Drug repositioning and indication discovery using description logics

Samuel Claude Jean Croset

Darwin College

A thesis submitted on January, 2014 for the Degree of Doctor of Philosophy

Abstract

My abstract balbalbalabalbal

Declaration

This thesis:

- is my own work and contains nothing which is the outcome of work done in collaboration with others, except where specified in the text;
- is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other university; and
- does not exceed the prescribed limit of 60,000 words.

Samuel Claude Jean Croset
Month, Year

Acknowledgements

My acknowledgements balbalbala

Contents

chapter 1

Review of computational drug repositioning approaches

Key points

- Drug repositioning is the discovery of new indications for approved or failed drugs.
- Repositioning opportunities exist because drugs perturb multiple biological entities (on and off-targets) themselves involved in multiple biological processes. As the drug discovery pipelines are focused on one disease of interest, a therapeutic application for a drug to other areas can be missed.
- A variety of predictive computational approaches have been developed to identify drug repositioning opportunities, each based around a biological concept of interest.
- For my thesis, I decided to investigate drug repositioning using the concept of mode of action. This notion is of interest as it can represent the various biological functions a drug can play in an organism and help to identify new therapeutic applications.

Authors comment

This chapter is a review of the computational methods developed in the last decade, presenting their relative strengths and weaknesses. This state-of-the-art analysis helped me to define the rationale for my thesis.

Living organisms, just as machines, are subject to dysfunction. Sometimes an internal abnormality can impair the correct functioning or sometimes external forces, such as the interaction with the environment, can damage a body. If not fixed, malfunctions accumulate and will eventually result in the death of the entity, namely the cessation of all functions. Since prehistoric times (Wikipedia, 2014f), humans have been interested in preventing and handling dysfunctions, in order to extend the lifespan of objects or to improve the quality of their own existence. When the aim is to fix living bodies, this practice goes by the name of medicine, or the art of preventing, diagnosing and treating diseases.

Alleviating an impairment is a daunting task, from both a biological and legal perspective. In this regard, doctors can rely on a number of tools, engineered throughout the years. Among the most commonly used ones such as medical devices, are the active small molecules, the drugs. Drugs are of primary interest, as they can impact the treatment of complex processes with molecular roots, such as cancer or pain for instance, in a relatively controllable and safe manner. However, in order to be usable in clinics, a candidate drug has first to go through a development phase which takes nowadays at least a dozen years and can cost up to a billion dollars (DiMasi, 2001). This process involves a myriad of different people, from biologists to law attorneys.

As my interests revolve primarily around pharmacology and computer sciences, I decided to focus my efforts on the molecular side of the problem. More precisely, I decided to systematically characterise and understand the multiple roles any drug can play in the human body, using computational means. My work will be extensively described throughout this manuscript and finds an application in a topic named drug repositioning, which is the object of this chapter.

1.1 Relevance of drug repositioning

Drug repositioning (also referred as drug repurposing, re-profiling, therapeutic switching and drug re-tasking) is the identification of new therapeutic indications for known drugs. These drugs can either be approved and marketed compounds used daily in a clinical setting, or they can be drugs that have been "shelved", namely molecules that did not succeed in clinical trials or for which projects have

been discontinued for various reasons. In one sentence, drug repositioning can be defined as *renewing failed drugs and expanding successful ones* (Barratt and Frail, 2012).

One motivation behind drug repositioning is the possibility to further market and extend the application line or patent life of a drug, therefore increasing the revenue stream generated from it. Another aim is the treatment of rare or neglected diseases; usually such conditions are difficult to address for financial reasons, yet there might exist some safe and active molecules already developed for other indications, deemed suitable for this scenario (Men et al., 2010). I refer the reader to some recent excellent reviews (Ashburn and Thor (2004), Dudley et al. (2011a), Hurle et al. (2013)) in order to fully appreciate the economical market behind this approach, as well as legal challenges coming along. I limited myself, in the context of this work, to exploring the subject from an academic perspective: Characterising the various roles approved drugs can play using computational means, without necessarily seeking business opportunities.

1.1.1 Opportunities for finding new indications

Many drugs have been successfully repositioned in the past; classical examples such as sildenafil (Viagra) and thalidomide will be presented in the coming sections. But first, the scientific legitimacy of the idea behind drug repurposing should be discussed: How is it possible for a drug to play multiple roles? What is the molecular rationale? Why does the drug discovery process not automatically identify such opportunities?

In order to be able to answer these questions, I will briefly present the traditional drug development pipeline, which produced most of the recent therapeutic chemicals (Swinney and Anthony, 2011).

The fundamental idea behind the discovery of a new medicine has not evolved much since prehistoric times (Wikipedia, 2014f); it still consists mostly of a trial and error process, hence the term "discovery". The operation starts by picking a disease of interest, selected in terms of market size or clinical needs. Then a large collection of chemicals is experimentally tried to see if any relevant effect in regards to the chosen condition can be produced. This procedure is called screening. There exist mostly two types of screening: target-based and phenotypic. The former optimises the selection of the chemicals on the ability to bind

a biological entity (usually a protein), called the target and that is relevant for the pathological process studied. The more specifically the chemical interacts with the active site of the target, the cleaner the action of the chemical will be (key-lock model - see Figure 1.1).

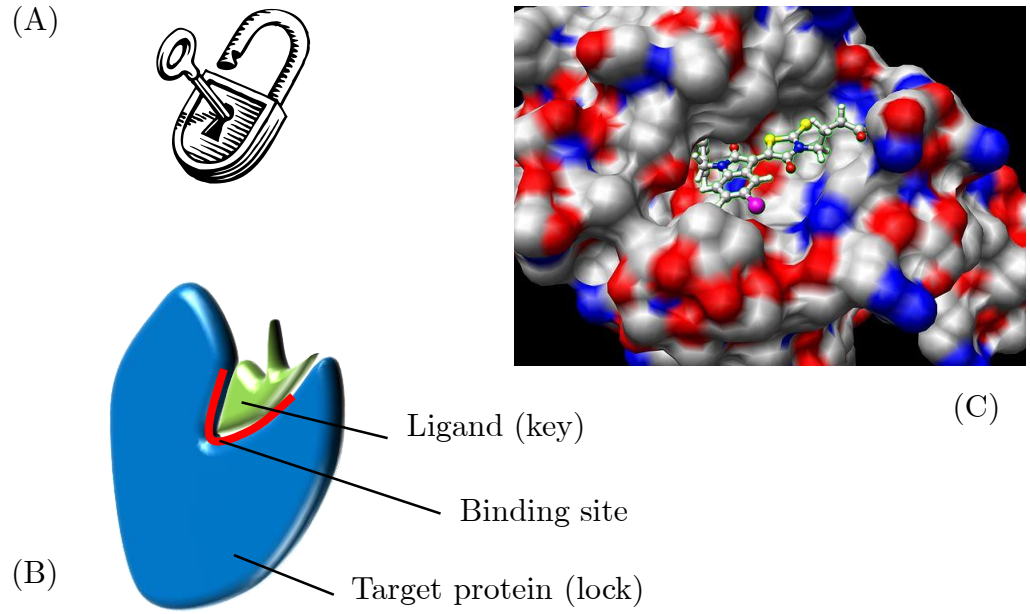


Figure 1.1: Key-lock model. (A) Illustration of the analogy. (B) Schematic representation of the interaction between a ligand and a protein (receptor, enzyme, transporter, etc). One aim of drug discovery is to find alternative molecules (i.e. drugs) capable of either mimicking the action of the natural ligand or reversing it. In theory, the more specific the molecule is to the binding site, the more accurate the pharmacological response will be. (C) Computer generated representation of the interaction between a ligand and a protein. (Source http://en.wikipedia.org/wiki/File:Docking.jpg)

On the other hand, a phenotypic screen does not make any assumptions about the underlying pathological mechanism and proteins involved. A cell line or model organism representative of the disease is directly used to read the results of the screening. Both methods help to efficiently discover active chemicals, whose structures are then further optimised for efficacy (so called *lead optimisation*). Now drug repurposing opportunities can be derived from three key observations (Barratt and Frail, 2012) about the traditional workflow presented above and summarised in Figure 1.2.

The first observation appreciates the limits of the key-lock model and the "magic bullet" concept (Wikipedia, 2014d). In practice, it is extremely challeng-

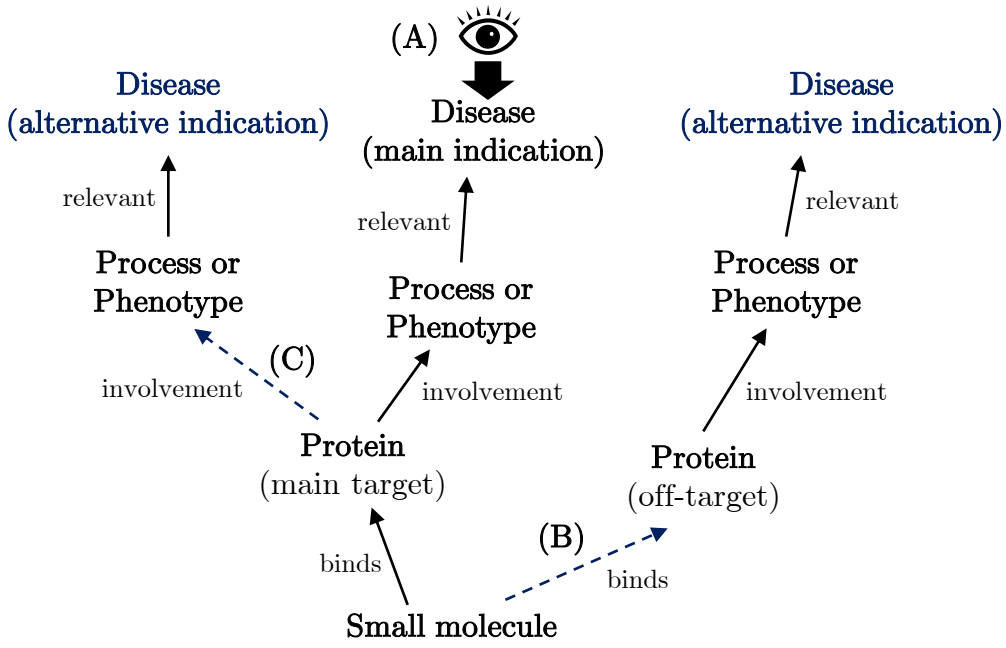


Figure 1.2: Drug discovery process and opportunities for drug repositioning (in blue; logical connection at the origin of the opportunity are dashed). (A) Traditional drug discovery workflow: From biological evidences showing the involvement of a protein in a process relevant to a particular disease of interest, a series of chemicals are screened for bioactivity. Potent and safe molecules will become drugs indicated for the disease. Paul Ehrlich postulated that chemicals should be as specific as possible against the target (organism or protein), in order to increase the control over the pharmacology. The concept of "magic bullet" describes such an ideal therapeutic compound. (B) Repositioning hypothesis: Often, a small molecule binds to other proteins (so called off-targets), themselves potentially involved in other pathologies. (C) Repositioning hypothesis: The main protein target can be involved in a series of alternative biological processes and relevant for another disease.

ing to design a molecule that will interact with a single protein only (Paolini et al., 2006) (Li et al., 2010). Most of the drugs will bind to multiple proteins in an organism and can therefore produce a variety of unwanted effects (concerns between 35 to 62% of the chemicals - sometimes described as "drug promiscuity"). This characteristic is well-known and referred to as off-target effect or polypharmacology. Identifying the off-target proteins interacting with known drugs can provide repurposing opportunities.

The second observation values evidence that molecular targets are themselves involved in multiple biological processes and capable of performing multiple functions. Historically speaking, one protein target was assumed to be responsible

for one biological role. Hitting the right protein, such as a receptor or an enzyme known to be directly involved in the disease was a straightforward way to generate relevant phenotypic outcomes, as illustrated in Figure 1.2. However, this approach overlooks the fact that proteins function in the context of signalling pathways and networks (Hopkins (2008), Barabasi and Oltvai (2004)), the perturbations experienced by one node cascading on to its neighbours due to molecular dynamism. Therefore understanding and appreciating the involvement of the protein target in the biological system can help to identify a new role for a drug.

The last observation derives from the very strategy used by drug discovery programs. As a disease or condition is the starting point of the screening, identified chemicals are by definition optimised in this regard. Compounds will be tested during the expensive clinical trials for the main indication mostly and obviously not for all possible diseases. It is therefore possible for a compound to be used for different purposes, yet not be identified as such. Related to this, new usages can be discovered during clinical practice, and the prescription of the drug can evolve accordingly even if the alternative indication is not legally approved. This practice is known as off-label prescription (Wikipedia, 2014b) and demonstrates the potential of a drug to address various indications.

To conclude, these three observations alone justify the potential existence of repositioning opportunities, especially for the drugs discovered via the target-based or phenotypic screening. Meaningful repurposing options appear to be strongly dependent on the available biomedical knowledge, as well as our understanding of the molecular system.

1.1.2 Drug repositioning faces legal and scientific challenges

In theory, it is possible for a chemical to be active for multiple therapeutic indications. Yet in practice, several obstacles can impair the development of a potential new usage.

The first factor to handle is serendipity. Biology is complex and capricious, exemplified by diseases such as cancers or dementias Ashburn and Thor (2004). A drug rarely keeps its original indication (Barratt and Frail, 2012); it gets reoriented throughout the years when more data becomes available and its in vivo

pharmacology better understood. Famous repurposing and discovery stories were mostly due to chance and unexpected results (e.g. the Viagra, as will be discussed), therefore it can be difficult to forecast any relevant opportunities.

Other challenges come from the corporate and legal aspect around drug discovery. Indeed, in order to be commercially valid, a molecule needs much more than just to exhibit powerful pharmacological features. If the intellectual properties around the molecule are expired or close to be, there might be no incentive to continue the research on an alternative indication, as no profit would derive from it. Re-indicating a drug is also not part of the standard regulatory procedures, therefore administrative problems can happen, delaying or preventing the new usage of the compound. Moreover, some unnecessary safety concerns could appear when the drug is tested for the new indication. Indeed, the repositioning would potentially target a different group of patients, with different physiological conditions, and it is not to be excluded that an unforeseen adverse event could happen during the trials over the new population, hence compromising the original indication.

The dosage at which the drug is administered could also be a potential obstacle; the molecule still needs to preserve a good efficacy and show some activity at low concentration. Depending on the anatomical part of the body targeted, the formulation should also be reconsidered for efficacy. These factors can alter the pharmacokinetic profile of the drug and compromise the safety of the patient.

To conclude, the molecular opportunities for drug repositioning are contrasted by practical challenges. Even if a compound is found to be active and safe for a new indication, additional factors, in particular legal issues and intellectual property, have to be considered in order to successfully bring the molecule to the market.

1.2 Drug repositioning and indication discovery: Success stories

I discussed briefly in the previous section the theoretical legitimacy of drug repositioning and its potential limitations. I will now present three success stories, exemplifying the theory and showing evidence that repositioning can happen in practice. These case-scenarios are of particular interest as they each illustrate

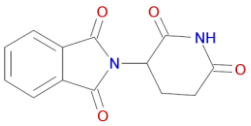
a different reason behind the repurposing. Understanding the logic backing the findings is paramount in order to build successful predictive methods later on.

1.2.1 Sildenafil: Repositioning from clinical side-effects

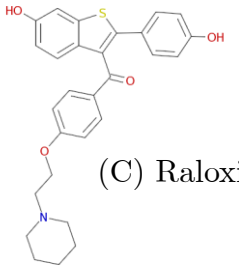
The National Health Service (NHS) defines angina as a chest pain that occurs when the blood supply to the muscles of the heart is restricted. It usually happens because the arteries supplying the heart become hardened and narrowed (NHS-Choices, 2014a). Sildenafil (see structure in Figure 1.3) was originally developed for this condition in the late 1980s. The working hypothesis was that an inhibition of the activity of the phosphodiesterase-5 (PDE5), an enzyme controlling the relaxation of the coronary arteries, should increase the blood flow and release the symptoms in the patient. Unfortunately, during the clinical trials, the drug lacked efficacy in regard to angina and development was discontinued, until patients shily started to report an unusual side-effect: prolonged erections (personal discussion with molecule's investigator). Pfizer's scientists therefore decided to investigate the drug for this indication and a worldwide study of 3700 men confirmed the effectiveness of the molecule (New-York-Times-Archives, 1998). As the pharmacokinetic profile was suitable, the drug was repurposed towards erectile dysfunction accordingly. Viagra became a blockbuster drug, with annual sales higher than 1.5 billion dollars (Renaud and Xuereb, 2002) during the first years of its release. The molecule was first-in-class for this indication and impacted the social life of millions of humans (Renaud and Xuereb, 2002). The story does not end here; the drug is currently used for pulmonary arterial hypertension too, after demonstrating a successful improvement in patients during clinical trials (Ghofrani et al., 2006).

What can be learned from this story? Firstly, the repositioning opportunities came from secondary functions of the enzyme targeted. PDE5 was known to be involved in the erection process (Krall et al., 1988), therefore the opportunity could have been logically identified rather than being discovered by pure chance. In this regard, characterising the systematic function of protein targets can therefore lead to repositioning hypotheses. Secondly, it was necessary to wait for clinical trials in order to observe the true behaviour of the drug. This observation stresses the known difficulty of moving from cell-based assays into the human body: physiology is complex and the expected outcomes are not necessarily met

(A) Sildenafil



(B) Thalidomide



(C) Raloxifene

Figure 1.3: Examples of drug successfully repurposed. (A) Sildenafil molecular structure and picture of tablets used for administration. (B) Thalidomide molecular structure. The teratogenic effect of the drug is illustrated on the picture. (C) Molecular structure of the raloxifene. Illustration are from Wikipedia, chemical structures from the ChEMBL database.

in practice. Lastly, the more is known about the compounds pharmacology, the more roles the molecule can play, as shown with the indication for the pulmonary arterial hypertension, which appeared years later. An understanding of the internal logic of biological organisms should therefore help to predict such cases.

1.2.2 Thalidomide: Repositioning a hazardous drug

The story of thalidomide illustrates how a drug can surprisingly come back from being a hazardous drug retracted from the market into a novel and unique therapeutic agent. The chemical started its life in 1957 in Europe, as a sedative, sleep-inducing agent, structurally close to barbiturates (see structure on Figure 1.3). Because of these pharmacological properties, thalidomide was marketed to treat morning sickness in pregnant women. The drug was assumed to be safe, based on an in vivo studies in rodents (Stephens and Brynner, 2009). Tragically, this was not the case for humans; the drug caused severe skeletal birth defects

in children born from women taking the drug. Over 15000 newborns were affected, suffering from anatomical malformations (see Figure 1.3). Because of this disastrous side-effect, the molecule was quickly withdrawn and triggered important reforms in the drug regulatory system (Stephens and Brynner, 2009). The story could have ended here, if it were not for an incidental discovery by Jacob Sheskin. The practitioner was trying to treat patients affected by erythema nodosum leprosum, a particularly painful inflammatory condition characterised by red nodules under the skin. An evening of 1964, an affected patient could not sleep as the pain was so intense. Sheskin decided to ultimately use some thalidomide, as the compound was known for its potent sleep-inductive properties and was available in this hospital. The drug worked and the patient was well rested in the morning. And as a general surprise, all pain and soreness disappeared overnight too. Intrigued by the idiosyncratic effect of the drug, Sheskin further studied the action of thalidomide in clinical trials (Barratt and Frail, 2012) and successfully showed that the drug can indeed treat erythema nodosum leprosum in two weeks' time in most subjects. Thalidomide found a new life and became the first and only drug approved for this indication. Just as for the sildenafil molecule, the potential of the chemical is still being unveiled: At the time of writing, thalidomide sales are reaching over 200 million dollars per year, mostly deriving from yet another off-label use for multiple myeloma, among other indications Ashburn and Thor (2004). The thalidomide story teaches us that a drug can be harmful in one patient population (pregnant women) and highly beneficial in another. Correctly identifying the molecular processes affected can help to predict adverse effects and reorient the drug accordingly. The root cause of the discovery is here again serendipity, yet a better understanding of the proteins interacting with the molecule could have been helpful to predict the opportunity (Sampaio et al., 1991).

1.2.3 Raloxifene: Expanding the application line

The last case discussed is less striking and unexpected than the two previous ones. Raloxifene (structure on Figure 1.3) is a selective estrogen receptor modulator (often abbreviated as SERM) marketed as Evista by Eli Lilly. Similarly to other estrogen modulators, such as tamoxifen, the original indication during preclinical developments was breast cancer Ashburn and Thor (2004). Despite early

studies showing the positive effect of antiestrogens on osteoporosis in rats in 1987 (Jordan et al., 1987), raloxifene's potential for this usage was not experimentally confirmed until 1994 (Black et al., 1994). Eventually, the molecule successfully passed clinical trials in 1999, with osteoporosis as a unique indication. However, the polypharmacology of the drug, particularly its action against breast cancer, was still under investigation. Finally, in 2007, the FDA approved raloxifene as a preventive agent for breast cancer in postmenopausal women (FDA, 2007), therefore extending the line of application of the drug back to its originally thought indication. In summary, raloxifene started its life as a breast cancer agent, was repositioned against osteoporosis, likely for strategic and commercial reasons, and eventually got approved for its breast cancer preventive properties. This drug is an example of smart and continuous development, expanding from one indication to another. The fundamental reasons behind the repositioning are grounded on early-stage experimental evidence and not due to a surprising effect appearing in clinical trials. The drug's polypharmacology was known and the indication of the molecule derived accordingly. Raloxifene is a good example of "educated repositioning" or indication discovery. The available information can help to appreciate and understand what the chemical might do from empirical evidence.

1.3 Computational approaches towards drug repositioning

The theory and the clinical cases presented the reality of drug repositioning. I briefly commented on the fundamental reasons enabling new usages and stressed the importance of serendipity in this process. Now the fantasy of many scientists working in the drug discovery domain is to be able to formally predict such repositioning scenarios and unveiling new pharmacology in an automated fashion. In order to reach this distant goal or at least get closer to it, several computational approaches have been developed throughout the years. This section summarises the previous work done on the topic and motivates the novel way I explored, based on a formal representation of the Mode of Action (MoA). I chose to classify the different computational approaches based on the biomedical concepts used as the centre of the methodology. Some recent reviews (Ashburn and Thor (2004), Dudley et al. (2011a) and Hurle et al. (2013)) of the field can also pro-

vide complementary information to the interested reader, as well as a different perspective and logical structure of the topic.

Abstracted to the extreme, the goal of a repositioning initiative is to establish a link between a drug and a disease. This edge represents the indication or a prescription possibility for the molecule. In order to computationally forward new indication hypotheses, it is possible to use the biomedical concepts depicted on Figure 1.4 as proxy. Usually, a similarity value is derived from the property studied (e.g. chemical structure or gene expression level), which serves as a descriptor to rank the information and predict the new indication, materialised by a new link between a drug and a disease or a molecular target. Approaches can be roughly divided into groups, named after the central property of the analysis. Some alternative methods rely on a combination of concepts and are presented at the end of this section.

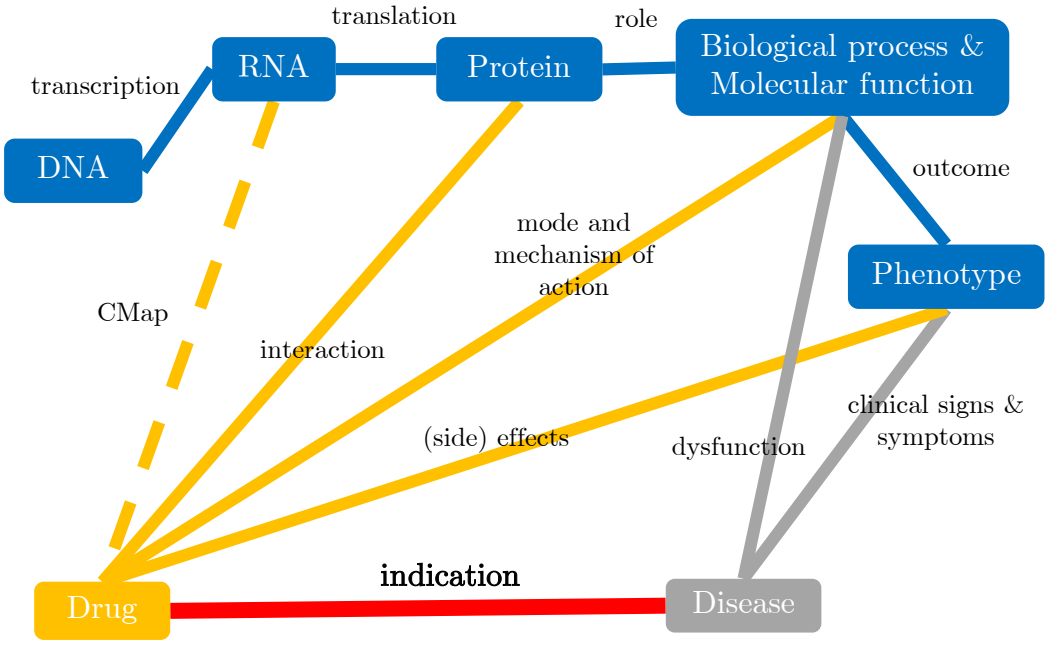


Figure 1.4: Conceptual map of the relationship between the different biomedical concepts. Relation related to the drug and its action are in orange, diseases in grey and biological concepts are in blue. Computational drug repositioning methods are based on either one or a series of such concepts in order to forward new indications for a drug, ultimate goal (red edge).

1.3.1 Chemical structure-based approaches

Traditionally speaking, orally active drugs are mostly small lipophilic molecules (Lipinski et al., 1997). It therefore intuitively makes sense to directly look at the chemical structure to compare the similarity among drugs: Similar structures are deemed to lead to similar biological outcomes. This rule of thumb goes by the name of "similar property principle" (Johnson and Maggiora, 1990) and is at the core of any quantitative structure-activity relationship (QSAR) study. A variety of methodologies exist to calculate the structural similarity between two chemicals, such as fingerprints or clustering algorithms (Eckert and Bajorath, 2007). These methods can be used to perform ligand-based virtual screenings: from a set of known active ligands, trying to find in a database of interest the structurally related molecules, supposedly bioactive too.

In the context of drug repositioning, one can search only among approved compounds for instance. This approach was successfully used by Noeske et al. (2006) implementing an unsupervised machine learning algorithm (self-organising map) in order to cluster chemicals based on their structure. Molecular scaffolds were represented as vectors and used as input for the clustering step. The authors identified cross-activities for the metabotropic glutamate receptor antagonists, on other protein targets such as the dopamine D2, histamine H1 and muscarinic acetylcholine receptors. The off-target predictions were experimentally validated in vitro and shown to be active, yet not necessarily pharmacologically relevant due to weak binding. The new knowledge on off-target binding from this study can lead the way to potential new usage for the drugs, by further modifying and optimising the molecular structure for instance.

Another interesting approach related to structural similarity for off-target identification comes from the work done by (Keiser et al., 2009). For this project, known ligands were grouped based on their known target binding partners and chemical features. The method is called "similarity ensemble approach" and "calculates whether a molecule will bind to a target based on the chemical features it shares with those of known ligands, using a statistical model to control for random similarity" (Lounkine et al. (2012) - adaptation of BLAST for chemical structures). In the case of drug repositioning, the molecules tested were only approved drugs. The results revealed a series of off-target cases from the similarity analysis. A retrospective investigation showed the validity of the approach;

then some predicted off-target bindings were experimentally validated, providing insightful clues about the pharmacological mechanism of some drugs. In some cases, such as for fabahistin, the off-target affinity (5-HT5A) was even better than for the known canonical receptor (H1), opening doors for meaningful alternative indications (see Figure 1.5).

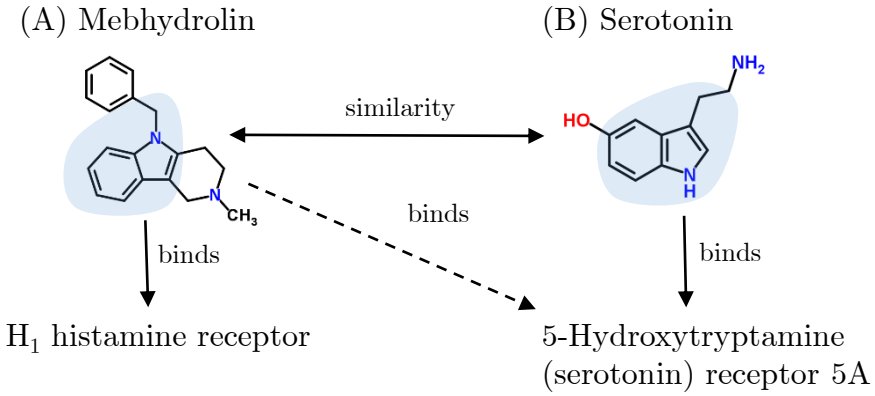


Figure 1.5: Drug repositioning using the chemical structure. Compounds with similar structures have similar biological activities (similarity principle). Molecule (A) shares some similarity with molecule (B), indicated by the blue areas. This observation leads to the conclusion that molecule (A) could be active on the canonical target of molecule (B), and indicated accordingly. See Keiser et al. (2009) for full details on example. Structures obtained from the ChemSpider database.

Chemical-based approaches are intuitive and build on the accepted "similar property principle" (see Figure 1.5 for summary). However in practice, small changes made to a molecular structure can lead to drastically different biological outcomes. Moreover, the predictions made by the various methodologies have little overlap between them, stressing the difficulty to select the adequate one for the right scenario (Eckert and Bajorath, 2007). Some compounds also undergo chemical modifications by the cell before being pharmacologically active, therefore the structure as recorded in databases can compromise the value of a predictive statistical model. Finally, by using potent and optimised molecules as starting point to infer new ligands or to train a model, there is a high risk that the predictions will manifest weak experimental pharmacology for the new indication. Indeed, such inferred compounds are not "dissimilar" enough to have an unexpected binding conformation with the protein and any differences in structure

against the endogenous ligand will only reduce potency.

1.3.2 Gene expression and functional genomics-based approaches

Living systems can be understood by looking at the behavior of their gene expression in a particular setting. Depending on the state of the system, certain genes are going to be over or under expressed, identifiable from the relative number of their messenger RNA (mRNA) molecules transcribed. Differentially expressed genes can serve as a proxy to characterise a molecular effect, so called the gene expression signature. This type of experiment is usually performed on a microarray, containing probes for the genes of interest (see Figure 1.6A). The approach provides a straightforward read of the condition studied and has been successfully used to find new indications for marketed drugs. In particular, the Connectivity Map (Lamb et al., 2006), was behind most recent repositioning stories (see Figure 1.6 for summary of the method).

The idea behind the CMap states that the action of a drug can be captured and compared by looking at the gene expression profile resulting from its administration onto a biological system. In this regard, the initiative recorded the molecular signatures of 164 FDA-approved molecules over 5 different cancer cell lines. The data is freely available and can help to perform various type analyses, for instance related to the understanding of the molecular mechanism of a drug.

Closer to my concerns, (Iorio et al., 2010) used the resource for drug repositioning, by comparing small molecules on the basis of their CMap gene expression signatures: Compounds with similar signatures were assumed to be functionally related, as they perturb the cell in a similar fashion. On this basis, the researchers identified communities of drugs sharing known protein targets and mechanism of actions. Interestingly, inside these hubs, most of the drugs were also sharing the same or similar therapeutic indications; yet outliers were present. Such cases were interpreted as repositioning opportunities, as these compounds appeared similar on the signature level (same gene expression profile), yet were clinically used for different purposes. Based on these considerations, the authors discovered that fasudil, a potent vasodilator, can also be used as an enhancer of cellular autophagy. The prediction was experimentally verified on standard cell assays. The novelty of the methodology lies in the way variable gene expression data from different

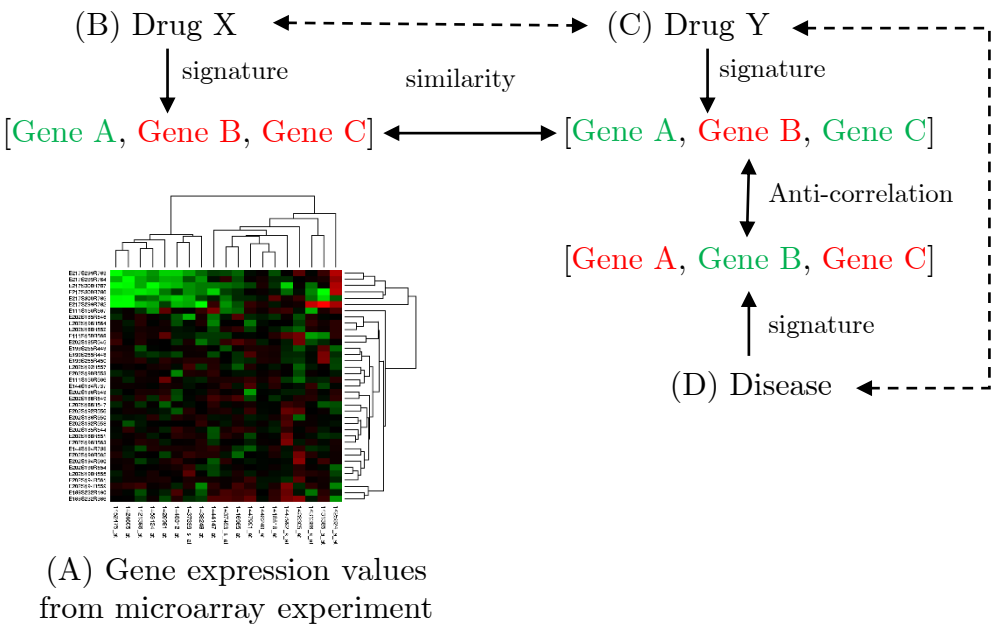


Figure 1.6: Drug repositioning using gene expression. (A) Example of result obtained from a gene expression experiment. Some of the probed genes are up-regulated (green), some of them down-regulated (red). (B) and (C) The gene expression data from the Connectivity Map provides a signature which can relate drugs on their functional aspect. For instance drug X and Y are considered similar because they share a significant amount of genes up and down related. (D) An analog reasoning can be made with the relation drug-disease: Disease signature can be treated by drugs with an anti-correlating signature.

cell lineages was integrated in order to derive meaningful metrics.

Messenger RNA expression can reflect the activity of a drug, but it can also be used to characterise disease states. Following this assumption, (Sirota et al., 2011) used a set of experiments from the Gene Expression Omnibus (GEO) in order to capture disease signatures from gene expression profiles. They further integrated this data with similarity values between drugs, derived from the CMap. The authors were able to find clusters of related diseases, appearing to negatively correlate with the signatures of the drugs currently used to treat them. The anti-correlation was used to predict the antiulcer drug cimetidine as a candidate for the treatment of lung cancer. The efficacy of the molecule for the new indication was demonstrated in vitro and in vivo on a mouse model. The analysis also identified topiramate, currently used as an anticonvulsant, as an active agent for the treatment of inflammatory bowel disease (Dudley et al., 2011b), for which no

cure currently exists. The relevance of the pharmacology in respect to the new usage was extensively confirmed in vivo in a rodent model. This study highlights a powerful feature of transcriptomics: Even when little molecular information is known about the exact underlying pathology, gene expression signatures can help to efficiently abstract away from mechanistic details and correctly identify potential treatments.

A homologous approach has been used to overcome cancer cell resistance Wei et al. (2006) in leukemia. Glucocorticoids are usually administered as treatment, yet sometimes the cancer cells of some patients develop a resistance to them. The gene expression profile of cells sensitive to the traditional treatment was compared against the signature of resistant cells, in order to characterise the molecular differences. Then using CMap data, the authors forecasted rapamycin as a novel agent to overcome the resistance, as the two gene expression profiles were correlated. The prediction was experimentally confirmed, providing further insight regarding the new mechanism of action. Rapamycin, an immunosuppressant agent indicated to prevent rejection in organ transplantations, might find here a new life for the treatment of lymphoid malignancies. A similar methodology, based again on anti-correlation of disease state against drug signature from the CMap, connected ursolic acid to skeletal muscle atrophy (Kunkel et al., 2011). The molecule improved the muscular mass and reduced adiposity in a clinical study on humans and mouse.

Gene expression analysis has led to numerous success stories, thanks to the valuable resource that is the CMap. The technique does not require much prior knowledge about the action of the drug or the pathology behind a phenotype; the creation of signatures from direct mRNA readout helps to simply retrieve unexpected drug-disease associations. Another important lesson from transcriptomics and the likely reason behind these successful results is the functional aspect: Drugs are solely characterised based on their role and action in the biological system, represented by a gene expression signature. The chemical structure is disregarded and almost irrelevant for such analyses.

Despite impressive results, the technique suffers from drawbacks, currently subject to improvements (LINCS, 2014). First, the expression profile of the drug or disease must be available. The CMap provides a relatively small list of molecules. This collection is far from being representative of all approved and experimental drugs, limiting the compounds that can be investigated. Second,

gene expression profiles can arguably define a disease state or a drug response; tissues are not considered in the CMap, the resource was only developed by recording the response of cancer cells, and is not necessarily relevant for all disease types. Finally, transcriptomics data also present considerable challenges in terms of statistical analysis, as recognised by the authors of the CMap project (Lamb et al., 2006).

1.3.3 Protein structure and molecular docking-based approaches

Most bioactive small molecules mediate their effects by interacting with proteins. This interaction can be analysed using computer software modelling the three dimensional (3D) structure of the target and the drug. This practice is known as molecular docking and commonly practiced within drug discovery pipelines; the method helps to identify and optimise binding affinities in the active site of the target (e.g. pocket of an enzyme) in order to increase the potency of the drug developed (Haupt and Schroeder, 2011). Because of the popularity of molecular docking, it is not surprising to find drug repositioning attempts based upon this approach. As most of the compounds are known to interact with more than one protein (see section 1.1.1), the aim is to identify these potential off-targets, by screening against the 3D structure of proteins present in a given database. If the predicted off-targets are disease relevant, then the drug could be repositioned accordingly.

In this respect, a series of recent studies focused on binding sites and compared their relative similarities (Haupt and Schroeder (2011) - see Figure 1.7 for a summary). Looking only at the structure of proteins active sites guarantees to be as close as possible to the biochemistry and physical reality of the interaction. From over 6000 binding site structures De Franchi et al. (2010) identified the synapsin I, protein involved in the regulation of neurotransmitter release as a new target of the drug staurosporine, known to bind the Pim-1 kinase (De Franchi et al., 2010). The finding was experimentally verified in vitro, yet the pharmacological relevance of the new target remains to be shown.

Zahler et al. (2007) performed an inverse screening (docking one compound over multiple binding sites) to characterise the off-target binding landscape of kinase inhibitors. This class of drugs, largely used in cancer therapy, has a no-

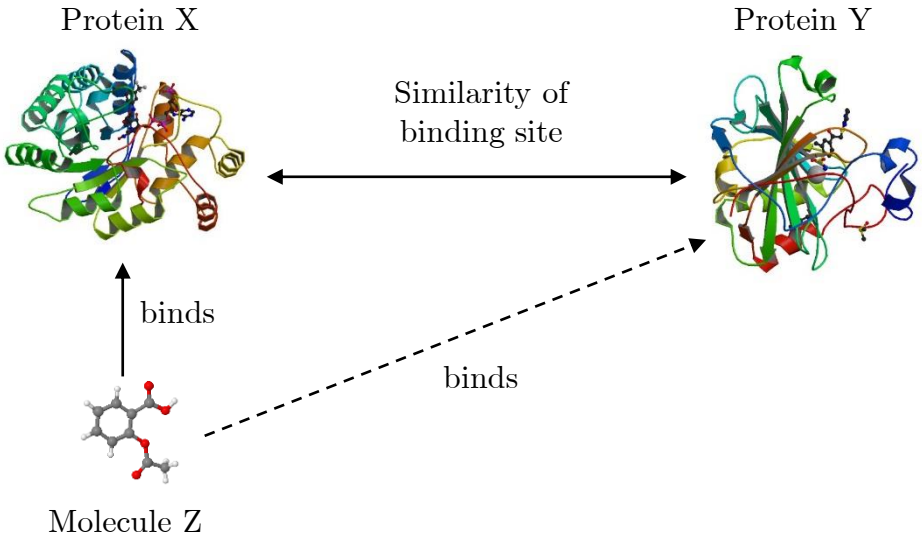


Figure 1.7: Drug repositioning using protein structure and binding site. The 3D structure of proteins and their respective binding sites can be compared using a scoring function. On this basis, it is assumed that similar binding sites can bind the same ligand. For instance, knowing that protein X has a similar binding site to the one in protein Y, and that molecule Z binds to protein X, one can forward the hypothesis stating that molecule Z should bind to protein Y too. Illustration from the Protein Data Bank.

torious promiscuous behaviour. The virtual screening revealed a new enzyme, PDK1, as an off-target of indirubin. Here again, the prediction was validated in vitro via a phenotypic cell proliferation assay demonstrating the validity of the approach and providing insights regarding kinase inhibitors side-effects.

Finally Kinnings et al. (2009) addressed drug resistant tuberculosis using molecular docking methods. The bacteria behind the condition can indeed sometimes resist the first-line drugs, and in such cases, the eradication of the pathogen becomes difficult. The authors followed a workflow called selective optimisation of side activities (SOSA), a technique developed to progressively move away from the original indication and optimise a compound across protein families (Wermuth, 2006). Briefly, the methodology is composed of the following steps: Binding site extraction from 3D structure of protein sequences, identification of similar binding sites across the proteome using a search algorithm, and finally manual docking analysis to make sure the physical interaction is possible. From this pipeline, the group predicted the approved drugs entacapone and tolcapone (prescribed for Parkinsons disease) to be potent against the enoyl-acyl carrier protein reductase,

an enzyme essential for the synthesis of fatty acid in Mycobacterium tuberculosis. The drugs were experimentally proven to be active in vitro, using commercially available tablets. Even more interestingly, the new mode of action introduced with these two compounds could bypass the drug resistance encountered in M. tuberculosis, and provide a valuable treatment for affected patients.

Despite the mentioned successes, molecular docking strategies for drug repositioning suffer from drawbacks. First, 3D structural data must be available. Databases such as the Protein Data Bank (PDB) contain numerous records, however they are still very far from covering the whole proteome (Haupt and Schroeder, 2011). Secondly, it can be challenging to automatically recognise a binding site, in particular when the protein structure was crystallized without the ligand. Finally, as all methodologies generate a large number of false positives, experimental and manual validations are the only solution to evaluate the predictions. Sometimes a single amino acid difference will totally change the pharmacology of the binding site (Kruger et al., 2012), a difficult problem to handle when the structures are analysed and aligned in an automated fashion.

In conclusion, protein-based approaches are arguably the closest methodology to the actual physical interaction between a drug and a protein target. Docking approaches provide a detailed low level picture of the biochemical complex, yet still suffer from challenges on the modelling side. The identification of off-target proteins does not necessarily always yield repositioning opportunities, and the results always have to be interpreted in a broader biological context.

1.3.4 Phenotype and side-effect-based approaches

The phenotype can be defined as the set of characteristics or traits attributed to an organism. Examples of phenotypes are the morphology, developmental, biochemical or physiological properties (Wikipedia, 2014e). This concept is widely used in biological sciences, to express the high level observations made when looking at a living organism. The phenotype is arguably the most primitive interaction between the biomedical scientist and its object of study: While travelling across the world, Darwin built the evidence for evolution from the phenotype of barnacles (Darwin and Bynum, 2009). Gregor Mendel first described inheritance based on the traits observed in pea plants (Mendel, 1866). None of these scientists had any idea about the actual molecular mechanism responsible for the

observed patterns, yet their phenotypic observations were strong enough to forward valid conclusions. This exercise is still very commonly practiced in clinical settings. Every time a doctor diagnoses a patient, he or she primarily relies on a phenotypic characterisation of the signs and symptoms present in the patient. As mentioned earlier in this chapter (section 1.1.1), phenotypic-driven screenings are also routinely performed within drug discovery pipelines. It actually appears to be the best technique to bring new medicine to market, according to a recent study (Swinney and Anthony, 2011). This successfulness can be attributed to the fact that a phenotypic observation is a more accurate representation of the underlying system; the physiological context is preserved, as opposed to target-based assays, therefore in vitro lead compounds have better chances to stay active when scaling to animal models and eventually clinical trials (Duran-Frigola et al., 2012).

Back to drug repositioning concerns, side-effects can also be seen as phenotypes. The sildenafil story emphasises their importance: No matter how potent a drug is in animal model or in in vitro assays, its true pharmacology will only appear during the clinical trials. Accurately characterising these side-effects can help to reposition a drug or reveal new interaction partners, as highlighted by two studies, summarised here (see Figure 1.8).

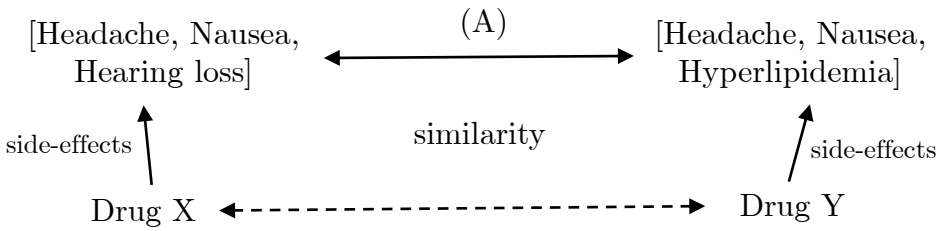


Figure 1.8: Drug repositioning using phenotype information. Knowledge about the phenotypic outcome triggered by a drug can be used in order to establish relative similarities. (A) The diagram illustrates a theoretical example using reported side-effects: The more side-effects are commonly shared by two drugs, the more similar these two drugs are. The similarity can be used to either derive potential off-targets or new indications.

Drugs with similar target binding profiles cause similar side-effects (Fliri et al.

(2005) and Fliri et al. (2007)). Starting from this rationale, (Campillos et al., 2008) defined the side-effect profiles for approved drugs, then used similarity among these to identify the drugs off-targets. The side-effects were first extracted using text-mining from package inserts (Campillos et al., 2008), in order to build a statistical model informing about the likelihood for two drugs to have a common target. The authors then focused on compounds from different therapeutic categories, yet with a high probability of sharing a target according to the model. They experimentally tested 20 of such predictions, validating 13 of them, 11 with an inhibition constant under 10 micromolars. The originality of the method demonstrates the molecular relevance of side-effects, and their potential to identify off-targets and reassign a therapeutic molecule to a new indication. An interesting aspect of this approach is the representation of side-effects. Just as any phenotypic manifestation, words or terms, derived from observations, are still the best way to express them. For this study, the authors used the Unified Medical Language System (UMLS - Bodenreider (2004)), a controlled vocabulary provided by the National Health Institute. From the experimental validation shown by the research group, one can infer that ontologies and controlled vocabularies indeed have the potential to generate sound predictions.

Another approach was presented by (Yang and Agarwal, 2011). The authors used side-effects from the database SIDER to link diseases and extract drug repositioning opportunities. Chemicals were linked to pathologies using the information available in pharmacogenomics knowledge base (PharmGKB - Whirl-Carrillo et al. (2012)). The approach leverages evidence showing that drugs used to treat similar diseases have similar side-effects. In this respect, side-effects can be indicators of a common underlying mode of action, and two drugs sharing a significant number of side-effects can be used to treat the same pathology. Sometimes a side-effect can also be deleterious in one case and beneficial in another. For instance hypotension is usually thought of as an unfavourable side-effect, yet the compounds producing such an effect could be used as antihypertensive agents. Following this hypothesis, a predictive model (Naive Bayes) was then built in order to assign drugs to diseases on the basis of side-effects. The high performance of the evaluation demonstrated the relevance of the methodology. The authors further developed a tool to predict new indications for compounds under development in a systematic fashion. No experimental validation was presented, yet a robust evaluation of the methodology was performed.

Fifty years ago, the phenotype was at the centre of drug discovery. With the advent of molecular biology, it was progressively replaced by the target-based paradigm Duran-Frigola et al. (2012). Phenotype-based approaches are still interesting for drug discovery, as they report the effect of a given substance on the level of whole organisms, which may be more representative for clinical applications. Thanks to the recent development of ontologies and methods (Hoehndorf et al. (2011b) and Hoehndorf et al. (2007)), it seems that the biomedical community has now better way to record, capture and align phenotypic information.

1.3.5 Genetic variation-based approaches

Closer to the molecular level, genetic variations can also provide valuable insights regarding drug repositioning opportunities. Due to the recent implementation of high-throughput DNA sequencing methods and analysis pipelines, it indeed becomes increasingly cheaper to sequence individuals and study their genotypes. From the information generated, one can isolate common mutations in the DNA that are significantly associated with a phenotypic trait. This method is known as genome-wide association study (GWAS) and is typically used to relate a single-nucleotide polymorphism (SNP) to a disease. The data about SNPs and their association to pathologies is indexed in databases, such as the one provided by National Human Genome Research Institute (http://www.genome.gov/gwastudies/). Using this resource, (Sanseau et al., 2012) performed an analysis to unveil potential new indications for protein targets from GWAS. The logic behind the approach is that the association between a SNP and a trait from a GWAS can be extrapolated as a relation between a gene and a disease (if the only traits considered are diseases - see Figure 1.9). Then knowing that a drug targets the given gene product, one would expect for the indication of the drug to be the same as the trait studied in the GWAS. For instance, in the gene encoding 3-hydroxy-3-methylglutaryl-CoA (HMGCR), a SNP was significantly associated with the trait LDL cholesterol (Kathiresan et al., 2008). A class of drugs, the statins, are known to target this gene product and are indicated as cholesterol lowering agents (hypercholesterolemia). The authors identified 97 of such cases, where the SNPs support the current drug indication and give more confidence to the biological role of the protein. On the contrary, for 123 associations, the authors reported a mismatch between the trait associated with the gene and the

current indication of the drug; these associations have been inferred as repositioning opportunities. For example, denosumab is a monoclonal antibody indicated for the treatment of osteoporosis and bone cancer. Its main target, the protein TNFSF11 (tumor necrosis factor superfamily, member 11) contains a SNP paired with Crohns disease (Franke et al., 2010). Based on this evidence, denosumab could be tested for this later condition. Another example reported by the authors is nepicastat, a small molecule indicated for cocaine addiction and post-traumatic stress disorder. The target of the compound, DBH (dopamine beta-hydroxylase), has been associated with the trait smoking cessation in a GWAS (Furberg et al., 2010). This result suggests a new and unreported use for nepicastat, as a drug for smokers willing to stop.

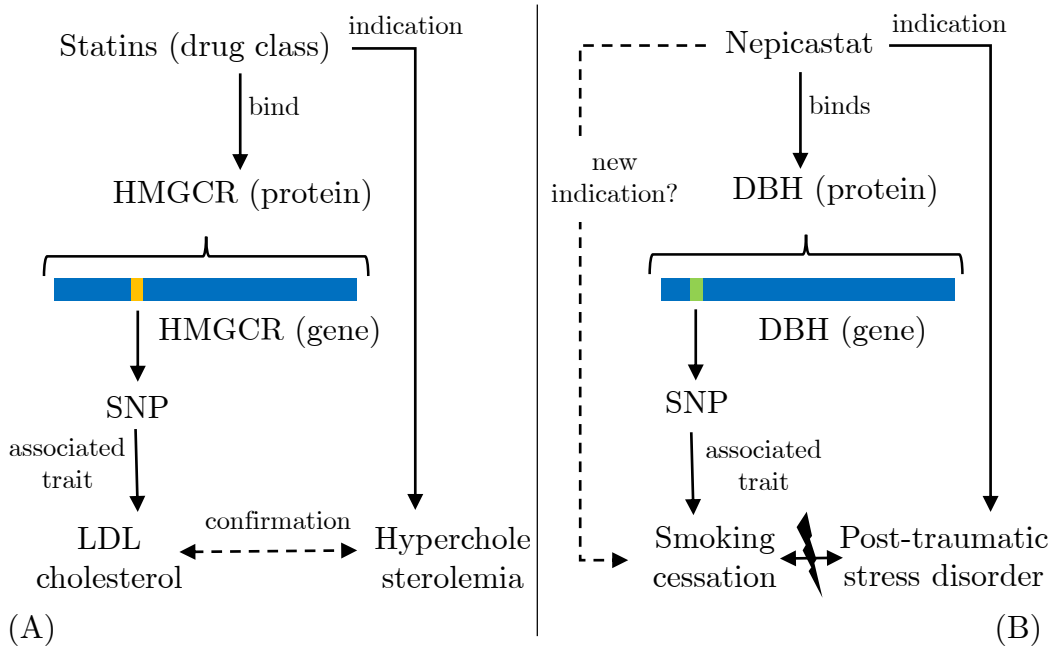


Figure 1.9: Drug repositioning using genetic information. (A) Single-nucleotide polymorphism (SNP) are associated with a phenotypic trait, here LDL cholesterol. The gene where the SNP is found (HMGCR) encodes for a protein, targeted by statins (drug class). Statins are indicated as cholesterol lowering agents, which is confirmed by the trait associated with the SNP. (B) Sometimes the trait associated with the SNP diverges from the indication of the drug, as shown on the diagram (post-traumatic stress disorder against smoking cessation). In such cases, a repositioning hypothesis can be generated. Examples are detailed in the text. See Sanseau et al. (2012) for more explanations.

The methodology presents some pitfalls, as shown by the prediction made for

NOS2 (nitric oxyde synthase 2) inhibitors to be active against psoriasis; clinical trials unfortunately failed to show any significant effects. The gene-disease relation is indeed complex in practice, and sometimes more information is needed to appreciate the potential effect of a drug. Moreover, since GWAS does not provide any information regarding the direction of the pharmacological effect and it is difficult to know for instance whether an agonist or antagonist should be used to produce an outcome. Despite these restrictions and because of the impressive recent progress made in genome sequencing, this approach, or a related methodology, might gain in importance in the coming years.

1.3.6 Disease network-based approaches

Traditionally, diseases have been grouped together, on the basis of the cause of the pathology (e.g. infection) or the biological dysfunction observed (e.g. uncontrolled cell growth) for instance. As similar diseases are treated in a similar fashion, a better characterisation of the relation holding between pathologies can generate drug repositioning hypotheses. I will briefly present some of the work done in this direction, on the construction of a diseasome or network of relationships between diseases (see Figure 1.10 for summary).

Chiang and Butte (2009) defined diseases from the list of drugs used in their therapies and off-label indications. Despite being fairly simplistic, the rationale is backed by successful examples and commonly practiced in clinical settings. The authors performed an associative indication transfer, namely, given two similar diseases, proposing to use a drug indicated only for one of them as a therapy for the other. From 700 diseases and 2000 drugs, over 150000 new associations were generated. Interestingly, the new indications are in agreement with clinical trials data, the predicted new usage has often been reported by doctors (12 fold enrichment against random). For instance, atorvastatine, a cholesterol lowering agent, was predicted to be active for asthma, Crohn's disease and myocardial infarction; all these associations have been positively reported in clinical trials, proving confidence in the methodology. For the same drug, some of the new associations have no clinical evidences, such as activity in breast cancer and osteosarcoma. Accordingly, it is possible to investigate the action of the drug for these pathologies. This work illustrates one possible approach relating diseases; two other methodologies presented now respectively construct the network from

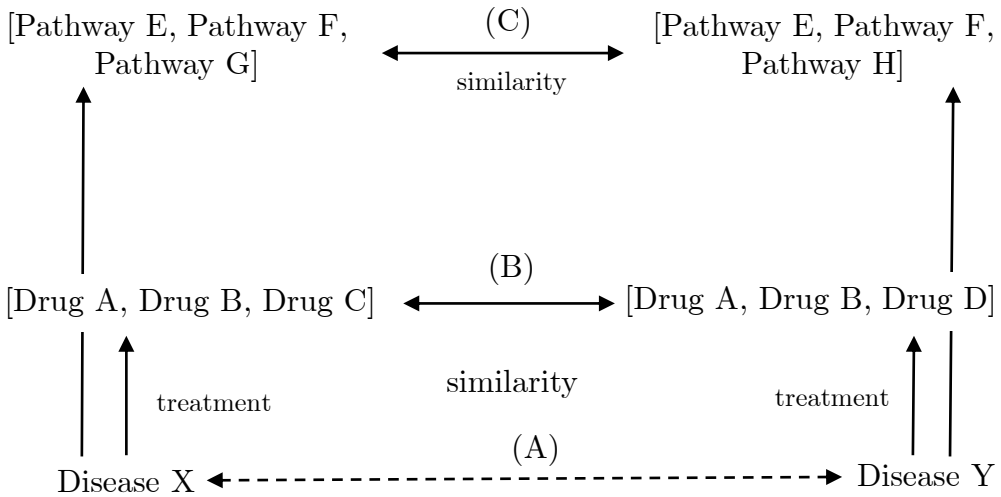


Figure 1.10: Drug repositioning using disease relationships (diseasome). The similarity between two diseases (A) can be calculated by looking at the shared drugs used for the treatment of these diseases (B) or at the commonly shared pathways (C). When applied to all diseases, one can build a diseasome or disease map, useful to relate indications and find drug repositioning opportunities.

shared pathways and functional modules. The first one (Li and Agarwal, 2009) built a map linking diseases from public resources (e.g. Reactome (Matthews et al., 2009), Kegg pathways (Goto et al., 1996) and text-mining. Diseases with commonly dysregulated pathways were deemed to be similar. The properties of the resulting graph were analysed and the authors showed how their work can provide new insights about disease relationships. No analysis for repositioning opportunities were performed, yet the map can serve as a starting point to identify similar conditions, on which one can perform an indication transfer as before. Finally, Suthram et al. (2010) constructed a disease graph from gene expression profiles and protein networks. An analysis revealed 59 functional modules, shared by half of the diseases studied. These modules relate pathologies on their molecular basis and help to better understand the internal wiring of the system. Similarly to other methods, once the disease network is created, drug repositioning hypotheses can be generated. As a conclusion, despite not directly addressing drug repositioning, disease maps can provide valuable insight regarding the usage of a drug. Such approaches also question the current way of classifying diseases, by considering molecular information as signature or definition.

1.3.7 Machine learning and concepts combination approaches

The approaches presented before mostly focus on one of the concepts of the map shown on Figure 1.4 and orient their analysis around it. It is also perfectly possible to use a combination of these biomedical descriptors to train a machine-learning algorithm and then generate predictions out of the statistical model (see Figure 1.11). Two recent studies address drug repositioning from this perspective. In both cases, first a series of biomedical heuristics is defined, then the model is trained on known data and predictions are made.

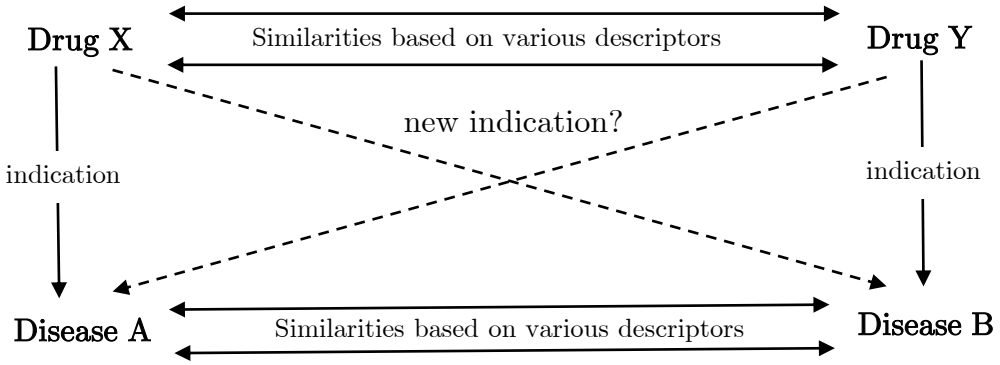


Figure 1.11: Drug repositioning using a combination of descriptors. A machine learning algorithm is trained over a series of features, such as chemical similarity, shared target proteins, etc After evaluation of the model, some repositioning predictions can be generated from the statistical learning.

The first method presented is called PREDICT (Gottlieb et al., 2011). In order to train the machine-learning algorithm, the authors decided to represent separately drug-drug and disease-disease associations. Drug-drug associations were characterised from fingerprinting of their chemical structure and the reported and predicted list of side-effects. These drug-drug associations were further enriched with information related to the targets of the drugs: the sequence similarity, distance in the protein-protein interaction network and semantic similarity of their GO annotations. The disease-disease associations were more simply characterised based on their semantic similarities calculated over the Human Phenotype Ontology (HPO) from the annotations present in the Online Mendelian Inheritance

in Man (OMIM) database. From these gold-standard associations, the authors trained a logistic regression classifier to recognise real associations from fake ones. The model was evaluated against the predictions made by other methodologies, such as guilt-by-association and CMap approaches (Lamb et al., 2006), presented earlier in this chapter. The evaluation shows little overlap between the various methodologies; it is difficult to align the various datasets, as often the diseases and drugs considered are different. Some drug repositioning predictions were then generated and evaluated from clinical trials data. Around a third of the predictions were reported as already investigated, giving confidence in the outcome of the methodology. In the last part of their work, the authors substituted the disease-disease associations based on phenotypic similarity with gene expression profiles. The idea behind this step was to test the developed method for personalised medicine: Assuming one has access to the gene expression profile of a patient, can PREDICT find the best drug to administer to the individual? Results were encouraging; the method reports high recall and specificity (area under curve of 0.92 obtained from receiver-emitter curve), providing a tangible proof-of-concept for the algorithm.

The second method presented (Napolitano et al., 2013) is very similar to PREDICT. The main difference comes from the machine-learning methodology, which was Support Vector Machine (SVM) in this study. The algorithm was used to predict therapeutic categories of the Anatomical Therapeutic Chemical Classification System (ATC), misclassifications being re-interpreted as drug repositioning hypotheses. The researchers also incorporated structural similarity, protein-protein interaction network distance and gene expression data as starting features to train the SVM. After standard machine-learning evaluation procedures, the authors derived repositioning predictions. The main hypotheses were anthelmintics compounds to be active as antineoplastic agents and antineoplastic drugs to be used as systemic antibacterials.

Machine-learning based approaches to drug repositioning provide a way to combine various descriptors into one statistical model, with the aim of increasing the accuracy of the predictions. The techniques presented however face some important pitfalls. One of them is the interpretation of the repurposing hypotheses: the statistical model is a black box, hiding the rational evidences of why a compound was chosen. Most of the hypotheses end up being obvious cases, easily explainable for a biologist by looking at the chemical structure or known

off-targets of the compounds. This result is maybe due to an over-training of the machine. Finally, one can question the biomedical meaningfulness of integrating a large number of descriptors; since diseases are subtle and unique, too much information risks blurring the outcome by overlooking important biological mechanistic details.

1.3.8 Summary

A variety of approaches have been tried in order to computationally repurpose drugs. The field is still in its infancy, as revealed by two factors.

First, it is still not clear which method provides the best results and why. The only absolute way to evaluate the predictions is when a drug will be routinely indicated in the clinic from a hypothesis generated in silico; as far as I know, there is no compelling story illustrating this yet. It is not surprising; developing a new drug is a long process, taking over twelve years to achieve and trapped with legal and economic hurdles. Knowing that the first study reported by PubMed for the keyword search computational drug repositioning dates to 2006 (An and Jones, 2006) (or 7 years before the time of writing), it seems realistic not to find any clinical examples yet.

Secondly, each method addresses the drug repositioning problem from a different angle or biomedical concept, which complicates the evaluation process. Objectively integrating the results from the various approaches is a difficult task, as the starting datasets are about different molecules and diseases and produce different type of outcomes. It would be beneficial for the community to have a standard dataset, covering the legal indications, as well as the known and confirmed alternative ones. Computational methods could use such a resource to benchmark their performance and evaluate their predictive power and perform error analysis. Immaturity also means creativity in this case, illustrated by the numerous methods implemented. Table 1.1 and 1.2 provide a summary of the approaches presented, alongside their respective strengths and weaknesses.

To conclude, computational drug repositioning appears as a topic of growing interest in the scientific community, as inferred from Figure 1.12. A series of methods have been developed in the last 5 years, summarised and discussed in this chapter. Drug repositioning is a subset of a larger problem type: Indication discovery and network biology (Hopkins, 2008). In can be summarised as taking

Biomedical concept	**Rationale**	**Biomedical advantages**	**Technical advantages**	**Biomedical pitfalls**	**Technical pitfalls**	**References**
Chemical structure	Similar chemical structures have similar biological outcomes (similar property principle)	Off-target identification, straigtforward interpretation, can be used on new chemical structures with unknown activities	Fast algorythm available to encode and search for chemical structures (fingerprint), large number of structures available	Prediction can have weak binding, not nessecary pharmacologically active, small changes made to a molecular structure can change a lot the biological outcome, compounds may undergo chemical modifications by the cell (pro-drugs), therefore original structure is not so helpful	Chemical structures information in databases is sometimes erronous	Noeske et al. (2006), Keiser et al. (2009)
Gene expression (mRNA)	A biological state can be defined by the list of genes under or over-expressed (signature) in the given state. It is possible to define disease states and drug states (Connectivity Map) and analyse their relative similarities.	Systematic characterisation of the biological function, no previous knowledge required about diseases or protein targets	Assay well understood and cheap, connectivity Map data freely available and extending	Too simplistic to characterise some states, important mechanistic aspect can be overlooked	Selection of the representative genes is challenging, connectivity Map data were recorded on cancer cell, might not feat all types of diseases or it can bias the signature	Iorio et al. (2010), Sirota et al. (2011), Dudley et al. (2011b), Wei et al. (2006), Kunkel et al. (2011), Lamb et al. (2006)
Protein	Computational modelling of the physical binding of a drug to a protein	Off-target identification, straigtforward interpretation, can be used with chemical structures with unknown activities, close to biochemical reality	Numerous tools available to perform docking studies	Predicting a binding is challenging because of molecular dynamics, high number of false positive predictions	Structural databases are not complete, not all proteins are considered	Ashburner et al. (2000), Haupt and Schroeder (2011), De Franchi et al. (2010), Zahler et al. (2007), Kinnings et al. (2009)

Table 1.1: Summary of drug repositioning approaches.

Biomedical concept	**Rationale**	**Biomedical advantages**	**Technical advantages**	**Biomedical pitfalls**	**Technical pitfalls**	**References**
Genetic	Genomic identification of phenotypic traits	Potentially closer to individual patients (SNP)	Large amount of SNP information already available and growing everyday, decreasing cost of sequencing	Does not provide mechanistic details nor context	Genome data analysis is still challenging	Sanseau et al. (2012)
Disease	Similar diseases receive similar treatment	Systemic characterisation and classification	Various methodologies can build the diseases map	Does not directly address drug repositioning	Highly rely on curated knowledge	Chiang and Butte (2009), Li and Agarwal (2009), Suthram et al. (2010)
Combination	Training of a machine-learning algorithm from a series of descriptors	Incorporation of multiple dimensions, providing a more consistent representation of the biological phenomenon	Numerous machine-learning program and methodologies available	Biological interpretation dificult (black box approach), risk of over training the system, does not provide mechanistic details	Choice of the heuristics is challenging	Gottlieb et al. (2011), Napolitano et al. (2013)
Biological process and molecular function	Formal representation of the mode and mecanism of action	Systemic characterisation of the biological role, mechanistic description	Large amount of information available, data integration	Predictions can have weak binding, not nessecary pharmacologically active	Highly rely on curated knowledge	N.A.

Table 1.2: Summary of drug repositioning approaches (continued).

advantage of our increasing knowledge about systemic behaviour to computationally design smart drugs. The conclusions drawn from this state-of-the-art review oriented my work and motivated me to explore drug repositioning from the perspective of biological processes and molecular functions.

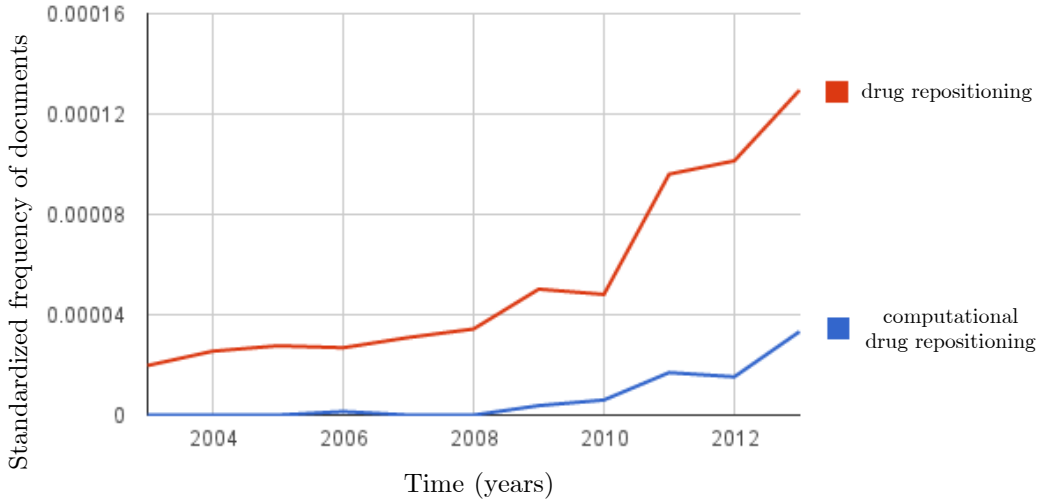


Figure 1.12: Evolution trend of the documents related to drug repositioning. Standardised frequency: Number of documents indexed on PubMed for a search divided by the total number of articles published the same year. The higher, the more popular a topic is. The frequency increases with the time for both searches, showing a growing interest in the domain.

1.4 Thesis: Biological process and molecular function for drug repositioning

The previous section (1.3) gave an overview of the various computational techniques developed for drug repositioning within the last decade. I believe that the most promising results appear to come from gene expression analyses (section 1.3.2), driven by the large amount of public data available and encouraging in vivo results. Additionally, gene expression provides a functional insight about a biological state, valuable in practice: From a drug indication perspective, it is indeed much more pertinent to know what a molecule does (role or function) rather than what it physically looks like (structure).

Following this philosophy, I argue in favour of the mode of action, as a means to formally define the function of a drug. The concept brings unique assets, such

as bridging from the molecular to the phenotype level, as well as providing a discrete mechanistic understanding of the role of a drug.

In this section, I will introduce my thesis, namely the vision and the rationale behind the approach of using curated knowledge of biological processes and molecular functions to characterise the drug repositioning landscape. Processes and functions provide a flexible abstraction layer in between biochemistry and systems biology. My thesis work is based on a computational characterisation of the mode of action of marketed drugs. This notion is particularly relevant to drug repositioning, as the concept provides mechanistic insights regarding the broad pharmacological action of a drug. Moreover, a compound can also have several modes of action, representing the various biological processes the molecule can perturb from its polypharmacology.

1.4.1 Rationale

Knowing the potential role a drug can play in a living organism such as a human body enables one to logically re-use the compound for a different indication. Moreover, a drug binds to multiple proteins, themselves involved into multiple biological processes (see section 1.1.1). Therefore a drug can potentially play a multitude of roles, or in other words can have several MoAs, which are accountable for its polypharmacology.

In practice, the biological roles of drugs are described as mechanism or mode of actions. The mechanism of action can be defined as the biochemical interaction that gives rise to the pharmacological effect of a drug. For instance, the term phosphodiesterase-5 binding is the mechanism through which the sildenafil molecule produces its action. Similarly, the mode of action (MoA) defines in a more abstract fashion the broad activity of the molecule on an organism. The terms pro-penile erection agent and vasodilator both are valid modes of action (MoA) of the same sildenafil molecule. Another example of MoA is anti-blood coagulant, representing the capacity of a drug to inhibit or decrease the extent of blood coagulation (see Figure 1.13 for example).

The concepts of mode and mechanism of action are broadly used in drug discovery; they help to classify drugs into therapeutic groups. Even more importantly, a chemical is assigned as a treatment to a disease because it exhibits a particular MoA. One example is the case of high blood pressure, a common medical

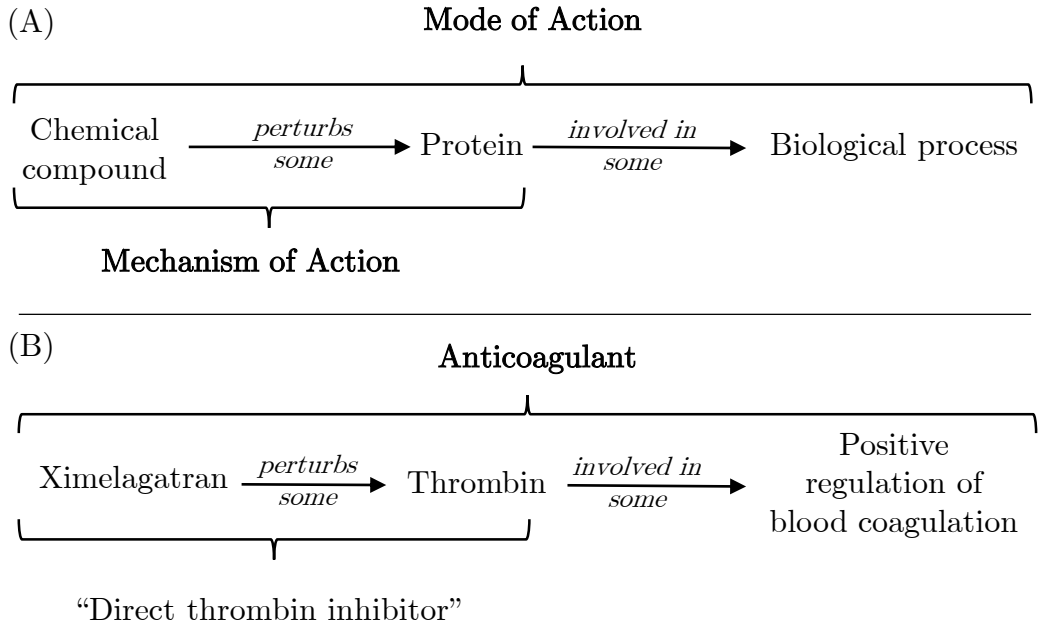


Figure 1.13: Schematic representation of the concept of mechanism and mode of action. (A) The mechanism of action can be defined as the physical activity of the ligand on a protein target. The mode of action characterises the pharmacology of the small molecule in the context of the organism. (B) Examples related to blood coagulation illustrating the usage of the mechanism and mode of action concepts.

condition leading to an increased risk of heart attacks and strokes (NHS-Choices, 2014b). In order to chemically decrease the blood pressure, several biological solutions can be considered. A first approach could remove the excess of salt from the body, thereby decreasing the tension in the blood vessels. An alternative solution could inhibit the vasoconstrictive signalling of a hormone. Finally, it is also possible to act directly on the cells physically narrowing the vessels and preventing their unwanted action this way (Ong et al., 2007). Now if a chemical was known to be capable of producing any of these biological actions, it would then be possible to indicate it for the treatment of the hypertension. This logic puts emphasis on the mechanistic details of the pharmacological effect. The MoA explicitly characterises the function of the drug in the cellular machinery, and helps to appreciate its overall effect. As opposed to statistical methods and black box techniques, the explicit reasons why a particular drug shows a particular outcome can be derived from curated knowledge in a systematic fashion.

A MoA-based approach comes with unique theoretical advantages. First, it

provides a systematic high-resolution picture of the function of drugs, just like gene expression experiments. Moreover, this characterisation of the function is discrete and granular. Unlike gene expression data (CMap), the function of the molecule is not vaguely summarised in a gene expression signature. MoAs are discrete categories, and a compound can be put in many of them, as illustrated with hypertension. In this respect, the MoA can be seen as the equivalent of a proteins functional annotations (Ashburner et al., 2000), but for drugs. Finally, the MoA provides a unique level of abstraction over biological systems: It logically links the biochemical mechanistic interaction, as studied with docking and chemical similarity, to a high level phenotypic process, such as a disease or a side-effect.

1.4.2 Towards the specification, implementation and analysis

The computational use of the MoA to repurpose drugs comes with obstacles. I gave in the previous section some informal examples of MoAs: vasodilator, anti-blood coagulant, anti-ageing agent, etc. However, in order to provide helpful information, all MoAs (or a large number at least) have first to be unambiguously and formally defined. From these examples and Figure 1.13, the reader can see that MoAs are terms, boxes or categories; classical mathematical frameworks used in biology and bioinformatics, such as statistics, do not provide any means to address this concern.

A MoA describes logical events happening inside the complex cellular machinery: Intuitively, an anti-blood coagulation agent is a type of compound capable of modifying the activity of a protein target somehow involved in the blood coagulation. Therefore representing the axioms or logical links underlying these statements can set the basis to derive new MoA categories, as shown in Figure 1.13. However, the definition of MoAs from this perspective will necessarily rely on existing knowledge or data. From the example before, it is indeed necessary yet sufficient to know first that a drug binds a particular protein and secondly that this protein is involved in the coagulation process in order to formally derive the MoA of the compound.

The implementation of the rationale is described in the rest of this document and organised as follow.

1.4.2.1 Chapter 2 - Description logics and biomedical knowledge (Specification)

Before representing modes of action and using them to classify drugs, the specifications of the system have first to be clearly determined. How are the logical connections between biological processes and molecules going to be represented? Will the system scale and work even with large biomedical input size? What is meant by biological knowledge? How does it fit currently available biomedical data? These questions will be addressed in Chapter 2, introducing description logics as a mathematical framework of choice to perform the MoA implementation later on. I propose an analogy considering living organisms as machines to describe a portion of the biomedical knowledge, relevant to drug discovery.

1.4.2.2 Chapter 3 - The Functional Therapeutic Chemical Classification System (Implementation)

The specification finds an implementation in Chapter 3, with the Functional Therapeutic Chemical Classification System (FTC). The classification features a representation of over 20000 modes and mechanisms of action categories, inside which approved drugs have been classified. The work done was evaluated against existing solutions and is publicly available via a web application. As expected, on average drugs are present in numerous MoA categories. This observation can be used to perform different types of analysis and generate drug repositioning hypotheses.

1.4.2.3 Chapter 4 - Systematic drug repositioning analysis

The content of the FTC is analysed in Chapter 4 from different perspectives. First, the relationship between the concepts of a drugs structure, function and indication is discussed in this chapter. Secondly a list of drug repositioning hypotheses is derived from this preliminary characterisation, covering a wide range of therapeutic areas. The hypotheses are used to analyse in a systematic fashion the drugs off-label uses. Finally hypertension and Alzheimer's disease have been selected as use-cases in order to further investigate drug repositioning opportunities.

1.4.2.4 Chapter 5 - Outlook

The outcomes of the thesis work are put into perspective: What part of the drug discovery process has been covered? What are the next logical steps for future work? Can the mode of action representation be further improved and how? The pitfalls encountered during the work are discussed and I present my vision towards a simpler knowledge representation system for the biomedical domain. Chapter 5 also discusses alternative analyses that can be performed with the content of the FTC, in particular against gene expression data.

chapter 2

Description logics and biomedical knowledge (Specification)

Key points

- Organisms can be compared to complex black box machines, namely devices with a hidden and unknown internal functioning.
- This analogy helps to define the study, needs and representation of biomedical knowledge.
- Description logics (DLs), part of a mathematical framework developed to formalise concepts, can be used to study the molecular black box machine and query over recorded biological knowledge. DLs provide the means to represent terms and logically link biological modules.
- The framework presented integrates with current life-science information, such as biomedical ontologies (OBO) and databases, and is theoretically highly scalable ($\mathcal{EL}^{++}$ profile).
- The notions introduced and discussed set the groundwork for the representation of the mode of action and the implementation of a new resource built on these principles: The Functional Therapeutic Chemical Classification System (FTC).

Authors comment

This chapter is a summary and record of my experience with the Web Ontology Language (OWL2) and knowledge representation for the biomedical domain. I present a theoretical analysis of the representation of the mode of action and its scalable implementation. A software library to assist the development of programmatic solutions (Brain) is briefly presented in this chapter. I invite the reader to refer to the original publication (Croset et al., 2013) for more detail if wanted.

2.1 Introduction

Science (from Latin scientia, meaning "knowledge") is a systematic enterprise that builds and organises knowledge in the form of testable explanations and predictions about the universe (Wikipedia, 2014h). More specifically, biomedical sciences handle the subset of knowledge related to living organisms. Understanding the biological world is relevant to society as it provides, among other things, some valuable insight to treat and cure diseases. In order for our knowledge to grow, discoveries and evidence have to be recorded, structured and shared with the community and society. Traditionally, biological knowledge is preserved in a narrative fashion, inside textual documents, such as a journal article for example. However, natural language is ambiguous, informal, and impairs an efficient reuse of the information. More recently, with the advent of computer systems, some biological knowledge is stored in a structured way inside databases or ontologies (Brooksbank et al., 2014). Biological entities and concepts have identifiers, enhancing the dissemination and reuse of the information, as well as enabling an efficient integration of multiple datasets. Despite the improvements made towards a more meaningful and consistent representation, I argue that the logic in biologic is still under-represented. Indeed, it would be valuable to formally derive and prove new facts from existing ones, the same way it is done in algebra for instance. This chapter presents an original approach towards this goal, using description logics (DLs) as a mathematical framework.

In this regard, I first illustrate how the study of organisms is analogous to the theoretical study of a black box machine. From this simple fundamental model

and its requirements, I then explain how DLs can, to some extent, capture the internal logic of the cellular machinery. Finally I discuss how this approach can be combined with existing biomedical data, in order to implement automated and scalable solutions. This theoretical work serves as the basis to formally define the concepts of mode and mechanism of action, central points in deriving drug repurposing hypotheses.

2.2 Biomedical knowledge

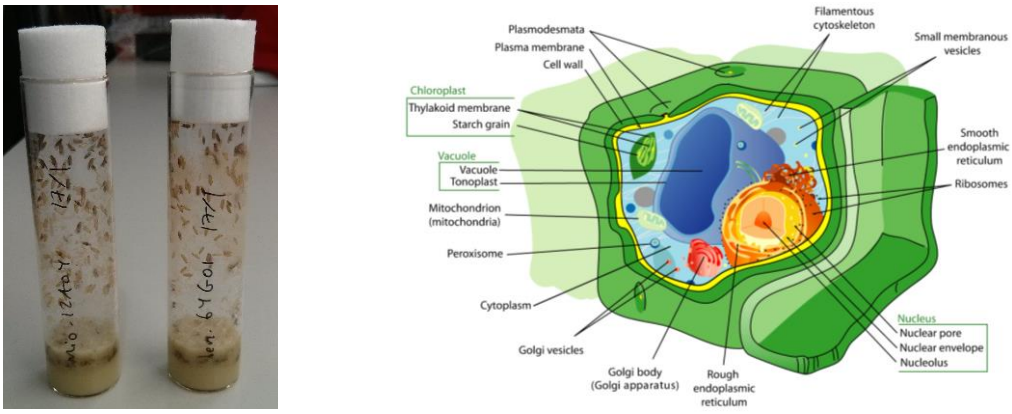
Life sciences addresses the study of living organisms. Just as in any other scientific discipline, biological researchers first collect data and evidence, which will then be turned into knowledge based on human interpretation (Antezana et al., 2009). In order to be efficiently reused, understood and shared with the scientific community, some of this knowledge can be represented by a process known as formalisation.

2.2.1 Contemporary formalism in biomedical sciences

Formalism can be described as an abstraction process by the human brain in order to model a system in mathematical terms. Formalising a problem helps to better analyse its complexity. Mathematical models (e.g, equations) are powerful, as they can eventually predict outcomes, once the system is understood well enough. Different scientific disciplines employ different formalisms, depending on the problem studied and research questions being asked.

2.2.1.1 Formalism varies among natural sciences

When I was a college student, I always found biology to differ from other natural sciences such as chemistry or physics for instance. The latter disciplines were built around a particularly strong mathematical scaffold; fundamental phenomena, like the speed of a body, were concisely captured by a function such as $v = d/t$. Chemical reactions were simplified in the form of an equation such as $H_2O = HO^- + H^+$ for instance. Despite being approximations of natural processes, these models and equations were useful to learn the discipline. Moreover, once the core concepts were understood, some supplementary layers of complexity were logically added on the top of the known things: Static representations of chemical reactions



(A) (B)

Figure 2.1: Examples of capture of information in biology. (A) A flask containing hundreds of individuals of the species drosophila melanogaster. The population is identified with a label on the container. (B) Schema of a canonical plant cell, similar to the one found in text-books. Parts of interest are annotated with terms, following a descriptive approach. Pictures from Wikipedia and courtesy of Sarah H Carl.

were transformed into dynamic ones, or it was possible to derive the trajectory of a projectile from its speed and direction, for example.

The formal representation of a system is important, as it clarifies the meaning of the concepts of interest for the community. For instance, speed has a very explicit and clear definition captured by its equation, which can be unambiguously understood and interpreted accordingly. Finally, the most important feature of a formal system is, I believe, that it enables predictions to be made. It becomes possible to infer behaviours and results on the sole basis of the theory. Complex systems can be built and fully understood, relying solely on fundamental principles.

The study of biomedical sciences was not guided by such strong mathematics. Core concepts such as evolution or gene were mostly described in a textual way, a sentence usually defining the meaning; it was impossible to combine concepts in order to create new ones. As biological organisms obey the law of physics and chemistry, one would therefore assume that these frameworks could assist in the study of the living world on their own. They do to some extent; it is for instance possible to represent and model the series of chemical reactions related to a biological pathway using standard chemistry (Le Novere et al., 2006). However, biomedical sciences present particular challenges that cannot adequately be

solved by the traditional molecular formalism. First of all, biological bodies are extremely complex from a chemical perspective, either in terms of size or types of compound present (3 x 10^{27} atoms in the adult body and a minimum of 25 atom types) (Nielsen, 1999). Secondly, some high level phenomena have an unknown molecular basis, and it becomes therefore impossible to capture formally the system while solely relying on chemistry. Phenotypes and diseases are good examples of this category.

Because of these obstacles among other things, biomedical sciences are traditionally less formalised than their counterparts, the biological knowledge being often captured by a textual description inside a document, or sometimes with the help of a conceptual schema (Lazebnik, 2002). Figure 2.1 and 2.2 present examples of media used to convey, record and communicate life sciences knowledge.

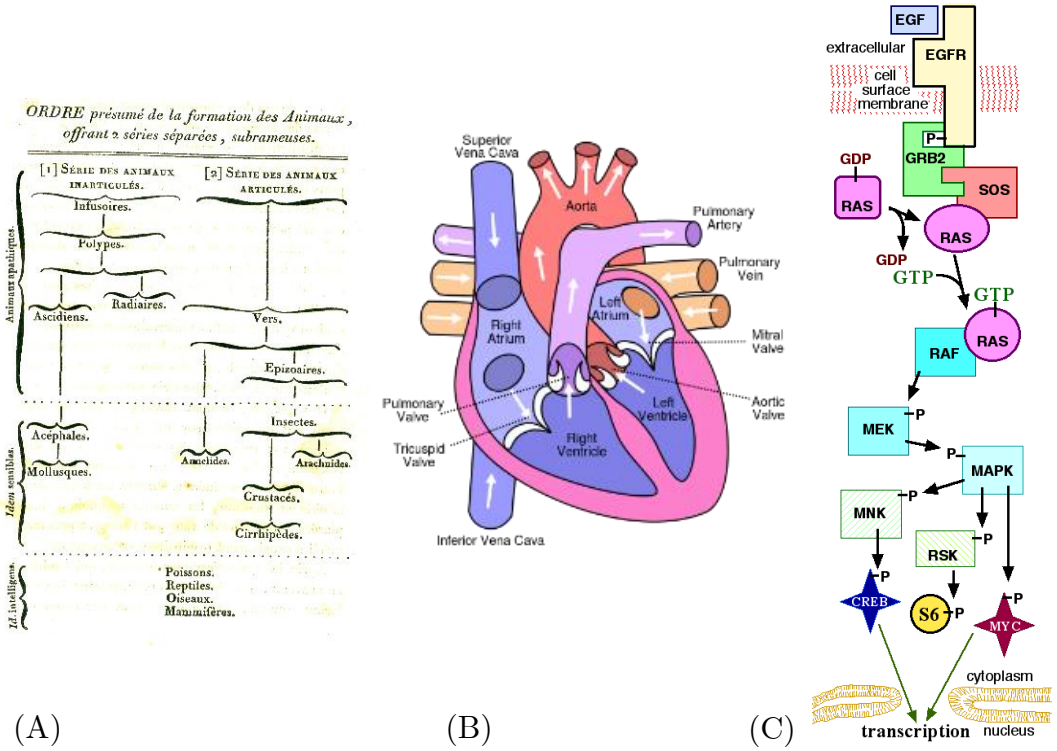


Figure 2.2: Formalisation in biomedical sciences. (A) Diagram showing the evolutionary taxonomy of invertebrates. Drawn by Jean-Baptiste Lamarcks in 1815. (B) Mechanistic illustration of the cardiovascular system. Arrows illustrate the flow of blood and the logical connection between the annotated parts. (C) MAPK/ERKsignaling pathway. Schema of the cascade of molecular events leading to activation of transcription factors. The logic of the system is informally captured using arrows, colours and shapes. Images from Wikipedia.

Nowadays, with the advent of computers and the Internet, known biological entities and concepts are further stored inside public databases. Often, a manual curation step over the published literature improves the consistency and correctness of the data (Brooksbank et al., 2014). The structured information makes it easier for the community to retrieve the data and to perform statistical analyses on it. Nonetheless, this framework is not fully formalised; for example it is not possible to mathematically prove why a drug could be useful for a disease, such as one can logically derive with a series of steps the value of the variable x out of the following equation: $2x = 4 + x$.

How can one further formalise biomedical knowledge to assist the development of new medicines and the study of the living world? A naive approach would try to simplify the system life scientists are working with into a meaningful analogy (Lazebnik, 2002). In this regard, I will present how the study of a living organism can be compared to the study of a black box machine, namely a device for which nothing about the internal workings is known (see Figure 2.3).

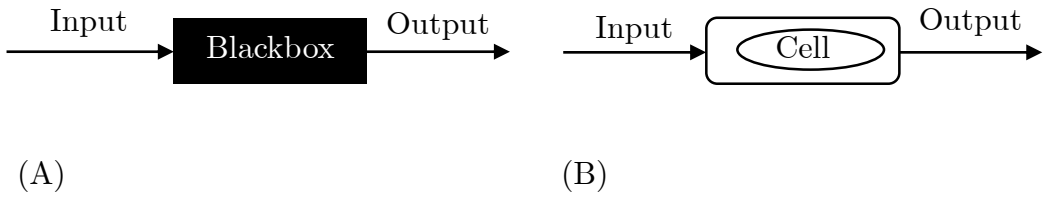


Figure 2.3: Blackbox model. (A) Schematic representation of a blackbox. Given an input, an observable output is produced. The internal workings are supposedly not understood. (B) Cells or organism are blackboxes: They can carry various and observable functions from an input, yet the internal workings are not necessarily fully understood.

2.2.1.2 Organisms as complex machines

From the perspective of drug discovery, an organism (related to organisation) can be broadly simplified as an assembly of molecules functioning as a stable whole. This characteristic makes organisms similar to a machine in the generic sense, which can be described as an assembly of parts functioning to meet a particular goal; in the case of the organism, the intrinsic goal is survival or reproduction.

Based on these definitions, biomedical sciences can be seen as the sciences of preventing and repairing dysfunctioning organisms - or molecular machineries.

According to this analogy, the study of a living organism can therefore be compared to the study of a complex black box machine, composed of a large number of physical parts carrying a certain number of internal functions and acting together in an organised fashion.

Now assuming that a certain community has access to such a machine and wants to study it for various reasons, how can it theoretically be done? The analogy becomes insightful at this stage, as there are plenty of complex man-made devices the reader can relate to, and which will serve to illustrate the thought process. I will consider an airplane as example, because it is a complicated device with a straightforward goal: safely flying in the air. A similar exercise can be done with a radio (Lazebnik, 2002).

Supposing that a fully functional airplane was found somewhere and the only thing known about it was that the device is capable of flying. The aim is to understand as accurately as possible how the machine works, in order to have enough knowledge to be able to fix it in case it gets broken or malfunctions in the future. In order to address this problem, I argue in favour of a straightforward descriptive approach, in line with the way biological sciences are performed, namely describing the device in as much detail as possible.

The first task done would be to schematise the device, as shown in Figure 2.4-A. Then the fundamental physical parts would get annotated with arbitrary names (2.4-B). With an increased understanding of the physical modules composing the airplane, it would then be possible to discover the roles played by the various parts of the machinery. Objects would receive functional annotations, namely an explanation of what they do in the overall flying process, as shown on Figure 2.4-C. Up to this step, the system would be characterised as a collection of discrete physical and functional modules, each isolated from one another. Finally, in order to appreciate the machine as a whole, it would become mandatory to link the modules based on their relation types. Figure 2.4-D shows the high level logical organisation of the machine, which integrates the parts to understand the overall process.

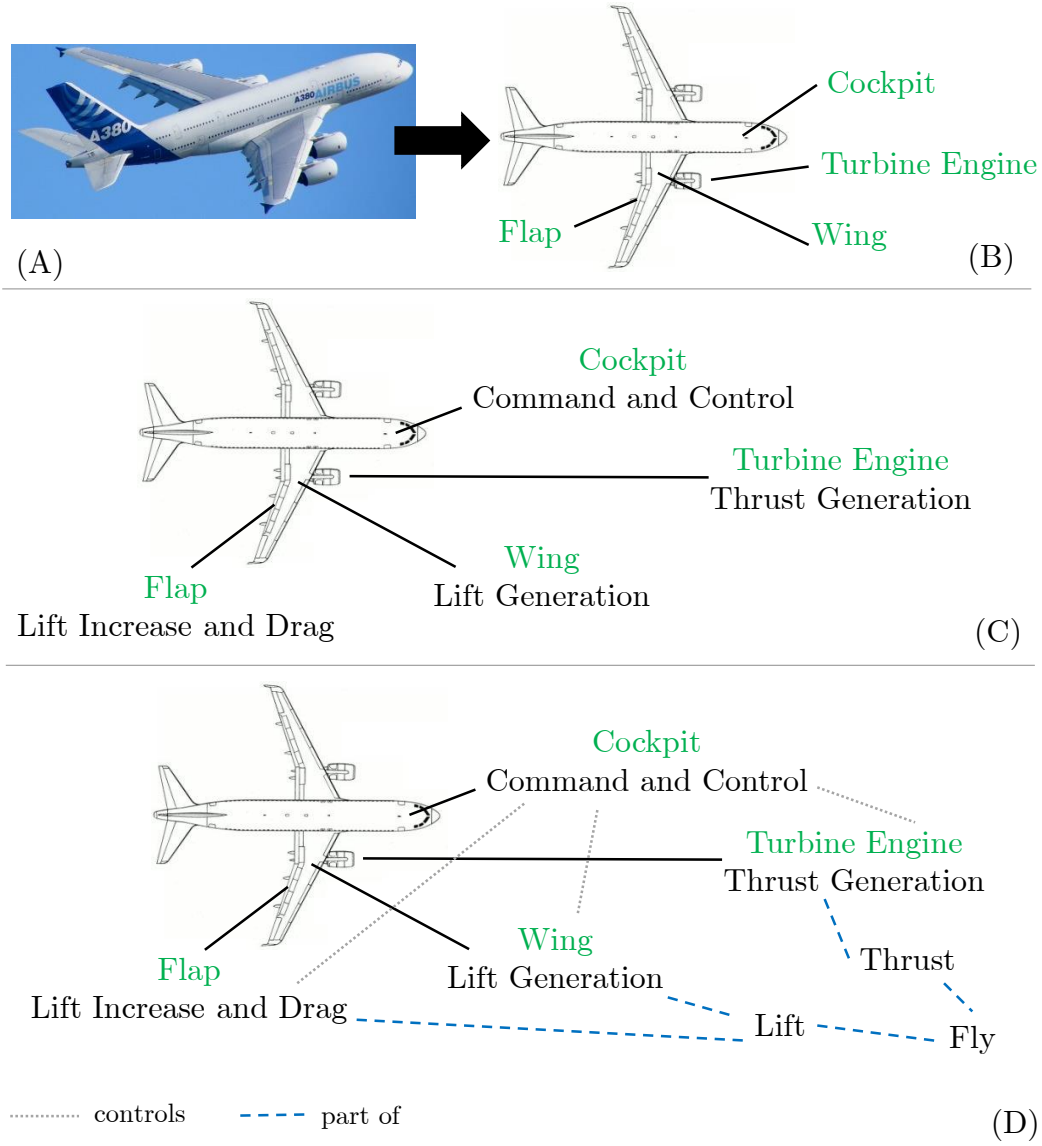


Figure 2.4: Caption.

Unveiling the internal logic of the black box machine using this descriptive methodology would later allow extensive querying of the model of the system and simulation of its behaviour. For instance, assuming a problem was identified with the lift of an airplane, preventing it from flying. From the model, it would be possible to retrieve all the known parts directly or indirectly involved in the lift process, derive a list of potentially faulty components and suggest ideas on how to logically repair it and restore its ability to fulfil its primary goal, flying.

The black box machine analogy allows one to better understand the require-

ments of a formal descriptive framework in order to capture biomedical knowledge. From a drug discovery point of view, organisms can be fundamentally reduced to machines. The descriptive process of studying such machines I presented is analogous to the approach researchers use to study organisms, understand diseases and find treatments for them.

In order to move from an informal characterisation into a defined framework, it is first necessary to determine what is required for a descriptive approach to successfully capture biology. The next section will identify some of the concrete needs for biomedical knowledge, derived from the theoretical model of the molecular black box machine and its study.

2.2.2 Requirements for biomedical knowledge formalisation

In order to be efficiently reused and shared, biomedical knowledge has to be formalised. One way to investigate the living world consists of considering organisms as complex machines; descriptions and annotations of the parts of the machine then help to represent the knowledge and understand better the functioning of the whole device. However, in order to be meaningful, this descriptive approach needs to fulfil a series of criteria introduced in the following section.

2.2.2.1 Mathematical framework

Extending a mathematical field is a key feature of any formal framework. Mathematics help to accurately formulate problems and provide the generic means to solve them. In the case of biomedical knowledge, the ultimate aim is to reduce the burden of diseases for society. To achieve this goal, a formal framework must first be able to capture biological information. Secondly and more importantly, it should be possible to further exploit the framework to deduce and prove assumptions over it. For instance, the mathematical branch of geometry handles questions related to shapes and space. Even if this framework provides only a simplified view of reality, geometry provides the formal means to attest to the validity of a buildings blueprints or optimise land exploitation. Similarly, in the biomedical domain, one should expect to be able to deduce implicit facts or formulate new hypotheses regarding potential treatments directly out of the mathematical

formalisation, which is currently not the case. Connecting descriptive biomedical knowledge with mathematics also insures that the framework can benefit from the latest progress in the field. It helps different communities to work together on the same issues from different angles. Formulating biomedical knowledge in mathematical terms also opens the door for computer sciences to assist later on with the implementation of a digital solution.

2.2.2.2 Definitions

The annotation of the parts of the machine requires the usage of a new and specific vocabulary. Words and concepts can be ambiguous, especially when the system analysed is complex such as a living body. A recent illustration of the importance of definitions in biology is the debate over the ENCODE project conclusions (Editorial, 2013); different scientists have different interpretations of the word function, which shift the explanation of the results. Semantics, the investigation of the meaning of symbols and words, can assist in this task and help to specify the intended meaning of a concept. Moreover, a formal framework for biomedical knowledge should be able to define any type of things: real-life objects, such as molecule, protein or cell for example. Abstract biological processes like blood coagulation or diseases such as cancer should also be part of the framework, as they are essential concepts in biomedicine. Finally, in order to appreciate the logic of the machine as a whole, it is mandatory to be able to link concepts and words, in order to show how parts of the machinery interact together. The meaning of such relations should be explicit and unambiguous, just as the definitions of concepts.

2.2.2.3 Hierarchies and abstraction

In practice, organisms are probably more analogous to Rube-Goldberg machines than airplanes (see Figure 2.5), with an internal logic sometimes difficult to understand on its own; this results in practice in a tangled network of chemical wiring, which can be abstracted and simplified into functional modules (Hartwell et al. (1999), Ravasz et al. (2002), Machado et al. (2011), Fisher and Henzinger (2007)). Biomedical knowledge has to deal with entities ranging from chemical drugs to high level concepts such as species or biological processes. All these layers have to be integrated and linked in order to understand the machine as a

whole. In this regard, it is critical for the formal framework to support abstraction and enable the representation of hierarchical information.

Taxonomies have always been at the heart of biological sciences; take for instance the work of Carl Linnaeus and the Systema Naturae (von Linné and Lange, 1770). Classifications are further used to organize species, protein and chemical families, to name a few examples. Historically speaking, categorical information has provided a good and intuitive framework to capture biomedical knowledge; therefore, any attempt for further formalisation must be able to handle this type of data, as well as to leverage its use.

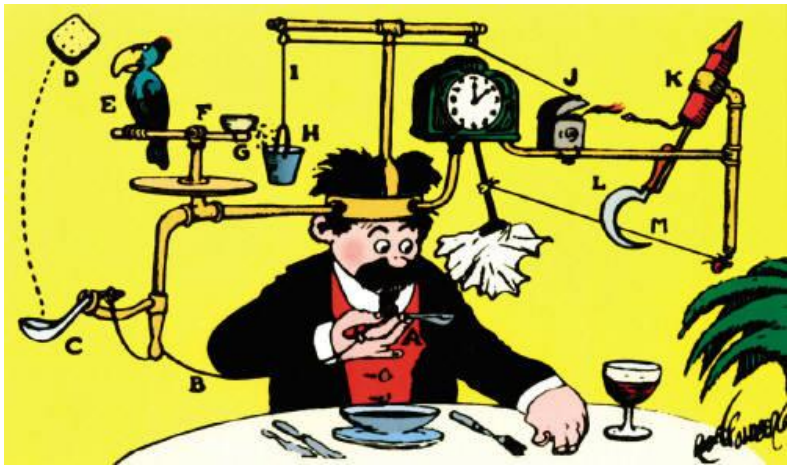


Figure 2.5: Rube Goldberg machine: Over-engineeredmachine that performs a very simple task in a very complex fashion, usually including achain reaction (Wikipedia, 2014g). The picture shows the "Self-Operating Napkin. When the spoon soup is raised, a cascade of events are triggered ending as the napkin coming toward the mans face. The task performed is relatively trivial, yet many steps are needed to execute it. Organisms are assumed to be analog to Rube Goldberg machines because of evolution; the internal wiring is not necessarily straightforward and progressively evolved and changed (Ravasz et al., 2002). Illustration from Wikipedia.

2.2.2.4 Distributed and scalable

Studying a system as elaborate as an organism requires the collaboration of a large number of persons, working in parallel on different facets of the problem. All these individuals must be able to access and share their knowledge, in order to communicate and be aware of the latest progress. Disseminating information has been usually performed via printed literature, which is being replaced at the time

of writing by the World Wide Web and the Internet. This shift of infrastructure allows the processing of more and more information in a digital context, where computers can perform an increasing amount of the work in an automated fashion. This characteristic is particularly interesting for biomedical information, as it is possible to use computers as part of the formalisation process. However, this approach leads to challenges, in particular scaling issues. As organisms are of high complexity, it is therefore important to opt for a formal solution able to cope with problems of large input size. It is assumed that biomedical knowledge will only grow bigger with time, so a mathematical framework should take care of this concern, in order to be future-proof. Finally, it is supposed that biomedical knowledge is and always will be incomplete; in order to be powerful enough, the mathematical formalisation has to able to handle missing information.

2.2.2.5 Molecular dynamism

One of the characteristics of living forms is their dynamism; as organisms are made of molecules, they are subject to the laws of chemistry and molecular dynamics. Organisms can be defined in chemical terms as semi open systems (Meng et al., 2004), which emphasises a strong relationship with the environment. Formalisms coming from chemistry, such as conservation of mass (Wikipedia, 2014a) or thermodynamics (Wikipedia, 2014i) can represent and solve such systems, yet they are not suited to handle more abstract concepts from the biomedical domain. An ideal formal framework for biomedical knowledge would appreciate the impact and interaction with the environment, yet this requirement is extremely challenging in regards to the number of chemical reactions to be considered (Meng et al., 2004). Moreover, the chemical formalisation strongly relies on kinetic parameters to capture behaviour, which makes them vulnerable to missing knowledge. Finally, another concern is the effect of chemical concentration in regards to the function. For instance, the action of a drug strongly depends on its administered dosage. Understanding in detail the biological machinery and deriving correct predictions out of it implies considering molecular dynamics.

Formalising biomedical sciences can be done using a descriptive methodology; this approach has been used since the origin of the field and is an intuitive way to represent biological systems and facts. I presented in this section the theoretical requirements deriving from the descriptive methodology and more generally

biomedical sciences. The coming section illustrates how DLs can address some of these requirements in order to mathematically formulate and further use biomedical knowledge.

2.3 Description logics for biomedical knowledge representation

DLs are part of a family of formal languages used to represent the knowledge of a domain of interest. The plural form of the word logic indicates that a multitude of languages exist, each one of them characterised by a certain type of expressivity, as it will be seen later in this chapter. The framework evolved from semantic networks (Allen and Frisch, 1982) and inherited its current name around 1980 with the advent of computer systems. DLs are well characterised from a theoretical point of view and implemented in the Web Ontology Language (OWL), a standard supported by the World Wide Web Consortium (W3C). For these reasons and because they address most of the requirements presented before, DLs are an ideal mathematical framework that can be used to formalise some biomedical knowledge.

2.3.1 Problems addressed by description logics

According to Gruber, DLs help to specify a conceptualisation (Gruber et al., 2009). In the case of life sciences, the conceptualisation is the molecular machinery, finding its specification being the task of the researcher. In this regard, DLs come with a series of tools to define words and concepts, and in particular their associated meaning or semantic. This feature allows terminologies to be built to describe the molecular black box machine. For instance, it is possible to capture the relative difference in meaning between these three following abstract biological concepts using DLs: positive regulation, negative regulation and regulation. This task appears trivial for humans, yet the main motivation behind DLs is to be able to express such things in an unambiguous and formal manner, in order to deduce less evident information later on. With DLs, the interpretation of the meaning of a concept comes from its relation to other concepts. For example, the concept mammal is more generic than the concept human; this is asserted in DLs

by stating that every instance of human is also a mammal. The mathematical interpretation is that the set of humans is a subset of mammals. An is a relation could also be drawn between the two concepts if they were represented as nodes (see Figure 2.6). This feature enables DLs to unambiguously define the meaning of words from a mathematical perspective. It is also possible to define in a similar way the meaning of relations and to use them to represent the logic and wiring of the molecular machine later on. Taken together, these theoretical properties alone address the needs for a mathematical framework (section 2.2.2.1), capable of handling definitions (section 2.2.2.2) and abstraction (section 2.2.2.3).

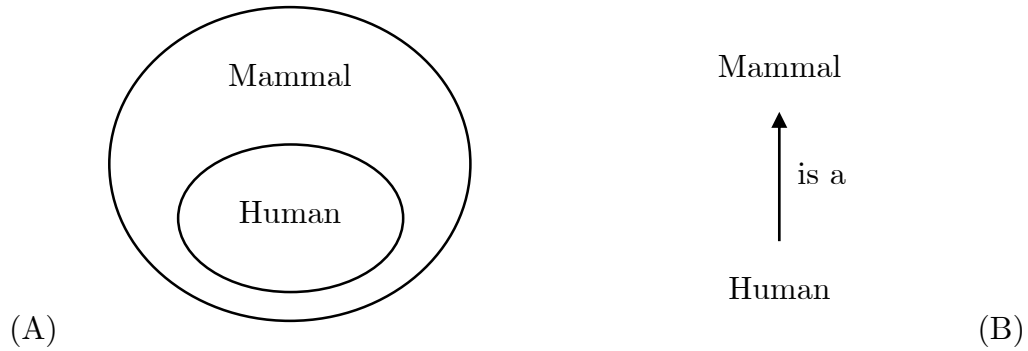


Figure 2.6: Example of problem addressed by description logics. (A) The concept Human is more specific than the concept Mammal (all humans are mammals), which can be represented as embedded sets. (B) Same logic captured, representing the concepts as nodes and the relation as edge. The mathematical meaning of (A) (sets of instances) is more accurate than the meaning of (B), yet both representation can exist in practice, in particular in biology.

The relations between entities and concepts are meant to formally express the known truth about a domain of interest (Stevens et al. (2007), Krötzsch (2012), Hitzler et al. (2009)). This type of construct is called an axiom. A set of axioms constitute a knowledge base, also called ontology. The main interest of a knowledge base is the possibility to use the meaning of the axioms to deduce implicit things. Consider as an example the question, What are the different types of mammals? From the example before, Human would be logically deduced from the connection between the two categories (Figure 2.6).

From the formal representation it becomes possible to generate proofs and ask queries over the knowledge base; this operation is called reasoning and consists of two tasks: subsumption and consistency checking. Briefly the structure of the

knowledge base arises from subsumption; concepts get classified into a taxonomy based on the axioms. Consistency checking can detect inconsistencies or contradictions that might be present in the knowledge base and report them. These services will be presented in details later in this chapter. It is important at this point to understand that the axioms of the knowledge base can be used to either generate more knowledge or to assess the validity of the current representation in a formal fashion. Reasoning tasks can be performed by humans and, more interestingly, by computers too. In fact, DLs have been developed for this very reason, to render domain knowledge computer understandable. The tight connection between computers and DLs is adequate for biomedical sciences, where it is expected to face large-scale data, beyond a sole human brains capability to handle at once. The interoperability with computers makes DLs adequate to address the requirement for a distributed and scalable framework (section 2.2.2.4).

DLs are efficient at representing terminologies and logical relations, yet they are unfortunately not appropriate to represent dynamic and temporal knowledge (Kim et al., 2008). This framework presents limits in regards to the molecular dynamicity requirement (section 2.2.2.5) and cannot handle well this facet of the machine. Work-around solutions will be presented later to the reader (Chapter 3).

Because DLs appear to address well a majority of the requirements for biomedical knowledge formalisation, I believe the framework to be suitable for the study of the molecular machinery. DLs can help to annotate and connect the logical parts in order to create a knowledge base, useful to understand the overall functioning of the system. Reasoning services allow for the querying and further use of the knowledge in a formal fashion, as expected from any mathematical framework. Finally the computer implementation of DLs is well studied, therefore part of the work can be automated, in order to come up with a scalable solution. Scalability guarantees the long-term success of the framework, but comes at the cost of expressivity.

2.3.2 Expressivity and complexity

Wikipedia defines expressivity as the breadth of ideas that can be represented and communicated. Lets consider for instance a molecule of methane; this concept can be represented in at least three different ways, more or less expressive and

detailed, as shown in Figure 2.7.

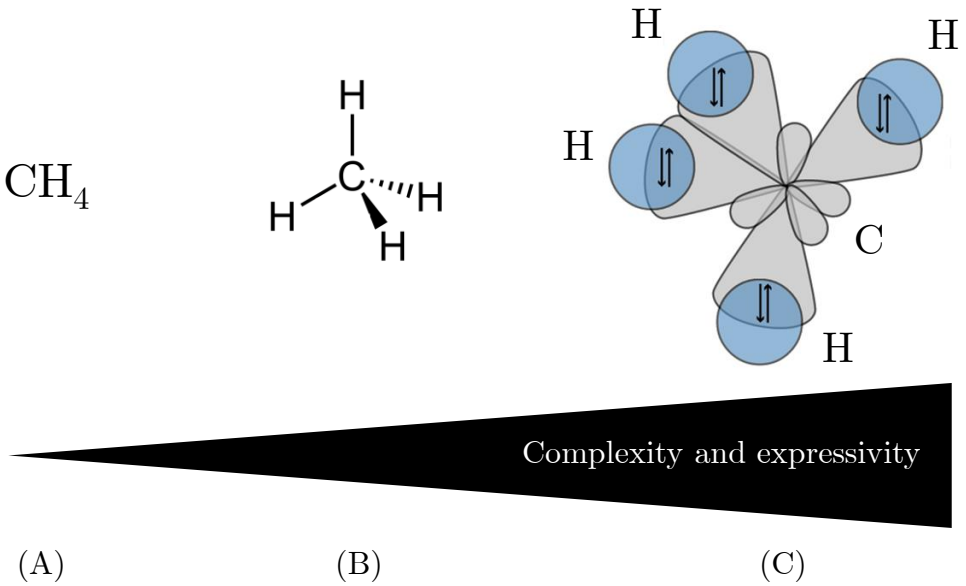


Figure 2.7: Example of problem addressed by description logics. (A) The concept Human is more specific than the concept Mammal (all humans are mammals), which can be represented as embedded sets. (B) Same logic captured, representing the concepts as nodes and the relation as edge. The mathematical meaning of (A) (sets of instances) is more accurate than the meaning of (B), yet both representation can exist in practice, in particular in biology.

What abstraction level is the correct one? It depends on the type of questions being asked on the model; all these three representations are legitimate and useful in practice for different purposes. The main difference between these models is their expressivity. Model 2.7-A for instance is less expressive than model 2.7-C; it conveys less detailed information, yet it is easier to understand and maybe more suited to study different types of problems. DLs are a family of logics, and just as with the methane molecule and depending on the type of axioms considered, they can model concepts in a more or less expressive fashion. As a rule of thumb, the more expressive the language is, the more complex it is. In the case of DLs, the complexity refers to the computational complexity, namely the time taken for the program to finish performing reasoning (Krötzsch, 2012). The more axiom types are available to the user, the more time it will take to deduce their entailments. Complex problems are to be avoided in computer science as it will take more time to return a deduction from the content of the knowledge

base, if a deduction is returned at all. This limitation comes from the hardware and the way computers are currently built (Turing (1950), Neumann and Burks (1966)). In this respect, a scalable solution is a trade-off between expressivity and complexity. One would like to be as expressive as possible in order to study the molecular machine, yet able to deal with the very large input size of biomedical information. Fortunately, one of the strength of DLs is that their computational complexity has been particularly well studied (W3C (2014a), ter Horst (2005)). It is possible to work with a restricted set of axiom types, guaranteeing an acceptable complexity, appropriate for large-scale implementation. These fragments of DLs are also called profiles or subsets; one of them, the $\mathcal{EL}^{++}$ profile, is particularly interesting for biomedical sciences (Baader et al. (2005), Baader et al. (2008), Hoehndorf et al. (2011a)). I will focus in the rest of this chapter on this very fragment, as it is possible to reason over an $\mathcal{EL}^{++}$ knowledge base in polynomial time. This characteristic makes reasoning tasks a tractable or so called easy problem (Cobham, 1965), which ensures that the framework can still work and scale for extremely large datasets. Despite offering a limited expressivity, the $\mathcal{EL}^{++}$ profile provides the means to address a good portion of the requirements for biomedical knowledge formalisation, as will be presented in the coming section.

In the quest to formalising biological information, DLs have the very clear advantage of a well characterised computational complexity, ensuring the creation of realistic practical solutions with the help of a computer, and not theoretically limited by the size of the data. On the contrary, the computational complexity of simulating and modelling biological systems at the level of the chemical reaction is less neatly defined (Meng et al. (2004), Gillespie (2007)) and appears much more complex and challenging; the fuzziness around the hardness of this task leaves an unanswered question as to whether modelling based only on chemical formalism will scale to systems as large as a cell.

2.3.3 DLs components and relation to life sciences

Because of scalability concerns, it is wise to stay within the boundaries of relatively low complexity 2.3.2. To address that matter, I have presented the EL++ profile as good candidate language, combining a decent expressivity for biomedical knowledge yet featuring constructs of an acceptable computational complexity. I will now present these constructs and components and explain how they relate

to the life science domain. DLs are a theoretical framework and find a computer implementation as the Web Ontology Language (OWL2) (W3C, 2014b). Despite being very close conceptually, these two frameworks have a few differences on the terminological level; I will therefore also mention the OWL2 equivalent names and symbols as a reference. OWL2 will be briefly discussed later in this chapter.

2.3.3.1 Description logics core entities

In order to represent the domain knowledge, DLs offer three kind of core entities, or building blocks: named individuals, roles and concepts. These entities are used to represent the world, and in this case, the components of the molecular machine. The first question that comes to mind is why these three types? The choice is purely arbitrary and likely derives from ancient Greece and the early work on categorisation done by philosophers such as Parmenides and Aristotle (Wikipedia, 2014c). It appears that these three types can represent quite a lot of information, and they are fairly intuitive for humans to understand and reason over. They are an acceptable way to abstract the world around us.

Named Individuals

OWL2 terminology: Individuals

The first building block handles individuals. Individuals refer to real-life instances and objects: this squirrel in a tree, this single molecule of water or glucose, this pen on a desk are all examples of individuals. Individuals are at the centre of DLs modelling. Everything else gravitates around them; all the further representation is done in regards to them. Every object can be considered as an individual.

Surprisingly, individuals are very rarely represented in life sciences. For example when one speaks about a particular protein, the reference is made towards the canonical version of the protein and not to the very one instance mixed with the millions of identical others. The same applies for diseases and species; the life scientist is concerned with extracting generic patterns and categories and reasoning at a more abstract level. This can be achieved by considering sets of individuals, so called classes or concepts.

Concepts

OWL2 terminology: Classes

The second building block type handles concepts and terminologies. DLs concepts are interpreted as mathematical sets, namely groups of objects or individuals. Concepts represent an abstraction over instances and fit biological reasoning well. For example, the concept human contains at least two individuals, the reader and myself. Most biomedical ideas can be described as a concept (or class), including not only material entities such as molecules (e.g, P53 or Paracetamol) but also immaterial processes or functions (e.g, Blood coagulation or Catalytic activity). This characteristic makes DLs concepts very suitable to describe the molecular machine. The difference between an OWL class and its members is illustrated in figure 2-8-A.

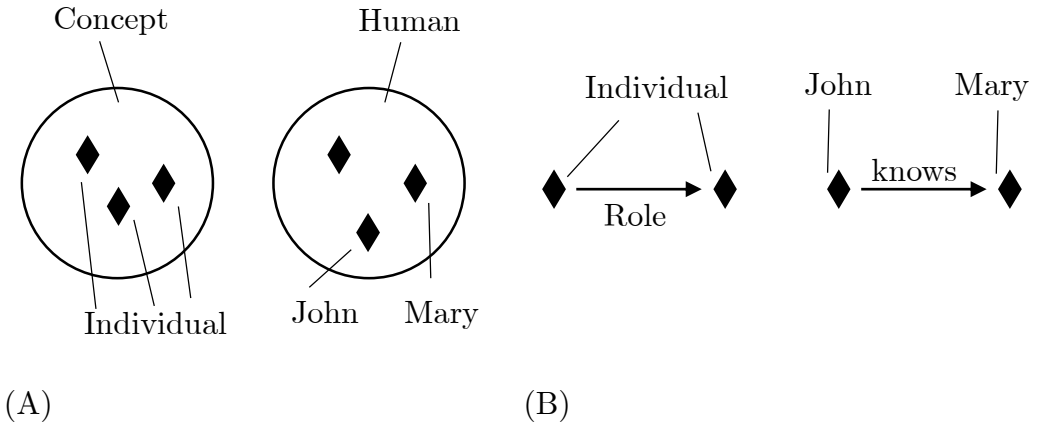


Figure 2.8: Description logics core entities: (A) Concepts and Individuals: A concept or class is a set containing some individuals. On the example shown, Human is a concept, John and Mary are both distinct individuals. (B) Roles and Individuals: Role are linking two individuals. In the example, John and Mary are still individuals linked by the role knows, specifying their relationship.

Roles

OWL2 terminology: Properties

The last building block deals with roles, namely how individuals relate one to another. Roles are more subtle and flexible to use than the two previous entity types. Basically they exist to capture a logical link between two individuals, and are so called binary in this respect (see Figure 2-8-B). In the study of the

molecular black box machine, roles connect the modules and processes; they can represent the logical connections of the machine. Examples of roles commonly used in the biomedical domain are: regulates, part-of, involved-in, etc. Note that roles are only linking individuals; they cannot be used to directly link concepts.

2.3.3.2 Axioms

In DLs, the axioms are the reflection of our current vision of the world; they express the mathematical truth and help both humans and machines to interpret the meaning conveyed by the entities of the knowledge base (concepts, individuals and roles). Axioms enable deductions to be made. For instance, lets assume that there is an enzyme called thrombin which is somehow involved in blood coagulation (first axiom) and that a compound named ximelagatran affects the activity of thrombin (second axiom). From these axioms, one could conclude that the ximelagatran might logically affect blood coagulation. The progression from a set of axioms to a conclusion is called reasoning, and, as discussed previously, this operation can be performed by humans or by computers with the help of a program, called a reasoner. Different types of axioms exist, depending on what needs to be modelled.

Assertional axioms (ABox)

The assertional axioms (or assertional box - ABox) are used to assert the truth about individuals. Although actual individuals are rarely represented as such in life sciences, it is still important to understand these fundamental types of axioms before looking at more complex ones.

Concept assertion

OWL2 terminology: Class assertion (or types). Concept assertion axioms link individuals to their concepts or types. For example, lets consider a knowledge base containing one named individual called john and one concept named Human. The axiom asserting that john is a human is written Human(john) or sometimes john : Human. It states that john is a member or instance of the class Human.

Role assertion

OWL2 terminology: Property assertion (or facts). This axiom captures the relationship between two individuals connected via a role. For example, consider two

named individuals, john and mary, and a role, knows. Asserting the fact that john knows mary as shown in figure 2-8-B is written knows(john, mary) or (john, mary) : knows in DLs. Role assertions are sometimes said to create triples or sentences; this axiom type helps to render logical networks or graphs.

Terminological axioms (TBox)

The terminological box contains the axioms related to concepts, which are therefore of primary interest for the biomedical domain. TBox axioms formalise the relationship among concepts in regards to the individuals they contain.

Concept inclusion ($\sqsubseteq$)

OWL2 terminology: SubClassOf. Concept inclusion expresses the relationship between two concepts, one being more specific than or subsumed by the other. Assuming that there are two concepts, Human and Mammal, the concept inclusion axiom Human Mammal entails that all instances of Human are also instances of Mammal, or in other words all humans are mammals. Note that even if the assertion was based on a mental reasoning about individuals, in practice we do not see any individual names, only concepts. The axiom can be visualised in figure 2-9, alongside concept assertions.

Concept equivalence ($\equiv$)

OWL2 terminology: EquivalentTo. It is sometimes interesting to define a set of individuals as equal to another set of individuals. This construct is mostly used to create new concepts from existing ones or to query a knowledge base. This axiom can state for example that the concept Human is equivalent to another concept or combination of concepts, as will be seen later in the chapter.

Relational axioms (RBox)

The meaning of a role can be further specified in regards to the other roles of the knowledge base, in a similar way as is achieved with terminological axioms.

Role inclusion ($\sqsubseteq$)

OWL2 terminology: SubPropertyOf. A role inclusion axiom defines the connection between two roles. Assuming that R1 and R2 are both roles in a knowledge base, and that the following role inclusion axiom holds: R1 $\sqsubseteq$ R2, it means that

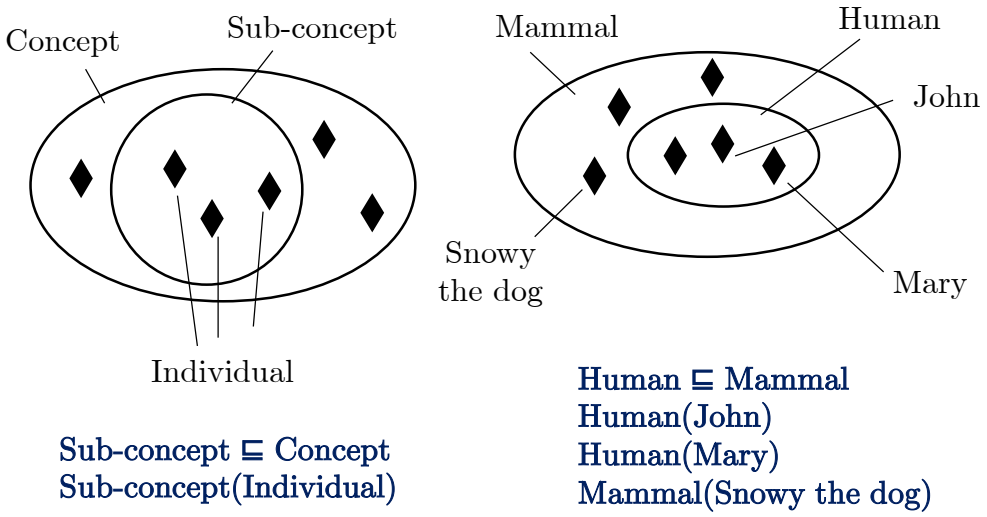


Figure 2.9: Examples of description logics axioms, in blue with their graphical representation (concept assertions and inclusion axioms). Axioms specify the semantics linking the basic entities (individuals, concepts and roles). From a series of axioms or knowledge base, it is possible to deduce information. The question What are the mammals present in the knowledge base? would return as a result Snowy, but also John and Mary, even if they are not directly declared as such, from the semantics encoded in the axioms.

all the time a pair of individuals is linked by the role R1, this pair is also linked by R2. The role inclusion axiom is analogous to the concept inclusion axiom for roles. A biological example concerns regulation: the role positively-regulates is subsumed by the role regulates: positively-regulates $\sqsubseteq$ regulates.

2.3.3.3 Constructors

It is possible to specify the semantics of the core entities using axioms. The expressivity seen so far is however fairly limited. Fortunately the DLs constructors provide new means of expression as illustrated in the coming sections. Constructors allow for the composition of complex types from simpler ones analogously to how symbols of cuneiform scripts appeared to be used in some situations (see figure 2-10-A). Just as with axioms and entities, there are different types of constructors.

Intersection ($\sqcap$)

OWL2 terminology: Intersection (and, that)

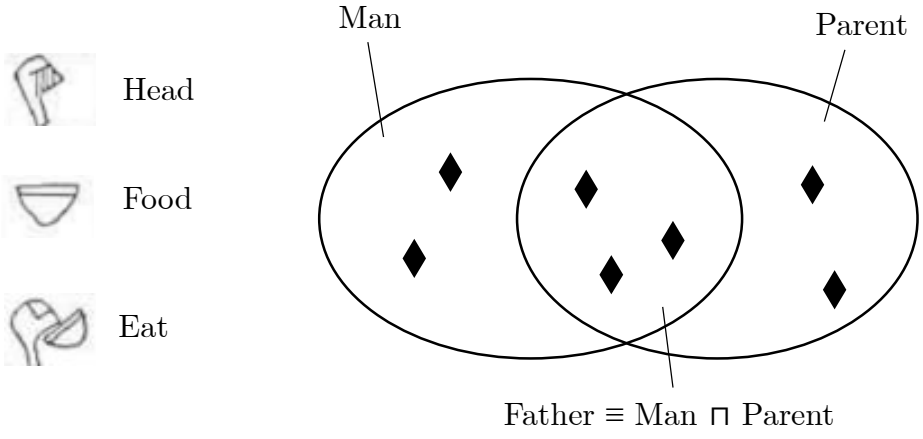


Figure 2.10: Description logics constructors. (A) Constructors help to compose with concepts. The logic behind constructors is analogous to some of the semantic found in cuneiform scripts (3000 BC) (personal interpretation). (B) Concept intersection: The individuals member in the same time of the concept Man and Parent are asserted to be of type Father on the axiom presented.

The intersection constructor corresponds to the basic Boolean and operator used to describe overlapping sets. The figure 2-10-B depicts the meaning of the construct. An intersection can be used to formulate new expressions, which are used to create more complicated axioms. The concept Father can for example be expressed as being equivalent (equivalent concept axiom) to the intersection of the concept Man and Parent. Both conditions have to be true in order to satisfy this expression. From this axiom, assuming one knows that an individual is both a parent and a man at the same time, one could deduce that this individual is also a father. Note that because of the equivalence axiom, the concept Father is also inferred as being a subconcept of Man and Parent; indeed if an individual is a father, it is therefore a man and a parent too from our definition.

Existential Restriction ($\exists$)

OWL2 terminology: Existential restriction (some)

Existential restrictions encode a semantic subtle to grasp at first encounter (personal teaching experience). Yet they are particularly suited for the biomedical domain, as a means to link concepts via roles, something not doable with the constructs introduced previously. An existential restriction captures the fact

that all individuals of a type X are necessarily linked by a property to some of the individuals of a type Y. For example, the expression ∃ part-of.Cell refers to the sets of individuals that are necessarily part of a cell (see Figure 2-11-A). The presence of one of these individuals implies that here exists an instance of cell inside which they are located, no matter what. This expression can be combined with a concept inclusion axiom; for example, the concept Nucleus can be expressed as follows: Nucleus part-of.Cell. The axiom entails that a nucleus is something that is always part of a cell. But only being part of a cell (∃ part-of.Cell) does not necessarily qualify something to be a nucleus. Moreover, the axiom does not entail that all the cells have a nucleus. The example is represented and labelled in figure 2-11. Existential restrictions are very useful to represent biomedical knowledge; they allow one to link concepts with roles, without explicitly referring to specific named individuals. Their entitlements are not too strong and often adequate for the biological world.

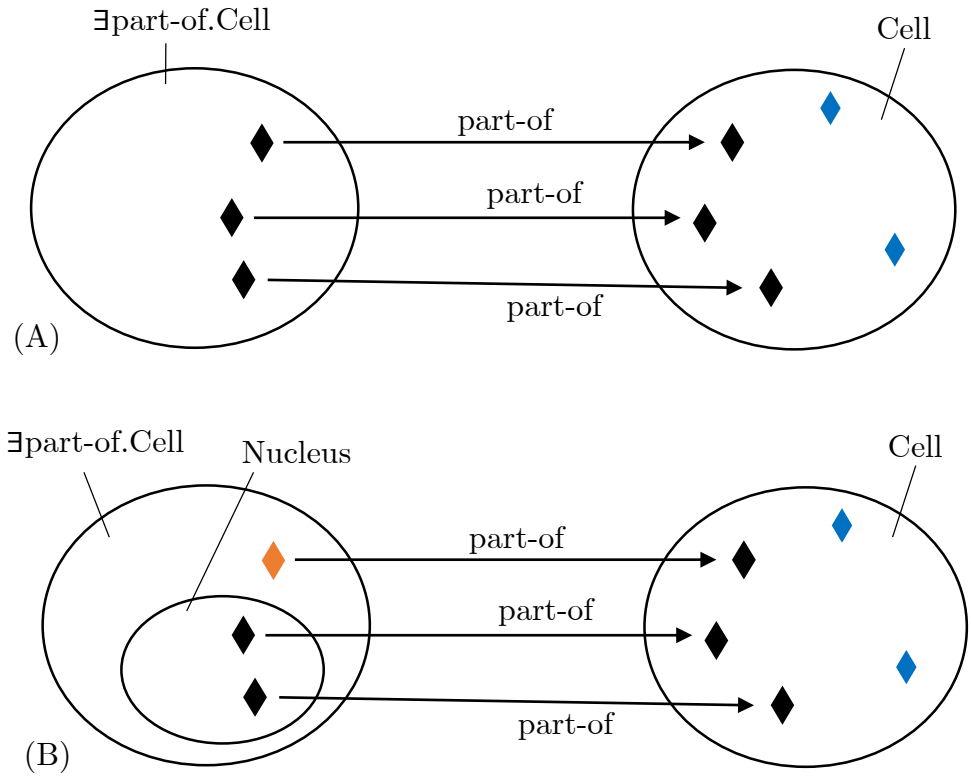


Figure 2.11: Description logics constructors: Existential restriction (A) Example of existential restriction expression with the implication. Note that all individuals of part-of.Cell are linked to a cell individual via a part-of role. Yet some cell instances exist without being linked (in blue for instance red blood cell). (B) Definition of the concept Nucleus from the existential restriction construct: Nucleus $\sqsubseteq \exists$ part-of.Cell, meaning that all instances of nucleus are necessarily linked to an instance of cell. There exist some instances of $\exists$ part-of.Cell not being nucleus instance (orange). This type of construct is commonly encountered in biology.

Role composition ($\circ$)

OWL2 terminology: Chained properties ($\circ$)

The last constructor presented is the role composition. This construct enables chaining two roles together in order to create a new one. The typical scenario for this constructor is the uncle role, defined as a role inclusion axiom: brother-of $\circ$ parent-of $\sqsubseteq$ uncle-of. It is interpreted as if an individual X is the brother of a person with a kid, then X is the uncle of the child (see Figure 2-12-A). Closer to biology, an example of role composition is the expression regulates $\circ$ part-of, which can be used to create the inclusion axiom: regulates $\circ$ part-of $\sqsubseteq$ regulates. This means

that when the role regulates is followed by the role part-of, it could be simplified into a single regulates role. There is a special type of complex property inclusion called transitivity. It means that the role is chained with itself. For example: part-of part-of part-of. The axiom is understood as if X is part of Y and Y part of Z then X is part of Z (see Figure 2-12-B).

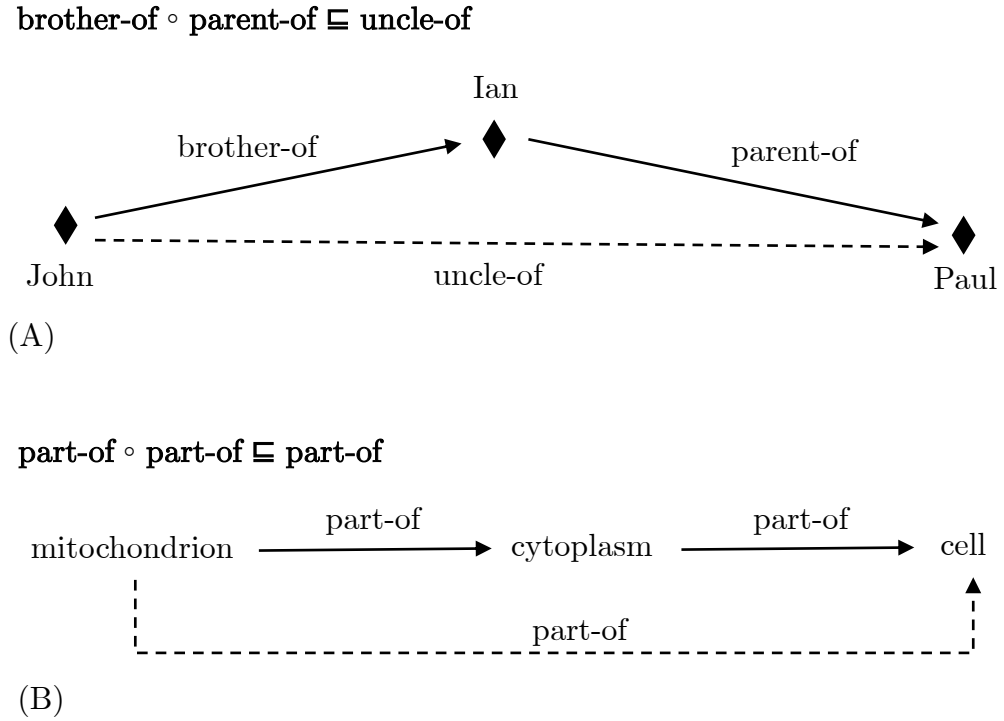


Figure 2.12: Description logics constructors: Role composition. (A) Example of composed property, the uncle relationship. When two instances are linked using a brother-of property, then followed by a parent-of property, then a reasoner can create a property between the first and the last individual, as shown on figure between John and Paul. (B) Informal illustration of a transitive property using the Gene Ontology specification (GO, 2014). When two terms are connected by a part-of relation and directly followed by another part-of property, then a part-of relation can be added between the first and last term. For instance here one can deduce that mitochondrion is part of cell, from the asserted facts. Note that this representation is not formal; OWL representation of OBO ontologies will be addressed later in this chapter.

2.3.4 Reasoning services

Combining basic entities, constructors and axioms gives a knowledge base or formal representation of a domain of interest whose content can be handled by a

reasoner. Considering the equation $x + 2 = 6$ as an analogy, axioms and entities helped to mathematically formulate the meaning of the symbols =, + and 21; now a reasoner can automatically solve the equation and find the value of x. In the biomedical case, the problem faced is different; the main task of the reasoner will be to classify the knowledge base and to answer queries about the functioning of the molecular machine.

As briefly mentioned earlier, reasoners perform two types of operations: subsumption and consistency checking of the knowledge base. I will not discuss consistency checking, as it is only meaningful if a certain type of axiom is present (disjunction axiom - not presented in this document, yet part of the EL++ profile). The subsumption service is also called classification. In this context, it means assigning the right concept to the correct place based on the meaning of the axioms. Classified concepts form a taxonomy; for example the section 2.3.3.3 explains how the concept Father can be asserted as equal to the intersection of the concepts Parent and Man. From this axiom, the reasoner can deduce that Father is subsumed by Man (all fathers are man), therefore Father can be classified as subconcept of Man and represented as such in a taxonomy (see Figure 2-13).

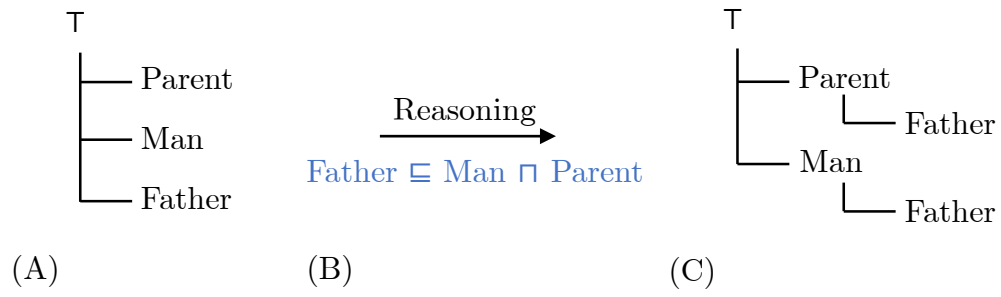


Figure 2.13: Description logics subsumption service example. (A) Taxonomy of concepts not classified. (B) From the axiom present in the knowledge base Father Man Parent (blue), the reasoning can deduce the taxonomy presented in (C). The concept Parent and Man subsumes Father, which appears deeper in the hierarchy, evidence of a more expressive meaning. The symbol represents the top concept (Thing in OWL) and is always present above every other concept of the knowledge base.

The subsumption service is also responsible for answering the queries formu-

lated over the knowledge base. A query is a standard concept expression. The list of concepts subsumed by the expression is the answer. DL queries are best orally formulated in the form of What are the things that.... For example, What are the things that are part of the cell?, expressed part-of.Cell in DLs. From the example of section 2.3.3.3, the reasoner would deduce nucleus (see Figure 2-11).

The complexity of reasoning services depends on the types of axioms present in the knowledge base. The ones I have discussed here are part of the EL++ profile, guaranteeing a tractable reasoning capable of handling large biomedical data input. Many more axiom types exists, featured by other DL families (Krötzsch, 2012). Deductions over the knowledge base are performed by the reasoner, and can be simplified as a classification where the taxonomy of concepts is generated.

2.3.5 The Web Ontology Language 2 (OWL2)

DLs are a theoretical mathematical framework. It is possible to manually exploit the meaning of axioms in order to perform reasoning services, but the goal is to eventually use a computer to perform this task, for time and data size concerns. In this regard, DLs are specified for computer implementation by the Web Ontology Language (OWL2) (W3C, 2014b). OWL2 was primarily designed to help data interoperability over the World Wide Web and is tightly linked to the semantic web. Yet as OWL2 derives directly from DLs, it is possible to use the language to deal with DL-related problems, such as the study of the molecular black box machine. OWL2 terminology is slightly different from that of DLs, and the language provides some extra functionalities relevant to software implementation: concepts are called classes, and rather than having a human readable name such as Father, entities are identified with a Uniform Resource Identifier (URI), for example http://www.example.org/Father. This feature guarantees the provenance of the information, as domain names are unique in the World Wide Web (Berners-Lee et al., 2001). I will not discuss in details here the semantic web principles and design; the reader can simply consider that OWL2 enables the implementation of DLs in a computing setting.

One of the distinctive features of OWL2 is the open world assumption. This statement implies that nothing can be deduced from missing information. As an example: If the fact that drug A perturbs protein B is present and someone asks Does drug A perturbs protein C?, the answer would be unknown. On the contrary,

in a closed world setting such as relational databases the answer would be drug A does not perturb protein C. The open world assumption fits the requirement of biomedical knowledge well; it is fair to assume that our knowledge of the living world will never be complete, therefore deductions can only be made over explicit evidence.

//TODO

2.4 Implementation with life-science information

In summary, the theory presented previously states that DLs are a suitable theoretical framework for the descriptive study of biological organisms. The reader now might be wondering how an actual knowledge base is built using DLs. First of all, multiple implementations of the descriptive framework to study the black box machine can exist. The implementation of the theory varies depending on the questions asked and granularity required. Chapter 3 will present an example of implementation, made with a particular set of axioms, addressing the mode of action and drug repositioning. Other possible implementations could use other DLs features to characterise the black box machine to study other biomedical topics. However, in any case and in order to be successfully implemented in practice, this descriptive approach needs to extend the current solutions used to store biomedical data as much as possible. Reusing the information already available is beneficial as it decreases the amount of work to be done and allows for a non-disruptive transition from existing and adopted technologies. I will discuss here how the theory, namely the descriptive approach based on DLs, relates to current biomedical databases and ontologies - traditional keepers of the knowledge. Brain, a programmatic library dedicated to the OWL2 EL profile, will finally be introduced to show how scalable real-life applications can be built using DLs.

Bibliography

Allen, J. F. and Frisch, A. M. (1982). What's in a semantic network? In *Proceedings of the 20th annual meeting on Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

An, Y. Y. L. J. and Jones, S. J. (2006). A large-scale computational approach to drug repositioning. *Genome Informatics*, 17(2):239–247.

Antezana, E., Kuiper, M., and Mironov, V. (2009). Biological knowledge management: the emerging role of the semantic web technologies. *Briefings in bioinformatics*, 10(4):392–407.

Ashburn, T. T. and Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, 3(8):673–683.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Baader, F., Brandt, S., and Lutz, C. (2005). Pushing the el envelope. In *IJCAI*, volume 5, pages 364–369.

Baader, F., Brandt, S., and Lutz, C. (2008). Pushing the el envelope further.

Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113.

Barratt, M. J. and Frail, D. E. (2012). *Drug repositioning: Bringing new life to shelved assets and existing drugs*. Wiley. com.

Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, 284(5):28–37.

Black, L. J., Sato, M., Rowley, E., Magee, D., Bekele, A., Williams, D., Cullinan, G., Bendele, R., Kauffman, R., Bensch, W., et al. (1994). Raloxifene (ly139481 hci) prevents bone loss and reduces serum cholesterol without causing uterine hypertrophy in ovariectomized rats. *Journal of Clinical Investigation*, 93(1):63.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

Brooksbank, C., Bergman, M. T., Apweiler, R., Birney, E., and Thornton, J. (2014). The european bioinformatics institutes data resources 2014. *Nucleic acids research*, 42(D1):D18–D25.

Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J., and Bork, P. (2008). Drug target identification using side-effect similarity. *Science*, 321(5886):263–266.

Chiang, A. P. and Butte, A. J. (2009). Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clinical Pharmacology & Therapeutics*, 86(5):507–510.

Cobham, A. (1965). The intrinsic computational difficulty of functions. In *Proceedings of the 1964 Congress for Logic, Methodology, and the Philosophy of Science*, pages 24–30.

Croset, S., Overington, J. P., and Rebholz-Schuhmann, D. (2013). Brain: biomedical knowledge manipulation. *Bioinformatics*, 29(9):1238–1239.

Darwin, C. and Bynum, W. F. (2009). *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life*. AL Burt.

De Franchi, E., Schalon, C., Messa, M., Onofri, F., Benfenati, F., and Rognan, D. (2010). Binding of protein kinase inhibitors to synapsin i inferred from pair-wise binding site similarity measurements. *PLoS One*, 5(8):e12214.

DiMasi, J. A. (2001). New drug development in the united states from 1963 to 1999. *Clinical pharmacology and therapeutics*, 69(5):286.

Dudley, J. T., Deshpande, T., and Butte, A. J. (2011a). Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics*, 12(4):303–311.

Dudley, J. T., Sirota, M., Shenoy, M., Pai, R., Roedder, S., Chiang, A. P., Morgan, A. A., Sarwal, M., Pasricha, P. J., and Butte, A. J. (2011b). Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science translational medicine*, 3(96):96ra76.

Duran-Frigola, M., Aloy, P., et al. (2012). Recycling side-effects into clinical markers for drug repositioning. *Genome Med*, 4(3).

Eckert, H. and Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug discovery today*, 12(5):225–233.

Editorial (2013). Form and function. *Nature*, 495(7440):141–142.

FDA (September 2007). Raloxifene hydrochloride. `http://www.fda.gov/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDER/ucm129243.htm`.

Fisher, J. and Henzinger, T. A. (2007). Executable cell biology. *Nature biotechnology*, 25(11):1239–1249.

Fliri, A. F., Loging, W. T., Thadeio, P. F., and Volkmann, R. A. (2005). Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nature chemical biology*, 1(7):389–397.

Fliri, A. F., Loging, W. T., and Volkmann, R. A. (2007). Analysis of system structure–function relationships. *ChemMedChem*, 2(12):1774–1782.

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nature genetics*, 42(12):1118–1125.

Furberg, H., Kim, Y., Dackor, J., Boerwinkle, E., Franceschini, N., Ardissino, D., Bernardinelli, L., Mannucci, P., Mauri, F., Merlini, P. A., et al. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature genetics*, 42(5):441–447.

Ghofrani, H. A., Osterloh, I. H., and Grimminger, F. (2006). Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond. *Nature Reviews Drug Discovery*, 5(8):689–702.

Gillespie, D. T. (2007). Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55.

GO (January 2014). Go relations. `http://www.geneontology.org/GO.ontology.relations.shtml`.

Goto, S., Bono, H., Ogata, H., Fujibuchi, W., Nishioka, T., Sato, K., Kanehisa, M., et al. (1996). Organizing and computing metabolic pathway data in terms of binary relations. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 175–186.

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. (2011). Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, 7(1).

Gruber, T., Ontology, I. L. L., and Özsu, M. T. (2009). Encyclopedia of database systems. *Ontology*.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402:C47–C52.

Haupt, V. J. and Schroeder, M. (2011). Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Briefings in bioinformatics*, 12(4):312–326.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., and Rudolph, S. (2009). Owl 2 web ontology language primer. *W3C recommendation*, 27(1):123.

Hoehndorf, R., Dumontier, M., Oellrich, A., Wimalaratne, S., Rebholz-Schuhmann, D., Schofield, P., and Gkoutos, G. V. (2011a). A common layer of interoperability for biomedical ontologies based on owl el. *Bioinformatics*, 27(7):1001–1008.

Hoehndorf, R., Loebe, F., Kelso, J., and Herre, H. (2007). Representing default knowledge in biomedical ontologies: Application to the integration of anatomy and phenotype ontologies. *BMC bioinformatics*, 8(1):377.

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2011b). Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic acids research*, 39(18):e119–e119.

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682–690.

Hurle, M., Yang, L., Xie, Q., Rajpal, D., Sanseau, P., and Agarwal, P. (2013). Computational drug repositioning: From data to therapeutics. *Clinical Pharmacology & Therapeutics*.

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., Murino, L., Tagliaferri, R., Brunetti-Pierri, N., Isacchi, A., et al. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, 107(33):14621–14626.

Johnson, M. A. and Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. Wiley New York.

Jordan, V. C., Phelps, E., and Lindgren, J. U. (1987). Effects of anti-estrogens on bone in castrated and intact female rats. *Breast cancer research and treatment*, 10(1):31–35.

Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N. P., Rieder, M. J., Cooper, G. M., Roos, C., Voight, B. F., Havulinna, A. S., et al. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature genetics*, 40(2):189–197.

Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., Jensen, N. H., Kuijer, M. B., Matos, R. C., Tran, T. B., et al. (2009). Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181.

Kim, S.-K., Song, M.-Y., Kim, C., Yea, S.-J., Jang, H. C., and Lee, K.-C. (2008). Temporal ontology language for representing and reasoning interval-based temporal knowledge. In *The Semantic Web*, pages 31–45. Springer.

Kinnings, S. L., Liu, N., Buchmeier, N., Tonge, P. J., Xie, L., and Bourne, P. E. (2009). Drug discovery using chemical systems biology: repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS computational biology*, 5(7):e1000423.

Krall, J., Fittingoff, M., and Rajfer, J. (1988). Characterization of cyclic nucleotide and inositol 1, 4, 5-trisphosphate-sensitive calcium-exchange activity of smooth muscle cells cultured from the human corpora cavernosa. *Biology of reproduction*, 39(4):913–922.

Krötzsch, M. (2012). Owl 2 profiles: An introduction to lightweight ontology languages. In *Reasoning Web. Semantic Technologies for Advanced Query Answering*, pages 112–183. Springer.

Kruger, F. A., Rostom, R., and Overington, J. P. (2012). Mapping small molecule binding data to structural domains. *BMC bioinformatics*, 13(Suppl 17):S11.

Kunkel, S. D., Suneja, M., Ebert, S. M., Bongers, K. S., Fox, D. K., Malmberg, S. E., Alipour, F., Shields, R. K., and Adams, C. M. (2011). mrna expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass. *Cell metabolism*, 13(6):627–638.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., et al. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935.

Lazebnik, Y. (2002). Can a biologist fix a radio?-or, what i learned while studying apoptosis. *Cancer cell*, 2(3):179–182.

Le Novere, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., et al. (2006). Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic acids research*, 34(suppl 1):D689–D691.

Li, Q., Cheng, T., Wang, Y., and Bryant, S. H. (2010). Pubchem as a public resource for drug discovery. *Drug discovery today*, 15(23):1052–1057.

Li, Y. and Agarwal, P. (2009). A pathway-based view of human diseases and disease relationships. *PloS one*, 4(2):e4346.

LINCS (January 2014). Connectivity map. `http://lincscloud.org/`.

Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1):3–25.

Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., Lavan, P., Weber, E., Doak, A. K., Côté, S., et al. (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, 486(7403):361–367.

Machado, D., Costa, R. S., Rocha, M., Ferreira, E. C., Tidor, B., and Rocha, I. (2011). Modeling formalisms in systems biology. *AMB Express*, 1(1):1–14.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., et al. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic acids research*, 37(suppl 1):D619–D622.

Men, A., Le, C., Charlab, R., Gobburu, J., and Lesko, L. (2010). 1: Sardana d, zhu c, zhang m, gudivada rc, yang l, jegga ag. drug repositioning for orphan diseases. brief bioinform. 2011 jul; 12 (4): 346-56. doi. *Policy*, 8(5):343–50.

Mendel, G. (1866). Versuche über pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brunn 4: 3*, 44.

Meng, T. C., Somani, S., and Dhar, P. (2004). Modeling and simulation of biological systems with stochasticity. *In silico biology*, 4(3):293–309.

Napolitano, F., Zhao, Y., Moreira, V. M., Tagliaferri, R., Kere, J., D'Amato, M., and Greco, D. (2013). Drug repositioning: A machine-learning approach through data integration. *Journal of Cheminformatics*, 5(1):30.

Neumann, J. v. and Burks, A. W. (1966). Theory of self-reproducing automata.

New-York-Times-Archives (March 1998). U.s. approves sale of impotence pill; huge market seen. http://www.nytimes.com/1998/03/28/us/us-approves-sale-of-impotence-pill-huge-market-seen.html?pagewanted=all&src=pm.

NHS-Choices (January 2014a). Angina. http://www.nhs.uk/conditions/Angina/Pages/Introduction.aspx.

NHS-Choices (January 2014b). High blood pressure. http://www.nhs.uk/conditions/Blood-pressure-(high)/Pages/Introduction.aspx.

Nielsen, F. H. (1999). Ultratrace minerals. *Modern nutrition in health and disease*, 9:283–303.

Noeske, T., Sasse, B. C., Stark, H., Parsons, C. G., Weil, T., and Schneider, G. (2006). Predicting compound selectivity by self-organizing maps: Cross-activities of metabotropic glutamate receptor antagonists. *ChemMedChem*, 1(10):1066–1068.

Ong, K. L., Cheung, B. M., Man, Y. B., Lau, C. P., and Lam, K. S. (2007). Prevalence, awareness, treatment, and control of hypertension among united states adults 1999–2004. *Hypertension*, 49(1):69–75.

Paolini, G. V., Shapland, R. H., van Hoorn, W. P., Mason, J. S., and Hopkins, A. L. (2006). Global mapping of pharmacological space. *Nature biotechnology*, 24(7):805–815.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555.

Renaud, R. C. and Xuereb, H. (2002). Erectile-dysfunction therapies. *Nature Reviews Drug Discovery*, 1(9):663–664.

Sampaio, E. P., Sarno, E. N., Galilly, R., Cohn, Z. A., and Kaplan, G. (1991). Thalidomide selectively inhibits tumor necrosis factor alpha production by stimulated human monocytes. *The Journal of experimental medicine*, 173(3):699–703.

Sanseau, P., Agarwal, P., Barnes, M. R., Pastinen, T., Richards, J. B., Cardon, L. R., and Mooser, V. (2012). Use of genome-wide association studies for drug repositioning. *Nature biotechnology*, 30(4):317–320.

Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., Sage, J., and Butte, A. J. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine*, 3(96):96ra77.

Stephens, T. D. and Brynner, R. (2009). *Dark remedy: The impact of thalidomide and its revival as a vital medicine*. Basic Books.

Stevens, R., Egaña Aranguren, M., Wolstencroft, K., Sattler, U., Drummond, N., Horridge, M., and Rector, A. (2007). Using owl to model biological knowledge. *International Journal of Human-Computer Studies*, 65(7):583–594.

Suthram, S., Dudley, J. T., Chiang, A. P., Chen, R., Hastie, T. J., and Butte, A. J. (2010). Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS computational biology*, 6(2):e1000662.

Swinney, D. C. and Anthony, J. (2011). How were new medicines discovered? *Nature reviews Drug discovery*, 10(7):507–519.

ter Horst, H. J. (2005). Completeness, decidability and complexity of entailment for rdf schema and a semantic extension involving the owl vocabulary. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2):79–115.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, pages 433–460.

von Linné, C. and Lange, J. J. (1770). *Systema Naturae: Per Regna Tria Naturae, Secundum Classes, Ordines, Genera, Species, Cum Characteribus, Differentiis, Synonymis, Locis*, volume 3. Curt.

W3C (January 2014a). Computational properties of owl2 profiles. `http://www.w3.org/TR/owl2-profiles/#Computational_Properties`.

W3C (January 2014b). Owl 2 web ontology language. `http://www.w3.org/TR/owl2-overview/`.

Wei, G., Twomey, D., Lamb, J., Schlis, K., Agarwal, J., Stam, R. W., Opferman, J. T., Sallan, S. E., den Boer, M. L., Pieters, R., et al. (2006). Gene expression-based chemical genomics identifies rapamycin as a modulator of mcl1 and glucocorticoid resistance. *Cancer cell*, 10(4):331–342.

Wermuth, C. G. (2006). Selective optimization of side activities: the sosa approach. *Drug discovery today*, 11(3):160–164.

Whirl-Carrillo, M., McDonagh, E., Hebert, J., Gong, L., Sangkuhl, K., Thorn, C., Altman, R., and Klein, T. E. (2012). Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417.

Wikipedia (January 2014a). Conservation of mass. `http://en.wikipedia.org/wiki/Conservation_of_mass`.

Wikipedia (January 2014b). Off-label use. `http://en.wikipedia.org/wiki/Off-label_use`.

Wikipedia (January 2014c). Ontology. `http://en.wikipedia.org/wiki/Ontology`.

Wikipedia (January 2014d). Paul ehrlich. `http://en.wikipedia.org/wiki/Paul_Ehrlich#Magic_bullet`.

Wikipedia (January 2014e). Phenotype. `http://en.wikipedia.org/wiki/Phenotype`.

Wikipedia (January 2014f). Prehistoric medicine. `http://en.wikipedia.org/wiki/Prehistoric_medicine`.

Wikipedia (January 2014g). Rube goldberg machine. `http://en.wikipedia.org/wiki/Rube_Goldberg_machine`.

Wikipedia (January 2014h). Science. `http://en.wikipedia.org/wiki/Science`.

Wikipedia (January 2014i). Thermodynamics. `http://en.wikipedia.org/wiki/Thermodynamics`.

Yang, L. and Agarwal, P. (2011). Systematic drug repositioning based on clinical side-effects. *PloS one*, 6(12):e28025.

Zahler, S., Tietze, S., Totzke, F., Kubbutat, M., Meijer, L., Vollmar, A. M., and Apostolakis, J. (2007). Inverse in silico screening for identification of kinase inhibitor targets. *Chemistry & biology*, 14(11):1207–1214.

appendix A

Extra Information

Some more text ...