry 40: 2176–2181.

[38] McGlynn KA, Quraishi SM, Graubard BI, Weber JPP, Rubertone MV, et al. (2009) Polychlorinated biphenyls and risk of testicular germ cell tumors. Cancer research 69: 1901–1909.

[39] Cooke PS, Zhao YD, Hansen LG (1996) Neonatal polychlorinated biphenyl treatment increases adult testis size and sperm production in the rat. Toxicology and applied pharmacology 136: 112–117.

[40] Knox C, Law V, Jewison T, Liu P, Ly S, et al. (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic acids research 39.

[41] Davis AP, King BL, Mockus S, Murphy CG, Saraceni-Richards C, et al. (2011) The Comparative Toxicogenomics Database: update 2011. Nucleic Acids Research 39: D1067–D1072.

[42] Kuhn M, Szklarczyk D, Franceschini A, Campillos M, von Mering C, et al. (2010) STITCH 2: an interaction network database for small molecules and proteins. Nucleic acids research 38: D552–556.

[43] Fields S, Song O (1989) A novel genetic system to detect protein-protein interactions. Nature 340: 245–246.

[44] Pandey A, Mann M (2000) Proteomics to study genes and genomes. Nature 405: 837–846.

[45] Gavin ACC, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631–636.

[46] Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637–643.

[47] Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141–147.

[48] Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5: 101–113.

[49] Agrawal R, Srikant R (1994) Fast Algorithms for Mining Association Rules in Large Databases. In: VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 487–499. URL `http://portal.acm.org/citation.cfm?id=645920.672836`.