UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA


# Knowledge Representation for Data Integration and Exploration in Translational Medicine

Cátia Maria Machado

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE BIOINFORMÁTICA

2013

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA


# Knowledge Representation for Data Integration and Exploration in Translational Medicine

**Cátia Maria Machado**

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE BIOINFORMÁTICA

Tese orientada pelo Prof. Doutor Francisco Moreira Couto e pela Prof.ª Doutora Ana Teresa Freitas

2013

# Resumo

A investigação biomédica evoluiu para uma ciência rica em dados, os quais podem ser recolhidos em enormes quantidades a partir de diversos recursos simultaneamente. No entanto, o valor dos dados está no conhecimento que deles se pode extraír através da sua análise. Em domínios como a medicina translacional, a integração e interoperabilidade dos dados são um requerimento fundamental para que estes possam ser analisados eficientemente.

A web semântica e as suas tecnologias foram propostas como uma solução para a integração e interoperabilidade de dados. Uma das ferramentas da web semântica é a utilização de ontologias, que permitem descrever o conhecimento de um domínio de uma maneira formal e estruturada.

A tese subjacente ao presente trabalho é que a representação em ontologias do conhecimento de um domínio pode ser explorada para melhorar o atual conhecimento sobre uma doença e os seus processos de diagnóstico e prognóstico. Os seguintes objetivos foram definidos para validar esta tese: 1) criar um modelo semântico que represente e integre as fontes de dados heterogéneos necessárias para a caracterização de uma doença e do seu prognóstico, explorando tecnologias da web semântica; 2) desenvolver uma metodologia que explore o conhecimento representado em ontologias por forma a melhorar os resultados obtidos com métodos de exploração aplicados a conjuntos de dados de medicina translacional.

O primeiro objetivo foi cumprido, tendo resultado nas seguintes contribuições: a metodologia para o desenvolvimento de um modelo semântico na linguagem OWL; um modelo semântico da doença cardiomiopatia hipertrófica; e uma revisão da exploração de recursos da web semântica em sistemas de medicina translacional.

O segundo objetivo, também cumprido, tem como contribuições a adaptação de uma análise de enriquecimento padrão ao uso de dados de doentes; e a aplicação dessa análise de enriquecimento no melhoramento das previsões feitas com conjuntos de dados de medicina translacional.

**Palavras Chave:** Representação de conhecimento, Medicina translacional, Web semântica, Ontologias, Integração de dados, Exploração de dados

# Abstract

Biomedical research has evolved into a data-intensive science, where prodigious amounts of data can be collected from disparate resources at any time. However, the value of data can only be leveraged through its analysis, which ultimately results in the acquisition of knowledge. In domains such as translational medicine, data integration and interoperability are key requirements for an efficient data analysis.

The semantic web and its technologies have been proposed as a solution for the problems of data integration and interoperability. One of the tools of the semantic web is the representation of domain knowledge with ontologies, which provide a formal description of that knowledge in a structured manner.

The thesis underlying this work is that the representation of domain knowledge in ontologies can be exploited to improve the current knowledge about a disease, as well as improve the diagnosis and prognosis processes. The following two objectives were defined to validate this thesis: 1) to create a semantic model that represents and integrates the heterogeneous sources of data necessary for the characterization of a disease and of its prognosis process, exploiting semantic web technologies and existing ontologies; 2) to develop a methodology that exploits the knowledge represented in existing ontologies to improve the results of knowledge exploration methods obtained with translational medicine datasets.

The first objective was accomplished and resulting in the following contributions: the methodology for the creation of a semantic model in the OWL language; a semantic model of the disease hypertrophic cardiomyopathy; and a review on the exploitation of semantic web resources in translation medicine systems. In the case of the second

objective, also accomplished, the contributions are the adaptation of a standard enrichment analysis to use data from patients; and the application of the adapted enrichment analysis to improve the predictions made with a translational medicine dataset.

**Keywords:** Knowledge Representation, Translational medicine, Semantic web, Ontologies, Data integration, Data exploration

# Resumo Estendido

A investigação biomédica tem vindo a tornar-se uma ciência rica em dados, os quais podem ser recolhidos em enormes quantidades a partir de diversos recursos simultaneamente. No entanto, o valor dos dados está no conhecimento que deles se pode extraír através da sua análise. Em domínios como a medicina translacional, a integração e interoperabilidade dos dados são um requerimento fundamental para que estes possam ser analisados eficientemente.

A web semântica foi proposta como uma nova abordagem para a Web, a qual, geralmente composta por documentos, passa a ser composta por dados. Esta abordagem permite uma mais fácil manipulação de dados entre domínios, e assim resolver a maioria dos problemas inerentes à integração de dados. A implementação de uma abordagem da web semântica resulta, em última instância, numa rede de dados e conhecimento que pode ser explorada tanto por computadores como por pessoas. Ao trabalhar para a criação desta rede, consegue-se melhorar a interoperabilidade entre aplicações, independentemente do seu domínio do conhecimento, assim como facilitar a integração e a exploração de dados.

Uma das ferramentas da web semântica é a representação do conhecimento de um domínio com ontologias, as quais fornecem uma descrição formal e estruturada desse conhecimento. A utilização de ontologias em medicina translacional é de extrema importância devido à necessidade de integrar e explorar dados para a sua implementação.

Os dados usados em medicina translacional são recolhidos no contexto da prática médica, e como tal são frequentemente caracterizados por um número reduzido de atributos clínicos e por um número elevado de valores em falta. Estas características dificultam significativamente o

uso destes conjuntos de dados em tarefas de exploração (e.g., mineração de dados), sendo consequentemente desejável enriquecê-los de algum forma.

A forma mais direta de enriquecimento de um conjunto de dados seria através da adição de mais instâncias ou de instâncias de melhor qualidade, mas devido à natureza restrita dos dados de doentes esta opção é raramente possível. Por outro lado, o conhecimento codificado em ontologias é público, e pode ser explorado para enriquecer conjuntos de dados da medicina translacional.

A tese subjacente ao presente trabalho é que a representação em ontologias do conhecimento de um domínio pode ser explorada para melhorar o atual conhecimento sobre uma doença, assim como melhorar os seus processos de diagnóstico e prognóstico. Os seguintes objetivos foram definidos para validar esta tese: 1) criar um modelo semântico que represente e integre as fontes de dados heterogéneos necessárias para a caracterização de uma doença e do seu prognóstico, explorando tecnologias da web semântica; 2) desenvolver uma metodologia que explore o conhecimento representado em ontologias por forma a melhorar os resultados obtidos com métodos de exploração aplicados a conjuntos de dados de medicina translacional.

O primeiro objetivo foi cumprido sob a forma de um modelo semântico da cardiomiopatia hipertrófica (CMH), uma doença que beneficia de uma abordagem como a da medicina translacional. O modelo representa dois domínios de conhecimento heterogéneos, o clínico e o genético, sendo composto por três módulos: *Clinical Evaluation*, o qual contém conceitos administrativos e os elementos clínicos necessários para o prognóstico de doentes com CMH; *Genotype Analysis*, contendo conceitos associados com a realização de testes genéticos em amostras biológicas; e *Medical Classifications*, um módulo auxiliar que contém padrões médicos usados na caracterização de elementos clínicos tais como sintomas. Este modelo semântico desempenha um papel importante na integração de dados de ambos os domínios

através das relações estabelecidas entre os módulos que compõem o modelo. Foi desenvolvido na linguagem OWL (uma das tecnologias da web semântica), reutiliza vocabulários já existentes, e tem mapeamentos definidos entre os seus conceitos e conceitos de vocabulários externos. Todos estes aspectos, em conjunto com o seu desenvolvimento em módulos, facilitam a sua utilização por terceiros. Apesar do uso de um caso de estudo, a metodologia para criar o modelo foi planeada de forma a poder ser utilizada para outras doenças, assim como o modelo em si pode ser usado e estendido para outras doenças.

O primeiro objetivo da tese resultou assim nas seguintes contribuições: a metodologia para o desenvolvimento de um modelo semântico na linguagem OWL; um modelo semântico da doença cardiomiopatia hipertrófica; e uma revisão da exploração de recursos da web semântica em sistemas de medicina translacional.

Do segundo objetivo resultou uma metodologia que identifica termos ontológicos enriquecidos num conjunto de dados de doentes. Esta metodologia foi adaptada a partir da técnica denominada análise de enriquecimento, a qual é extensivamente usada na análise funcional de grandes conjuntos de genes identificados com técnicas de alto rendimento tais como *microarrays* de expressão. A análise de enriquecimento explora o uso de métodos estatísticos para analisar os termos ontológicos com os quais um conjunto de genes está anotado. O propósito desta análise é a identificação de atributos biológicos que estejam representados no conjunto de genes em estudo em maior quantidade do que seria esperado devido ao acaso. Esses atributos biológicos são considerados como estando enriquecidos, ou sobre-representados, no conjunto de estudo e são usados na formulação de uma interpretação biológica acerca desse conjunto.

A metodologia de enriquecimento desenvolvida neste trabalho foi adaptada para analisar dados clínicos e genéticos de doentes, em vez de genes. Os dados clínicos incluem atributos como sintomas e medições

(e.g., peso, pressão arterial), enquanto os genéticos se referem à presença/ausência de mutações específicas da doença. Esta metodologia insere-se numa abordagem de prognóstico cujo propósito é auxiliar médicos na avaliação de doentes em relação à possibilidade de virem a sofrer um evento associado à doença ou de apresentarem uma manifestação característica da doença. A abordagem de prognóstico explorará os resultados obtidos com a metodologia de enriquecimento numa subsequente tarefa de classificação com algoritmos de mineração de dados, de forma a obter o prognóstico dos doentes.

Foram usados dois conjuntos de dados nesta parte do trabalho: um de doentes com CMH, focando a ocorrência de um evento associado à doença - paragem cardíaca súbita; e outro de doentes com doença pulmonar obstrutiva crónica (DPOC), focando uma das manifestações características da doença - enfisema.

Os termos ontológicos identificados como enriquecidos pela metodologia de enriquecimento são usados como perfil de anotação dos doentes através da sua incorporação no conjunto de dados sob a forma de atributos. Esta utilização dos termos enriquecidos foi testada com diferentes conjuntos de atributos e diferentes classificadores, e os resultados mostram que resulta numa incorporação de novo conhecimento no conjunto de dados que conduz a uma melhoria das previsões do prognóstico.

O segundo objetivo da tese resultou assim nas seguintes contribuições: a adaptação de uma análise de enriquecimento padrão ao uso de dados de doentes; e a aplicação dessa análise de enriquecimento no melhoramento das previsões feitas com conjuntos de dados de medicina translacional.

O trabalho aqui apresentado consiste no primeiro passo para o desenvolvimento de um sistema de análise de doenças para assistir médicos nos processos de diagnóstico e prognóstico, o qual contribuirá para o avanço do conhecimento sobre a doença em análise. Este sistema contribuirá igualmente para o avanço da medicina translacional, uma

vez que facilitará a integração de dados do domínio da investigação básica com dados clínicos, e a sua transformação em conhecimento que virá a ser usado na prática clínica.

# Acknowledgements

Firstly, I would like to thank my supervisor Doctor Francisco Moreira Couto and my co-supervisor Doctor Ana Teresa Freitas for their constant guidance and support throughout these four years of work.

I would like to thank Doctor Alexandra Fernandes, Doctor Susana Santos, and Dr. Nuno Cardim for providing the hypertrophic cardiomyopathy data for my study and for sharing their knowledge on the disease.

I am grateful to Doctor Deborah Penque and Doctor Bruno Alexandre for promptly accepting to share their data and knowledge on chronic obstructive pulmonary disease.

I am very grateful to Doctor Dietrich Rebholz-Schuhmann for his invaluable help and advice, both work-related and otherwise.

I would like to thank the Fundação para a Ciência e Tecnologia for my PhD scholarship (SFRH/BD/65257/2009) which made this work possible.

To my colleagues and friends in the Informatics department: Tiago Grego, Hugo Bastos, Ana Teixeira, João Ferreira, Cátia Pesquita, Juliana Duque, and Jeferson Sousa - thank you all for being such a good company every single day!

To my dear friends who were so important throughout my life: Ana Mafalda Fonseca, Paula Vieira, Leonardo Mendes, Ana Oliveira, Marta Fiúza, Rita Ferreira, and Cassilda Reis - thank you for being the way you are and for sharing your life with me!

To my family: José and Rosa Amorim, and Helena, Carlos, Ricardo, João and Hugo Lourenço - thank you for all the love and support you gave me, even before I was aware of it!

To my mother and father, thank you for giving me life and so much to think about, despite everything.

To my more recent family: Adelaide Pedro de Jesus, Filipe Faria, and David Faria - thank you for accepting me so warmly in your life!

Finally, to Daniel Faria, who has shared his life with me for the last 10 years - thank you for all the things we shared, for the love and care, and for helping me grow (as well as for cooking such delicious meals)!

Live as if you were to die tomorrow.

Learn as if you were to live forever.

- Mahatma Gandhi

# Contents

# List of Figures

## LIST OF FIGURES

# List of Tables

# Chapter 1

# Introduction

## 1.1 A data intensive world

Biomedical research has evolved into a data-intensive science, where prodigious amounts of data can be collected from disparate resources at any time (Hey *et al.*, 2009). However, the value of data can only be leveraged through its analysis, which ultimately results in the acquisition of knowledge. In domains such as translational medicine, where multiple types of data are involved, often from different sources and kept in different formats (e.g., hospital data, genetic data and pharmaceutical data), data integration and interoperability are key requirements for an efficient data analysis.

Translational medicine focuses on the improvement of human health by bridging the gap between basic science research and the clinical practice (Albani & Prakken, 2009; Webb & Pass, 2004; Woolf, 2008). It is unquestionable that translational medicine is a multidisciplinary research domain that relies both on public and protected data. Public data includes resources such as medical guidelines, scientific literature and biomedical databases, while protected data is composed of private data from patients and proprietary data from pharmaceutical and publishing companies. Translational medicine thus requires appropriate technologies for the correct interpretation of distributed and disparate data resources, and it is easy to conceive that such a large scale endeavor will eventually require a versatile infrastructure that preserves data semantics at all integration levels.

## 1.2 Using the semantic web for data integration and exploration

The need for data integration and data interoperability has a long-standing history. The Committee on Models for Biomedical Research proposed in 1985 a structured and integrated view of biology to cope with the available data (Committee on Models for Biomedical Research, 1985). Ten years later, in 1995, Davidson *et al.* questioned the feasibility of data integration, since the resulting data structure has to follow changes in the data itself and individual research groups fail to comply with the integration structure (Davidson *et al.*, 1995). In 2007, the challenges identified for data integration in genomic medicine were the lack of clinical data sources; the privacy issues linked to clinical data; the inherent complexity of medical records; and finally, the lack of data representation standards in the clinical domain (Louie *et al.*, 2007). These selected examples show that data integration remains an open research topic and that its complexity escalates with the increase in number of the heterogeneous domains to be integrated.

The World Wide Web is the key information channel for the communication of public data, particularly for the scientific community, since it enables a fast publication of methods, results and opinions, and it is easily reached by virtually anyone, anywhere. This information channel fulfills the requirements for efficient data exchange between scientific communities and data repositories, and thus should also be explored in translational medicine for optimal progress. However, its usefulness in this context is counterbalanced by the lack of data standards across domains, of explicit data representations, and of interoperability among data resources, which hinder the sharing of data between the biomedical and the clinical domains (Sagotsky *et al.*, 2008).

Berners-Lee *et al.* (2001) proposed the vision of the semantic web, where the Web of documents is replaced by the Web of data, thus enabling the manipulation of data over disparate domains and solving most of the problems previously stated for data integration. The manipulation of data is achieved by substituting the links connecting Web pages (i.e., the documents) with links connecting the data elements themselves, through the representation of the data domain of knowledge with structured semantic representations, and through the use of the standard

technologies. If built upon this infrastructure, many of the technical challenges faced by translational medicine are thus prevented.

The implementation of a semantic web approach ultimately results in a network of data and knowledge that can be exploited by computers and not only humans. Working towards the creation of this network improves the interoperability among applications, independently of their domain of knowledge, facilitates data integration and exploration. The mappings defined between data elements enable the issuing of queries over otherwise independent datasets, and the formal representation of the domain knowledge can be leveraged through the use of reasoning for the identification of new implicit connections between data elements (Rebholz-Schuhmann *et al.*, 2012).

The domain knowledge can be represented through resources that reliably abstract their real-world objects and the interactions existing amongst those objects. Such representations exist in the form of ontologies, where an ontology is "an explicit specification of a conceptualization" (Gruber, 1993).

Ontologies provide a means to formally describe the domain knowledge in a structured manner that can be shared among people and computers alike. If this description of the knowledge is accepted as a reference by the community (e.g., the Gene Ontology (Ashburner *et al.*, 2000)), its representation of the reality becomes a standard, and data integration is facilitated. This is true even if different abstraction levels are provided from unrelated datasets, since the hierarchical structure of ontologies supports the identification of a common ancestor for any two related concepts, by traversing the ontology graph (Stein, 2003).

Another common use of ontologies is for data annotation: by annotating an instance with ontology terms it is possible to obtain a description of that instance according to the knowledge encoded in the ontology. Different instances thus annotated can be compared based on their set of annotations, in order to calculate their degree of (semantic) similarity (Pesquita *et al.*, 2009). A set of instances can also be compared against the population from which it was collected to identify the set of ontology terms that can be used to explain what differentiates it from the rest of the population (i.e., enrichment analysis) (Khatri *et al.*, 2012; Robinson & Bauer, 2011).

## 1.3 Domain knowledge representation in translational medicine

The use of ontologies is invaluable in translational medicine due to the need of data integration and exploration for its realization. As indicated before, the use of semantic web technologies and of ontologies in particular facilitates data integration. Additionaly, the use of OWL (a semantic web technology) as the ontology encoding language enables the realization of inference over the explicitly represented data in order to identify non-explicitly represented knowledge. The exploitation of these aspects is of great interest in the data and knowledge pipeline at the core of the diagnosis and prognosis processes, two of the possible outcomes of a translational medicine approach. In those two processes, clinical and genetic data need to be integrated, consumed by medical and molecular biology experts alike, and the use of inference over both types of data can be fundamental in the identification of specific data and patients for subsequent analysis.

Since the data used in the diagnosis/prognosis process is collected in the context of medical practice, the resultant datasets are frequently characterized by a small number of clinical features and a high number of missing values. These characteristics can significantly hamper the use of such datasets for knowledge exploration purposes (e.g., data mining), and consequently it is greatly desirable to enrich them in some manner.

The most straightforward form of dataset enrichment would be through the addition of more data instances or of higher quality data instances, but due to the restricted nature of patients' data this option is seldom available. On the other hand, the knowledge encoded in ontologies is public, and can be exploited to enrich translational medicine datasets.

## 1.4 Objectives

The thesis underlying the present work is that the representation of domain knowledge in ontologies can be exploited to improve the current knowledge about a disease, as well as improve the diagnosis and prognosis processes.

In order to validate this thesis, the following two objectives were defined:

1. To create a semantic model that represents and integrates the heterogeneous sources of data necessary for the characterization of a disease and of its prognosis process, exploiting semantic web technologies and existing ontologies.

2. To develop a methodology that exploits the knowledge represented in existing ontologies to improve the results of knowledge exploration methods obtained with translational medicine datasets.

### 1.4.1 Contributions

The contributions of my work are the following:

- The methodology for the creation of a semantic model in the OWL language describing a genetic disease (Machado *et al.*, 2010, 2012a). This methodology is composed by numerous steps such as the identification of the data to represent, the definition of the hierarchical relations between the data elements, and the identification of external resources to use. All the steps included several iterations and decision points that can be extrapolated to any model facing the same decisions.

- A semantic model of the disease hypertrophic cardiomyopathy (used as case-study), for which no public semantic representation existed upon the beginning of this work. The model is composed by three modules: *Clinical Evaluation*, *Genotype Analysis*, and *Medical Classifications*. The concepts in the two first modules are mapped to external resources, mappings that can be used to traverse between resources. The three modules can be used both together and independently, and can be extended to represent any disease. The model is available from `https://sites.google.com/site/hcmsemanticmodel/home-1`.

- A review on the exploitation of semantic web resources in translation medicine systems, which analyzes 11 non-commercial systems (developed from 2007 to 2013) integrating genetic and clinical data (Machado *et al.*, 2013b).

- The adaptation of a standard enrichment analysis (based on the hypergeometric distribution) to directly use clinical and mutation data from patients (Machado *et al.*, 2012b, 2013a).

- The application of the adapted enrichment analysis to the improvement of a translational medicine dataset. The methodology was tested with the diseases hypertrophic cardiomyopathy and chronic obstructive pulmonary disease, and three ontological resources. The evaluation of the enrichment analysis was done based on the performance of the enriched dataset in data mining algorithms. The methodology can be used with any structured vocabulary and for any dataset, medical or otherwise, provided that there are vocabularies representing the domain of knowledge of the dataset.

## 1.5 Document organization

The rest of the document is organized as follows:

- Chapter 2 - a state of the art chapter divided in the four following subjects: biomedical topics; biomedical vocabularies used in this work; knowledge representation for data integration; and knowledge representation for data exploration.

- Chapter 3 - a description of the first part of my work, which is the representation of knowledge for data integration. In this chapter is described the methodology for the creation of the semantic model and the composition of the model.

- Chapter 4 - a description of the second part of my work, which is the exploitation of knowledge representations for data exploration. In this chapter is described the adapted enrichment methodology and the results obtained in terms of dataset quality improvement.

- Chapter 5 - the conclusions and future possibilities of the present work.

# Chapter 2

# State of the Art

This chapter is divided in four sections.

The first section details relevant biomedical concepts such as translational medicine and personalized medicine.

The second section introduces the biomedical vocabularies used in the present work.

The third section presentes the main semantic web standard technologies and tools, how domain knowledge is represented with ontologies, as well as eleven translational medicine implementations of a semantic web approach. These implementations are detailed in terms of their biomedical goal and their use of the semantic web tools, which is compared with how the semantic web tools are best exploited.

The fourth and final section describes the concepts and work relevant for the second part of my work, which is the exploration of domain knowledge represented in existing ontologies for the quality improvement of translational medicine datasets.

## 2.1 Biomedical background

### 2.1.1 Translational medicine

The bridging between basic science research and the clinical practice done by translational medicine approaches works at two distinct levels: at the level of

basic science research, translating it into new devices or treatments ("from the bench to the bedside"); and at the level of clinical practice, transferring the new treatments into its daily routine (Figure 2.1) (Wei, 2012; Woolf, 2008). Additionally, knowledge in translational medicine can also flow in the contrary direction, resulting in the initiation of new basic research based on the clinical observations of a disease development.

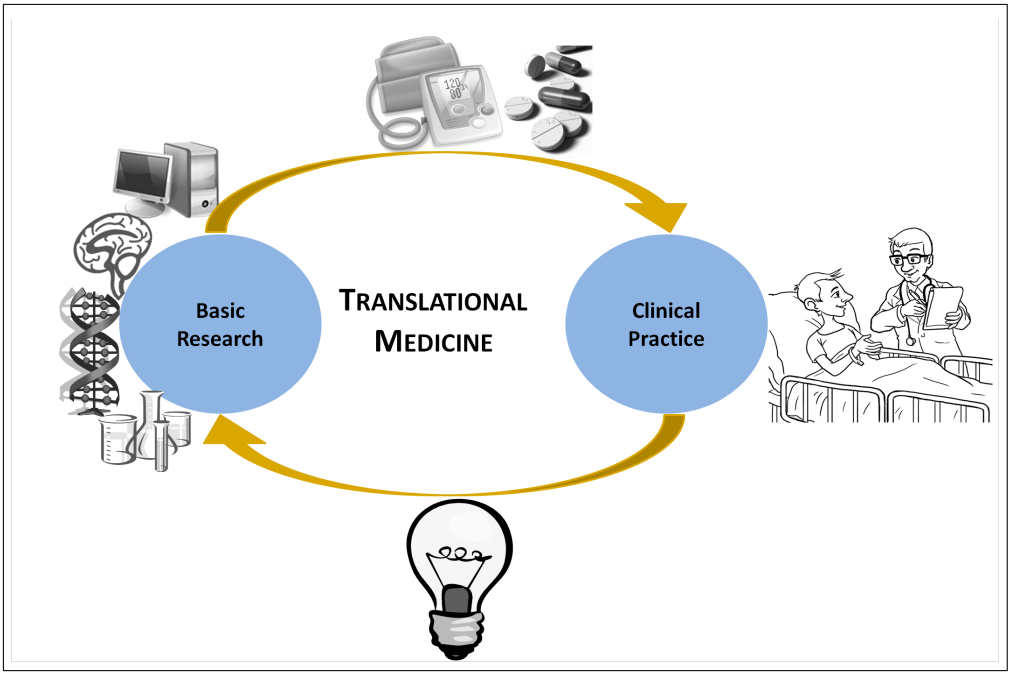


Figure 2.1: Knowledge workflow in translational medicine. Translational medicine improves the knowledge on human diseases by translating basic science research results into new exams, devices and treatments, which are then incorporated into the clinical practice. It also explores the knowledge collected during patient care to identify new research topics and topics that need further research.

Delivering solutions from the "bench to the bedside" and their incorporation into the health care practice requires that data flows from research in molecular biology, genetics and pharmacology into the clinical domain and in reverse. Within this flow of data and knowledge, research on the molecular mechanisms of diseases and drugs can be translated more quickly into novel treatment approaches, and observations about patients can as well lead to novel hypotheses and experimental conditions.

### 2.1.2 Personalized medicine

Many human diseases are influenced by both environmental and genetic factors (Manolio *et al.*, 2010). One of the action fields of translational medicine is precisely genomic medicine, which consists in exploring the molecular genetics knowledge of diseases and translating it into: personalized treatments with more beneficial responses and with reduced undesired effects; the prediction of disease susceptibility; new drug targets for diseases until now untreatable (Louie *et al.*, 2007). For example, a clinician may analyze a patient's mutations to explain observed drug side effects or may retrieve the list of biomarkers and their functions that have been associated with a specific cancer type.

A disease biomarker is any measurable characteristic that can be used as an indicator of a disease (Atkinson *et al.*, 2001), such as blood-based proteins and genetic variants. Just as the name suggests, genetic variants are variations in the genetic code that can occur throughout the human population. With the completion of the Human Genome Project, our potential knowledge on the molecular pathways underlying human diseases and on the treatment of these diseases has increased exponentially due to the possibility to associate those genetic variants with the observable manifestations of diseases. This is an invaluable accomplishment of modern medicine that can result in the association of individual genotypes (i.e., the genetic information of a person) with individual phenotypes (i.e., the observable characteristics of a person) (Frazer *et al.*, 2009).

The identification of genotype-phenotype relationships is consequently an important result in translational medicine approaches that can result in timely diagnostics (i.e., the identification of a disease in a person) and increasingly accurate prognostics (i.e. the prediction of the likely outcome of a disease in a person).

## 2.2 Biomedical vocabularies

Several biomedical ontologies from the clinical and the genetic domains of knowledge were used in the two parts of the present work. On the first part, the development of a semantic model, the following vocabularies were considered: SNOMED-

CT[1]; National Cancer Institute Thesaurus (NCIt) (Sioutos *et al.*, 2007); Ontology of Clinical Research (OCRe) (Sim *et al.*, 2010); Gene Regulation Ontology (Beisswanger *et al.*, 2008); and Sequence Ontology (Eilbeck *et al.*, 2005). The first three are clinical vocabularies, whereas the last two are genetic. On the second part, the enrichment of translational medicine datasets, the following vocabularies were considered: SNOMED-CT; NCIt; and the Gene Ontology (a genetic ontology) (Ashburner *et al.*, 2000).

All of these vocabularies are organized in hierarchies with multiple levels of granularity and, with the exception of the Sequence Ontology, are available on BioPortal (an ontology repository from the National Center for Biomedical Ontology[2]). A brief description of each of these ontologies is provided below.

SNOMED-CT is a healthcare terminology created to record clinical information in Electronic Health Records, covering topics such as clinical finding/disorder, body structure, pharmaceutical/biologic product, and social context. This terminology currently contains more than 397,000 concepts[3] (as of October, 2013).

The NCIt is a controlled terminology created by the National Cancer Institute to integrate molecular and clinical cancer-related information within a unified biomedical informatics framework. It covers topics that include cancers, drugs, therapies, anatomy, cellular and subcellular processes, and experimental organisms. It currently contains more than 103,000 concepts[4] (as of May, 2013).

OCRe was created in the context of the Human Studies Database Project to model features such as design type, interventions, and outcomes to support scientific query and analysis. It currently contains more than 380 concepts[5] (as of June, 2013).

The Gene Regulation Ontology was designed to model complex events that are part of the gene regulatory processes, with the purpose of meeting the needs of advanced information extraction and text mining systems targeting the identification of event representations in scientific literature. In its more recent version,

[1] http://www.ihtsdo.org/snomed-ct/
[2] http://bioportal.bioontology.org/
[3] http://bioportal.bioontology.org/ontologies/SNOMEDCT
[4] http://bioportal.bioontology.org/ontologies/NCIT
[5] http://bioportal.bioontology.org/ontologies/OCRE

it contains more than 500 classes[1] (as of April, 2010).

The Sequence Ontology was created by a group of scientists and developers from the model organism databases (FlyBase (Marygold *et al.*, 2013), WormBase (Harris *et al.*, 2010), Ensembl (Flicek *et al.*, 2013), SGD (Cherry *et al.*, 2012) and MGI (Eppig *et al.*, 2012)) for the annotation of genomic data, which is invaluable for sequencing experiments, bioinformatics analysis and molecular biology. This is the only ontology not currently available on BioPortal.

Finally, the Gene Ontology resulted from a coordinated effort to create a common description for gene and protein functions across species, since until then each species-specific database used its own terminology. It is composed by three orthogonal branches that enable the annotation of biological products with terms describing: the molecular functions they perform, the biological processes in which they are involved, and the cellular components where they are located or of which they are a component. This ontology currently contains more than 40,000 concepts[2] (as of November, 2013).

## 2.3 The semantic web in translational medicine

### 2.3.1 Technological standards in the semantic web

Over the past decade, the semantic web community, and in particular the World Wide Web Consortium (W3C)[3], has been developing a set of core technologies to realize the vision of the semantic web. Some of these technologies have since become de facto standards[4], and have brought the semantic web to life.

The existing technologies have been defined for purposes ranging from data and knowledge representation to data querying and data transformation. In this section are presented the technological aspects of the semantic web that are more relevant for data integration and knowledge representation.

[1] http://bioportal.bioontology.org/ontologies/GRO
[2] http://bioportal.bioontology.org/ontologies/GO
[3] http://www.w3.org/
[4] http://www.w3.org/standards/semanticweb/

## 2. STATE OF THE ART

The Resource Description Framework (RDF) is a standard language for data representation and interchange on the Web[1]. It uses the Universal Resource Identifier (URI) to identify each data element represented[2]. The basic structure of RDF is the triple, a statement composed of a subject connected with an object through a predicate, similar to narrative statements in English (e.g., "HomoSapiens isA mammal.", "Dopamin treats ParkinsonSyndrome ."). Since either of these elements can be part of different statements, data in RDF is best visualized through a directed graph, where the nodes represent the subjects and objects, and the arcs represent the predicates (or relations).

Due to its very basic and simple format, RDF restricts the representation of data to low levels of expressiveness (e.g., it does not allow the union of concepts, the definition of hierarchic relations between concepts, or the definition of cardinality in non-hierarchical relations). To overcome this limitation, two other technologies have been proposed: the RDF Schema (RDFS)[3], a specification language for data properties based on RDF; and the Web Ontology Language (OWL)[4], a language to formally define semantics, which also enables reasoning based on Description Logics (Baader *et al.*, 2003). Both formal languages extend RDF and enable the inference of new knowledge. As a result, knowledge can be shared and at the same time assessed for formal semantic consistency.

The representation of ontologies in RDFS or OWL provides additional advantages, namely: novel interpretations of the existing data against the ontological knowledge enabled by the mapping of data elements in RDF representation ("instances") to the ontological concepts ("classes" or "types"); and more detailed semantic comparisons of concepts that exploit the expressiveness of these formats (Couto & Pinto, 2013).

The Open Biomedical Ontologies (OBO) format[5] also exists for ontology representation, although it is not a standard semantic web technology. Due to its popularity in the health care and life sciences domains, extensive work has been

[1] http://www.w3.org/TR/2004/REC-rdf-primer-20040210/
[2] http://www.w3.org/standards/techs/uri#w3c_all
[3] http://www.w3.org/TR/rdf-schema/
[4] http://www.w3.org/standards/techs/owl#w3call
[5] http://www.geneontology.org/GO.format.obo-1_2.shtml

done in the conversion of ontologies in this format to OWL[1,2] (Tirmizi *et al.*, 2011).

SPARQL, a self-referencing acronym for SPARQL Protocol and RDF Query Language[3], is a query language to access RDF data. Since RDF data may be distributed over disparate data sources (including data stores exporting RDF from non-RDF relational databases), SPARQL has to retrieve data from all these resources. Due to the graph structure of RDF, SPARQL queries are transformed into graph pattern searches that rely only on the knowledge about the relations between concepts but not on a particular data model. SPARQL is also able to query RDFS and OWL provided that the graph pattern matching of the SPARQL query is defined with semantic entailment relations instead of the explicit graph structures[4]. Although other query languages exist for RDF (e.g., RDQL[5]), the availability of a SPARQL endpoint (i.e., an interface that provides access to a dataset through SPARQL queries) guarantees the independence from software and implementation specifications.

In addition to the technological standards, the definition of mappings between resources is another key element in the semantic web, enabling interlinked structured data according to the principles defined by Tim Berners-Lee: (1) use Uniform Resource Identifiers (URIs) as names for things; (2) use resolvable URIs (e.g., based on the HTTP protocol) so that those names can be looked up (either by people or machines); (3) provide useful information for look-up through the URI, using the standards (e.g., RDF, SPARQL); and (4) include links to other URIs, so that they can discover more things[6,7,8]. The URI can then be used to define any real-world entity (or "thing"), be it an object or an abstract concept (Dodds & Davis, 2012).

Examples of real-world entities in the biomedical domain are diseases, drugs, facts related to genes and protein functions, patient symptoms, biological mea-

[1] http://www.cs.man.ac.uk/~horrocks/obo/
[2] http://www.bioontology.org/wiki/index.php/OboInOwl:Main_Page
[3] http://www.w3.org/standards/techs/sparql#w3c_all
[4] http://www.w3.org/TR/2013/REC-sparql11-entailment-20130321/
[5] http://www.w3.org/Submission/RDQL/
[6] http://www.w3.org/DesignIssues/LinkedData.html
[7] https://dvcs.w3.org/hg/gld/raw-file/default/bp/index.html
[8] http://www.w3.org/TR/2013/NOTE-ld-glossary-20130627/

surements and family history. Ideally, each individual entity should have only one URI, so that every application points to the same source, regardless of its domain. This means that if the entity is altered in the original source, all applications pointing to it will be automatically updated. Additionally, the correct definition of URIs ensures that mappings between resources do not lead to semantic inconsistencies.

### 2.3.2 Knowledge representation with ontologies

In computer science the term "ontology" encompasses several types of conceptual models such as thesauri and glossaries, in addition to formal ontologies, which are logical structures defined by axioms and represented in formal languages (e.g., RDFS and OWL).

The use of ontologies is widespread in the biomedical domain due to the need of common representations among databases and research groups, but they can serve several other purposes (Noy & McGuinness, 2001):

- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge

Following the definition of the same authors (Noy & McGuinness, 2001), an ontology is a formal explicit description of concepts (called *classes*) in a domain of discourse; of properties defined for each concept that describe various features and attributes of the concept (called *properties*); and of restrictions on properties that, for example, define which classes the properties can be associated with.

The development of an ontology can be a complex task. Among other aspects, its complexity increases with the complexity of the domain to model. It also requires knowledge on the ontology developing tools; for complex domains, it takes time to analyze which part to model; it requires the evaluation of different modeling options; and that a consensus is reached between developers and users.

The design of an ontology should follow three rules and six steps (Noy & McGuinness, 2001). The rules are the following:

- There is no one correct way to model a domain - there are always viable alternatives. The best solution almost always depends on the application and on anticipated extensions.
- Ontology development is necessarily an iterative process.
- Concepts in the ontology should be close to objects (physical or logical) and relationships in the domain of interest. These are most likely to be nouns (objects) or verbs (relationships) in sentences that describe the domain.

The six ontology design steps are the following:

1. Determine the domain and scope of the ontology. The domain is the knowledge to be represented in the ontology (e.g., a disease), and the scope is the purpose of the ontology (e.g., the diagnosis of a disease).
2. Consider reusing existing ontologies.
3. Enumerate important terms in the ontology.
4. Define the classes and the class hierarchy.
5. Define the properties of classes.
6. Define the restrictions on the properties.

The properties can be of two types: data properties, and object properties. The first are properties for which the value in the RDF triple is a data literal, whereas the second are properties for which the value is an individual. For both types of properties it is possible to define a domain and a range. Domain axioms state that the subjects of the property have to belong to the class(es) set as domain, and range axioms state that the objects of the property have to be of the type of literal specified (for data properties) or have to belong to the class(es) specified as range (for object properties).

The definition of restrictions is the explicit representation of data or object properties as class descriptors.

### 2.3.3 Translational medicine systems

According to my analysis, eleven systems have been reported in the scientific literature from 2007 to 2013 that present translational medicine solutions dealing with medical conditions as disparate as cardiovascular diseases, cancer and diabetes.

Three systems focused on the cardiovascular system: one on the identification and prioritization of candidate genes for cardiovascular diseases (Gudivada *et al.*, 2008); another one on genetic association studies for hypercholesterolemia (Coulet *et al.*, 2008); and the third one also addressing association studies but for cerebrovascular diseases (Colombo *et al.*, 2010).

Two systems targeted cancer and its causes: one exploring genetic association studies for cervical cancer (ASSIST) (Agorastos *et al.*, 2009); and the other one identifying personalized treatments for colon cancer patients (MATCH) (Siddiqi *et al.*, 2008).

Two other systems targeted type 2 diabetes mellitus: one focused on the understanding of its causes to discover novel treatment hypotheses (SESL) (Rebholz-Schuhmann *et al.*, 2013); and the other one on genetic association studies (Pathak *et al.*, 2012). The latter covered hypothyroidism in addition to type 2 diabetes.

Each of the remaining four solutions tackled different biomedical tasks: the repurposing of drugs (Qu *et al.*, 2007); Traditional Chinese Medicine (TCM) (Chen *et al.*, 2007); neuroscience research (Receptor Explorer) (Cheung *et al.*, 2009); and congenital muscular dystrophy (Sahoo *et al.*, 2007).

Seven of these systems integrate public resources, while the remaining four consider only private data. In Figure 2.2 is shown the distribution of public resources integrated in each solution.

#### 2.3.3.1 Technical implementation

The translational medicine systems can be divided in four groups regarding their technical goal: exploitation of RDF as a data structure; data integration with the use of semantic web technologies; data integration through the representation of formal semantics; and data integration for inference purposes. Gudivada *et al.* (2008) developed the only system in the first group, with the goal of exploring the

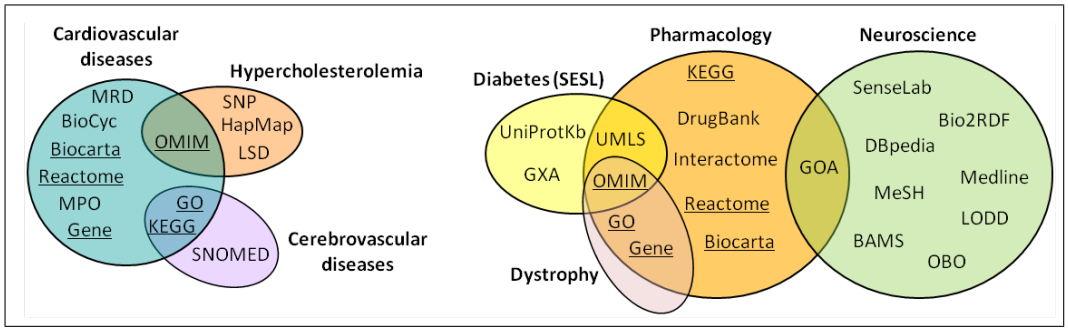


Figure 2.2: Public resources integrated by the translational medicine systems surveyed. The resources shown on the left are those integrated by the three systems targeting the cardiovascular system, whereas the resources shown on the right side are those integrated by the remaining four systems. The resources integrated in the cardiovascular system sub-domain that were also considered in at least one of the other sub-domains are underlined.
MRD - Mental Retardation Database; MPO - Mammalian Phenotype Ontology; Gene - NCBI Gene Database; OMIM - Online Mendelian Inheritance in Man; GO - Gene Ontology; KEGG - Kyoto Enclycopedia of Genes and Genomes; SNP - SNP Database; LSD - Locus Specific Databases; GXA - Gene Expression Atlas; UMLS - Unified Medical Language System; GOA - Gene Ontology Annotation; OBO - Open Biomedical Ontologies; LODD - Linked Open Drug Data; BAMS - Brain Architecture Management System; MeSH - Medical Subject Headings.

graph structure of RDF (see Tables 2.1 and 2.2 for a compilation of the semantic web resources used in the systems). The integrated resources were stored locally in relational databases and instantly converted to RDF when necessary for a specific disease and gene set. The conversion of the resources was mediated by the Disease Card Ontology (also referred to as Disease-Drug Correlation Ontology by Qu *et al.* (2007)), which was developed in OWL and included the reuse of external resources. The authors applied a graph-theory measure to the resulting RDF graph to score the importance of data elements (the graph nodes) in the data network.

Four systems exploit the semantic web technologies for data integration: Cheung *et al.* (2009), Sahoo *et al.* (2007), Rebholz-Schuhmann *et al.* (2013), and Pathak *et al.* (2012).

Cheung *et al.* (2009) present Receptor Explorer, developed by the W3Cs BioRDF task force group to demonstrate the use of the technologies and respective enabling tools in the implementation of a data federation (i.e., the data integrated is not locally stored but rather maintained at its original location). The system integrates resources already in RDF, mapping them at ID level, some of

Table 2.1: Standard semantic web technologies used in the surveyed translational medicine systems.

| Medical domain | Designation | RDF | OWL | SPARQL |
|---|---|---|---|---|
| Pharmacology | | + | + | + |
| Cardiovascular diseases | | + | + | + |
| Diabetes mellitus Type II | SESL | + | + | + |
| Diabetes & hypothyroidism | Mayo Clinic | + | | + |
| Neuroscience | Receptor Explorer | + | | + |
| Hypercholesterolemia | | + | + | + |
| Cerebrovascular diseases | | | + | |
| Cervical cancer | ASSIST | | + | |
| Traditional Chinese Medicine | TCM | + | | + |
| Muscular dystrophy | | + | | + |
| Colon cancer | MATCH | | + | |

This table lists the use of three semantic web standard technologies: RDF, OWL and SPARQL.

Table 2.2: Technical description of the surveyed translational medicine systems.

| Medical domain | Designation | Use of Ontologies | Reuse of Ontologies | Links |
|---|---|---|---|---|
| Pharmacology | | + | + | |
| Cardiovascular diseases | | + | + | |
| Diabetes mellitus Type II | SESL | | + * | + |
| Diabetes & hypothyroidism | Mayo Clinic | | | + |
| Neuroscience | Receptor Explorer | | | + |
| Hypercholesterolemia | | + | + | |
| Cerebrovascular diseases | | + | | + |
| Cervical cancer | ASSIST | + | | |
| Traditional Chinese Medicine | TCM | + | | |
| Muscular dystrophy | | | | + |
| Colon cancer | MATCH | + | | |

This table shows the use of ontologies for knowledge representation, the reuse of ontologies for both knowledge representation and data annotation (marked with *), and the use of mappings between resources.

which are part of the Linked Open Data cloud, namely DBpedia. It is through this resource that the authors are able to exploit other resources in the Open Data cloud, such as the NCBI's Gene Database and Linked Clinical Trials. The data can be accessed through the links between datasets and through a SPARQL endpoint.

The muscular dystrophy system presented by Sahoo *et al.* (2007) was developed at the US National Library of Medicine as a contribution to the Biomedical Knowledge Repository, and demonstrates data integration with RDF. It integrates the Gene Ontology and the NCBI's Gene Database, which were stored locally and integrated through mappings at ID level. It is also through those mappings that the Gene Database is connected with the Online Mendelian Inheritance in Man knowledgebase (OMIM) (Hamosh *et al.*, 2005), thus resulting in an indirect integration with the other two resources. The system exploits the use of inference rules over RDF to perform reasoning.

Rebholz-Schuhmann *et al.* (2013) describe SESL, a system integrating scientific literature from several publishers, namely Elsevier and Nature Publishing Group, as well as the following resources: the UniProt Knowledgebase (UniProtKB) (The UniProt Consortium, 2007), the Gene Expression Atlas (GXA) (Kapushesky *et al.*, 2012), and OMIM. The resources were stored locally in RDF. In terms of controlled vocabularies, the data in the GXA is annotated with the Experimental Factor Ontology (EFO) (Malone *et al.*, 2010), which for this work was mapped to the Disease Ontology (DO) (Schriml *et al.*, 2012) in order to exploit the existing mappings to the Unified Medical Language System (UMLS) (Bodenreider, 2004). Mappings were also defined between the integrated resources and external resources such as the Wikipedia[1].

Pathak *et al.* (2012) present a system that integrates solely private data from the Mayo Clinic Biobank. The integration is performed through the creation of virtual RDF graphs, and the original relational schemas are mapped to several existing ontologies such as the Translational Medicine Ontology (TMO) (Luciano *et al.*, 2011) and the Sequence Ontology (SO) (Eilbeck *et al.*, 2005). Additionally, the authors reused the Ontology for Biomedical Investigations (Brinkman *et al.*, 2010) and the Prostate Cancer Ontology (Min *et al.*, 2009) to create new concepts

[1] http://www.wikipedia.org/

and properties for the TMO, concepts that were subsequently mapped to the National Cancer Institute Thesaurus (NCIt) (Sioutos *et al.*, 2007).

The goal of data integration through the representation of formal semantics was pursued in three systems: Colombo *et al.* (2010), Siddiqi *et al.* (2008), and Chen *et al.* (2007).

Colombo *et al.* (2010) were interested in exploiting the potential of ontological modeling to solve the interoperability difficulties associated with integrating resources stored in different platforms, and according to different semantic representations. The authors created the Neuroweb Reference Ontology, in OWL, for that purpose, having established mappings between the ontology and external resources. The access to the data is restricted, and is done through SQL queries formulated based on the reference ontology.

The approach followed by Siddiqi *et al.* (2008) in the MATCH system consists solely in the virtual integration of the data sources through the Colon Cancer Ontology, developed in OWL. According to the authors, it enables them to explicitly enunciate the domain knowledge and to aggregate and extract information from disparate sources. The ontology models the medical domain and the analytical methods used throughout the decision support process, and was used to create the database schemas for each individual data repository maintained in relational format.

Chen *et al.* (2007) implement a virtual integration of the TCM domain through the TCM Ontology (Fikes & Kehler, 1985), developed in RDFS. For the authors, the ontology allows them to model the domain with richer semantics, which is necessary for the integration of the distributed resources. The data is maintained in its original location, in relational format, but is exposed as if in RDF.

Finally, the systems developed for inference purposes were those created by Coulet *et al.* (2008), Qu *et al.* (2007), and Agorastos *et al.* (2009).

The goal of Coulet *et al.* (2008) in the hypercholesterolemia system was to exploit the inference capabilities of OWL to guide the selection of data (through the identification of instances of classes) to further analyze with data mining algorithms. To that end, all the integrated resources were converted to RDF with the assistance of two ontologies: the SNP-Ontology[1] and SO-Pharm (Coulet

[1] http://bioportal.bioontology.org/ontologies/39215

*et al.*, 2006). Both ontologies were developed in OWL by the authors, the first for the representation of genomic variations and the second for pharmacogenomics. SO-Pharm is articulated with several external vocabularies through the reuse of concepts. In addition to their role in the selection of patients, inference was also used to perform consistency evaluations of the ontologies.

The purpose of Qu *et al.* (2007) was the identification of novel therapeutic applications for drugs and of novel disease mechanisms implicitly represented in the resulting RDF network of integrated resources. The Gene Ontology, one of the resources integrated, was obtained already in RDF/XML format, but all the other resources were converted to RDF. The integration and conversion of the data was done with the assistance of the Disease-Drug Correlation Ontology (already described for the system developed by Gudivada *et al.* (2008)). In order to perform reasoning, the authors had to define their own set of inference rules, since the triple store used did not provide support for direct inferencing over OWL.

The interest of Agorastos *et al.* (2009) (ASSIST) to exploit the inference capabilities of OWL resided in the possibility to automatically evaluate the disease severity for individual patients, and to assist in the unification of the database schemas of the participating hospitals. Inference was performed based on the ASSIST Cervical Cancer Ontology and on sets of medical rules. The data was converted to RDF and stored locally. The access was made available through a tree-like hierarchical visualization of the concepts defined in the ontology.

#### 2.3.3.2 Overview of the usage of semantic web technologies

The analysis of the eleven translational medicine examples permitted the identification of possible approaches to the problem of data integration and knowledge representation with semantic web standard technologies, with focus on the aspects of data and knowledge representation, use and reuse of ontologies, and definition of mappings with external resources.

As referred in Section 2.3.1, RDF is the standard language for data representation in the semantic web. If the resources to be integrated are already in this format then no transformation is required, but if they are stored in other

formats, such as relational or XML, a transformation is necessary. A third option is possible for data originally maintained in relational databases, which consists in exposing it as if in RDF, based on the mapping of the database schema to an ontology in RDFS or OWL. This last option is very interesting since it avoids the replication of the data, which is desirable to avoid different locations where it has to be independently updated, the duplication of object identifiers, and the need of extra storage space. The data can still be browsed and queried with SPARQL, however, the translation of the mapping between RDF and the schemas has to be done every time the data is accessed.

From the eleven systems described, eight use RDF (Table 2.2), of which five use it as data storage format. The cardiovascular diseases system did not store data in this format, but converted to RDF only the data necessary for an analysis. Based on the advantages presented by the authors associated with the use of semantic web technologies, and in particular RDF, it is possible to ascertain that they were interested in benefiting from the graph-like topography of RDF, but without converting all the data contained in their data sources of interest. In the case of the TCM and the diabetes/hypothyroidism systems, RDF was used to expose the integrated relational databases. In the TCM system, the results of the queries posed to the databases were also presented in RDF.

The use of ontologies during the integration process facilitates the integration itself by representing the domain knowledge in a formalized manner. When the ontologies are developed in RDFS or OWL, the translation of the resources to RDF (when such is necessary) is also facilitated, since a mapping is established between a data element in its original format and the concept that represents it in the ontology.

As referred in Section 2.3.2, the development of an ontology from scratch is a non-trivial task. As such, ontologies should be reused whenever possible (ideally ontologies widely accepted by the scientific community), since this improves the level of integration with other systems. However, the process of reutilization is not always straightforward: the ontology of interest can insufficiently represent the knowledge required (under-coverage); it can model more knowledge than is necessary (over-coverage); or it can model the knowledge not quite how we wish to

convey it in our application. Three forms of ontology reutilization have been identified that help to deal with these shortcomings: in the case of under-coverage, it is possible to reuse the whole ontology and add new concepts and relations; in the case of over-coverage, it is possible to extract only portions of the ontology that satisfy our needs (i.e., modules); if the ontology does not provide the necessary representation, it is possible to consider just some of the concepts and relations (thus ignoring hierarchical relations). It is important to stress out that this last option is not suitable if one wishes to infer knowledge based on the hierarchical organization of the ontology reused.

Seven systems use controlled vocabularies, three of which reusing existing vocabularies: pharmacology, cardiovascular diseases, and hypercholesterolemia (Table 2.2). This means that the remaining four created their own vocabulary, with a representation of the domain knowledge not shared by any other researchers. Despite the use of the standard technologies, this approach reduces significantly the interoperability among applications. The developers of the cerebrovascular diseases system refer that they considered the reuse of existent resources, namely SNOMED-CT and the Disease Ontology. However, they identified some shortcomings: an unsuitable formulation of concepts in SNOMED-CT; and the adoption by the Disease Ontology of a taxonomy that is different from the one used by the clinicians participating in the project. SESL is the only system that reuses an ontology solely for data annotation purposes.

Two other advantages of the use of ontologies are the possibility to reason over the knowledge explicitly represented and to infer new knowledge, and to evaluate the consistency of an ontology (i.e., verify if there are no contradictory statements that can lead to incorrect logic assumptions). This is automatically performed by reasoning engines over RDFS and OWL, but also over inference rules (typically *if-then* clauses) that can be used either instead of RDFS/OWL or in conjunction with it.

The use of OWL instead of RDFS enables a more expressive representation of data, although with the possible cost of undecidability, which is the inexistence of guaranties that any algorithm will be able to provide complete reasoning when using complex ontologies and large knowledgebases. In order to deal with this problem, a set of profiles was initially defined for OWL: Lite, DL and Full. Each

of these sublanguages has increasing expressiveness over the previous one, and consequently decreasing computational efficiency. With OWL 2, three more profiles have been defined: OWL EL, OWL QL and OWL RL [1]. If OWL does not convey the necessary expressiveness, inference rules can be defined to complement it.

Another form of improving interoperability between applications is trough the definition of mappings. They enable the explicit definition of connections between data resources, stating for example the likeliness of resources, or between data resources and controlled vocabularies through which data are linked with their describing concepts. Mappings between data instances result in the disambiguation of identifiers (URIs) by enabling the explicit indication of different identifiers (URIs) as referring to the same data instance in different datasets. If such mappings are not done, those instances are treated as if referring to different things, and cannot be automatically merged (a more detailed discussion on this topic was done by Cheung *et al.* (2009)). All these types of mappings improve the access to resources that have them, hence increasing the interoperability among applications that use those resources and the impact that those resources can have in the knowledge discovery process. Despite these advantages, only five systems exploited the use of mappings (see Table 2.2).

It is possible to have an idea of the importance of sharing resources, reusing existing vocabularies and defining links between resources when considering that some of the translational medicine systems already have the base work done for a future integration with one another through the resources used. Figure 2.2 shows the data resources integrated by seven of the systems, indicating several resources that are integrated by more than one of the systems. Additionally, the cardiovascular system and the pharmacology system reuse the NCI Thesaurus and SNOMED-CT[2], resources that were mapped to the diabetes/hypothyroidism system and to the cerebrovascular diseases system, respectively (Sioutos *et al.*, 2007).

[1] http://www.w3.org/TR/2009/REC-owl2-profiles-20091027/
[2] http://www.ihtsdo.org/snomed-ct/

## 2.4 Data exploration with ontologies

Enrichment analysis is a technique extensively used for the functional analysis of large lists of genes identified with high-throughput technologies, such as expression microarrays. It exploits the use of statistical methods over ontological gene annotations to identify biological features that are represented in a gene set under analysis more than would be expected by chance. Such biological features are said to be enriched, or overrepresented, in the study set and are then used to formulate a biological interpretation about it (Bauer *et al.*, 2010; Lu *et al.*, 2008; Robinson & Bauer, 2011).

Enrichment analyses are normally divided in three categories: Singular Enrichment Analysis (SEA), Modular Enrichment Analysis (MEA) and Gene Set Enrichment Analysis (GSEA). SEA works with a user-selected gene set and iteratively tests the enrichment of each individual ontology concept in a linear mode. MEA builds upon the enrichment calculation made in SEA and incorporates network discovery algorithms by considering the relationships between terms, which evolves the analysis from a term-centric approach into a biological module-centric approach. Finally, GSEA evaluates the terms individually but considering all the genes in the experiment and not just a user-selected gene set (Huang *et al.*, 2009). While in all these three categories the final result is a list of ontology terms considered to be enriched, a fourth approach has been proposed that searches for the optimal combination of terms that better explain the observed data in the user-selected gene set. This fourth approach is a model-based approach that calculates a score (or probability) for the entire set of ontology terms (Bauer *et al.*, 2010; Lu *et al.*, 2008; Robinson & Bauer, 2011). Several tools have been developed that implement one or more of the three main enrichment categories, such as Ontologizer[1] (Bauer *et al.*, 2008), Onto-express (Khatri *et al.*, 2002) and GSEA (Subramanian *et al.*, 2005).

SEA is the most commonly used category of enrichment analysis (Robinson & Bauer, 2011). Its underlying statistic test is normally the Fisher's exact test, and the distribution considered when working with small datasets is the hypergeometric distribution. This distribution is applied to situations of sampling

[1] http://compbio.charite.de/contao/index.php/ontologizer2.html

without replacement from a finite population when considering that the population elements are in one of two possible states. Translating this to the enrichment analysis, the goal is to evaluate if the genes in the population set are annotated with a term $t$, which means that the two possible states for a gene are: being annotated with the term, and not being annotated with the term. When drawing a sample of genes from the population (thus forming the study set), the objective is then to evaluate if the probability of annotation with term $t$ is higher in this sample than would be expected by chance. The expected frequency of annotation is given by the knowledge of the population set, and if the frequency of annotation in the sample is higher than in the population, then term $t$ might be used to explain the study set. In this type of analysis, what is being calculated is the probability of observing at least $n$ genes in the study set annotated with term $t$, given the knowledge of: the size of the study set, the size of the population set, and the number of genes in the population set annotated with $t$ (Robinson & Bauer, 2011). For a term to be considered enriched in the study set, the $p$-value obtained from the Fisher's test has to be lower than a significance level, which is normally considered to be 0.05 or 0.1.

In terms of ontologies, the Gene Ontology is the most commonly used (Ashburner *et al.*, 2000; Robinson & Bauer, 2011; Zhang *et al.*, 2010). Other resources that have also been explored in enrichment analysis are the Kyoto Enclycopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2012), the Medical Subject Headings (MeSH)[1], and the Human Disease Ontology (Gong *et al.*, 2012; Hoehndorf *et al.*, 2012; LePendu *et al.*, 2011; Schriml *et al.*, 2012).

A fundamental aspect in any enrichment analysis is the existence of a background set of annotations, against which the enrichment is calculated. While this background exists for resources such as the Gene Ontology (which is one of the main reasons for its popularity) and KEGG, when dealing with clinical ontologies this is frequently not the case. Tirrell *et al.* (2010) present a solution for this limitation by exploiting two sets of automatically created annotations: a set obtained by annotating the corpus of MEDLINE abstracts, and a set obtained by annotating public biomedical databases (the National Center for Biomedical Ontology's (NCBO) Resource Index). This approach uses any of the ontologies available

[1] http://www.nlm.nih.gov/mesh/

through BioPortal, a repository currently containing more than 360 ontologies (as of October 2013). LePendu and colleagues adopt a similar methodology, in which they use the title and abstract of publications that originated manual Gene Ontology annotations instead of the corpus of MEDLINE abstracts (LePendu *et al.*, 2011). In these publications are identified disease terms from the Human Disease Ontology, terms that are thus associated with the genes annotated with the Gene Ontology terms.

The terms tested in these analysis are not only those that directly annotate the genes, but also their ancestors. Given the high number of tests that are performed with resources such as the Gene Ontology (with more than 38,000 terms on January, 2013), a multiple-testing correction is necessary to reduce the possibility of false-positive results. The most conservative multiple-testing correction is the Bonferroni method, which is obtained simply by multiplying the calculated $p$-value by the number of tests performed. The fact that it is a very conservative correction means that some true-positives can be considered as false-positives. In order to deal with this limitation, several other methods have been developed that still provide the adequate control over the occurrence of false-positives (Robinson & Bauer, 2011).

# Chapter 3

# Knowledge representation for data integration

In this chapter is described the methodology underlying the development of the OWL semantic model that serves as a data integration framework for heterogeneous data collected in a translational medicine context.

As happens with an ontology, a semantic model is defined by axioms and is represented in a formal language; it is evaluated in terms of consistency; and it is evaluated by the domain experts. However, contrarily to ontologies, a semantic model has not been subjected to a formal evaluation as proposed by Gangemi *et al.* (2006) is terms of its structural, functional, and usability-profiling dimensions.

## 3.1 Methods

The semantic model was developed in OWL to comply with the semantic web standards and to take advantage of external resources published in the semantic web. It was chosen over RDFS due to its increased vocabulary[1], namely in terms of property definition (e.g., *Symmetric*, *Inverse*) and of class property restriction (e.g., *AllValuesFrom*, *SomeValuesFrom*).

[1] http://www.w3.org/TR/2012/REC-owl2-primer-20121211/

### 3.1.1 Semantic model development

The model was developed according to the guidelines presented in Section 2.3.2, through the following steps:

1. Definition of the domain and the scope of the model.
2. Search for existing ontologies representing the defined domain and scope.
3. Identification of the concepts to represent in the model.
4. Representation in OWL Lite of the concepts and respective properties, including hierarchical and non-hierarchical relations.
5. Identification of new concepts to represent in the model from existing ontologies.
6. Validation of the concepts and relations continuously done by the biomedical experts.
7. Consistency evaluation of the model (i.e., confirmation that no syntactic or semantic errors exist) periodically done with the reasoner HermiT[1] available in Protégé.

The identification of ontologies of interested was performed using BioPortal, from the National Center for Biomedical Ontology (Noy *et al.*, 2009). The concepts initially identified in collaboration with the biomedical experts were used as search terms. Given the translational medicine case-study, I searched for ontologies from the medical and molecular biology domains that contained the concepts of interest, and that represented these concepts in a hierarchical organization in accordance with the vision of the disease domain conveyed by the experts. The adequacy of the ontologies was evaluated based on their scope, and the initial list was narrowed down based on the number of concepts of interest the ontology contained.

After the identification of the ontologies to reuse, these were processed in the following manner:

[1] http://hermit-reasoner.com/

1. The regions of interest in each ontology were identified.
2. The hierarchical structure of the semantic model was refined in accordance with the ontology considered.
3. The concepts in the model were renamed in accordance to the ontology.
4. The concepts in the model were manually mapped to the equivalent concept in the ontology, through a *hasDbXRef* property [1].
5. When the ontology provided a definition for the mapped concept, it was added to the model.

Considering that the ontologies were also exploited to identify new concepts to include in the model, they served the dual purpose of aiding in the development of the model and providing mappings.

The development of the semantic model followed a combination development process Noy & McGuinness (2001), in the sense that both a top-down and a bottom-up approach were used: first a top-down, when defining with the domain experts the concepts to consider; and afterwards a bottom-up, when identifying generalizations for some of the concepts.

The Protégé-OWL editor (version 3.4.2)[2] was used to create the model.

### 3.1.2 The hypertrophic cardiomyopathy semantic model

Hypertrophic cardiomyopathy (HCM) is an autosomal dominant genetic disease that may afflict as many as 1 in 500 individuals, and is the most frequent cause of sudden cardiac death among apparently healthy young people and athletes (Alcalai *et al.*, 2008; Maron *et al.*, 2009). It can manifest itself either in a sporadic form or in a familial form, and in the latter case the first-degree relatives of the patient may also be at risk.

Since the disease is characterized by a variable clinical presentation and onset, its clinical diagnosis is difficult prior to the development of severe or even fatal

[1] http://www.geneontology.org/formats/oboInOwl#hasDbXref
[2] http://protege.stanford.edu/

symptoms (Alcalai *et al.*, 2008; Maron *et al.*, 2009). Therefore, its early diagnosis is extremely important.

Approximately 900 mutations in more than 30 different genes are currently know to be associated with the disease [1] (Ho, 2010), and the existence of a single mutation is sufficient for a positive diagnosis. However, the severity of HCM may not be the same for two individuals, even if direct relatives, since the presence of a given mutation can have a benign pattern in one individual and result in sudden cardiac death in another (Alcalai *et al.*, 2008; Brito *et al.*, 2003; Maron *et al.*, 2009).

The domain of the semantic model was identified as the disease, HCM, and the scope as the representation of the data necessary for the diagnosis and the prognosis of HCM. No public existing ontology was found that corresponded to these specifications.

Upon consultation with domain experts (that included a medical doctor and molecular biologists working with HCM patients and patients' samples, respectively), the set of concepts to represent in the model was identified based on the HCM characterization workflow shown in Figure 3.1.

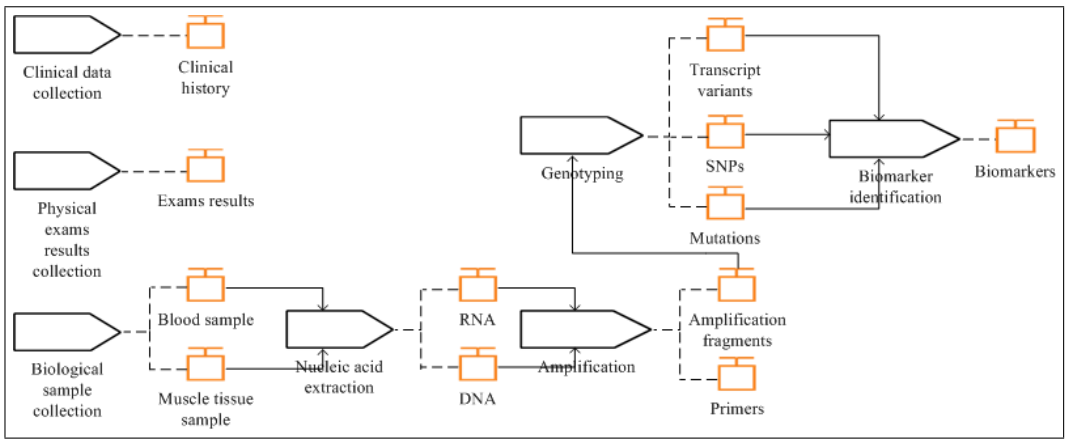


Figure 3.1: HCM characterization workflow. Schematic representation of the activities (semi-rectangular shapes) that characterize HCM and of the corresponding generated data elements (box-like shapes). Dashed lines connect activities with generated data elements, whereas full lines connect data elements with the activities in which they are used. The symbols used in this representation are based on the work developed by Constantine (2009).

[1] http://www.hgmd.org

The following elements of Figure 3.1 were represented as sibling classes in the semantic model:

- *Clinical History* (examples of clinical history elements are present and past symptoms).
- *Exams' Results* (e.g., electrocardiogram and echocardiogram).
- *Biological sample*, with *Blood sample* as subclass.
- *Nucleic acid*, with an associated property *hasType* to indicate if it is DNA or RNA.
- *Mutation*, which includes the elements *SNPs* and *Biomarkers*. The difference between mutations and SNPs is the frequency of occurrence in the population, which is indicated by an associated property *hasFrequency*. Biomarkers, in the case of HCM, are the mutations associated with the disease.
- *Amplification fragment*. This class represents segments of the genome to be screened for mutations.
- *Primers*, represented as *Amplification primer*. Represents auxiliary laboratory elements used in the obtention of amplification fragments.

The concepts *Muscle tissue sample* and *Transcript variant* were not included in the model (since they are not included in the scope of the model), and two additional concepts were identified with the experts, which were represented also as sibling classes: *General Characterization* of the patient (e.g., height and weight) and *Treatments* (e.g., prescription of drugs).

All the possible elements of each class were identified at this stage. However, their representation was not straightforward since three approaches were possible: to represent them as properties, as instances, or as subclasses of the respective main class. Neither of the first two options was the best solution since in that format it was not easy to maintain semantic coherence, nor to directly relate each individual data element with, for instance, a data value or a collection date. The last option implied the existence of instances for those subclasses, which meant

that it was necessary to define the lowest level of granularity to be considered in the model. The option chosen was the last one, the representation as subclasses, and considering that the patients are the central element of the HCM model, all instances of all classes in the model represent a measurement of or a statement concerning a patient.

After this decision, it was clear that the classes relative to the molecular biology analysis could not be represented as sibling of the remaining classes, since they represent data elements concerning a laboratory procedure performed in a sample from a patient. This resulted in the division of the model in two modules: *Clinical Evaluation* and *Genotype Analysis*. A third module was added to the semantic model to represent auxiliary medical information, *Medical Classifications*, which contains medical standards used to characterize clinical elements. The module *HCM Clinical Evaluation* imports the two other modules, *Genotype Analysis* and *Medical Classifications*.

The transition from a non-modular to a modular model corresponded to the first major milestone in the development of the model. The second one was the incorporation in the model of the knowledge represented in the following ontologies, selected for the module *Clinical Evaluation*: SNOMED-CT (version 2010_01_31)[1], NCIt (version 10.03) (Sioutos *et al.*, 2007), and the Ontology of Clinical Research (OCRe; version 0.95) (Sim *et al.*, 2010).

SNOMED CT and NCIt were used in the reorganization of the clinical data elements, which resulted in modifications such as the following: the *General Characterization* class was considered as a subclass of *Clinical History*; the *Treatments* and *Exams' Results* classes were considered as siblings under a parent class *Procedure*, represented as sibling of *Clinical History*. The classes in the *Clinical Evaluation* module were also renamed according to the ontologies reused: the class *Clinical History* according to *Clinical history and observation findings* from SNOMED CT; and the class *Procedure* according to *Intervention or Procedure* from NCIt.

A third and final milestone in the development of the model was a second phase of ontology reuse, with extensive alterations in the model hierarchical structure and the addition of new concepts, in particular to the module *Genotype Analysis*.

[1] http://www.ihtsdo.org/snomed-ct/

In this iteration of the model the mappings to OCRe were removed due to the deprecation of the respective concepts (e.g., *Health care site*), and two new ontologies were considered: the Gene Regulation Ontology (version 0.5, released on 04_20_2010) (Beisswanger *et al.*, 2008), and the Sequence Ontology (released on 11_22_2011) (Eilbeck *et al.*, 2005). Both ontologies were reused in the *Genotype Analysis* module.

The *Medical Classifications* module does not contain mappings to ontologies, but its concepts are linked to Web pages where their definition can be found.

In addition to the two major milestones indicated, the semantic model suffered several rounds of adjustments.

## 3.2 Results

The HCM semantic model is composed by three modules:

- *Clinical Evaluation* - containing administrative concepts and clinical data elements necessary for the prognosis of HCM patients.
- *Genotype Analysis* - containing concepts associated with the genetic testing of biological samples.
- *Medical Classifications* - an auxiliary module containing medical standards used in the characterization of clinical elements such as patient symptoms.

The following sections contain a detailed description of each module and of the mappings established between the main module, *Clinical Evaluation*, and the other two modules. The description of the modules is separated in the two iterations corresponding to the milestones of external ontologies reuse.

### 3.2.1 First phase of ontology reuse

#### 3.2.1.1 Module *Clinical Evaluation*

The main module of the HCM model comprises six high-level classes (see Table 3.1 for a complete characterization), two of which pertaining to administrative

Table 3.1: Composition of the modules *Clinical Evaluation*, *Genotype Analysis*, and *Medical Classifications*.

| Module | Iteration | Top-level concepts | Total concepts | Properties |
|---|---|---|---|---|
| *Clinical Evaluation* | 1 | 6 | 55 | 61 |
| | 2 | 5 | 63 | 60 |
| *Genotype Analysis* | 1 | 6 | 7 | 39 |
| | 2 | 7 | 19 | 39 |
| *Medical Classifications* | 1 & 2 | 2 | 4 | 2 |

For each module is shown the number of top-level concepts, the total number of concepts, and the number of data and object properties. The row marked Iteration 1 shows the composition after the first phase of ontology reuse, whereas Iteration 2 after the second phase.

elements, *Health Care Site* and *Clinician*, and the remainder four to the subjects and their clinical data: *Subject*, *Clinical History*, *Procedure* and *Heart Disease*.

Fig. 3.2 provides a visual representation of the high-level classes and their direct subclasses. The non-hierarchical relations between classes are not represented.

The class *Subject* corresponds to a central concept in this model, and is related to all the other concepts. It includes three subclasses: *Patient*(s) - individuals diagnosed with HCM; *Family Member*(s) - direct relatives of *Patient*(s); and *Control*(s) - individuals that do not suffer from HCM.

The classes *Health Care Site* and *Clinician* do not have subclasses. *Health Care Site* refers to the institutions where the subjects receive health care services and *Clinician* to the medical doctors involved in the assessment or administration of treatment to a *Subject*.

The class *Clinical History* has five subclasses that refer to clinical elements collected upon questioning or direct examination of the subject, namely: *Cardiovascular Measurement*, *Cardiovascular Finding*, *Body Measurement*, *Resuscitated Sudden Death* and *Death*. The subclass *Cardiovascular Measurement* contains the elements *Blood Pressure* and *Pulse Rate*. *Cardiovascular Finding* contains six elements: *Angina*, *Congestive Heart Failure*, *Cardiac Auscultation Finding*, *Palpitations* and *Syncope*. *Body Measurement* includes *Weight* and *Height*. While

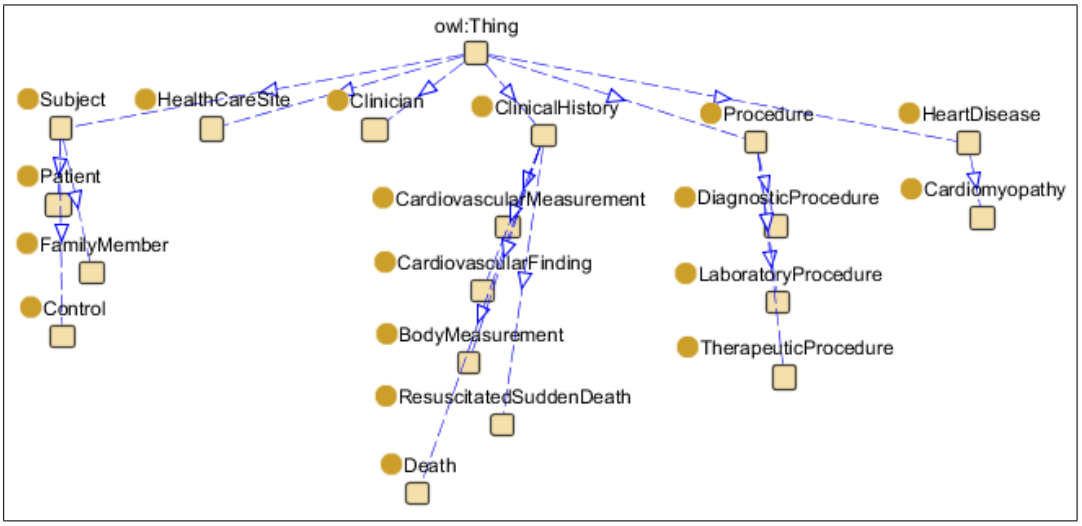


Figure 3.2: Graphical representation of the module *Clinical Evaluation*. The figure shows the top classes of the module and their direct subclasses (when existent). It was obtained with Jambalaya plug-in of Protégé Editor.

*Resuscitated Sudden Death* has no subclasses, *Death* has two: *Sudden Death* and *Non Sudden Death*.

Regarding the class *Procedure*, its subclasses represent the different types of procedures the subject can be subjected to, namely: *Diagnostic*, *Laboratory* and *Therapeutic Procedure*(s). *Diagnostic Procedure*(s) include *Cardiac Magnetic Resonance Imaging*, *Echocardiography* and *Electrocardiographic Monitoring*. *Laboratory Procedure*(s) consist of tests carried out in biological samples, such as blood, and those considered are the *Biomarker Analysis* (in particular *Genetic Marker Analysis*) and the *Hematology Test*(s). *Therapeutic Procedure*(s) comprise the subcategories *Prescription Of Drug* and *Cardiac Procedure*, the latter including *Medical Device Implantation*, *Septal Ablation* and *Septal Myectomy*.

The class *Heart Disease* contains cardiomyopathies (i.e., diseases of the heart's muscle), in particular *Hypertrophic Cardiomyopathy* and *Dilated Cardiomyopathy*.

For the classes *Clinical History*, *Procedure* and *Heart Disease*, the instances are records of that clinical element pertaining to a *Subject*. Considering, for example, the classes *Pulse Rate* and *Dilated Cardiomyopathy*, the instances are, respectively, a pulse rate measurement for a given *Subject* and the *Subject* to

which the disease was diagnosed.

#### 3.2.1.2 Module *Genotype Analysis*

The design of the *Genotype Analysis* module was oriented to the maintenance of data related to biological specimens and laboratorial activities, rather than of *Subject*'s records. It contains six high-level classes and a total of 39 properties (see Figure 3.3 and Table 3.1). All classes are related to the process of identifying genetic markers associated with HCM in biological samples.

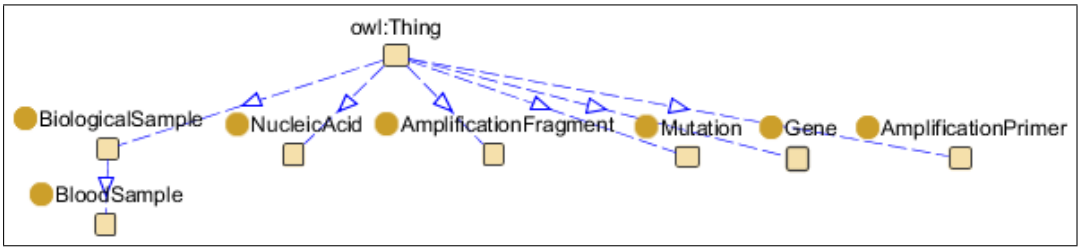


Figure 3.3: Graphical representation of the module *Genotype Analysis*. The figure shows the top classes of the module, namely: *Biological Sample* (with subclass *Blood Sample*), *Nucleic Acid*, *Amplification Fragment*, *Mutation*, *Gene* and *Amplification Primer*. The figure was obtained with Jambalaya plug-in of Protégé Editor.

The activities underlying a genotype analysis involve the manipulation of *Biological Sample*(s), from which *Nucleic Acid*(s) are extracted. From the latter it is possible to obtain *Amplification Fragment*(s), which correspond to the segments of the genome to be screened for HCM-related *Mutation*(s). Each of these *Mutation*(s) is located in a specific *Gene*. All these non-hierarchical relations are represented in the model under the form of restrictions applied to the classes.

#### 3.2.1.3 Module *Medical Classifications*

The module *Medical Classifications* is intended for the maintenance of data necessary for the characterization of clinical elements. Such data can be either standards or guidelines, developed to provide some degree of uniformity in the description of medical observations made by medical practitioners.

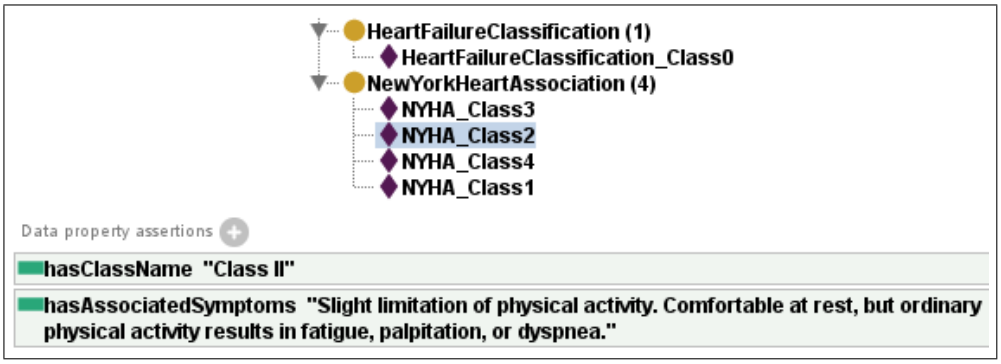


Figure 3.4: Concept *Heart Failure Classification*, from the module *Medical Classifications*. All the instances of this concept are shown, as well as its single sub-concept (*New York Heart Association*). The properties of the instance *NYHA_Class2* are also shown.

It contains two high-level classes (Table 3.1), *Angina Classification* and *Heart Failure Classification*, referring to functional classification systems created to assess the degree of severity of two *Cardiovascular Finding*(s), respectively angina and heart failure. Both classes have one subclass: *Angina Classification* has the classification system for angina created by the *Canadian Cardiovascular Society*[1], and *Heart Failure Classification* has the classification system for heart failure created by the *New York Heart Association* (Criteria Committee of the New York Heart Association, 1994).

Each classification system relates the onset of the symptoms to everyday activities of the patients. In the case of angina, the Canadian Cardiovascular Society defined 5 degrees of severity, from Class 0 to Class 4. According to this classification system, a *Subject* with angina *CSS_Class1* feels chest pain associated only with strenuous exercise, while other with angina *CSS_Class4* feels chest pain at any level of physical exertion, even at rest. *CSS_Class1* and *CSS_Class4* are instances of the class *Canadian Cardiovascular Society*.

As an example of this module, Figure 3.4 shows the concept *Heart Failure Classification* and the data properties for one of its instances, *NYHA_Class2*.

This module was not changed in the second phase of ontology reuse.

[1] http://www.ccs.ca/home/index_e.aspx

#### 3.2.1.4 Mapping between modules and with external ontologies

The mapping between the module *Clinical Evaluation* (ce:) and the modules *Genotype Analysis* (ga:) and *Medical Classifications* (mc:) is done through the following non-hierarchical relationships (here represented as triples, where the central elements are object properties):

- ce:*Subject* ce:*hasBiologicalSample* ga:*Biological Sample*
- ce:*Laboratory Procedure* ce:*performedInBiologicalSample* ga:*Biological Sample*
- ce:*Angina* ce:*hasAnginaClassification* mc:*Angina Classification*
- ce:*Congestive Heart Failure* ce:*hasHeartFailureClassification* mc:*Heart Failure Classification*

Patients' mutations can be identified through this relationship between *Clinical Evaluation* and *Genotype Analysis* since in the latter module a *Biological Sample* is connected with the mutations identified therein.

In terms of mappings to external ontologies, a total of 78% of the concepts in the module *HCM Clinical Evaluation* were mapped, 6% of which to OCRe (see Table 3.2). Regarding the latter, the mapped terms originate from its superclass *clinical:Role*, specifically *Subject*, *Clinician* and *Health Care Site*. In respect to the module *Genotype Analysis*, it contains only one link to OCRe occurring between the concept *Biological Sample* and the term *clinical:Sample*, a subtype of *clinical:Role*.

In all iterations of the semantic model, new terms were considered only when no ontology contained a suitable representation.

### 3.2.2 Second phase of ontology reuse

The main differences between the previous version of the model and the one shown in this section are the structural organization of the module *Clinical Evaluation* and the increased number of concepts in the module *Genetic Analysis*.

Table 3.2: Percentage of concepts in the HCM semantic model mapped to external ontologies.

| Module | Iteration | Vocabulary (%) | | | | | Total (%) |
|---|---|---|---|---|---|---|---|
| | | SNOMED-CT | NCIt | OCRe | SO | GRO | |
| *Clinical Evaluation* | 1 | 44 | 27 | 6 | - | - | 78 |
| | 2 | 42.9 | 42.9 | - | - | - | 85.8 |
| *Genotype Analysis* | 2 | - | 63.2 | - | 26.3 | 5.3 | 94.8 |

The percentages are indicated for the modules *Clinical Evaluation* and *Genotype Analysis*, and for each individual ontology. SNOMED-CT and the NCI Thesaurus (NCIt) were considered in both phases of ontology reuse, while OCRe only in the first (Iteration 1), the Sequence Ontology (SO) and the Gene Regulation Ontology (GRO) only in the second (Iteration 2). In Iteration 1, the module *Genotype Analysis* had only one concept mapped to OCRe.

Figures 3.5 and 3.6 show the current version of the module *Clinical Evaluation*, and Figure 3.7 of the module *Genotype Analysis*.

In this version of the module *Clinical Evaluation*, the class *Subject* was renamed *Person*, which has two subclasses: *Patient* and *Physician* (previously named *Clinician*). The class *Clinical History* was divided in two sibling classes: *Clinical Finding* and *Observable Entity*. The first contains clinical elements obtained either upon questioning or observation of the patient (e.g., *Angina* and *Cardiac Auscultation Finding*), while the second only upon observation (e.g., *Body Mesurement*(s) such as *Weight*, and *Pulse Rate*). The class *Procedure* has three new subclasses: *Management Procedure* (e.g., *Precription of Drug*); *Surgical Procedure* (e.g., *Myectomy*, which is the excision of a portion of muscle); and *Biomarker Analysis*, which was previously represented under *Laboratory Procedure*.

The module *Genetic Analysis* has a new class *Protein*, several new subclasses of *Mutation*, and the class *Amplification Primer* was renamed to *Primer*, with two new subclasses added.

As a result of the renaming of some of the classes, the mappings between the modules *Clinical Evaluation* and *Genetic Analysis* are now represented by the following relationships:

- ce:*Patient* ce:*hasBiologicalSample* ga:*Biological Sample*

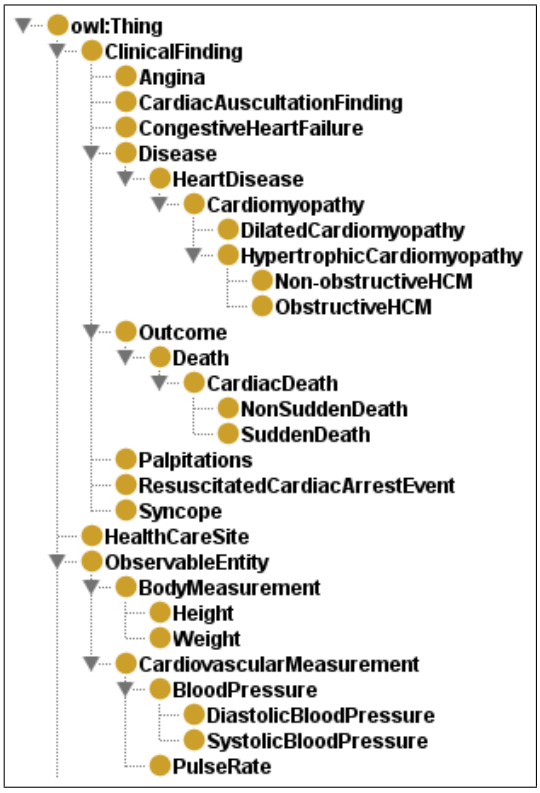


Figure 3.5: Hierarchical structure of the module *Clinical Evaluation* (Part I). Three of the top-level concepts (*Clinical Finding*, *Health Care Site* and *Observable Entity*) are shown, with all their sub-concepts visible.

- ce:*Biomarker Analysis* ce:*performedInBiologicalSample* ga:*Biological Sample*.

In terms of mappings to ontologies, SNOMED-CT was used in the *Clinical Evaluation* module, the NCIt in the *Clinical Evaluation* and *Genotype Analysis* modules, the Gene Regulation Ontology and the Sequence Ontology in the *Genotype Analysis* module (see Table 3.2). More precisely, each vocabulary was considered in the following top-level concepts:

- SNOMED CT: *Clinical Finding* and *Observable Entity*
- NCIt: *Health Care Site*, *Person* and *Procedure* (from *Clinical Evaluation*); *Biological Sample*, *Gene*, *Mutation* and *Protein* (from *Genotype Analysis*)
- Gene Regulation Ontology: *Nucleic Acid Molecule*

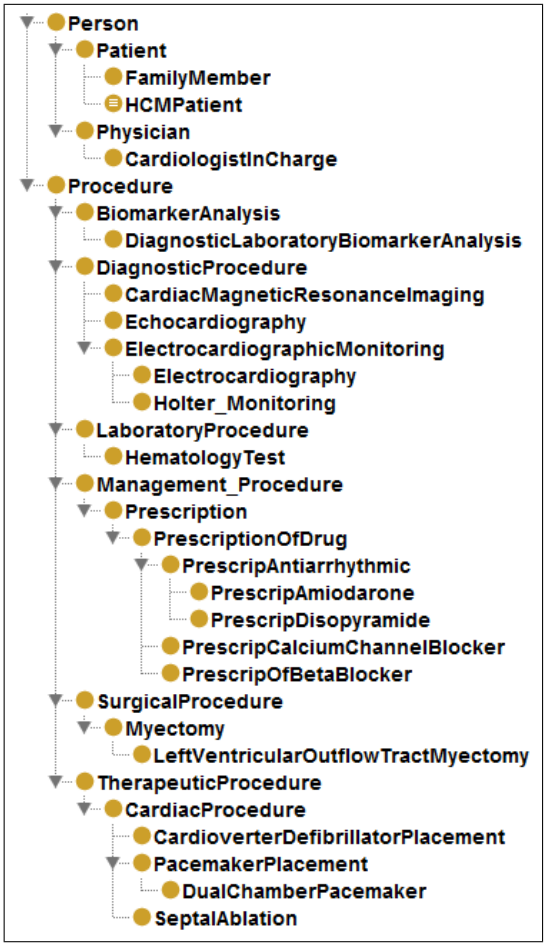


Figure 3.6: Hierarchical structure of the module *Clinical Evaluation* (Part II). Two of the top-level concepts (*Person* and *Procedure*) are shown, with all their sub-concepts visible. *Defined* classes (i.e., classes containing necessary and sufficient conditions) are indicated with the symbol ≡.

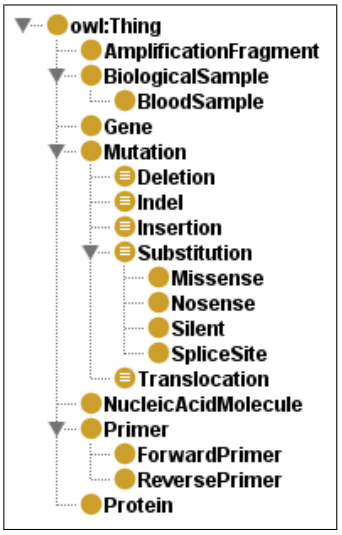


Figure 3.7: Hierarchical structure of the module *Genotype Analysis*. All seven top-level concepts are shown, with all their sub-concepts visible. *Defined* classes (i.e., classes containing necessary and sufficient conditions) are indicated with the symbol ≡.

- Sequence Ontology: *Primer*

The total number of properties shown in Table 3.1 for the modules *Clinical Evaluation* and *Genotype Analysis* includes both Data Properties (see Figures 3.8 and 3.11) and Object Properties (see Figures 3.9 and 3.10 for the module *Clinical Evaluation*, and Figure 3.12 for the module *Genotype Analysis*).

All the properties were defined in terms of domain and range, and were all used in the definition of class restrictions.

Figure 3.13 shows a detailed example of the annotations and the description of a class, *HCM Patient*.

## 3.3 Discussion

### 3.3.1 Modeling decisions

The HCM model was initially designed as a stand-alone module containing all concepts necessary to characterize a patient in terms of the disease, based on the activities involved in its prognosis. However, this approach was not compatible with the integration of the clinical elements with the molecular biology elements

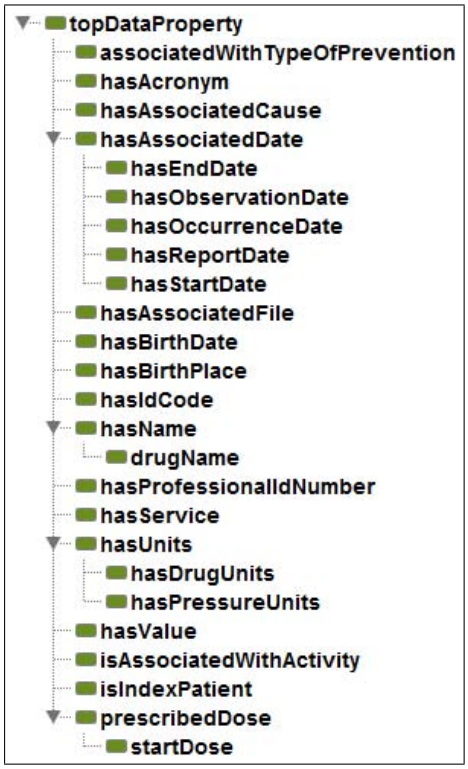


Figure 3.8: Data properties of the module *Clinical Evaluation*.

of the biomarker analysis. On the one hand, this analysis is considered as an exam by a medical doctor, and so it should be included under the concept *Procedure*. On the other hand, it has several associated concepts (e.g., *Gene* and *Mutation*) with information to be maintained, which are not used by the clinician. These different views and characterization needs of the data resulted in the division of the model in two modules: one comprising the data elements needed by a medical doctor to evaluate a patient in terms of HCM (*Clinical Evaluation*); and another one comprising the data elements needed by a molecular biologist to perform a biomarker analysis in a *Subject* sample (*Genotype Analysis*). In this manner, the latter data elements are suitably integrated with the clinical evaluation of the disease, and at the same time maintained as laboratory elements that can be managed independently of the *Subject* medical data.

The division of the model in modules results in three additional advantages: it facilitates the reuse of the model as a whole and of its modules individually;

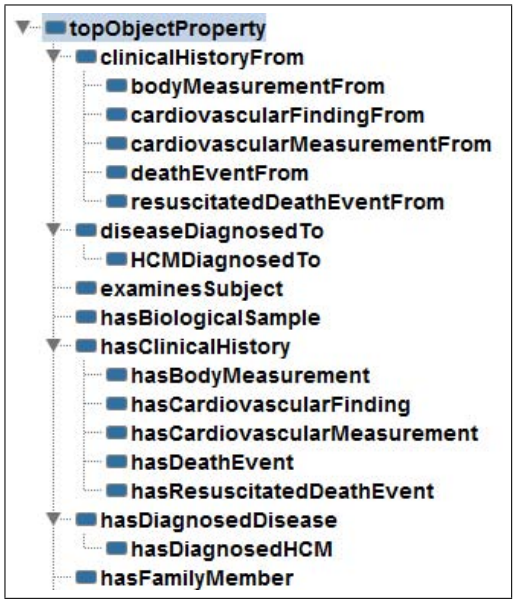


Figure 3.9: Object properties of the module *Clinical Evaluation* (Part I).

it facilitates the extension of the model, either with concepts pertaining to the same disease or any other; it permits different views of the data without added effort, useful to the different types of end-user - medical doctors and molecular biologists.

In respect to the reuse of existing ontologies, the first difficulty was the selection of which to use. The concepts of interest were searched in all the vocabularies available in BioPortal, and this was a challenging task since several vocabularies exist that fulfilled the requirement. After evaluating the most promising options, the initial list was progressively narrowed down.

During the ontology selection phase, it was necessary to decide whether to use one or more ontologies, and upon their selection, whether to reuse entire modules or individual concepts. I opted to use more than one ontology for each module for two reasons: (*i*) none of the vocabularies contained a complete list of the concepts of interest; (*ii*) the representation of the concepts in the vocabularies was not always the most suitable for my purposes. I opted to reuse individual concepts since my goal was not to convey the most complete representation of the disease, but rather to represent the concepts necessary for its prognosis, as well as include a minimum set of concepts that would facilitate the mapping between the semantic

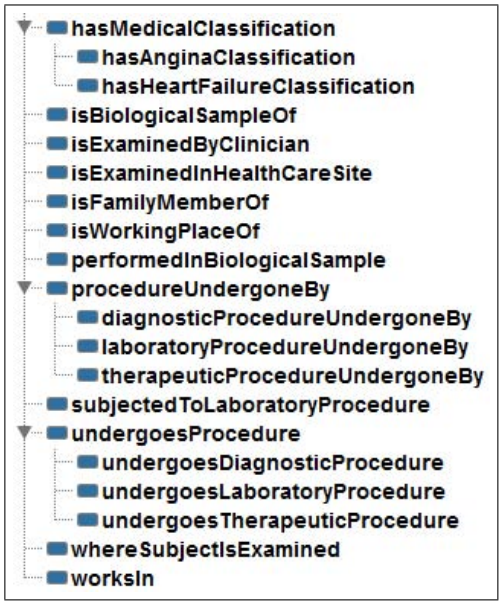


Figure 3.10: Object properties of the module *Clinical Evaluation* (Part II).

model and the vocabularies. One of the concerns throughout the development of the model was precisely to maintain it as simple as possible, in order to avoid overwhelming the biomedical end-users with superfluous information.

The reuse of existing ontologies proved to be advantageous at more than one level: it assisted in the identification of additional concepts and relations of interest; and it will facilitate the future addition of concepts since they can be searched in the vocabularies and easily integrated in their hierarchy.

The selection of the ontologies was based on their content and also on their structural organization. I searched for structures similar to the representation intended for the semantic model and that better conveyed the vision of the domain experts. However, the adoption of these structures was not always straightforward, namely in the case of the classes *Laboratory Procedure* and *Diagnostic Procedure*. In the first phase of ontology reuse, these classes were defined as siblings rather than the first as subclass of the second, even though the procedures considered under *Laboratory Procedure* can, in fact, be considered under *Diagnostic Procedure*(s). I advocate the organization proposed by NCIt insomuch as it separates procedures that involve the manipulation of a biological sample (*Laboratory Procedure*) from those that do not and are performed directly upon the

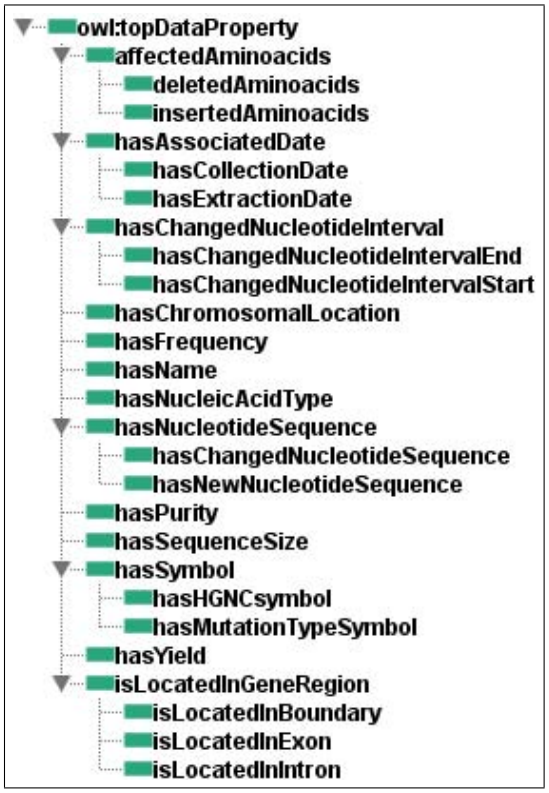


Figure 3.11: Data properties of the module *Genotype Analysis*.

subject as a whole (*Diagnostic Procedure*).

In addition to previous issues, several others emerged during the development of the model, as shown in the following list:

- **Absent concepts**: Inexistence of a concept of interest in the vocabulary.
- **Complexity**: Excess of concepts and of level of detail in general.
- **Placement**: Different possibilities concerning the placement of a concept in the hierarchy of the model.
- **Overlapping regions**: Existence of overlapping concepts/regions of interest on different vocabularies.
- **Absent textual definitions**: Inexistence of textual definitions for concepts of interest.

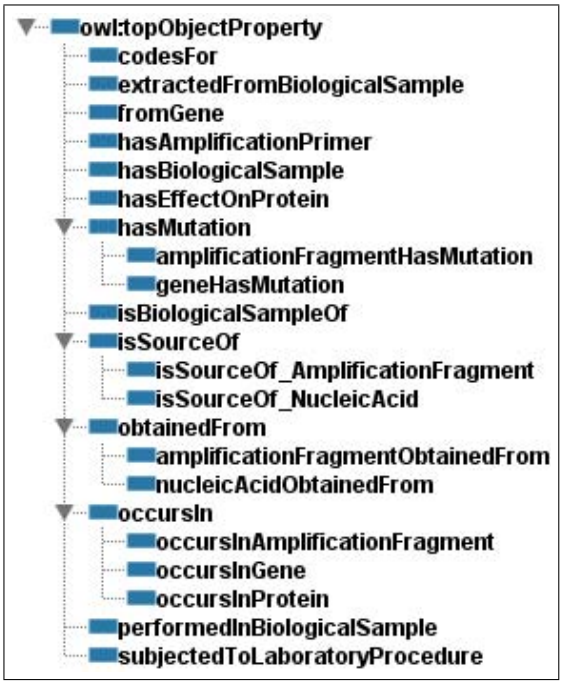


Figure 3.12: Object properties of the module *Genotype Analysis*.

The **Absent concepts** issue occurred in the modules *Clinical Evaluation* and the *Genotype Analysis*. In the former module, a concept *Cardiologist in charge* was needed to represent the cardiologist that is primarily responsible for the HCM patient. According to the specifications of the biomedical experts guiding the development of the model, this cardiologist is the only medical doctor associated with the patient for this disease, and is responsible for every data element and evaluation represented in the model. Neither SNOMED CT nor NCIt provide such a representation, and the notion of "Physician" and of specific medical specialties such as "Cardiologist" are represented under *Occupation*, which can be interpreted as a label rather than a representation of a person. In this situation, I opted to use the concept *Person* from NCIt to aggregate *Patient* and *Physician*, and added *Cardiologist in Charge* as a sub-concept of *Physician*. In the *Genotype Analysis* module, it was necessary to represent the *Translocation* and *Indel* sub-concepts of *Mutation* (as shown in Figure 3.7). While *Mutation* was mapped to the NCIt, this vocabulary does not include the indicated sub-concepts, and thus they were mapped to the Sequence Ontology.

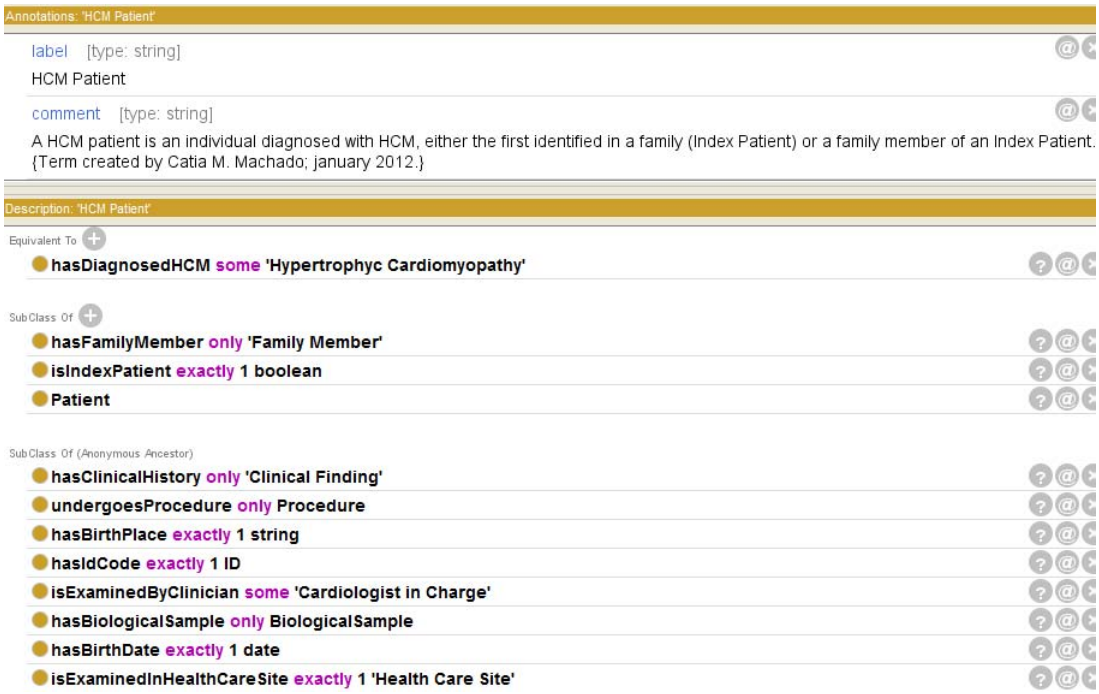


Figure 3.13: Class *HCM Patient* from the module *Clinical Evaluation*. All the details of the class are shown, including annotation and description statements.

The solution followed to deal with the **Complexity** of the controlled vocabularies, both in the form of number of concepts and detail of representation, was to consider only the concepts necessary for the description of the disease and for the structure of the model. The structure is particularly important for the mapping of the HCM model to external resources and for the future addition of concepts. An example of the complexity issue occurred with the concept *Procedure*. This concept is mapped to *Intervention or Procedure* from the NCIt, which contains thirteen sub-concepts, of which only five were considered. If all thirteen were used, the level of complexity of the model would be increased without any benefit for the end-users.

The **Placement** issue derived from the decision of not representing more than one parent per concept (i.e. multiparenting), even at the expense of a possible loss of detail. This decision was motivated by the intention of creating a model that would provide a straightforward experience to the biomedical experts when inputing or retrieving data, and thus avoid possible uncertainties due to multiple

options. As such, I was occasionally forced to evaluate different possibilities for the placement of a concept in the hierarchy of the model. This occurred with concepts from SNOMED CT, in which situations I recurred to the NCIt to help me identify a solution common to both vocabularies. One such case occurred with *Syncope*, a *Clinical Finding* that is represented in SNOMED CT as a sub-concept of three different concepts: *Clinical history and observation finding*, *Finding by site* and *Disease*. In the HCM model are considered the concept *Clinical Finding* and its sub-concept *Disease*, and the decision was whether to place *Syncope* directly under the first-level *Clinical Finding* or the second-level *Disease*. In NCIt the concept is represented directly under the concept *Finding* and not under its sibling *Disease or disorder*, and consequently the choice made was to place it under *Clinical Finding* in the HCM model. Similar decisions were made for the concepts *Angina* and *Congestive heart failure*, which are sub-concepts of *Finding by site* and *Disease* in SNOMED CT, and of *Finding* in NCIt.

The **Overlapping regions** issue results from the existence of more than one vocabulary describing the same domain of knowledge. According to the accepted OBO Foundry Smith *et al.* (2007) principle named "clearly delineated content" (FP005[1]), ontologies should be orthogonal to each other in order to enable the utilization of two different ontologies to define complementary perspectives on the same entities. In essence, I agree with this principle since the existence of a single ontology for a given domain would mean that anyone wanting to reuse it in an application semantic model would just have to follow it and consider the necessary knowledge. On the other hand, in light of my experience with the development of the HCM model, I consider that the availability of more than one vocabulary can be positive when no vocabulary is accepted as the single reference by the community.

An example of the overlapping regions in the *Clinical Evaluation* module occurred with the concept *Outcome*, a *Clinical Finding* with possible examples of outcomes being decreased pain and death. *Clinical Finding* and its sub-concepts are mapped to SNOMED CT, but this vocabulary represents *Death* in a high-level class *Event*, which is not necessary for the HCM model. Moreover, NCIt

[1] http://www.obofoundry.org/wiki/index.php/FP_005_delineated_content

has a concept *Outcome* under *Finding*, which also includes several sub-concepts relevant for the HCM model: *Death*, *Cardiac death*, *Sudden cardiac death* and *Non sudden cardiac death*. In this situation the decision was to consider *Outcome* and its sub-concepts from NCIt in the *Clinical Finding* concept, which is otherwise mapped to SNOMED CT.

Two other examples of the overlapping regions issue in the *Genotype Analysis* module involved the concepts *Primer* and *Nucleic acid molecule*. *Primer* is represented in the NCIt under *Drug, Food, Chemical or Biomedical Material* and without sub-classes. However, in the Sequence Ontology, a *Primer* is a *Sequence feature* with the two sub-classes *Forward Primer* and *Reverse Primer*, which were included in the HCM model. In the second situation, the concept *Nucleic Acid* was intended to represent actual nucleic acid molecules extracted from biological samples. While both the NCIt and the Sequence Ontology include the concept *Nucleic Acid*, neither define it suitably for our purposes: the former defines *Nucleic acids* as "A family of macromolecules", whereas the latter defines *Nucleic acid* as "An attribute describing a sequence consisting of nucleobases bound to repeating units". Consequently, I opted to use the Gene Regulation Ontology exclusively for its concept *Nucleic acid molecule*, which is more suitably defined as a "A complex, high-molecular-weight biochemical macromolecule composed of nucleotide chains that convey genetic information".

The overlapping regions issue is of particular importance given that using a domain representation that is unfamiliar to the end-users of the HCM prognosis framework may hinder significantly their acceptance of the framework.

The **Absent textual definitions** issue was perceived as a significant burden to the reuse of the affected concepts, since there where situations in which their intended use was not readily understandable. This was a common problem when using SNOMED CT, as this vocabulary lacks definitions for most of its concepts. For example when representing the concept *Cardiologist in Charge*, it was only possible to interpret the intended use of the concept *Cardiologist* based on the hierarchical organization of the vocabulary. By contrast, the NCIt has available detailed descriptions for the majority of its concepts, which provides a greater assistance when more complex decisions have to be made. This issue is not new,

and has already been the subject of an OBO Foundry principle (FP 006 textual definitions[1]).

An issue particular to the development of the *Genotype Analysis* module occurred with the concepts *Nucleic acid molecule*, *Gene* and *Protein*. As represented in the Gene Regulation Ontology, these concepts are related with each other: *Gene* is represented under *DNA*, which in turn is a *Nucleic acid*; and *Nucleic acid* and *Protein* are both *Information biopolymer*(s) ("macromolecules that harbor biological information in their structures"). However, these relationships could not be conveyed in the HCM because what we want to represent under each concept is conceptually different: *Nucleic Acid Molecule*, the physical molecules; *Gene*, the list of genes associated with HCM (not the physical genes); and *Protein*, the list of proteins encoded by the genes associated with HCM (not the physical proteins).

### 3.3.2 Participation in the semantic web

The semantic model was built to be a part of the semantic web. On the one hand, it was built in accordance to the guidelines defined for ontologies (one of the pillars of a semantic web approach) and with the same rigor, lacking only a formal evaluation. This evaluation was not done solely because it was not one of the objectives of this work, but rather a future work. On the other hand, it was built in accordance to four principles that accelerate data integration and its exploration, in particular when following a semantic approach (Machado *et al.*, 2013b):

1. Represent data and knowledge with technologies that serve as a standard across the entire community.

2. Define mappings between resources.

3. Provide access to the resources so they can be integrated.

4. Share the effort of resource integration among data providers and data users.

[1] http://www.obofoundry.org/wiki/index.php/FP_006_textual_definitions

The model was developed in OWL, and any data mapped to or modeled according to this model is in the RDF format.

The two main modules were mapped to external resources, which permits that both tools and people traverse from one resource to the other.

The model is available to the general public from https://sites.google.com/site/hcmsemanticmodel/home-1. The actual data modeled is not available since it is private data from patients. Since the model has open-access, anyone can use it, as well as alter it to add more concepts, relations or mappings to additional resources.

### 3.3.3 Integration of translational medicine data

The creation of a new semantic model was necessary in my work since no publicly available ontology existed for the case-study.

The model fulfills the first objective of my thesis as it effectively represents and integrates the clinical and genetic data necessary for the characterization of a disease and its prognosis process, exploiting existing ontologies.

The representation of the data according to the model results in the following advantages:

- The data is automatically converted to a language that is independent of the original domain of knowledge. This facilitates the integration of heterogeneous data sources, and the integration with data from any other application that uses the same language.

- The data is automatically integrated through the relations defined between concepts, relations that also exist between the modules representing clinical and genetic data.

In terms of the role the semantic model has in the improvement of the current knowledge about a disease and its diagnosis and prognosis, it occurs essentially at two levels: at the level of the knowledge explicitly stated; and at the level of the inferred knowledge.

Knowledge explicitly stated is any declared information. Examples of declared information are the following statements (written according to the model):

"DC001 isA HCM Patient"; "DC001 undergoesDiagnosticProcedure P11"; "P11 isA Ecochardiography". Based on this type of statement, it is possible to query the modeled data to retrieve the mutations a patient has; if there is any relation between certain mutations and any of the elements tested in a *Hematology Test* (e.g., total cholesterol); if there is any relation between certain mutations and the type or dosage of the prescribed drugs. Since the basic unit of the module *Clinical Evaluation* is the patient and the basic unit of the module *Genotype Analysis* is the genetic test performed in a patient's biological sample, a query can be made to test the relation between virtually any concept in the two modules.

On the other hand, inferred knowledge is any knowledge that is not explicitly stated but that can be automatically inferred based on the knowledge explicitly stated. Inference normally results in the attribution of a class to an instance, which is only possible due to the explicit definition of the concepts, their properties and their relations with other concepts. One of the situations in which inference is possible is when a class has a necessary and sufficient condition, that is, when to be part of that class an instance has to absolutely verify that condition. In the module *Clinical Evaluation*, the class *HCM Patient* is an example of such a situation, in which the condition is to be diagnosed with *Hypertrophic Cardiomyopathy* (stated in "hasDiagnosedHCM some HypertrophicCardiomyopathy", see Figure 3.13). Based on this condition, the class *HCM Patient* is automatically attributed to the subject "DC001" in the following explicit statement: "DC001 hasDiagnosedHCM ObstructiveHCM" (note that "ObstructiveHCM isA HypertrophicCardiomyopathy"). Another situation is when a property is defined in terms of domain and range. In the HCM semantic model, the domain of the property *undergoesProcedure* was defined as the class *Patient*, whereas the range as the class *Procedure*. Based on this definition, the statment "DC001 undergoesProcedure LaboratoryProcedure" results in the inference that "DC001" is of type *Patient* and "LaboratoryProcedure" is of type *Procedure*.

Performing inference is a powerful tool. Since it was not one of my objectives in this work, its capabilities were not thoroughly explored, which leaves ample space for future work possibilities. Nonetheless, the consistency checks done throughout the development of the model ensure that it can be used for inference purposes in its current state.

Whether the model is exploited in terms of explicit or inferred knowledge, ultimately any advancement in the prognosis and diagnosis processes translates itself in new knowledge about the disease.

The decisions made throughout the development of the model took into consideration that the final model should be usable for diseases other than its case-study, and consequently the methodology as well as the individual decisions can also be exploited in other translational medicine contexts.

Despite the usefulness of the semantic model in all the aspects discussed, it cannot itself mitigate the effects of problematic factors in the analysis of the modeled data. Examples of those factors are a small number of instances, a small amount of data collected, or the existence of missing values, all of which are present in the dataset of the model. These factors are not uncommon on translational medicine datasets, and it was thus interesting to study if the enrichment of a dataset with knowledge from ontologies with a greater range of information could improve its quality for data exploration purposes.

# Chapter 4

# Knowledge representation for data exploration

This chapter describes the exploitation of ontologies for the improvement of translational medicine datasets.

At the center of the methodology developed to perform that exploitation is the technique called enrichment analysis, which in this work was adapted from the standard implementation in enrichment analysis tools, to be able to used data from patients instead of genes.

The enrichment methodology presented in this chapter is the first step in the development of a two-part prognosis approach, conceived to assist medical doctors in the evaluation of patients in respect to the likelihood of suffering a disease-related event or having a specific disease manifestation (see Figure 4.1). The second part of the prognosis approach consists in performing a classification step with the results obtained in the first to perform a prognosis prediction.

In this work, the use of classification algorithms serves a dual purpose: from a medical perspective, it assists in the evaluation of the patients; and from an informatics perspective, it evaluates if the inclusion of the enriched ontology terms in the datasets improves their predictive capabilities.

Two datasets were used as case-study: one of patients with the disease hypertrophic cardiomyopathy (HCM, presented in Chapter 3), with focus on the occurrence of a specific event - sudden cardiac death (SCD); another of patients

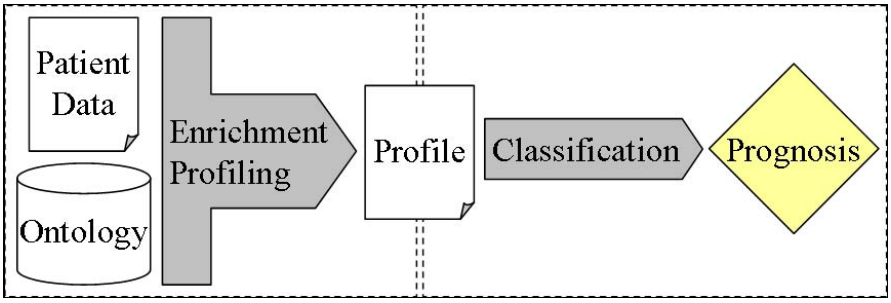


Figure 4.1: Schematic representation of the prognosis methodology. The prognosis methodology has two parts. The first part (on the left-side), receives as input data from patients that is mapped to biomedical ontologies. It performs an enrichment analysis to identify a list of ontology terms considered to be enriched, which are used to create profiles for individual patients. These profiles are then subjected to an evaluation step (the second part of the methodology, on the right-side) that results in the evaluation of the prognosis for the patients.

with chronic obstructive pulmonary disease, with focus in one of the manifestations of the disease - emphysema. Both diseases are characterizable with clinical and genetic data, and this data is analyzed in terms of the enrichment of ontology terms. The clinical data in both datasets includes features such as symptoms and measurements, whereas the HCM genetic data refers to the presence or absence of mutations. The genetic data associated with chronic obstructive pulmonary disease was not used in this work since it was obtained for groups of patients instead of individual patients.

The rest of the chapter is divided in two parts: in the first is described the work pertaining to the enrichment analysis; and in the second, the evaluation of the enrichment results done with data mining algorithms.

## 4.1 Enrichment analysis

In this section is described the enrichment methodology designed to analyze genetic and clinical data from patients (adapted from a Singular Enrichment Analysis approach), followed by the application of the methodology to the analysis of the two datasets of patients: HCM, in respect to the occurrence of SCD; and COPD, in respect to the presence of emphysema.

### 4.1.1 Methods

#### 4.1.1.1 Genetic data analysis

Two enrichment analyses were devised for analysis of genetic data from patients: an enrichment profiling and a differential enrichment.

Both patient analyses differ from the standard enrichment analysis in the two following aspects:

- In the standard analysis, only one set of genes is analyzed, such as the genome of an organism. In the patient analysis, several sets of genes are taken into consideration, exactly one set for each patient.
- In the standard analysis, the frequency of annotation of a term is given by the number of genes annotated with that term. In the patient analysis, the frequency of annotation is given by the number of mutated genes annotated with the term.

**Enrichment profiling**

The purpose of this analysis is to characterize the genotype of a group of patients (e.g., the patients positive for a disease-related event), based on the set of mutations they have. Since the knowledge of these mutations is normally not available for the complete genome of a patient but only for a set of genes associated with the disease under analysis, the characterization is performed by comparing the information of the genes mutated in the patients with the complete set of genes in the genome. Genes associated with the disease but not tested are treated as genes without mutations just as happens with the genes not associated with the disease.

Given a group of patients, for each of which is known his/her set of mutations, and the set of Human protein-coding genes, the enrichment profiling analysis is performed as follows:

- Define the population set as the union of the genes in the genome of all the patients.

- Define the study set as the union of the genes mutated in all the patients (see Figure 4.2).

- Find all ontology terms annotating at least one gene mutated in the patients.

- Calculate the population set frequency of annotation (PFreq) of term $t$ as follows:

$$PFreq(t) = \sum_{1}^{n} count(gene(t)) \tag{4.1}$$

  where $n$ is the total number of patients, and *gene(t)* is a gene annotated with $t$ (see Figure 2).

- Calculate the study set frequency of annotation (SFreq) of term $t$ as follows:

$$SFreq(t) = \sum_{1}^{n} count(mut\_gene(t)) \tag{4.2}$$

  where *mut_ gene(t)* is a mutated gene annotated with $t$.

- Apply Fisher's exact test to calculate the probability of enrichment of term $t$.

- Perform a multiple-testing correction (e.g., Bonferroni) over the $p$-values obtained.

- Consider term $t$ as enriched in the study set if $p$-value$(t) < \alpha$ (e.g. 0.05 or 0.1).

In this methodology, the inputs are: the set of genes in the Human genome; the set of genes mutated in the patients; and the set of ontology terms annotating the genes. The output is the list of ontology terms and their respective $p$-values (not corrected and/or corrected).

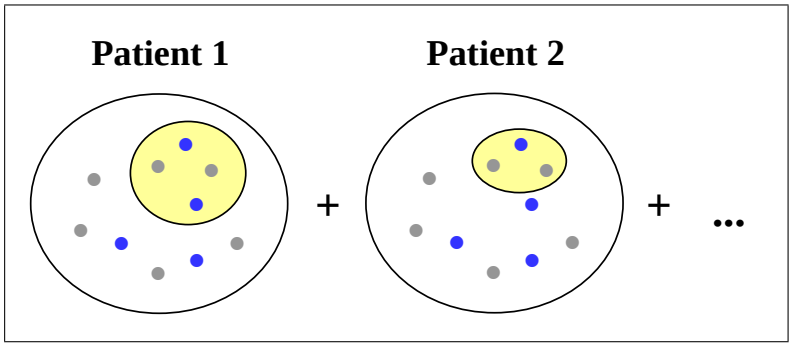


Figure 4.2: Representation of the population and study sets in the enrichment profiling analysis. The two sets of dots represent the genome of two patients, from the same group (e.g., with SCD). The smaller, yellow set of dots, corresponds to the genes mutated in the patient; the larger, white set of dots, corresponds to the entire genome of the patient: genes not mutated (outside the yellow set) and genes mutated. In these sets of genes, blue dots represent genes annotated with a term of interest ($t$); gray dots represent genes not annotated with $t$. In the profiling analysis, the study set is the union of the genes mutated in all the patients; the population set is the union of the genome of all the patients. The annotation frequency is then calculated by counting the total number of genes annotated with the term in the study set (study frequency) and in the population set (population frequency).

### Differential enrichment

The purpose of this analysis is to identify differentiating features between a group of patients with a particular characteristic, for example being positive for a disease-related event, and all the patients with the disease. This analysis is also based in the set of mutations the patients have, considering the mutations in the study group vs. the mutations in all the patients.

Given a group of patients with a disease, a sub-group of those patients with a study characteristic, and the set of mutations in each group:

- Define the population set as the union of the genes mutated in the group of patients with the disease.
- Define the study set as the union of the genes mutated in the sub-group of patients with the study characteristic.
- Find all ontology terms annotating at least one gene mutated in the sub-group of patients.

- Calculate the population set frequency of annotation (PFreq) of term $t$ as follows:

$$PFreq(t) = \sum_{1}^{n} count(mut\_gene(t)) \tag{4.3}$$

  where $n$ is the total number of patients with the disease, and *mut_ gene(t)* is a mutated gene annotated with $t$.

- Calculate the study set frequency of annotation (SFreq) of term $t$ as follows:

$$SFreq(t) = \sum_{1}^{n} count(mut\_gene(t)) \tag{4.4}$$

  where $n$ is the number of patients in the sub-group with the study characteristic.

- Apply Fisher's exact test to calculate the probability of enrichment of term $t$.

- Perform a multiple-testing correction (e.g., Bonferroni) over the $p$-values obtained.

- Consider term $t$ as enriched in the study set if $p$-value$(t) < \alpha$ (e.g. 0.05 or 0.1).

In this methodology, the inputs are: the set of genes mutated in the patients; and the set of ontology terms annotating the genes. As before, the output is the list of ontology terms and their respective $p$-values.

### Genetic data analysis of HCM

Four enrichment experiments were performed, two enrichment profiling analyses and two differential enrichment analyses, with the Gene Ontology. These experiments are described below in accordance with the steps previously indicated for each analysis.

Enrichment profiling for the group of patients positive for SCD:

- Population set: 18,759 genes x 14 patients
- Study set: 16 mutated genes, corresponding to 4 distinct genes
- GO terms obtained for the previous 4 genes
- PFreq: number of genes annotated with $t$ in the 14 patients
- SFreq: number of mutated genes annotated with $t$ in the 14 patients

Enrichment profiling for the group of patients negative for SCD (from now on referred to as no-SCD):

- Population set: 18,759 genes x 69 patients
- Study set: 100 mutated genes, corresponding to 7 distinct genes
- GO terms obtained for the previous 7 genes
- PFreq: number of genes annotated with $t$ in the 69 patients
- SFreq: number of mutated genes annotated with $t$ in the 69 patients

Differential enrichment for the HCM patients and the sub-group of SCD patients:

- Population set: 116 mutated genes, corresponding to the 7 distinct genes mutated in the 83 patients
- Study set: 16 mutated genes, corresponding to the 4 distinct genes mutated in the 14 SCD patients
- GO terms obtained for the 7 genes
- PFreq: number of mutated genes annotated with $t$ in the 83 patients
- SFreq: number of mutated genes annotated with $t$ in the 14 SCD patients

The differential enrichment for the HCM patients and the sub-group of no-SCD patients differs only from that of the SCD patients in the two following points:

- Study set: 100 mutated genes, corresponding to the 7 distinct genes mutated in the 69 no-SCD patients
- SFreq: number of mutated genes annotated with $t$ in the 69 no-SCD patients

In all the analyses a Bonferroni correction was performed, and 0.1 was the confidence level considered.

#### 4.1.1.2 Clinical data analysis

The enrichment analysis devised for the clinical data corresponds to the differential analysis described above for the genetic data. As such, its purpose is the identification of differentiating features between a group of patients with a particular characteristic and all the patients with the disease. The analysis is based on the clinical features themselves in the case of boolean and numeric features, and on the values of the features in the case of categorical features.

This clinical differential analysis differs from the standard enrichment analysis in the form how each instance is identified to be annotated with a term: in the standard analysis, the genes are directly annotated with the ontology term; in the patient analysis, the patients are indirectly annotated with the ontology term through their clinical features. More concretely, a patient is considered to be annotated with a given term if at least:

- One Boolean feature annotated with that term has a positive value
- One value of a categorical feature is annotated with the term
- One numerical feature annotated with that term has a known value

Given a group of patients with a disease, a sub-group of those patients with a study characteristic, and the set of clinical features in each group, the clinical differential analysis is performed as follows:

- Define the population set as the group of patients with the disease.
- Define the study set as the sub-group of patients with the study characteristic.

- From the set of clinical features, exclude those used to define the classification of the patients (i.e., the features used to verify if they are positive or negative for the study characteristic).
- Find all ontology terms annotating at least one feature or one value of a feature present in the set defined in the previous point.
- Calculate the population set frequency of annotation (PFreq) of term $t$ as follows:

$$PFreq(t) = count(patient(t)) \tag{4.5}$$

  where *patient(t)* is a patient with the disease annotated with $t$.
- Calculate the study set frequency of annotation (SFreq) of term $t$ in the same manner as PFreq, but only for the sub-group of patients with the study characteristic.
- Apply Fisher's exact test to calculate the probability of enrichment of term $t$.
- Perform a multiple-testing correction (e.g. Bonferroni) over the $p$-values obtained.
- Consider term $t$ as enriched in the study set if $p$-value$(t) < \alpha$ (e.g. 0.05 or 0.1).

In this methodology, the inputs are: the set of patients, with their respective values for each clinical feature; and the set of ontology terms annotating the clinical features. As happened on the genetic analyses, the output is the list of ontology terms and their respective $p$-values.

### Clinical analysis of the disease datasets

A total of four clinical enrichment experiments was performed, conidering the following datasets and the respective group of patients:

- HCM and the sub-group of patients with SCD.
- HCM and the sub-group of patients without SCD.
- COPD and the sub-group of patients with emphysema.
- COPD and the sub-group of patients without emphysema.

The differential enrichment of the group of HCM patients and the sub-group of SCD patients was performed as follows:

- Population set: 83 patients with HCM
- Study set: 14 patients with HCM and positive for SCD
- Ontology terms obtained for all the clinical features with known value for at least one patient in the population set
- PFreq: number of patients in the population annotated with term $t$
- SFreq: number of patients in the study set annotated with term $t$

The other differential enrichments were performed in the same manner, but considering the respective population and study set sizes: 83 and 69 for HCM and the sub-group of patients with SCD; 155 and 32 for COPD and the sub-group of patients with emphysema; 155 and 123 for COPD and the sub-group of patients without emphysema.

#### 4.1.1.3 Datasets

**Hypertrophic cardiomyopathy**

The HCM dataset is composed by clinical and genetic features characterizing 83 patients, which was previously collected from Portuguese hospitals and molecular biology research laboratories. From these 83 patients, 14 are positive for SCD and the remaining 69 are negative for SCD. Table 4.1 characterizes the clinical dataset in terms of the distribution of features in the two classes of patients, Table 4.2 indicates the percentage of patients with known values for each

Table 4.1: Characterization of the HCM clinical features in terms of their distribution in the two patient classes: SCD and no-SCD.

| Clinical feature | Feature value | SCD (%) | no-SCD (%) |
|---|---|---|---|
| **Sudden death (SD)** | True | 5 (36) | 0 |
| | False | 9 (64) | 69 (100) |
| **Resuscitated SD** | True | 3 (21) | 0 |
| | False | 8 (57) | 69 (100) |
| **Cardioverter defibrillator** | True | 9 (64) | 0 |
| | False | 2 (14) | 69 (100) |
| **Non-sudden death** | True | 0 | 0 |
| | False | 14 (100) | 69 (100) |
| **Obstructive HCM** | True | 4 (29) | 8 (12) |
| | False | 1 (7) | 17 (25) |
| **Non-obstructive HCM** | True | 1 (7) | 17 (25) |
| | False | 4 (29) | 8 (12) |
| **SD family history** | True | 3 (21) | 1 (1) |
| | False | 2 (14) | 25 (36) |
| **HCM form** | Familial | 9 (64) | 32 (46) |
| | Sporadic | 2 (14) | 37 (54) |
| **Hypertrophy morphology** | Apical | 4 (29) | 8 (12) |
| | Septal | 5 (36) | 16 (23) |
| | Concentric | 0 | 3 (4) |
| | Concentric/Septal | 0 | 5 (7) |
| **Blood pressure** | Normal | 4 (29) | 22 (32) |
| | Hypotension | 0 | 1 (1) |
| | Hypertension | 0 | 5 (7) |
| **Gender** | Male | 6 (43) | 41 (59) |
| | Female | 5 (36) | 25 (36) |
| **Age** | [0,20] | 0 | 5 (7) |
| | ]20,40] | 2 (14) | 11 (16) |
| | ]40,60] | 3 (21) | 15 (22) |
| | >60 | 3 (21) | 10 (14) |

Indicated for each feature are its possible values, the number of SCD and no-SCD patients for each value and the respective percentages.

feature, and Table 4.3 shows the number of SNOMED-CT and NCIt annotations for each feature.

Table 4.2: Characterization of the HCM clinical features in terms of the percentage of SCD and no-SCD patients with known values.

| Clinical feature | SCD | no-SCD |
|---|---|---|
| **Sudden death (SD)** | 100 | 100 |
| **Resuscitated SD** | 79 | 100 |
| **Cardioverter defibrillator** | 79 | 100 |
| **Non-sudden death** | 100 | 100 |
| **Obstructive HCM** | 36 | 36 |
| **Non-obstructive HCM** | 36 | 36 |
| **SD family history** | 36 | 38 |
| **HCM form** | 79 | 100 |
| **Hypertrophy morphology** | 65 | 46 |
| **Blood pressure** | 29 | 41 |
| **Gender** | 79 | 96 |
| **Age** | 57 | 61 |

From the total set of clinical features, the following three were used to define which patients are positive for SCD: *sudden death*, *resuscitated sudden death*, and *cardioverter defibrillator*. The first two indicate if the patient suffered a sudden cardiac death, either resuscitated or not, whereas the third indicates if the patient has an implanted cardioverter defibrillator. This device prevents the occurrence of SCD by delivering an electric charge when cardiac arrhythmia is detected, and it is implanted after a resuscitated sudden death occurred or when there is a very high risk of SCD occurrence. Patients are then considered positive for SCD if they are positive for at least one of the three features. Considering the three features instead of just two resulted in an increase of 4 SCD positive patients.

The genetic features are the mutations associated with the disease, which are represented as Boolean variables. From the total 569 mutations associated with the disease, only 78 were found in the HCM patients. These occur in 7 distinct genes (shown in Table 4.4), all of which are mutated in at least one no-SCD patient. The SCD patients show mutations in only 4 of those 7 genes: MYBPC3, MYH7, CSRP3, and TNNT2. The number of mutations identified per patient ranges from 1 to 5, with an average of 1.8.

Table 4.3: Characterization of the HCM clinical features in terms of their annotation with SNOMED-CT and NCIt.

| Clinical feature | Feature value | Annotations | |
|---|---|---|---|
| | | Snomed CT | NCI Thesaurus |
| Sudden death (SD) | - | 11 | 0 |
| Resuscitated SD | - | 11 | 0 |
| Cardioverter defibrillator | - | 6 | 0 |
| Non-sudden death | - | 23 | 0 |
| Obstructive HCM | - | 0 | 0 |
| Non-obstructive HCM | - | 0 | 12 |
| SD family history | - | 26 | 0 |
| HCM form | Familial | 0 | 19 |
| | Sporadic | 0 | 0 |
| Hypertrophy morphology | Apical | 12 | 0 |
| | Septal | 4 | 0 |
| | Concentric | 7 | 0 |
| | Concentric/Septal | 0 | 0 |
| Blood pressure | Normal | 18 | 0 |
| | Hypotension | 27 | |
| | Hypertension | 31 | |
| Gender | Male | 22 | 0 |
| | Female | 22 | |
| Age | - | 0 | 0 |

The genotyping of the patients was done in two manners: with a microarray able to detect 508 mutations associated with HCM, and high-resolution melting analysis (HRM) (Wittwer *et al.*, 2003) followed by sequencing. The HRM analysis was used to analyze individual exons to identify the presence of mutations, whereas the sequencing permits the identification of the exact mutation. Some of the patients were analyzed with both techniques, whereas others with only one of the techniques. HRM can be used to test for mutations not present in the microarray and/or to confirm the results obtained with the microarray. One of the reasons to use HRM instead of the microarray is that when patients are tested after a family member was diagnosed, only the mutations found in this one are searched for. Additionally, the identification of only one mutation is sufficient for

Table 4.4: Genes used for the genetic characterization of the HCM patients.

| Gene | SCD | no-SCD | GO annotations |
|---|---|---|---|
| MYBPC3 | 4 | 25 | 202 |
| MYH7 | 9 | 36 | 192 |
| CSRP3 | 1 | 4 | 138 |
| TNNT2 | 2 | 20 | 178 |
| TNNI3 | 0 | 13 | 173 |
| MYL2 | 0 | 1 | 133 |
| MYH6 | 0 | 1 | 251 |

Indicated for each gene are the number of SCD and no-SCD patients with at least one mutation in it, and the number of GO annotations.

a positive diagnosis, and the overall process is cheaper.

### Chronic obstructive pulmonary disease

Chronic obstructive pulmonary disease (COPD) is a common disease characterized by persistent airflow limitation. It is usually progressive and associated with an enhanced chronic inflammatory response in the airways and the lung to noxious particles or gases (such as cigarette smoke) (GOLD). Approximately 2.7 million deaths were caused by COPD in 2000, establishing this disease as the fourth leading cause of death in the world (Alexandre, 2011).

The chronic airflow limitation characteristic of COPD is caused by a mixture of small airways disease and emphysema. The latter is one of the several structural abnormalities present in patients with COPD, and consists in the destruction of the gas-exchanging surfaces of the lungs (alveoli) (GOLD).

COPD results from gene-environment interactions, being cigarette smoking the best studied risk factor. Studies show that among people with the same smoking history, not all will develop COPD due to differences in genetic predisposition to the disease, or due to how long the person lives (since living longer will allow a greater lifetime exposure to the risk factors). Additionally, these factors may also be related in more complex and subtle ways. For example, gender may

influence whether a person starts smoking or is subjected to specific occupational or environmental exposures (GOLD).

The COPD dataset contains 155 patients, 32 positive for emphysema and 123 negative. For this analysis, only the clinical set of features was considered, since it was the only one available for individual patients. Table 4.5 characterizes the clinical dataset in terms of the distribution of the features in the two classes of patients, Table 4.6 shows the percentage of patients with known values for each feature, and Table 4.7 shows the number of SNOMED-CT and NCIt annotations for each feature.

Table 4.5: Characterization of the COPD clinical features in terms of their distribution in the two patient classes: emphysema and no-emphysema.

| Clinical feature | Feature value | Emphysema (%) | no-Emphysema (%) |
|---|---|---|---|
| Gender | Male | 13 (41) | 78 (63) |
| | Female | 19 (59) | 45 (37) |
| GOLD stage | I | 10 (31) | 30 (24) |
| | II | 17 (53) | 37 (30) |
| | III | 0 | 0 |
| | IV | 0 | 0 |
| Race | Black or African American | 3 (9) | 5 (4) |
| | Asian | 0 | 1 (1) |
| | Caucasian/White | 29 (91) | 117 (95) |
| | American Indian or Alaska Native | 0 | 0 |
| Smoker | True | 11 (34) | 62 (50) |
| | False | 21 (66) | 60 (49) |
| Emphysema | True | 32 (100) | 0 |
| | False | 0 | 123 (100) |

In this table are shown only categorical and Boolean features. For each feature are indicated its possible values, the number of patients with and without emphysema for each value and the respective percentages. The total number of patients with emphysema and without emphysema is 32 and 123, respectively.

From the features shown on Table 4.6, FEV1 stands for forced expiratory volume in one second (%) and FVC for forced vital capacity (%), which is the total

Table 4.6: Characterization of the COPD clinical features in terms of the percentage of patients with and without emphysema with known values.

| Clinical feature | Emphysema | no-Emphysema |
|---|---|---|
| Gender | 100 | |
| GOLD stage | 84 | 54 |
| Race | 100 | |
| Smoker | 100 | 99 |
| Emphysema | 100 | |
| Age | 100 | |
| Height | 100 | |
| Weight | 100 | |
| Pre-FEV1 | 100 | 99 |
| Post-FEV1 | 100 | |
| Pre-FEV1/FVC | 100 | 99 |
| Post-FEV1/FVC | 100 | |
| W_Aperc_Mean_Large | 100 | |
| W_Aperc_Mean_Medium | 100 | |
| W_Aperc_Mean_Small | 100 | |
| W_Aperc_RAB | 100 | |
| Frac_950 | 100 | |
| Frac_910 | 100 | |
| TNF_A_40 | 100 | |
| IL_6_12 | 100 | |
| IL_13_26 | 100 | |
| W1_Distance | 100 | 99 |
| CCP | 99 | 99 |
| SPD | 100 | 100 |
| NT_Pro_BNP | 91 | 80 |
| PRED_DLCO | 100 | |
| Best_DLCO | 100 | |
| BEST_PRED_DLCO | 100 | |

amount of air blown out in one breath. The prefixes "Pre" and "Post" indicate that the respiratory test was done, respectively, before and after the administration of a bronchodilator. The features Frac_950 and Frac_910 are measures of the destruction of the pulmonary alveoli obtained from computed tomography scans, and TNF_A_40, IL_6_12 and IL_13_26 are inflammation indicators. W1_Distance is a measure of the distance the patient can walk per unit of time; CCP stands for clara-cell protein 16, a protein that appears to protect the respiratory tract against oxidative stress and inflammation (Broeckaert & Bernard, 2000); SPD stands for pulmonary clara cell, a circulating surfactant protein; and D_LCO is the carbon monoxide diffusion capacity, which is a measure of gas transfer across the alveolar capillary membrane (Fitting, 2000).

Table 4.7: Characterization of the COPD clinical features in terms of their annotation with SNOMED-CT and NCI Thesaurus.

| Clinical feature | Feature value | Annotations | |
|---|---|---|---|
| | | Snomed | NCIt |
| Gender | Male | 22 | 0 |
| | Female | 22 | |
| Race | Black or African American | 4 | 0 |
| | Asian | 4 | 0 |
| | Caucasian/White | 13 | 0 |
| | American Indian or Alaska Native | 4 | 0 |
| Smoker | | 21 | 0 |
| Emphysema | | 23 | 0 |
| Height | | 0 | 3 |
| Weight | | 0 | 4 |
| Pre-FEV1 | | 10 | 0 |
| Post-FEV1 | | 10 | 0 |
| TNF_A_40 | | 0 | 5 |
| IL_6_12 | | 33 | 0 |
| IL_13_26 | | 0 | 6 |
| PRED_DLCO | | 15 | 0 |
| Best_DLCO | | 15 | 0 |
| BEST_PRED_DLCO | | 15 | 0 |

Features without annotations are not shown.

#### 4.1.1.4 Ontology annotations

A total of three ontologies were used to test the adaptation of enrichment analysis to translational medicine datasets: one genetic, the Gene Ontology (as of the release of October 4th, 2012) (Ashburner *et al.*, 2000); and two clinical, the National Cancer Institute Thesaurus (as of the release of June 6th, 2013) (Sioutos *et al.*, 2007) and the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT, as of the release of May 8th, 2013)[1].

All these ontologies are available in BioPortal, and its Web service Annotator was used to retrieve the ontology terms annotating the features in the dataset. The service was used with the default settings[2], with the exception of the following: *levelMax*, set to '999'; *longestOnly* and *isVirtualOntologyId*, both set to 'true'. The first setting indicates that both direct and indirect annotations (up to the root of the ontology) are retrieved. The second, *longestOnly*, indicates that the annotations retrieved match the longest term matching phrase. In the example of the query term "concentric hypertrophy" this means that annotations for "concentric" or "hypertrophy" are not retrieved. The last setting, *isVirtualOntologyId*, is recommended to be set to 'true', and specifies that the virtual ontology identifier (stable across ontology versions) is used instead of the ontology version identifier.

In the case of the genetic features, the query terms given to the Annotator were the official gene names (according to the HUGO Gene Nomenclature Committee[3]). In the case of the clinical features, the query terms used were their labels, or the label of the feature values for categorical features. Since the original labels can differ from the label or the synonyms of the terms in the ontology, and in order to ensure that the Annotator retrieved annotations for the higher number possible of features, the following measures were taken: a) a manual verification of the labels and synonyms was made; and b) when there were discrepancies, the original label of the feature was replaced by the label in the ontology.

[1] http://www.ihtsdo.org/snomed-ct/

[2] http://www.bioontology.org/wiki/index.php/Annotator_User_Guide#Annotator_Web_Service_Parameters

[3] http://www.genenames.org/

The clinical features used in the definition of the patients' class (i.e., SCD or no-SCD) were not considered in the enrichment analysis nor in the subsequent data mining analysis.

The set of genes in the Human genome was obtained from the GeneCards Database[1] and the set of GO annotations from the GOA database (Barrell *et al.*, 2009), as of the releases of October 4th, 2012. From the total set of Human protein-coding genes, only 18,759 were annotated with GO terms. All types of GO annotations were considered, including inferred from electronic annotation.

The genetic enrichment analysis was performed for the three types of GO terms: biological process, molecular function, and cellular component. In order to filter out uninformative GO terms, we considered only terms with information content (IC) above 60%. The IC of a term $t$ is given by the expression (Resnik, 1995):

$$IC(t) = -\log_2 \frac{f(t)}{f(root)} \tag{4.6}$$

where *f(t)* is the annotation frequency of the term (i.e. the number of distinct gene products it annotates) and *f(root)* is the frequency of annotation of the root term of the GO (which corresponds to the total number of annotated gene products). In this work, we used the annotations to Human genes to compute the IC, including annotations with all evidence codes. In order to obtain a normalized IC, we divided the IC values by the scale maximum ($\log_2 f(root)$).

## 4.1.2 Results

In this section are presented the results of the enrichment analyses performed with the genetic data from HCM patients, and with the clinical data from HCM and COPD patients.

### 4.1.2.1 Genetic data analysis

**Enrichment profiling**

[1] http://www.genecards.org/

For SCD patients, the study set contains 16 mutated genes (total for the 14 SCD patients) and the population set contains 18,759 x 14 genes (the number of GO annotated protein-coding Human genes multiplied by the number of SCD patients). For no-SCD patients, the study set contains 100 mutated genes (total for the 69 no-SCD patients) and the population set contains 18,759 x 69 genes (see Table 4.8 for a compilation of the number of genes analyzed in both enrichment analyses).

Table 4.8: Number of genes considered in the profiling and the differential enrichment analyses.

| Enrichment test | | Study set | Population set |
|---|---|---|---|
| Enrichment profiling | SCD | 16 | 18,759 x 14 |
| | no-SCD | 100 | 18,759 x 69 |
| Differential enrichment | SCD | 16 | 116 |
| | no-SCD | 100 | 116 |

Table 4.9: Number of enriched terms in each of the genetic analyses performed.

| Number of enriched terms | | Analysis | | | |
|---|---|---|---|---|---|
| | | Profiling | | Differential | |
| | | SCD | no-SCD | SCD vs. no-SCD | no-SCD vs. SCD |
| Bio.Proc. | noCorr | 30 | 39 | 0 | 2 |
| | Bonf | 19 | 33 | 0 | 0 |
| Mol.Func. | noCorr | 13 | 21 | 1 | 1 |
| | Bonf | 11 | 19 | 0 | 0 |
| Cel.Comp. | noCorr | 10 | 10 | 0 | 2 |
| | Bonf | 10 | 10 | 0 | 0 |
| **Total** | noCorr | 53 | 70 | 1 | 5 |
| | Bonf | 40 | 62 | 0 | 0 |

For each enrichment analysis the table indicates the number of terms of each GO type (biological process, molecular function and cellular component), with $p$-value below 0.1, when considering no multiple-testing correction (noCorr) and with Bonferroni correction (Bonf).

As shown in Table 4.9 (in the column *Total*), the enrichment profiling analysis identified the following number of enriched terms ($p$-value $< 0.1$): 53 for SCD and 70 for no-SCD, without multiple-testing correction; 40 for SCD and 62 for no-SCD, with Bonferroni correction.

Tables 4.10, 4.11 and 4.12 show, respectively, the top 10 enriched biological process (BP), molecular function (MF) and cellular component (CC) terms for the SCD patients. The full set of results is available in Appendix A. The results obtained with no-SCD differ from those obtained with SCD only in the 18 terms shown in Table 4.13. The full set of results for no-SCD is available in Appendix B.

Table 4.10: Top 10 enriched biological process terms (Gene Ontology), obtained in the profiling analysis of SCD patients.

| Name | $p$-value | $p$-Bonf | SFreq | PFreq |
|---|---|---|---|---|
| muscle filament sliding | 7.7E-40 | 4.1E-38 | 94% | 0.21% |
| actin-myosin filament sliding | 7.7E-40 | 4.1E-38 | 94% | 0.21% |
| ventricular cardiac muscle tissue morphogenesis | 7.7E-40 | 4.1E-38 | 94% | 0.21% |
| ventricular cardiac muscle tissue development | 2.4E-39 | 1.3E-37 | 94% | 0.22% |
| actin-mediated cell contraction | 6.8E-39 | 3.6E-37 | 94% | 0.24% |
| regulation of heart rate | 1.3E-31 | 6.9E-30 | 81% | 0.26% |
| adult heart development | 6.8E-25 | 3.6E-23 | 56% | 0.07% |
| positive regulation of ATPase activity | 2.9E-15 | 1.5E-13 | 38% | 0.09% |
| regulation of ATPase activity | 2.6E-14 | 1.4E-12 | 38% | 0.12% |
| regulation of muscle filament sliding | 1.0E-12 | 5.4E-11 | 25% | 0.02% |

$p$-value corresponds to the value without multiple-testing correction, whereas $p$-Bonf is the corresponding value with Bonferroni correction; SFreq and PFreq are the annotation frequencies in the study and population set respectively.

Table 4.11: Top 10 enriched molecular function terms (Gene Ontology), obtained in the profiling analysis of SCD patients.

| Name | $p$-value | $p$-Bonf | SFreq | PFreq |
|---|---|---|---|---|
| structural constituent of muscle | 2.9E-35 | 1.6E-33 | 88% | 0.25% |
| actin-dependent ATPase activity | 1.1E-26 | 6.1E-25 | 56% | 0.05% |
| microfilament motor activity | 1.8E-23 | 9.4E-22 | 56% | 0.11% |
| myosin heavy chain binding | 3.3E-11 | 1.8E-09 | 25% | 0.04% |
| ATPase activator activity | 9.1E-11 | 4.8E-09 | 25% | 0.05% |
| titin binding | 2.1E-10 | 1.1E-08 | 25% | 0.06% |
| ATPase regulator activity | 4.0E-10 | 2.1E-08 | 25% | 0.07% |
| myosin binding | 7.6E-09 | 4.0E-07 | 25% | 0.14% |
| troponin C binding | 5.3E-06 | 2.8E-04 | 13% | 0.02% |
| troponin I binding | 8.4E-06 | 4.4E-04 | 13% | 0.03% |

$p$-value corresponds to the value without multiple-testing correction, whereas $p$-Bonf is the corresponding value with Bonferroni correction; SFreq and PFreq are the annotation frequencies in the study and population set respectively.

Table 4.12: Top 10 enriched cellular component terms (Gene Ontology), obtained in the profiling analysis of SCD patients.

| Name | $p$-value | $p$-Bonf | SFreq | PFreq |
|---|---|---|---|---|
| muscle myosin complex | 2.4E-37 | 1.3E-35 | 81% | 0.10% |
| myosin filament | 2.4E-37 | 1.3E-35 | 81% | 0.10% |
| myosin II complex | 1.1E-35 | 5.8E-34 | 81% | 0.13% |
| stress fiber | 4.1E-20 | 2.2E-18 | 56% | 0.25% |
| actin filament bundle | 7.3E-20 | 3.8E-18 | 56% | 0.27% |
| C zone | 9.2E-15 | 4.9E-13 | 25% | 0.01% |
| striated muscle myosin thick filament | 1.0E-12 | 5.4E-11 | 25% | 0.02% |
| A band | 4.7E-09 | 2.5E-07 | 25% | 0.13% |
| troponin complex | 2.2E-05 | 1.1E-03 | 13% | 0.04% |
| striated muscle thin filament | 6.6E-05 | 3.5E-03 | 13% | 0.07% |

$p$-value corresponds to the value without multiple-testing correction, whereas $p$-Bonf is the corresponding value with Bonferroni correction; SFreq and PFreq are the annotation frequencies in the study and population set respectively.

Table 4.13: Enriched terms in the profiling analysis of no-SCD patients (genetic data), not identified in the SCD patients.

| Name | $p$-value | $p$-Bonf | SFreq | PFreq |
|---|---|---|---|---|
| **Biological Process** | | | | |
| regulation of systemic arterial blood pressure by ischemic conditions | 6.0E-41 | 4.4E-39 | 13% | 0.00 |
| neurological system process involved in regulation of systemic arterial blood pressure | 3.4E-25 | 2.5E-23 | 13% | 0.00 |
| regulation of smooth muscle contraction | 6.6E-19 | 4.9E-17 | 13% | 0.00 |
| visceral muscle development | 5.3E-03 | 3.9E-01 | 1% | 0.00 |
| muscle cell fate specification | 5.3E-03 | 3.9E-01 | 1% | 0.00 |
| atrial cardiac muscle tissue morphogenesis | 2.6E-02 | 1.9E+00 | 1% | 0.00 |
| atrial cardiac muscle tissue development | 2.6E-02 | 1.9E+00 | 1% | 0.00 |
| cardiac muscle fiber development | 4.2E-02 | 3.1E+00 | 1% | 0.00 |
| muscle cell fate commitment | 5.7E-02 | 4.2E+00 | 1% | 0.00 |
| **Molecular Function** | | | | |
| troponin T binding | 9.9E-33 | 7.3E-31 | 13% | 0.00 |
| calcium channel inhibitor activity | 1.9E-31 | 1.4E-29 | 13% | 0.00 |
| ion channel inhibitor activity | 1.4E-23 | 1.0E-21 | 13% | 0.00 |
| channel inhibitor activity | 1.4E-23 | 1.0E-21 | 13% | 0.00 |
| calcium channel regulator activity | 4.9E-23 | 3.6E-21 | 13% | 0.00 |
| calcium-dependent protein binding | 1.1E-19 | 8.1E-18 | 13% | 0.00 |
| actinin binding | 5.6E-06 | 4.2E-04 | 4% | 0.00 |
| calcium-dependent ATPase activity | 1.6E-02 | 1 | 1% | 0.00 |
| actin monomer binding | 7.2E-02 | 1 | 1% | 0.00 |

$p$-value corresponds to the value without multiple-testing correction, whereas $p$-Bonf is the corresponding value with Bonferroni correction; SFreq and PFreq are the annotation frequencies in the study and population set respectively.

### Differential enrichment

A total of one term for SCD and five terms for no-SCD were identified as enriched ($p$-value $< 0.1$, not considering multiple-testing correction), as shown on Table 4.14.

Table 4.14: Complete results of the two differential enrichment analyses performed with genetic data: HCM patients and the sub-group of SCD patients; HCM patients and the sub-group of no-SCD patients.

| Sub-group | Name | $p$-value | $p$-Bonf | SFreq | PFreq |
|---|---|---|---|---|---|
| SCD | **Molecular Function** | | | | |
| | structural constituent of muscle | 8.1E-02 | 1 | 88% | 70% |
| no-SCD | **Biological Process** | | | | |
| | negative regulation of ATPase activity | 8.1E-02 | 1 | 33% | 30% |
| | regulation of ATPase activity | 9.1E-02 | 1 | 59% | 56% |
| | **Molecular Function** | | | | |
| | troponin C binding | 8.1E-02 | 1 | 33% | 30% |
| | **Cellular Component** | | | | |
| | striated muscle thin filament | 8.1E-02 | 1 | 33% | 30% |
| | troponin complex | 8.1E-02 | 1 | 33% | 30% |

$p$-value corresponds to the value without multiple-testing correction, whereas $p$-Bonf is the corresponding value with Bonferroni correction; SFreq and PFreq are the annotation frequencies in the study and population set respectively.

#### 4.1.2.2 Clinical data analysis

The enrichment analysis was tested with two sets of clinical data: HCM and COPD.

The enrichment of the clinical HCM data resulted in the identification of 17 enriched terms in the sub-group of SCD patients (Table 4.15), and 14 enriched terms in the sub-group of no-SCD (Table 4.16), with no terms in common between the two sub-groups. These results were obtained considering no multiple-testing correction and $p$-values $< 0.1$. No terms were identified with Bonferroni correction.

The enrichment of the clinical COPD data resulted in the identification of 8 enriched terms in the sub-group of patients with emphysema (Table 4.17) and 24 enriched terms in the sub-group of healthy patients (Table 4.18), with the two sub-groups having no terms in common. These results were obtained considering no multiple-testing correction and $p$-value $< 0.1$. Again, no terms were identified with Bonferroni correction.

Table 4.15: Terms enriched in the clinical analysis of the group of HCM patients and the sub-group of SCD patients.

| ID | Name | $p$-value | SFreq | PFreq |
|---|---|---|---|---|
| 106227002 | General information qualifier | 1.4E-02 | 21.4% | 4.8% |
| 26636000 | Sudden death | | | |
| 272379006 | Event | | | |
| 392521001 | History of | | | |
| 419620001 | Death | | | |
| 51042001 | History of (present illness) | | | |
| C18772 | Personal Medical History | | | |
| C19332 | Personal Attribute | | | |
| C28389 | NCI Administrative Concept | | | |
| C42687 | Concept History | | | |
| C42698 | Terminology Entity | | | |
| C53787 | Adverse Event Associated with Death | | | |
| C53814 | Death Adverse Event Not Associated with More Specific CTCAE Term | | | |
| C54625 | History | | | |
| C55285 | Sudden Death Adverse Event Not Associated with More Specific CTCAE Term | | | |
| 272099008 | Descriptor | 5.1E-02 | 64.3 | 41.0 |
| 362981000 | Qualifier value | | | |

Numeric IDs correspond to terms from SNOMED-CT, whereas IDs starting with 'C' correspond to terms from NCIt; $p$-value corresponds to the value without multiple-testing correction; SFreq and PFreq are the annotation frequencies in the study and population set respectively. *CTCAE* stands for Common Terminology Criteria for Adverse Events.

Table 4.16: Terms enriched in the clinical analysis of the group of HCM patients and the sub-group of no-SCD patients.

| ID | Name | $p$-value | SFreq | PFreq |
|---|---|---|---|---|
| 404684003 | Clinical finding | 4.0E-03 | 100% | 96.4% |
| 123037004 | Body structure | 3.2E-02 | 97.1% | 94.0% |
| 362955004 | Inactive concept | 5.7E-02 | 95.7% | 92.8% |
| 363662004 | Duplicate concept | | | |
| 365860008 | General clinical state finding | | | |
| 370115009 | Special concept | | | |
| 429019009 | Finding related to biological sex | | | |
| 442083009 | Anatomical or acquired body structure | | | |
| 57312000 | Sex structure | | | |
| 91722005 | Physical anatomical entity | | | |
| 91723000 | Anatomical structure | | | |
| C20189 | Property or Attribute | | | |
| C27993 | General Qualifier | | | |
| C41009 | Qualifier | | | |

Numeric IDs correspond to terms from SNOMED-CT, whereas IDs starting with 'C' correspond to terms from NCIt; $p$-value corresponds to the value without multiple-testing correction; SFreq and PFreq are the annotation frequencies in the study and population set respectively.

Table 4.17: Terms enriched in the clinical analysis of the group of COPD patients and the sub-group of patients with emphysema.

| ID | Name | $p$-value | SFreq | PFreq |
|---|---|---|---|---|
| 1086007 | Female structure | 1.7E-02 | 19% | 64% |
| 139867007 | Female | | | |
| 162600001 | Female | | | |
| 248152002 | Female | | | |
| C16576 | Female | | | |
| C46108 | Female, Self-Report | | | |
| C46110 | Female Gender | | | |
| C46113 | Female Phenotype | | | |

Numeric IDs correspond to terms from SNOMED-CT, whereas IDs starting with 'C' correspond to terms from NCIt; $p$-value corresponds to the value without multiple-testing correction; SFreq and PFreq are the annotation frequencies in the study and population set respectively.

Table 4.18: Terms enriched in the clinical analysis of the group of COPD patients and the sub-group of patients without emphysema.

| ID | Name | $p$-value | SFreq | PFreq |
|---|---|---|---|---|
| 10052007 | Male structure | 1.7E-02 | 78% | 91% |
| 139866003 | Male | | | |
| 162599004 | Male | | | |
| 248153007 | Male | | | |
| C20197 | Male | | | |
| C46107 | Male, Self-Report | | | |
| C46109 | Male Gender | | | |
| C46112 | Male Phenotype | | | |
| 110483000 | Tobacco user | 6.6E-02 | 63% | 74% |
| 118228005 | Functional finding | | | |
| 138008005 | Smoker | | | |
| 160622001 | Smoker (& cigarette) | | | |
| 225786009 | Smoker | | | |
| 250171008 | Clinical history and observation findings | | | |
| 363660007 | Ambiguous concept | | | |
| 365949003 | Health-related behavior finding | | | |
| 365980008 | Tobacco use and exposure - finding | | | |
| 365981007 | Tobacco smoking behavior - finding | | | |
| 77176002 | Smoker | | | |
| 844005 | Behavior finding | | | |
| C19332 | Personal Attribute | | | |
| C19796 | Smoking Status | | | |
| C67147 | Current Smoker | | | |
| C68751 | Smoker | | | |

Numeric IDs correspond to terms from SNOMED-CT, whereas IDs starting with 'C' correspond to terms from NCIt; $p$-value corresponds to the value without multiple-testing correction; SFreq and PFreq are the annotation frequencies in the study and population set respectively.

### 4.1.3 Discussion

From the four approaches of enrichment analysis presented in Section 2.4, only those working with a study set are relevant for my work. This is due to the medical objective of this part of the work, which is the evaluation of patients in respect to the likelihood of suffering a disease-related event or having a specific disease manifestation. This medical objective thus results in the existence of a differentiating factor that separates the patients in positive and negative groups for the study characteristic.

From the three enrichment analyses working with a study set, the Singular Enrichment Analysis (SEA) was the one selected because it is the most commonly used and was thus a good starting point.

#### 4.1.3.1 Genetic data analysis

The first test with the adapted enrichment analysis was done with the genetic set of features. Since all the patients share the same genome, it is through their individual mutations that we can find differentiating features. However, information regarding a patient's mutations, when available, exists only for a few genes. In the case of the HCM patients, the genetic data used in this analysis is precisely the presence/absence of the mutations in the genes associated with the disease.

An oversimplified way to define the study set when analyzing, for example, the SCD patients, would be to consider the list of genes mutated in at least one of these patients. However, this would only be accurate if all the SCD patients had a mutation in those genes, which might not be the truth. In order to maximize the use of the available genetic information, the best option was to consider the set of mutations each patient has, individually.

Two enrichment analyses were performed exploiting the patients' mutations data: a profiling analysis, where the total number of genes mutated in each group of patients (with SCD and without SCD) was compared with all (protein-coding) genes in the same group of patients; and a differential analysis, where the total number of genes mutated in each group of patients (with SCD and without SCD) was compared with the total number of genes mutated in all the HCM patients.

Terms identified as enriched by the profiling analysis can be used to characterize the genotype of patients with and without SCD, since they correspond to specific functional aspects that are mutated in the patients. These functional aspects, in turn, correspond to phenotypical traits expected to be altered. While terms identified as enriched both in SCD and no-SCD patients can be interpreted as associated with the disease, terms enriched differently can be interpreted as associated with the occurrence of SCD. This profiling analysis is more directly comparable with the application of enrichment analysis to gene expression data, where a set of genes (e.g., overexpressed) can be analyzed against the whole genome.

Analyzing the enriched terms identified in the profiling analysis, their relation with HCM can be confirmed. According to the biological process terms enriched, the patients analyzed suffer from cardiac alterations (e.g. *regulation of heart rate*, *adult heart development*), in particular in the ventricle (*ventricular cardiac muscle tissue morphogenesis* and *ventricular cardiac muscle tissue development*), and some of those alterations affect the contraction of striated muscles, in which group the cardiac muscle is included (e.g. *actin-myosin filament sliding* and *actin-mediated cell contraction*). HCM is indeed a cardiac disease, in which the main anatomical manifestation is the thickening of the interventricular septum, and the occurrence of a sudden cardiac arrest can be a consequence of the malfunctioning of the heart contraction. Considering the molecular function terms, several *binding* terms are enriched, namely *myosin heavy chain binding*, *titin binding*, *troponin C* and *troponin I binding*. All of these terms refer to proteins that participate in the contraction of the filaments that compose striated muscles, and thus the HCM patients present alterations in the normal function of this type of muscle. Finally, the cellular component terms confirm the previous observations that the alterations in HCM patients occur at the level of striated muscle functioning, namely through the following terms: *striated muscle myosin thick filament*, *striated muscle thin filament*, *troponin complex*, *A band* (a component of the sarcomere) and *C zone* (a component of the A band).

The difference between SCD and no-SCD consists in a set of 18 terms identified in the no-SCD patients and not in the SCD patients. These differences can be explained by the fact that the number of no-SCD patients is considerably larger

than the number of SCD patients (69 vs. 14) and consequently there are more distinct genes mutated (7 vs. 4). However, there is also the possibility of a biological explanation. On the one hand, it may not be correct to interpret that when a function or process is altered the patients will not suffer a SCD episode. On the other hand, dominant mutations (i.e., mutations that can have a manifestation by affecting only one of the gene's copies) may increase the activity of a gene product, which can ultimately result in the prevention of a SCD episode (Lodish *et al.*, 2000). A more detailed analysis of the type of mutation that led to these results is thus necessary in the future.

The purpose of the differential enrichment analysis was the identification of the differences between SCD and no-SCD, and thus compare each, in turn, with the complete set of HCM patients. Since this set is divided in SCD and no-SCD patients, the comparison is basically between one group and the other.

For the purpose of prognosis, the most interesting terms are those identified as enriched in SCD. Thus, the term *structural constituent of muscle* may have potential for prognosis, given that it occurs more frequently in SCD patients than in no-SCD patients. Nevertheless, the fact that the corrected $p$-value is above the significance level and that the term is not particularly informative in respect to HCM, limits the confidence with which this term can be used for that purpose.

The results obtained with the genetic data clearly show that, by itself, this data is not sufficient to clearly separate SCD patients from no-SCD patients. This may be due to the dataset tested, which has a small number of instances and is not balanced in terms of positive and negative instances. It may also be due to the event tested, since currently it cannot be predicted solely based on genetic data. Additionally, it is also possible that the genetic data is not yet being fully exploited. In this preliminary test, I considered only the presence or absence of mutations in the genes, but not their type nor number. For example, it is known that some mutations are associated with a benign outcome (i.e., no occurrence of SCD) whereas others with a malignant outcome. It has also been reported that the occurrence of mutations in some genes is associated with a higher incidence of SCD than in others (Bos *et al.*, 2009). All of these aspects can be taken into consideration when calculating the frequencies of annotation or even be added as features to the dataset. It is important to note, however, that I was not concerned

with pleiotropic effects. It is known that some HCM mutations have different phenotypic manifestations in different patients, and different manifestations should also be expected if a patient has multiple mutations. Nevertheless, the goal of this analysis is to obtain profiles that provide a global characterization of the patients in respect to an event, and not to perform a precise evaluation of each patient in terms of his mutations. Regarding the genetic enrichment analysis, that global characterization should be in terms of the functions and processes most frequently affected in the event-positive patients

In terms of missing values, in the genetic dataset these are mutations associated with HCM that were not tested. Due to the sparseness of the dataset, it was not feasible to simply eliminate the mutations not tested or the patients with mutations not tested. Consequently, we considered these mutations as having a negative value, as if they were not present in the patient. This approach allowed me to exploit all the available data and to obtain an informative characterization of the patients. It is important to stress out that an evaluation of all the patients for all the mutations is almost never done. More mutations tested might result, on the one hand, in an increase in the number of genes analyzed, possibly leading to an increment in the number of terms tested and, consequently, in the terms found enriched. On the other hand, it might result in an increase in the frequency of annotation of the terms in the study set of the enrichment profiling analysis, and in both the study and the population sets in the differential enrichment analysis. In the profiling analysis, this increase would result in the strengthening of the confidence in the results since we would increase the difference of annotation frequency between the study set and the population set. In the differential analysis, the results might be more strongly altered, since both sets of annotation frequencies would have to be recalculated.

#### 4.1.3.2 Clinical data analysis

The enrichment analysis of the two disease datasets resulted in sets of terms with a considerable number of high-level terms, which means that analyzing them simply by their names does not permit an obvious interpretation of the case-study. Another evident observation is that there are repeated terms, which results not

only from the consideration of two clinical vocabularies but also from the existence of terms with the same label (i.e., the term name) and different identifiers in the same vocabulary.

In the SCD sub-group of HCM patients (see Table 4.15), all but one of the enriched terms annotate the feature *Sudden death family history*, reflecting the higher number of of SCD patients positive for this feature (3 to 1 no-SCD). The term that annotates a different feature is *Descriptor*, which annotates two of the values of the feature *Hypertrophy morphology*: *Apical* and *Septal*. In the no-SCD sub-group (see Table 4.16), all of the terms annotate the two values of the feature *Gender*, reflecting the higher number of no-SCD patients with known gender (96% to 79% SCD). In both sub-groups, there are terms that annotated more than one feature. In SCD, the term *Qualifier value* annotates *Apical* and *Septal* in addition to *Sudden death family history*. In the no-SCD, the following four terms annotated "gender" in addition to the features indicated:

- *Body structure* - annotates *Concentric*, a value of the feature *Hypertrophy morphology*.
- *Clinical finding* - annotates the three values of *Blood pressure*: *Normal*, *Hypotension*, and *Hypertension*.
- *Property or Attribute* - annotates *Apical* (*Hypertrophy morphology*) and *Sudden death family history*.
- *Qualifier* - annotates *Apical* (*Hypertrophy morphology*).

In the healthy sub-group of COPD patients, the terms enriched annotate either the value *Male* of *Gender* or the feature *Smoker*. Only one term annotates non-related features, *Ambiguous concept*, which annotates *O2 pressure*, *CO2 pressure* and *Alfa1 antitrypsin deficiency* in addition to *Smoker*. In the emphysema sub-group, all the terms enriched annotate the value *Female* of *Gender*. The terms enriched simply reflect the fact that the healthy sub-group has a higher number of males and smokers, and the emphysema sub-group has a higher number of females. The lower number of smokers in the emphysema sub-group is expected since this disease results in a poor respiratory capacity[1].

[1] http://www.lung.ca/diseases-maladies/a-z/emphysema-emphyseme/index_e.php

The enrichment analysis performed with the clinical data corresponds only to a differential profiling since an enrichment profiling implies the need of a background set of annotations beyond the datasets tested. As referred in Section 2.4, approaches have been proposed to overcome this limitation associated with patient datasets that retrieve the clinical annotations either from a corpus of MEDLINE abstracts (Tirrell *et al.*, 2010), or the publications that originated manual Gene Ontology annotations (LePendu *et al.*, 2011). Both approaches can be adopted to improve the clinical enrichment profiling of patient datasets.

#### 4.1.3.3 Study limitations

In respect to the tool Annotator, there are two aspects of its functioning that might have had an effect on the annotations retrieved:

- Contrarily to what is indicated in the tool's description, at least in one case the definition of the setting *longestOnly* did not prevent that partial matches were retrieved. In the clinical dataset, the feature *Sudden death history* was directly annotated with terms such as *Sudden death* and *Death*.
- The order of the words in the query text is relevant. For example, searching for *Obstructive hypertrophic cardiomyopathy* yields no results, whereas *Hypertrophic obstructive cardiomyopathy* does.

Additionally, there are a few aspects of the use of ontologies that have to be taken into consideration. By applying a methodology that relies on controlled vocabularies, it is possible that we are working with incomplete annotations, as well as with a set of ontology terms that might not provide the level of detail necessary to fully characterize the patients. In respect to the possibility of incomplete annotations, I tried to deal with it by considering all types of annotation, including inferred from electronic annotation in the Gene Ontology, even with the risk of introducing some annotation errors. In respect to the possibility of an insufficient level of detail, it can be overcome by considering more than one vocabulary for the same domain of knowledge, which was already done for the clinical data and can be done also for the genetic data. The Gene Ontology was chosen for the preliminary implementation of the methodology since it is the most well studied

application of enrichment analysis, but any ontology can be exploited provided the existence of a background set of annotations against which the study set can be compared.

The use of SNOMED in particular presented some hindrances. This vocabulary contains more than one term with the same label, which means that the terms represent the same real-world entity but are attributed different identifiers; it contains in its hierarchy terms pertaining to the management of the vocabulary itself, terms that were found enriched in my analysis (e.g., *Inactive concept* and *Ambiguous concept*); and it contains terms representing the same entity but with different labels (and identifiers). Neither of these aspects were found with the Gene Ontology or the NCIt.

Finally, another relevant factor in the methodology is the enrichment analysis approach followed. The results obtained with the Singular Enrichment Analysis are interesting, but there are two other approaches worth exploiting: the Modular Enrichment Analysis, and the model-based approach. The first because it takes into consideration the existence of relations between the genes, and the second because it does not analyze individual terms but rather aims to obtain the best set of terms annotating the data.

## 4.2 Evaluation of the enrichment analysis with data mining algorithms

In this section is described the evaluation of the results of the previous enrichment analyses with data mining algorithms. For this purpose were considered the results obtained in the differential enrichment of the HCM dataset (genetic and clinical data), and in the differential enrichment of the COPD clinical data.

### 4.2.1 Methods

In the final prognosis methodology (shown in Figure 4.1), the terms identified as enriched in each group of patients will be used as an annotation profile of that group. New patients will then be assessed against these profiles in order to predict their prognosis.

## 4.2 Evaluation of the enrichment analysis with data mining algorithms

The group profiles are used in the assessment of the prognosis by incorporating their respective enriched terms as features in the disease dataset, which is then subjected to a classification step. In order to evaluate if the incorporation of the terms improved the predictive capability of the datasets, several classification algorithms were tested with the following sets of features:

- Original dataset: clinical and genetic features
- Original dataset + Annotated terms
- Annotated terms
- Original dataset + Enriched terms
- Enriched terms

In all the sets containing ontology terms, these were added as Boolean features. The values of the terms were obtained from the values of the original features annotated with the term. For a term to be considered present in a patient, at least:

- One original Boolean feature annotated with that term has a positive value
- One value of an original categorical feature is annotated with the term
- One original numerical feature annotated with that term has a known value

The terms added as features are those annotating the original features (the annotated terms) and those resulting from the enrichment analyses performed with the two groups of patients in each dataset (the enriched terms). In the case of the HCM dataset the terms are those enriched for the SCD and the no-SCD patients, and in the case of the COPD dataset the terms are those enriched for the patients with and without emphysema. The patients considered as the instances of the positive class are those in the SCD group and those in the emphysema group, for HCM and COPD, respectively.

The only features, and respective terms, not considered in this data mining evaluation were also not considered in the enrichment analysis, and correspond to the features used to define the classification of the patients.

Five classifiers were tested, all available through the tool Weka (version 3.7.5) (Hall *et al.*, 2009): J48 (decision trees) (Quinlan, 1993), Random Forest (Breiman, 2001), Naive Bayes (John & Langley, 1995), Bayesian networks (Bouckaert, 2004), and K-nearest neighbors (Aha & Kibler, 1991). All classifiers were run with their default parametrizations (detailed in Appendix D), and with a 10-fold cross-validation.

The performance of the data mining methods was evaluated in terms of precision, recall, and F-measure. The precision measures the proportion of instances classified as positive that are indeed positive, ranging from 0 when there are no true positives to 1 when there are no false positives. The recall measures the proportion of instances with positive class that are correctly classified as positives, ranging from 0 when there are no true positives, to 1 when there are no false negatives. Considering the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), the precision and the recall of a binary classifier are calculated by the following expressions:

$$Precision = \frac{TP}{TP + FP} \tag{4.7}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.8}$$

The F-measure is the weighted harmonic mean of precision and recall, calculated by the following expression:

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.9}$$

### 4.2.2 Results

#### 4.2.2.1 Hypertrophic cardiomyopathy

The results of the data mining tests performed with the SCD class of HCM patients and the five different sets of features are shown in terms of F-measure (Table 4.19), precision (Table 4.20), and recall (Table 4.21).

Table 4.19: F-measure results obtained with the event-positive class (SCD) of HCM patients.

| Dataset | | Method | | | | |
|---|---|---|---|---|---|---|
| | | J48 | RF | NB | BNet | KNN |
| Original | | 0 | 0.235 | 0 | 0.273 | 0.37 |
| Original + Annotations | | 0.5 | 0.25 | 0.385 | 0.412 | 0.467 |
| Annotations | | **0.571** | 0.32 | 0.323 | 0.424 | 0.385 |
| Original + Enriched | $\alpha$=0.1 | 0.5 | 0.417 | **0.5** | **0.5** | 0.37 |
| | $\alpha$=0.05 | 0.5 | 0.455 | 0.333 | 0.316 | 0.429 |
| Enriched | $\alpha$=0.1 | **0.571** | **0.615** | **0.5** | **0.5** | 0.364 |
| | $\alpha$=0.05 | **0.571** | 0.571 | 0.333 | 0.333 | **0.571** |

J48 corresponds to the decision tree learning algorithm, RF corresponds to Random Forest, NB to Naive Bayes, BNet to Bayes Network and KNN to K-nearest neighbors. The highest values obtained for each method are shown in bold, and the highest overall value amongst all methods is shown in red.

Table 4.20: Precision results obtained with the event-positive class (SCD) of HCM patients.

| Dataset | | Method | | | | |
|---|---|---|---|---|---|---|
| | | J48 | RF | NB | BNet | KNN |
| Original | | 0 | 0.667 | 0 | 0.375 | 0.385 |
| Original + Annotations | | 0.833 | 0.3 | 0.417 | 0.35 | 0.438 |
| Annotations | | **0.857** | 0.364 | 0.294 | 0.368 | 0.417 |
| Original + Enriched | $\alpha$=0.1 | 0.833 | 0.5 | 0.6 | 0.6 | 0.385 |
| | $\alpha$=0.05 | 0.833 | 0.625 | **0.75** | 0.6 | 0.429 |
| Enriched | $\alpha$=0.1 | **0.857** | 0.667 | 0.6 | 0.6 | 0.5 |
| | $\alpha$=0.05 | **0.857** | **0.857** | **0.75** | **0.75** | **0.857** |

J48 corresponds to the decision tree learning algorithm, RF corresponds to Random Forest, NB to Naive Bayes, BNet to Bayes Network and KNN to K-nearest neighbors. The highest values obtained for each method are shown in bold, and the highest overall value amongst all methods is shown in red.

Table 4.21: Recall results obtained with the event-positive class (SCD) of HCM patients.

| Dataset | | Method | | | | |
|---|---|---|---|---|---|---|
| | | J48 | RF | NB | BNet | KNN |
| Original | | 0 | 0.143 | 0 | 0.214 | 0.357 |
| Original + Annotations | | 0.357 | 0.214 | 0.357 | **0.5** | **0.5** |
| Annotations | | **0.429** | 0.286 | 0.357 | 0.5 | 0.357 |
| Original + Enriched | $\alpha$=0.1 | 0.357 | 0.357 | **0.429** | 0.429 | 0.357 |
| | $\alpha$=0.05 | 0.357 | 0.357 | 0.214 | 0.214 | 0.429 |
| Enriched | $\alpha$=0.1 | **0.429** | **0.571** | **0.429** | 0.429 | 0.286 |
| | $\alpha$=0.05 | **0.429** | 0.429 | 0.214 | 0.214 | 0.429 |

J48 corresponds to the decision tree learning algorithm, RF corresponds to Random Forest, NB to Naive Bayes, BNet to Bayes Network and KNN to K-nearest neighbors. The highest values obtained for each method are shown in bold, and the highest overall value amongst all methods is shown in red.

The most notable result obtained with the original dataset is that J48 and Naive Bayes are unable to correctly classify the positive instances. It is also evident that the inclusion of annotations or enriched terms lead to improved results in terms of F-measure, precision and recall for all methods tested.

Regarding F-measure, the best result obtained with the original dataset was 0.37 (using K-nearest neighbors). By contrast, all methods produced an F-measure of at least 0.5 with the addition or the exclusive use of annotated terms (J48), the addition of enriched terms (J48, Naive Bayes and Bayes Network), and the exclusive use of enriched terms (all the 5 classifiers). The best result overall (0.615) was obtained with Random Forest using enriched terms exclusively ($\alpha$=0.1).

With respect to precision, the best result obtained with the original dataset was 0.667 (Random Forest). As observed for the F-measure, all methods improved with the addition of ontology terms, obtaining precisions of at least 0.75. Overall, the best result obtained was 0.857, with J48 (with annotations and with enriched terms) and with Random Forest and K-nearest neighbors (both with enriched terms). The recall results followed a similar pattern, with improved results for all methods with the addition of annotations and enriched terms. The best result

obtained with the original dataset was 0.357 (K-nearest neighbors), whereas the best result overall was 0.571, obtained with Random Forests and enriched terms ($\alpha$=0.1).

Table 4.22 shows the F-measure results of the data mining tests performed with the no-SCD class of HCM patients and the five different sets of features. These results show that the classification of the negative instances is also improved by the addition of ontology terms, and in particular enriched terms.

Table 4.22: F-measure results of the event-negative class (no-SCD) of HCM patients.

| Dataset | | Method | | | | |
|---|---|---|---|---|---|---|
| | | J48 | RF | NB | BNet | KNN |
| Original | | 0.908 | 0.913 | 0.893 | 0.889 | 0.878 |
| Original + Annotations | | 0.932 | 0.873 | 0.886 | 0.848 | 0.882 |
| Annotations | | **0.938** | 0.879 | 0.844 | 0.857 | 0.886 |
| Original + Enriched | $\alpha$=0.1 | 0.932 | 0.901 | 0.915 | 0.915 | 0.878 |
| | $\alpha$=0.05 | 0.932 | 0.917 | **0.919** | 0.912 | 0.884 |
| Enriched | $\alpha$=0.1 | **0.938** | 0.929 | 0.915 | 0.915 | 0.903 |
| | $\alpha$=0.05 | **0.938** | **0.938** | **0.919** | **0.919** | **0.938** |

J48 corresponds to the decision tree learning algorithm, RF corresponds to Random Forest, NB to Naive Bayes, BNet to Bayes Network and KNN to K-nearest neighbors. The highest values obtained for each method are shown in bold, and the highest overall value amongst all methods is shown in red.

Table 4.23 shows the number of instances correctly classified when considering the original dataset and when considering the best sets of features per method (i.e., those that originate the highest number of correctly classified positive instances) as detailed below:

- J48 - Enriched terms, without correction, $\alpha$=0.1 and $\alpha$=0.05; and annotated terms.
- NB - Enriched terms, without correction, $\alpha$=0.1; and original set of features + enriched terms, without correction, $\alpha$=0.1.
- BNet - Annotated terms.

- KNN - Original set of features + annotated terms.
- RF - Enriched terms, without correction, $\alpha$=0.1.

Table 4.23: Number of instances correctly classified in the data mining tests with the HCM dataset.

| Dataset | | Method | | | | |
|---|---|---|---|---|---|---|
| | | J48 | RF | NB | BNet | KNN |
| Original dataset | Pos | 0 | 2 | 0 | 3 | 5 |
| | Neg | 69 | 68 | 67 | 64 | 61 |
| Best set of features | Pos | 6 | 8 | 6 | 7 | 7 |
| | Neg | 68 | 65 | 65 | 57 | 60 |

J48 corresponds to the decision tree learning algorithm, RF corresponds to Random Forest, NB to Naive Bayes, BNet to Bayes Network and KNN to K-nearest neighbors. The best set of features was the set that produced the best results for the data mining method, from among the feature sets previously tested. The total number of positive instances is 14, whereas the total number of negative instances is 69.

Out of the 14 total positive instances, the methods ranged from 0 (J48 and Naive Bayes) to 5 (K-nearest neighbors) correctly classified instances with the original dataset. With the best sets, the methods ranged from 6 (J48 and Naive Bayes) to 8 (Random Forest) correctly classified instances.

#### 4.2.2.2 Chronic obstructive pulmonary disease

The results of the data mining tests performed with the emphysema class of COPD patients and the five different sets of features are shown in terms of F-measure (Table 4.24), precision (Table 4.25), and recall (Table 4.26).

In the case of this dataset, only J48, Random Forest and K-nearest neighbors showed improvement in F-measure over the original dataset with the use of ontology terms. For all three methods, the best results were obtained with the addition of enriched terms to the original dataset. The best result overall was obtained for K-nearest neighbors (0.779), which was significantly higher than the best result obtained with the original dataset (0.62 with Bayes Network).

In terms of precision, only J48 and K-nearest neighbors showed improved results with ontology terms, with the former producing the best results when

Table 4.24: F-measure results of the COPD patients with emphysema.

| Dataset | | Method | | | | |
|---|---|---|---|---|---|---|
| | | J48 | RF | NB | BNet | KNN |
| Original | | 0.648 | 0.469 | **0.6** | **0.62** | 0.407 |
| Original + Annotations | | 0.667 | 0.51 | 0.548 | 0.514 | 0.4 |
| Annotations | | 0 | 0 | 0.412 | 0.394 | 0 |
| Original+ Enriched | $\alpha$=0.1 | **0.676** | **0.542** | **0.6** | 0.513 | **0.779** |
| | $\alpha$=0.05 | 0.648 | 0.478 | 0.523 | 0.513 | 0.407 |
| Enriched | $\alpha$=0.1 | 0 | 0 | 0.342 | 0.342 | 0 |
| | $\alpha$=0.05 | 0 | 0 | 0.396 | 0.396 | 0 |

J48 corresponds to the decision tree learning algorithm, RF corresponds to Random Forest, NB to Naive Bayes, BNet to Bayes Network and KNN to K-nearest neighbors. The highest values obtained for each method are shown in bold, and the highest overall value amongst all methods is shown in red.

Table 4.25: Precision results of the COPD patients with emphysema.

| Dataset | | Method | | | | |
|---|---|---|---|---|---|---|
| | | J48 | RF | NB | BNet | KNN |
| Original | | 0.59 | **0.833** | **0.553** | **0.564** | 0.444 |
| Original + Annotations | | **0.622** | 0.684 | 0.488 | 0.474 | 0.429 |
| Annotations | | 0 | 0 | 0.389 | 0.359 | 0 |
| Original + Enriched | $\alpha$=0.1 | 0.615 | 0.813 | 0.5 | 0.435 | **0.667** |
| | $\alpha$=0.05 | 0.59 | 0.786 | 0.411 | 0.435 | 0.444 |
| Enriched | $\alpha$=0.1 | 0 | 0 | 0.295 | 0.295 | 0 |
| | $\alpha$=0.05 | 0 | 0 | 0.297 | 0.297 | 0 |

J48 corresponds to the decision tree learning algorithm, RF corresponds to Random Forest, NB to Naive Bayes, BNet to Bayes Network and KNN to K-nearest neighbors. The highest values obtained for each method are shown in bold, and the highest overall value amongst all methods is shown in red.

using all annotations and the latter producing the best results when using the original dataset plus enriched terms. The best result overall was obtained with the original dataset and Random Forest.

With respect to recall, J48, Naive Bayes and K-nearest neighbors all show

Table 4.26: Recall results of the COPD patients with emphysema.

| Dataset | | Method | | | | |
|---|---|---|---|---|---|---|
| | | J48 | RF | NB | BNet | KNN |
| Original | | 0.719 | **0.469** | 0.656 | **0.688** | 0.375 |
| Original + Annotations | | 0.719 | 0.406 | 0.625 | 0.563 | 0.375 |
| Annotations | | 0 | 0 | 0.438 | 0.438 | 0 |
| Original + Enriched | $\alpha$=0.1 | **0.75** | 0.406 | **0.75** | 0.625 | **0.938** |
| | $\alpha$=0.05 | 0.719 | 0.344 | 0.719 | 0.625 | 0.375 |
| Enriched | $\alpha$=0.1 | 0 | 0 | 0.406 | 0.406 | 0 |
| | $\alpha$=0.05 | 0 | 0 | 0.594 | 0.594 | 0 |

J48 corresponds to the decision tree learning algorithm, RF corresponds to Random Forest, NB to Naive Bayes, BNet to Bayes Network and KNN to K-nearest neighbors. The highest values obtained for each method are shown in bold, and the highest overall value amongst all methods is shown in red.

improved results with the addition of enriched terms to the original dataset, with the overall best result (0.938) obtained with the K-nearest neighbors.

Table 4.27 shows the F-measure results of the data mining tests performed with the healthy class of COPD patients and the five different sets of features. In the case of this dataset, J48 and Naive Bayes showed improved results with the addition of annotated terms to the original dataset, and Naive Bayes also improved with the addition of enriched terms to the original dataset. The overall best result (0.927) was obtained with the K-nearest neighbors with enriched terms added to the original dataset.

Table 4.28 shows the number of instances correctly classified when considering the original dataset and when considering the sets of features that originate the highest number of positive instances correctly classified. Out of the 32 total positive instances in this dataset, the data mining methods correctly classify from 12 (K-nearest neighbors) to 23 (J48) of them. In the case of Random Forest and Bayes Network, the best results were obtained with the original dataset, whereas the remaining methods had improved results with the addition of enriched terms to the original dataset. The overall best result was obtained with K-nearest neighbors, which correctly classified 30 positive instances.

Table 4.27: F-measure results of the healthy class of COPD patients.

| Dataset | | Method | | | | |
|---|---|---|---|---|---|---|
| | | J48 | RF | NB | BNet | NN |
| Original | | 0.895 | **0.923** | 0.883 | **0.887** | 0.861 |
| Original + Annotations | | **0.905** | 0.903 | **0.861** | 0.858 | 0.856 |
| Annotations | | 0.885 | 0.873 | 0.835 | 0.82 | 0.877 |
| Original + Enriched | $\alpha$=0.1 | 0.904 | 0.916 | **0.861** | 0.836 | **0.927** |
| | $\alpha$=0.05 | 0.895 | 0.909 | 0.811 | 0.836 | 0.861 |
| Enriched | $\alpha$=0.1 | 0.885 | 0.885 | 0.786 | 0.786 | 0.885 |
| | $\alpha$=0.05 | 0.885 | 0.885 | 0.727 | 0.729 | 0.885 |

J48 corresponds to the decision tree learning algorithm, RF corresponds to Random Forest, NB to Naive Bayes, BNet to Bayes Network and KNN to K-nearest neighbors. The highest values obtained for each method are shown in bold, and the highest overall value amongst all methods is shown in red.

Table 4.28: Number of instances correctly classified in the data mining tests with the COPD dataset.

| Dataset | | Method | | | | |
|---|---|---|---|---|---|---|
| | | J48 | RF | NB | BNet | NN |
| Original dataset | Pos | 23 | 15 | 21 | 22 | 12 |
| | Neg | 107 | 120 | 106 | 106 | 108 |
| Best set of features | Pos | 24 | - | 24 | - | **30** |
| | Neg | 108 | - | 99 | - | 108 |

J48 corresponds to the decision tree learning algorithm, RF corresponds to Random Forest, NB to Naive Bayes, BNet to Bayes Network and KNN to K-nearest neighbors. The best set of features was the original set with the addition of enriched terms ($\alpha$=0.1), for all methods except RF and BNet, in which the best results were obtained with the original dataset.

### 4.2.3 Discussion

Given the biomedical goals of analyzing the patients in terms of a disease-related event or condition, the number of positive instances correctly classified were considered as the most important improvement in the data mining results.

A preliminary evaluation of the enrichment methodology was done considering only the enriched genetic terms obtained for the disease HCM (results not shown), which did not yield good results.

However, the inclusion of the clinical information produced very interesting results with both datasets, despite the fact that the enriched terms were not, apparently, very informative by themselves.

In the HCM case-study, all the tested classifiers show an improvement in the classification of the positive instances with the inclusion of the enriched terms. This is readily apparent when analyzing the results in terms of F-measure and precision.

It is interesting to note that some of the best precision results were obtained with enriched terms with p-value $< 0.05$, whereas some of the best recall results were obtained with p-value $< 0.1$. This is expected since the first set of terms contains fewer terms, which were considered as enriched with more confidence, resulting in more positive instances correctly classified. On the other hand, the second set of terms contains more terms, possibly encompassing more information albeit less specific, which results in the identification of a higher number of the positive instances.

The inclusion of the whole set of annotations generally produced better results than the original dataset, however the best classifiers were obtained with the enriched annotations. The only exception occurred with the J48 algorithm, which produced the best precision result with annotations and enriched terms alike. These results demonstrate the importance of including knowledge from ontologies in the original dataset, and the importance of filtering that knowledge to consider only the more significant terms.

In the case of the COPD dataset, the improvement in the data mining results is not as evident and as disseminated by all the algorithms as with the HCM dataset. Instead, only one of the classifiers shows clear improvements: the

K-nearest neighbors. This classifier yielded the worst results with the original dataset and the best results in terms of F-measure and recall with the set of features plus enriched terms (p-values $< 0.1$). However, it is important to note that these results were obtained considering only the clinical data.

The reason why the inclusion of the enriched terms leads to improvements in the classification of positive instances might be related with how the information is encoded. We saw that some clinical terms annotate more than one feature, and this might provide more information for the data mining algorithm to work with. In the case of the HCM dataset, this new encoding by itself provides the best results, whereas in the COPD dataset the original features are still needed. This is interesting to note since HCM has a higher number of terms annotating distinct original features than COPD.

## 4.3 Exploration of translational medicine data

The data mining results validate the application of the enrichment methodology to improve the prediction capability of translational medicine datasets.

However, further research still needs to be done to deliver the prognosis framework. On the one hand, it is important to understand with the disease experts if the enriched terms bear any relevance beyond what can be ascertained simply by their names, since it is possible that the combination of features under a single term may shed additional insights about the disease. On the other hand, it is essential to test the methodology with other ontologies, genetic and clinical. We saw that SNOMED-CT annotates more features than NCIt, but that vocabulary has also several limitations that might be hindering the quality of the classification results.

Furthermore, it is important to bear in mind that the knowledge about a disease might not be all encoded in vocabularies at a given time. However, it is currently nearly impossible to ascertain automatically how much knowledge about a disease is encoded in vocabularies. In order to do so, it would be necessary the existence of relationship mappings between concepts from distinct vocabularies, but the most frequently established mappings are those defining equivalence. The former mappings define a relation other than hierarchical or of equivalence

(i.e., stating that two concepts have the same set of instances) (Machado *et al.*, 2013b), and would thus allow a search with a disease name to identify all the concepts related with the disease. BioPortal has a service that performs such type of search, but testing it for the diseases HCM and COPD yielded no results in terms of relationship mappings.

# Chapter 5

# Conclusions

The work presented herein is the first step in the creation of a disease analysis framework intended to assist medical doctors in the diagnosis and prognosis processes of a disease, eventually resulting in the advancement of the current knowledge about that disease. This framework will comprise two individual components: one for the representation and integration of patient data; and the other for the exploration of the patient data, aiming at the identification of new knowledge to assist in the diagnosis and/or the prognosis of a disease (Fig. 5.1).

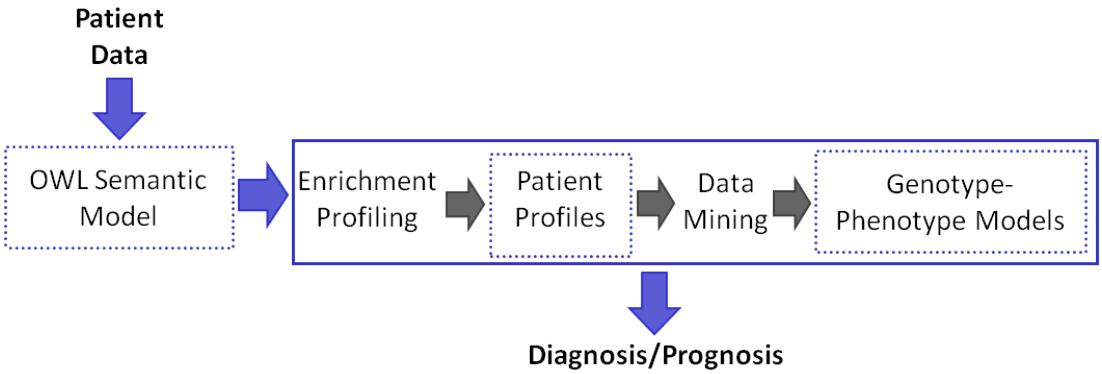


Figure 5.1: Disease analysis framework. This framework comprises a data representation and integration component, and a data enrichment and analysis component. The core of the first component is a semantic model that provides the conceptual representation of disease knowledge, which assists in the integration of heterogeneous data from patients. The second component involves an enrichment analysis to identify overrepresented ontology terms that can be used to profile the patients, and the use of these profiles by data mining algorithms to identify genotype-phenotype association models. These models can then be used to predict the diagnosis (or prognosis) of new patients.

The disease framework contributes to advance the translational medicine domain since it will enable the integration of data from basic science research with the clinical data, and its transformation into knowledge that can be used in the clinical practice.

The thesis underlying the present work consisted in exploring the use of ontologies in the implementation of the two components of the disease analysis framework and thus evaluate their role in the advancement of the knowledge about a disease. The first objective of this thesis was to create an ontological representation of the knowledge pertaining to a disease, exploiting semantic web technologies and existing ontologies. The second objective was to evaluate if the inclusion of knowledge from ontologies into translational medicine datasets would improve the quality of the obtained data exploration results.

The first objective was accomplished in the form of a semantic model of hypertrophic cardiomyopathy, a disease that benefits from a translational medicine approach. The model represents two heterogeneous domains of knowledge, the clinical and the genetic, and plays an important role in the integration of data from both domains through the relations established between the modules comprising the model. The development of the model in OWL, the reuse of existing vocabularies and the definition of mappings are all factors that facilitate the use of the model by third parties. Despite the use of a case-study, the methodology to create the model was planned so that it can be used for other diseases, as well as the model itself.

My contributions in the first objective of this thesis are threefold: the methodology for the creation of a semantic model in the OWL language; a semantic model of the disease hypertrophic cardiomyopathy; and a review on the exploitation of semantic web resources in translation medicine systems.

The second objective of this thesis was also accomplished: a methodology was devised that identifies enriched ontology terms associated with the genetic and clinical features of a dataset of patients; and the inclusion in the dataset of the enriched terms was evaluated with data mining algorithms. The complete analysis was tested with two distinct datasets, and the results show that using the enriched terms as features brings new knowledge into the dataset that results in the improvement of the predictions made with it.

My contributions in the second objective of this thesis are twofold: the adaptation of the standard enrichment analysis to use data from patients; and the application of the adapted enrichment analysis to improve the predictions made with a translational medicine dataset.

Overall, the work presented corresponds to a step forward in the use of knowledge representations to improve the current knowledge about a disease and its diagnosis and prognosis.

In order to promote the use of the semantic model, it can be subjected to a formal evaluation such as is normally done for ontologies, a task that was outside the goals of this work.

Additionally, further work should be done in both components of the disease analysis framework, before it can be used by medical doctors. In respect to the semantic model, it can be explored for the inference of new knowledge so that it can be included in the second step, the data exploration. In respect to the enrichment analysis, other enrichment approaches can be tested, namely the Modular Enrichment Analysis and the model-based approach; additional genetic ontologies can be considered, as well as vocabularies other than NCIt (due to the low number of annotations) and SNOMED-CT (due to its intrinsic problems, namely the representation of the same concept with more than one term).

Once the disease analysis framework is completed, its use can go beyond its original purpose. For instance, it will be possible to exploit the integrated data to construct risk prediction models for any number of factors, without burdening the medical doctors with the need to input the data themselves and without making pre-assumptions on the data that is more relevant to build the model.

# Appendix A

# Complete results of the enrichment profiling of SCD patients (HCM dataset)

For each term is indicated: GO accession number (Acc), term name, $p$-value without multiple-testing correction, $p$-value with Bonferroni correction ($p$-Bonf), annotation frequency in the study set (SFreq), annotation frequency in the population set (PFreq), and information content (IC).

Table A.1: Complete set of biological process enriched terms (Gene Ontology), obtained in the profiling analysis of SCD patients.

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| GO:0030049 | muscle filament sliding | 7.7E-40 | 4.1E-38 | 94% | 0.21% | 63% |
| GO:0033275 | actin-myosin filament sliding | 7.7E-40 | 4.1E-38 | 94% | 0.21% | 63% |
| GO:0055010 | ventricular cardiac muscle tissue morphogenesis | 7.7E-40 | 4.1E-38 | 94% | 0.21% | 63% |
| GO:0003229 | ventricular cardiac muscle tissue development | 2.4E-39 | 1.3E-37 | 94% | 0.22% | 62% |

(continued on next page)

Table A.1: (continued)

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| GO:0070252 | actin-mediated cell contraction | 6.8E-39 | 3.6E-37 | 94% | 0.24% | 61% |
| GO:0002027 | regulation of heart rate | 1.3E-31 | 6.9E-30 | 81% | 0.26% | 60% |
| GO:0007512 | adult heart development | 6.8E-25 | 3.6E-23 | 56% | 0.07% | 73% |
| GO:0032781 | positive regulation of ATPase activity | 2.9E-15 | 1.5E-13 | 38% | 0.09% | 72% |
| GO:0043462 | regulation of ATPase activity | 2.6E-14 | 1.4E-12 | 38% | 0.12% | 68% |
| GO:0032971 | regulation of muscle filament sliding | 1.0E-12 | 5.4E-11 | 25% | 0.02% | 89% |
| GO:0031034 | myosin filament assembly | 3.4E-12 | 1.8E-10 | 25% | 0.02% | 86% |
| GO:0031033 | myosin filament organization | 8.4E-12 | 4.5E-10 | 25% | 0.03% | 84% |
| GO:0030239 | myofibril assembly | 1.4E-10 | 7.6E-09 | 31% | 0.20% | 63% |
| GO:0060048 | cardiac muscle contraction | 4.2E-10 | 2.2E-08 | 31% | 0.25% | 61% |
| GO:0031032 | actomyosin structure organization | 5.1E-10 | 2.7E-08 | 31% | 0.26% | 60% |
| GO:0045214 | sarcomere organization | 6.5E-09 | 3.5E-07 | 25% | 0.14% | 67% |
| GO:0006942 | regulation of striated muscle contraction | 5.8E-08 | 3.1E-06 | 25% | 0.24% | 61% |
| GO:0051764 | actin crosslink formation | 3.0E-06 | 1.6E-04 | 13% | 0.02% | 89% |
| GO:0032780 | negative regulation of ATPase activity | 5.3E-06 | 2.8E-04 | 13% | 0.02% | 86% |
| GO:0035995 | detection of muscle stretch | 2.6E-03 | 1.4E-01 | 6% | 0.02% | 89% |
| GO:0035994 | response to muscle stretch | 3.4E-03 | 1.8E-01 | 6% | 0.02% | 86% |
| GO:0055003 | cardiac myofibril assembly | 1.4E-02 | 7.2E-01 | 6% | 0.09% | 72% |

(continued on next page)

Table A.1: (continued)

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| GO:0002026 | regulation of the force of heart contraction | 1.9E-02 | 9.9E-01 | 6% | 0.12% | 69% |
| GO:0014897 | striated muscle hypertrophy | 2.1E-02 | 1 | 6% | 0.13% | 67% |
| GO:0003300 | cardiac muscle hypertrophy | 2.1E-02 | 1 | 6% | 0.13% | 67% |
| GO:0050982 | detection of mechanical stimulus | 2.2E-02 | 1 | 6% | 0.14% | 67% |
| GO:0014896 | muscle hypertrophy | 2.3E-02 | 1 | 6% | 0.14% | 67% |
| GO:0055013 | cardiac muscle cell development | 2.9E-02 | 1 | 6% | 0.18% | 64% |
| GO:0055006 | cardiac cell development | 3.1E-02 | 1 | 6% | 0.20% | 63% |
| GO:0001974 | blood vessel remodeling | 3.4E-02 | 1 | 6% | 0.21% | 63% |

Table A.2: Complete set of molecular function enriched terms (Gene Ontology), obtained in the profiling analysis of SCD patients.

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| GO:0008307 | structural constituent of muscle | 2.9E-35 | 1.6E-33 | 88% | 0.25% | 61% |
| GO:0030898 | actin-dependent ATPase activity | 1.1E-26 | 6.1E-25 | 56% | 0.05% | 78% |
| GO:0000146 | microfilament motor activity | 1.8E-23 | 9.4E-22 | 56% | 0.11% | 70% |
| GO:0032036 | myosin heavy chain binding | 3.3E-11 | 1.8E-09 | 25% | 0.04% | 80% |
| GO:0001671 | ATPase activator activity | 9.1E-11 | 4.8E-09 | 25% | 0.05% | 78% |
| GO:0031432 | titin binding | 2.1E-10 | 1.1E-08 | 25% | 0.06% | 76% |
| GO:0060590 | ATPase regulator activity | 4.0E-10 | 2.1E-08 | 25% | 0.07% | 74% |

(continued on next page)

Table A.2: (continued)

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| GO:0017022 | myosin binding | 7.6E-09 | 4.0E-07 | 25% | 0.14% | 67% |
| GO:0030172 | troponin C binding | 5.3E-06 | 2.8E-04 | 13% | 0.02% | 86% |
| GO:0031013 | troponin I binding | 8.4E-06 | 4.4E-04 | 13% | 0.03% | 84% |
| GO:0005523 | tropomyosin binding | 7.6E-05 | 4.0E-03 | 13% | 0.08% | 72% |
| GO:0031433 | telethonin binding | 3.4E-03 | 1.8E-01 | 6% | 0.02% | 86% |
| GO:0042805 | actinin binding | 1.8E-02 | 9.4E-01 | 6% | 0.11% | 69% |

Table A.3: Complete set of cellular component enriched terms (Gene Ontology), obtained in the profiling analysis of SCD patients.

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| GO:0005859 | muscle myosin complex | 2.4E-37 | 1.3E-35 | 81% | 0.10% | 71% |
| GO:0032982 | myosin filament | 2.4E-37 | 1.3E-35 | 81% | 0.10% | 71% |
| GO:0016460 | myosin II complex | 1.1E-35 | 5.8E-34 | 81% | 0.13% | 68% |
| GO:0001725 | stress fiber | 4.1E-20 | 2.2E-18 | 56% | 0.25% | 61% |
| GO:0032432 | actin filament bundle | 7.3E-20 | 3.8E-18 | 56% | 0.27% | 60% |
| GO:0014705 | C zone | 9.2E-15 | 4.9E-13 | 25% | 0.01% | 100% |
| GO:0005863 | striated muscle myosin thick filament | 1.0E-12 | 5.4E-11 | 25% | 0.02% | 89% |
| GO:0031672 | A band | 4.7E-09 | 2.5E-07 | 25% | 0.13% | 68% |
| GO:0005861 | troponin complex | 2.2E-05 | 1.1E-03 | 13% | 0.04% | 79% |
| GO:0005865 | striated muscle thin filament | 6.6E-05 | 3.5E-03 | 13% | 0.07% | 73% |

# Appendix B

# Complete results of the enrichment profiling of no-SCD patients (HCM dataset)

For each term is indicated: GO accession number (Acc), term name, $p$-value without multiple-testing correction, $p$-value with Bonferroni correction ($p$-Bonf), annotation frequency in the study set (SFreq), annotation frequency in the population set (PFreq), and information content (IC).

Table B.1: Complete set of biological process enriched terms, obtained in the profiling analysis of no-SCD patients.

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| GO:0030049 | muscle filament sliding | 2.3E-252 | 1.7E-250 | 96% | 0.21% | 63% |
| GO:0033275 | actin-myosin filament sliding | 2.3E-252 | 1.7E-250 | 96% | 0.21% | 63% |
| GO:0055010 | ventricular cardiac muscle tissue morphogenesis | 2.3E-252 | 1.7E-250 | 96% | 0.21% | 63% |
| GO:0003229 | ventricular cardiac muscle tissue development | 3.2E-249 | 2.4E-247 | 96% | 0.22% | 62% |

(continued on next page)

## B. COMPLETE RESULTS OF THE ENRICHMENT PROFILING OF NO-SCD PATIENTS (HCM DATASET)

Table B.1: (continued)

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| GO:0070252 | actin-mediated cell contraction | 2.7E-246 | 2.0E-244 | 96% | 0.24% | 61% |
| GO:0043462 | regulation of ATPase activity | 1.1E-144 | 8.0E-143 | 59% | 0.12% | 68% |
| GO:0002027 | regulation of heart rate | 2.1E-133 | 1.6E-131 | 62% | 0.26% | 60% |
| GO:0032781 | positive regulation of ATPase activity | 1.8E-110 | 1.4E-108 | 45% | 0.09% | 72% |
| GO:0032780 | negative regulation of ATPase activity | 2.8E-96 | 2.1E-94 | 33% | 0.02% | 86% |
| GO:0007512 | adult heart development | 3.2E-89 | 2.4E-87 | 37% | 0.07% | 73% |
| GO:0060048 | cardiac muscle contraction | 1.1E-81 | 7.9E-80 | 42% | 0.25% | 61% |
| GO:0032971 | regulation of muscle filament sliding | 6.6E-73 | 4.9E-71 | 25% | 0.02% | 89% |
| GO:0031034 | myosin filament assembly | 1.3E-69 | 9.6E-68 | 25% | 0.02% | 86% |
| GO:0031033 | myosin filament organization | 4.3E-67 | 3.2E-65 | 25% | 0.03% | 84% |
| GO:0030239 | myofibril assembly | 1.5E-58 | 1.1E-56 | 31% | 0.20% | 63% |
| GO:0051764 | actin crosslink formation | 2.5E-56 | 1.8E-54 | 20% | 0.02% | 89% |
| GO:0031032 | actomyosin structure organization | 4.1E-55 | 3.0E-53 | 31% | 0.26% | 60% |
| GO:0045214 | sarcomere organization | 2.6E-51 | 1.9E-49 | 26% | 0.14% | 67% |
| GO:0006942 | regulation of striated muscle contraction | 4.0E-45 | 3.0E-43 | 26% | 0.24% | 61% |
| GO:0001980 | regulation of systemic arterial blood pressure by ischemic conditions | 6.0E-41 | 4.4E-39 | 13% | 0.01% | 100% |

(continued on next page)

Table B.1: (continued)

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| GO:0001976 | neurological system process involved in regulation of systemic arterial blood pressure | 3.4E-25 | 2.5E-23 | 13% | 0.08% | 72% |
| GO:0006940 | regulation of smooth muscle contraction | 6.6E-19 | 4.9E-17 | 13% | 0.25% | 61% |
| GO:0035995 | detection of muscle stretch | 2.5E-09 | 1.8E-07 | 4% | 0.02% | 89% |
| GO:0035994 | response to muscle stretch | 7.8E-09 | 5.8E-07 | 4% | 0.02% | 86% |
| GO:0055003 | cardiac myofibril assembly | 3.1E-08 | 2.3E-06 | 5% | 0.09% | 72% |
| GO:0002026 | regulation of the force of heart contraction | 1.5E-07 | 1.1E-05 | 5% | 0.12% | 69% |
| GO:0055013 | cardiac muscle cell development | 1.3E-06 | 9.4E-05 | 5% | 0.18% | 64% |
| GO:0055006 | cardiac cell development | 1.9E-06 | 1.4E-04 | 5% | 0.20% | 63% |
| GO:0003300 | cardiac muscle hypertrophy | 1.1E-05 | 8.2E-04 | 4% | 0.13% | 67% |
| GO:0014897 | striated muscle hypertrophy | 1.1E-05 | 8.2E-04 | 4% | 0.13% | 67% |
| GO:0050982 | detection of mechanical stimulus | 1.3E-05 | 9.6E-04 | 4% | 0.14% | 67% |
| GO:0014896 | muscle hypertrophy | 1.5E-05 | 1.1E-03 | 4% | 0.14% | 67% |
| GO:0001974 | blood vessel remodeling | 6.9E-05 | 5.1E-03 | 4% | 0.21% | 63% |
| GO:0007522 | visceral muscle development | 5.3E-03 | 3.9E-01 | 1% | 0.01% | 100% |
| GO:0042694 | muscle cell fate specification | 5.3E-03 | 3.9E-01 | 1% | 0.01% | 100% |
| GO:0055009 | atrial cardiac muscle tissue morphogenesis | 2.6E-02 | 1 | 1% | 0.03% | 84% |

(continued on next page)

## B. COMPLETE RESULTS OF THE ENRICHMENT PROFILING OF NO-SCD PATIENTS (HCM DATASET)

Table B.1: (continued)

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| GO:0003228 | atrial cardiac muscle tissue development | 2.6E-02 | 1 | 1% | 0.03% | 84% |
| GO:0048739 | cardiac muscle fiber development | 4.2E-02 | 1 | 1% | 0.04% | 79% |
| GO:0042693 | muscle cell fate commitment | 5.7E-02 | 1 | 1% | 0.06% | 76% |

Table B.2: Complete set of molecular function enriched terms, obtained in the profiling analysis of no-SCD patients.

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| GO:0008307 | structural constituent of muscle | 1.7E-149 | 1.3E-147 | 67% | 0.25% | 61% |
| GO:0030898 | actin-dependent ATPase activity | 1.8E-96 | 1.3E-94 | 37% | 0.05% | 78% |
| GO:0030172 | troponin C binding | 2.8E-96 | 2.1E-94 | 33% | 0.02% | 86% |
| GO:0000146 | microfilament motor activity | 2.1E-83 | 1.6E-81 | 37% | 0.11% | 70% |
| GO:0032036 | myosin heavy chain binding | 2.5E-66 | 1.9E-64 | 26% | 0.04% | 80% |
| GO:0001671 | ATPase activator activity | 1.5E-60 | 1.1E-58 | 25% | 0.05% | 78% |
| GO:0031432 | titin binding | 2.5E-58 | 1.9E-56 | 25% | 0.06% | 76% |
| GO:0060590 | ATPase regulator activity | 1.7E-56 | 1.3E-54 | 25% | 0.07% | 74% |
| GO:0031013 | troponin I binding | 9.8E-52 | 7.3E-50 | 20% | 0.03% | 84% |
| GO:0017022 | myosin binding | 6.9E-51 | 5.1E-49 | 26% | 0.14% | 67% |
| GO:0005523 | tropomyosin binding | 4.8E-42 | 3.5E-40 | 20% | 0.08% | 72% |
| GO:0031014 | troponin T binding | 9.9E-33 | 7.3E-31 | 13% | 0.02% | 86% |
| GO:0019855 | calcium channel inhibitor activity | 1.9E-31 | 1.4E-29 | 13% | 0.03% | 84% |
| GO:0008200 | ion channel inhibitor activity | 1.4E-23 | 1.0E-21 | 13% | 0.11% | 70% |

(continued on next page)

Table B.2: (continued)

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| GO:0016248 | channel inhibitor activity | 1.4E-23 | 1.0E-21 | 13% | 0.11% | 70% |
| GO:0005246 | calcium channel regulator activity | 4.9E-23 | 3.6E-21 | 13% | 0.12% | 69% |
| GO:0048306 | calcium-dependent protein binding | 1.1E-19 | 8.1E-18 | 13% | 0.21% | 63% |
| GO:0031433 | telethonin binding | 7.8E-09 | 5.8E-07 | 4% | 0.02% | 86% |
| GO:0042805 | actinin binding | 5.6E-06 | 4.2E-04 | 4% | 0.11% | 69% |
| GO:0030899 | calcium-dependent ATPase activity | 1.6E-02 | 1 | 1% | 0.02% | 89% |
| GO:0003785 | actin monomer binding | 7.2E-02 | 1 | 1% | 0.07% | 73% |

Table B.3: Complete set of cellular component enriched terms, obtained in the profiling analysis of no-SCD patients.

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| GO:0032982 | myosin filament | 9.0E-161 | 6.7E-159 | 62% | 0.10% | 71% |
| GO:0005859 | muscle myosin complex | 9.0E-161 | 6.7E-159 | 62% | 0.10% | 71% |
| GO:0016460 | myosin II complex | 7.3E-153 | 5.4E-151 | 62% | 0.13% | 68% |
| GO:0014705 | C zone | 2.5E-86 | 1.8E-84 | 25% | 0.01% | 100% |
| GO:0005861 | troponin complex | 6.6E-86 | 4.9E-84 | 33% | 0.04% | 79% |
| GO:0005865 | striated muscle thin filament | 1.0E-77 | 7.7E-76 | 33% | 0.07% | 73% |
| GO:0005863 | striated muscle myosin thick filament | 6.6E-73 | 4.9E-71 | 25% | 0.02% | 89% |
| GO:0001725 | stress fiber | 1.4E-69 | 1.0E-67 | 37% | 0.25% | 61% |
| GO:0032432 | actin filament bundle | 1.4E-68 | 1.0E-66 | 37% | 0.27% | 60% |
| GO:0031672 | A band | 8.7E-50 | 6.5E-48 | 25% | 0.13% | 68% |

# Appendix C

# Complete results of the differential enrichment analysis of SCD and no-SCD patients (HCM dataset)

For each term is indicated: GO accession number (Acc), term name, $p$-value without multiple-testing correction, $p$-value with Bonferroni correction ($p$-Bonf), annotation frequency in the study set (SFreq), annotation frequency in the population set (PFreq), and information content (IC).

Table C.1: Complete results of the differential enrichment analysis obtained with the HCM dataset and the sub-group of SCD patients.

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| **Molecular Function** | | | | | | |
| GO:0008307 | structural constituent of muscle | 8.1E-02 | 1 | 88% | 70% | 61% |

## C. COMPLETE RESULTS OF THE DIFFERENTIAL ENRICHMENT ANALYSIS OF SCD AND NO-SCD PATIENTS (HCM DATASET)

Table C.2: Complete results of the differential enrichment analysis obtained with the HCM dataset and the sub-group of no-SCD patients.

| Acc | Name | $p$-value | $p$-Bonf | SFreq | PFreq | IC |
|---|---|---|---|---|---|---|
| **Biological Process** | | | | | | |
| GO:0032780 | negative regulation of ATPase activity | 8.1E-02 | 1 | 33% | 30% | 86% |
| GO:0043462 | regulation of ATPase activity | 9.1E-02 | 1 | 59% | 56% | 68% |
| **Molecular Function** | | | | | | |
| GO:0030172 | troponin C binding | 8.1E-02 | 1 | 33% | 30% | 86% |
| **Cellular Component** | | | | | | |
| GO:0005865 | striated muscle thin filament | 8.1E-02 | 1 | 33% | 30% | 73% |
| GO:0005861 | troponin complex | 8.1E-02 | 1 | 33% | 30% | 79% |

# Appendix D

# Default settings of the data mining algorithms

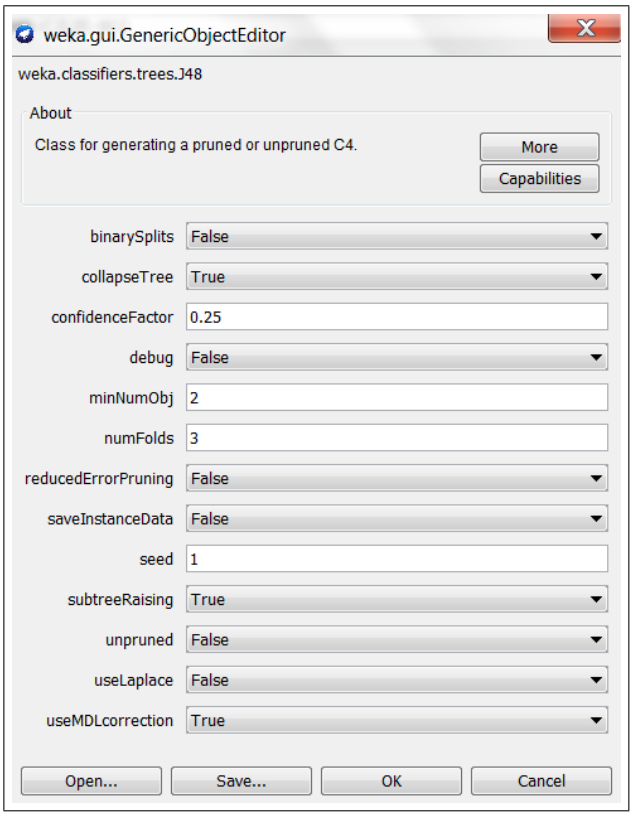


Figure D.1: Default settings of the J48 algorithm (decision trees).

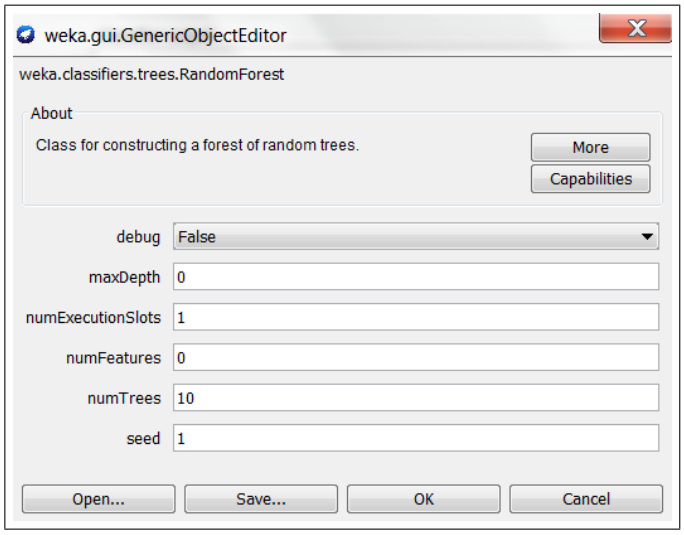


Figure D.2: Default settings of the Random Forest algorithm (decision trees).

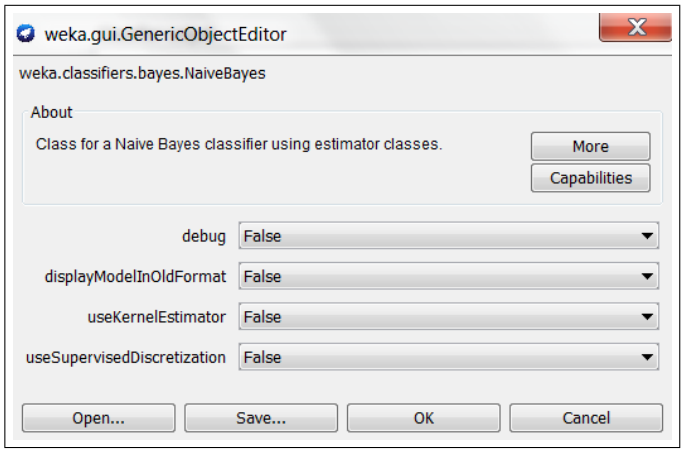


Figure D.3: Default settings of the Naive Bayes algorithm.

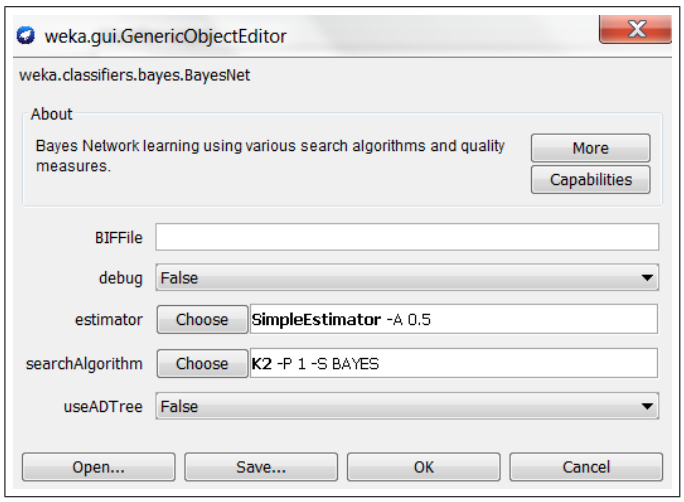


Figure D.4: Default settings of the Bayes Network algorithm.

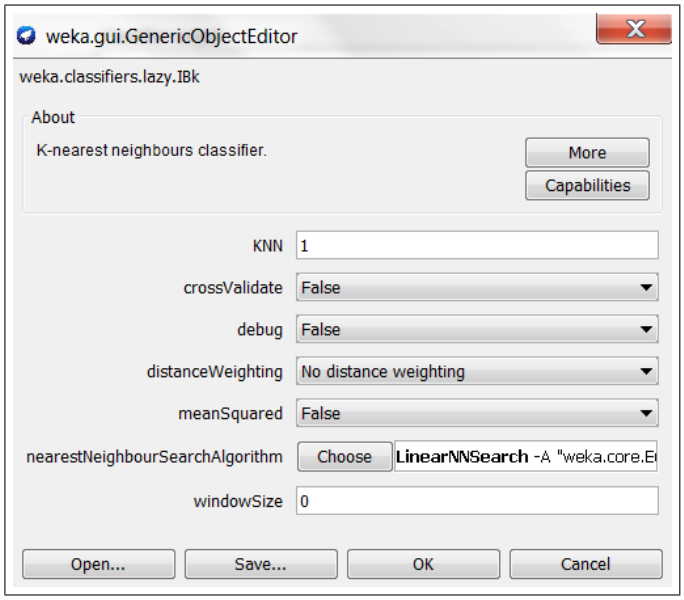


Figure D.5: Default settings of the K-nearest neighbors algorithm.

# References


Agorastos, T., Koutkias, V., Falelakis, M., Lekka, I., Mikos, T., Delopoulos, A., Mitkas, P.A., Tantsis, A., Weyers, S., Coorevits, P., Kaufmann, A.M., Kurzeja, R. & Maglaveras, N. (2009). Semantic integration of cervical cancer data repositories to facilitate multicenter association studies: the ASSIST approach. *Cancer Informatics*, **8**, 31–44. 16, 20, 21

Aha, D. & Kibler, D. (1991). Instance-based learning algorithms. *Machine Learning*, **6**, 37–66. 92

Albani, S. & Prakken, B. (2009). The advancement of translational medicine from regional challenges to global solutions. *Nature Medicine*, **15**, 1006–1009. 1

Alcalai, R., Seidman, J.G. & Seidman, C.E. (2008). Genetic basis of hypertrophic cardiomyopathy: from bench to the clinics. *Journal of Cardiovascular Electrophysiology*, **19**, 104–110. 31, 32

Alexandre, B.M.C. (2011). *Chronic obstructive pulmonary disease: a proteomics approach*. Ph.D. thesis, Faculdade de Ciências, Universidade de Lisboa. 70

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., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29. 3, 10, 26, 74

## REFERENCES


Atkinson, A.J., Colburn, W.A., DeGruttola, V.G., DeMets, D.L., Downing, G.J., Hoth, D.F., Oates, J.A., Peck, C.C., Schooley, R.T., Spilker, B.A., Woodcock, J. & Zeger, S.L. (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, **69**, 89–95. 9

Baader, F., Calvanese, D., McGuinness, D., Nardi, D. & Patel-Schneider, P., eds. (2003). *The description logic handbook: theory, implementation, and applications*. Cambridge University Press, New York. 12

Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C. & Apweiler, R. (2009). The GOA database in 2009 - an integrated gene ontology annotation resource. *Nucleic Acids Research*, **37**, D396–D403. 75

Bauer, B., Gagneur, J. & Robinson, P. (2010). GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Research*, **38**, 3523–3532. 25

Bauer, S., Grossmann, S., Vingron, M. & Robinson, P.N. (2008). Ontologizer 2.0 - a multifunctional tool for go term enrichment analysis and data exploration. *Bioinformatics*, **24**, 1650–1651. 25

Beisswanger, E., Lee, V., Kim, J., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., Schulz, S. & Hahn, U. (2008). Gene Regulation Ontology (GRO): design principles and use cases. *Studies in Health Technology and Informatics*, **136**, 9–14. 10, 35

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The semantic web. *Scientific American*, **284**, 34–43. 2

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, **32**, D267–D270. 19

Bos, J.M., Towbin, J.A. & Ackerman, M.J. (2009). Diagnostic, prognostic, and therapeutic implications of genetic testing for hypertrophic cardiomyopathy. *Journal of the American College of Cardiology*, **54**, 201–211. 86

Bouckaert, R.R. (2004). Bayesian networks in weka. Tech. Rep. 14/2004, University of Waikato. 92

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32. 92

Brinkman, R., Courtot, M., Derom, D., Fostel, J., He, Y., Lord, P., Malone, J., Parkinson, H., Peters, B., Rocca-Serra, P., Ruttenberg, A., Sansone, S., Soldatova, L., Stoeckert, C., Turner, J. & Zheng, J. (2010). Modeling biomedical experimental processes with OBI. *Journal of biomedical semantics*, **1**, S7. 19

Brito, D., Richard, P., Isnard, R., Pipa, J., Komajda, M. & Madeira, H. (2003). Familial hypertrophic cardiomyopathy: the same mutation, different prognosis. comparison of two families with a long follow-up. *Revista Portuguesa de Cardiologia*, **22**, 1445–1461. 32

Broeckaert, F. & Bernard, A. (2000). Clara cell secretory protein (CC16): characteristics and perspectives as lung peripheral biomarker. *Clinical & Experimental Allergy*, **30**, 469–475. 73

Chen, H., Mao, Y., Zheng, X., Cui, M., Feng, Y., Deng, S., Yin, A., Zhou, C., Tang, J., Jiang, X. & Wu, Z. (2007). Towards semantic e-Science for traditional chinese medicine. *BMC Bioinformatics*, **8**, S6. 16, 20

Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., T.Chan, E., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Karra, K., Krieger, C.J., Miyasato, S.R., Nash, R.S., Park, J., Skrzypek, M.S., Simison, M., Weng, S. & Wong, E.D. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research*, **40**, D700–D705. 11

Cheung, K., Frost, H., Marshall, M., Prud'hommeaux, E., Samwald, M., Zhao, J. & Paschke, A. (2009). A journey to semantic web query federation in the life sciences. *BMC Bioinformatics*, **10**, S10. 16, 17, 24

Colombo, G., Merico, D., Boncoraglio, G., Paoli, F.D., Ellul, J., Frisoni, G., Nagy, Z., van der Lugt, A., Vassányi, I. & Antoniotti, M. (2010). An ontological modeling approach to cerebrovascular disease studies: the NEUROWEB case. *Journal of Biomedical Informatics*, **43**, 469–484. 16, 20

Committee on Models for Biomedical Research (1985). *Models for biomedical research: a new perspective*. National Academy Press, Washington, D.C. 2

Constantine, L.L. (2009). Interaction design and model-driven development. In *Model Driven Engineering Languages and Systems*, Lecture Notes in Computer Science, 377, Springer. 32

Coulet, A., Smail-Tabbone, M., Napoli, A. & Devignes, M. (2006). Suggested ontology for pharmacogenomics (SO-Pharm): modular construction and preliminary testing. In *Proceedings of International Workshop on Knowledge Systems in Bioinformatics*. 20

Coulet, A., Smail-Tabbone, M., Benlian, P., Napoli, A. & Devignes, M. (2008). Ontology-guided data preparation for discovering genotype-phenotype relationships. *BMC Bioinformatics*, **9**, S3. 16, 20

Couto, F.M. & Pinto, H.S. (2013). The next generation of similarity measures that fully explore the semantics in biomedical ontologies. *Journal of Bioinformatics and Computational Biology*, **11**, 1–12. 12

Criteria Committee of the New York Heart Association (1994). *Nomenclature and criteria for diagnosis of diseases of the heart and great vessels*. Boston. 39

Davidson, S.B., Overton, C. & Buneman, P. (1995). Challenges in integrating biological data sources. *Journal of Ccomputational Biology*, **2**, 557–572. 2

Dodds, L. & Davis, I. (2012). *Linked Data Patterns: A pattern catalogue for modelling, publishing, and consuming Linked Data*. http://patterns.dataincubator.org/book/. 13

Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R. & Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, **6**, R44. 10, 19, 35

Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E. & the Mouse Genome Database Group (2012). The mouse genome database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Research*, **40**, D881–D886. 11

Fikes, R. & Kehler, T. (1985). The role of frame-based representation in reasoning. *Communications of the ACM*, **28**, 904–920. 20

Fitting, J. (2000). Transfer factor for carbon monoxide: a glance behind the scene. *Swiss Medical Weekly*, **134**, 413–418. 73

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A.K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W.M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T.J.P., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A. & Searle, S.M.J. (2013). Ensembl 2013. *Nucleic Acids Research*, **41**, D48–D55. 11

Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, **10**, 241–251. 9

Gangemi, A., Catenacci, C., Ciaramita, M. & Lehmann, J. (2006). Modelling ontology evaluation and validation. In *Proceedings of the 3rd European Semantic Web Conference (ESWC2006)*, Springer. 29

GOLD (2013). *Global Strategy for the Diagnosis, Management and Prevention of COPD*. Global Initiative for Chronic Obstructive Lung Disease (GOLD), http://www.goldcopd.org/guidelines-global-strategy-for-diagnosis-management.html. 70, 71

Gong, X., Yu, H., Yang, C.B. & Li, Y.F. (2012). Knowledge enrichment analysis for human tissue-specific genes uncover new biological insights. *Journal of Integrative Bioinformatics*, **9**. 26

Gruber, T.R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5**, 199–220. 3

Gudivada, R.C., Qu, X.A., Chen, J., Jegga, A.G., Neumann, E.K. & Aronow, B.J. (2008). Identifying disease-causal genes using semantic web-based representation of integrated genomic and phenomic knowledge. *Journal of Biomedical Informatics*, **41**, 717–729. 16, 21

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, **11**. 92

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, **33**, D514–517. 19

Harris, T.W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W.J., De La Cruz, N., Davis, P., Duesbury, M., Fang, R., Fernandes, J., Han, M., Kishore, R., Lee, R., Müller, H., Nakamura, C., Ozersky, P., Petcherski, A., Rangarajan, A., Rogers, A., Schindelman, G., Schwarz, E.M., Tuli, M.A., Auken, K.V., Wang, D., Wang, X., Williams, G., Yook, K., Durbin, R., Stein, L.D., Spieth, J. & Sternberg, P.W. (2010). WormBase: a comprehensive resource for nematode research. *Nucleic Acids Research*, **38**, D463–D467. 11

Hey, T., Tansley, S. & Tolle, K., eds. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington. 1

Ho, C.Y. (2010). Genetics and clinical destiny: Improving care in hypertrophic cardiomyopathy. *Circulation*, **122**, 24302440. 32

Hoehndorf, R., Dumontier, M. & Gkoutos, G. (2012). Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. *Bioinformatics*, **28**, 2169–75. 26

Huang, D., Sherman, B. & Lempicki, R. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, **37**, 1–13. 25

John, G.H. & Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo*, 338–345. 92

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, **40**, D109–D114. 26

Kapushesky, M., Adamusiak, T., Burdett, T., Culhane, A., Farne, A., Filippov, A., Holloway, E., Klebanov, A., Kryvych, N., Kurbatova, N., Kurnosov, P., Malone, J., Melnichuk, O., Petryszak, R., Pultsin, N., Rustici, G., Tikhonov, A., Travillian, R.S., Williams, E., Zorin, A., Parkinson, H. & Brazma, A. (2012). Gene Expression Atlas update a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Research*, **40**, D1077–D1081. 19

Khatri, P., Draghici, S., Ostermeier, G. & Krawetz, S. (2002). Profiling gene expression using onto-express. *Genomics*, **79**, 266–270. 25

Khatri, P., Sirota, M. & Butte, A.J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, **8**, e1002375. 3

LePendu, P., Musen, M. & Shah, N. (2011). Enabling enrichment analysis with the human disease ontology. *Journal of Biomedical Informatics*, **44**, S31–8. 26, 27, 89

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D. & Darnell, J. (2000). *Molecular Cell Biology*. W. H. Freeman, New York, 4th edn. 86

Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A. & Tarczy-Hornoch, P. (2007). Data integration and genomic medicine. *Journal of Biomedical Informatics*, **40**, 5–16. 2, 9

Lu, Y., Rosenfeld, R., Simon, I., Nau, G.J. & Bar-Joseph, Z. (2008). A probabilistic generative model for go enrichment analysis. *Nucleic Acids Research*, **36**, e109. 25

Luciano, J.S., Andersson, B., Batchelor, C., Bodenreider, O., Clark, T., Denney, C.K., Domarew, C., Gambet, T., Harland, L., Jentzsch, A., Kashyap, V., Kos, P., Kozlovsky, J., Lebo, T., Marshall, S.M., McCusker, J.P., McGuinness, D.L., Ogbuji, C., Pichler, E., Powers, R.L., Prud'hommeaux, E., Samwald, M., Schriml, L., Tonellato, P.J., Whetzel, P.L., Zhao, J., Stephens, S. & Dumontier, M. (2011). The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside. *Journal of Biomedical Semantics*, **2**, S1. 19

Machado, C.M., Couto, F., Fernandes, A.R., Santos, S., Cardim, N. & Freitas, A.T. (2010). Semantic characterization of hypertrophic cardiomyopathy disease. In *First Workshop on Knowledge Engineering, Discovery and Dissemination in Health (KEDDH10)*. 5

Machado, C.M., Couto, F.M., Fernandes, A.R., Santos, S. & Freitas, A.T. (2012a). Toward a translational medicine approach for hypertrophic cardiomyopathy. In *Lecture Notes In Computer Science - 3rd Information technology in bio and medical informatics International Conference*. 5

Machado, C.M., Freitas, A.T. & Couto, F.M. (2012b). Enrichment analysis applied to disease prognosis. In *Proceedings of the 4th Workshop of the GI workgroup "Ontologies in biomedicine and life sciences" (OBML) 2012*. 6

Machado, C.M., Freitas, A.T. & Couto, F.M. (2013a). Enrichment analysis applied to disease prognosis. *Journal of Biomedical Semantics*, **4**, 21. 6

Machado, C.M., Rebholz-Schuhmann, D., Freitas, A.T. & Couto, F.M. (2013b). The semantic web in translational medicine: current applications and future directions. *Briefings in Bioinformatics*. 5, 53, 102

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A. & Parkinson, H. (2010). Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**, 1112–1118. 19

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F.C., McCarroll, S.A. & Visscher, P.M. (2010). Finding the missing heritability of complex diseases. *Nature*, **461**, 747753. 9

Maron, B.J., Maron, M.S., Wigle, E.D. & Braunwald, E. (2009). The 50-year history,controversy, and clinical implications of left ventricular outflow tract obstruction in hypertrophic cardiomyopathy: from idiopathic hypertrophic subaortic stenosis to hypertrophic cardiomyopathy. *Journal of the American College of Cardiology*, **54**, 191–200. 31, 32

Marygold, S., Leyland, P., Seal, R., Goodman, J., Thurmond, J., Strelets, V., Wilson, R. & the FlyBase Consortium (2013). FlyBase: improvements to the bibliography. *Nucleic Acids Research*, **41**, D751–D757. 11

Min, H., Manion, F.J., Goralczyk, E., Wong, Y.N., Ross, E. & Beck, J.R. (2009). Integration of prostate cancer clinical data using an ontology. *Journal of Biomedical Informatics*, **42**, 1035 – 1045. 19

Noy, N.F. & McGuinness, D.L. (2001). Ontology development 101: A guide to creating your first ontology. Tech. Rep. KSL-01-05, Knowledge Systems, AI Laboratory, Stanford University. 14, 15, 31

Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M., Chute, C.G., & Musen, M.A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, **37**, W170W173. 30

Pathak, J., Kiefer, R.C., Bielinski, S.J. & Chute, C.G. (2012). Applying semantic web technologies for phenome-wide scan using an electronic health record linked biobank. *Journal of Biomedical Semantics*, **3**, 10. 16, 17, 19

Pesquita, C., Faria, D., ao, A.O.F., Lord, P. & Couto, F.M. (2009). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, **5**, e1000443. 3

Qu, X., Gudivada, R., Jegga, A., Neumann, E. & Aronow, B. (2007). Semantic web-based data representation and reasoning applied to disease mechanism and pharmacology. In *Bioinformatics and Biomedicine Workshops, 2007. BIBMW 2007. IEEE International Conference on*, 131–143. 16, 17, 20, 21

Quinlan, J.R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Mateo, CA, USA. 92

Rebholz-Schuhmann, D., Oellrich, A. & Hoehndorf, R. (2012). Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics*, **13**, 829–839. 3

Rebholz-Schuhmann, D., Grabmuller, C., Kavaliauskas, S., Croset, S., Kapushesky, M., Stott, I., Woollard, P., Westaway, M., Wilkinson, N., Marshall, C., Clark, D., Backofen, R., Filsell, W. & Harrow, I. (2013). Semantic integration of gene-disease associations

for type 2 diabetes mellitus from literature and 2 biomedical data resources (accepted for publication). *Drug Discovery Today*. 16, 17, 19

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 448–453. 75

Robinson, P. & Bauer, S. (2011). Overrepresentation analysis. In N. Britton, X. Lin, H.M. Safer, M. Singh & A. Trammontano, eds., *Introduction to Bio-Ontologies*, Mathematical and Computational Biology Series, 181–218, CRC Press, Taylor and Francis Group. 3, 25, 26, 27

Sagotsky, J.A., Zhang, L., Wang, Z., Martin, S. & Deisboeck, T.S. (2008). Life sciences and the web: a new era for collaboration. *Molecular Systems Biology*, **4**, 201. 2

Sahoo, S.S., Zeng, K., Bodenreider, O. & Sheth, A. (2007). From "glycosyltransferase" to "congenital muscular dystrophy": integrating knowledge from NCBI Entrez Gene and the Gene Ontology. *Studies in Health Technology and Informatics*, **129**, 1260–1264. 16, 17, 19

Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., Feng, G. & Kibbe, W.A. (2012). Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, **40**, D940–D946. 19, 26

Siddiqi, J., Akhgar, B., Gruzdz, A., Zaefarian, G. & Ihnatowicz, A. (2008). Automated diagnosis system to support colon cancer treatment: MATCH. In *Fifth International Conference on Information Technology: New Generations, IEEE Press.*, 201–205. 16, 20

Sim, I., Carini, S., Tu, S., Wynden, R., Pollock, B.H., Mollah, S.A., Gabriel, D., Hagler, H.K., Scheuermann, R.H., Lehmann, H.P., Wittkowski, K.M., Nahm, M. & Bakken, S. (2010). The human studies database project: Federating human studies design data using the ontology of clinical research. In *AMIA Summits Translational Science Proceedings*, 51–55. 10, 34

## REFERENCES


Sioutos, N., Coronado, S., Haber, M.W., Hartel, F.W., Shaiu, W.L. & Wright, L.W. (2007). NCI thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, **40**, 30–43. 10, 20, 24, 34, 74

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., The OBI Consortium, Leontis, N., Philippe R-S, Ruttenberg, A., Sansone S-A, Scheuermann, R.H., Shah, N., Whetzel, P.L. & Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, **25**, 1251–1255. 51

Stein, L.D. (2003). Integrating biological databases. *Nature Reviews Genetics*, **4**, 337–345. 3

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. & Mesirov, J. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, **102**, 15545–15550. 25

The UniProt Consortium (2007). The universal protein resource (UniProt). *Nucleic Acids Research*, **35**, D193–D197. 19

Tirmizi, S., Aitken, S., Moreira, D., Mungall, C., Sequeda, J., Shah, N. & Miranker, D. (2011). Mapping between the OBO and OWL ontology languages. *Journal of Biomedical Semantics*, **2011**, S3. 13

Tirrell, R., Evani, U., Berman, A.E., Mooney, S.D., Musen, M.A. & Shah, N.H. (2010). An ontology-neutral framework for enrichment analysis. *AMIA Annual Symposium Proceedings*, **2010**, 797–801. 26, 89

Webb, C.P. & Pass, H.I. (2004). Translation research: from accurate diagnosis to appropriate treatment. *Journal of Translational Medicine*, **2**, 35. 1

Wei, W. (2012). Global health and translational medicine: New drivers for medicine and medical sciences. *Journal of Medicine and Medical Sciences*, **3**, 126–127. 8

Wittwer, C.T., Reed, G.H., Gundry, C.N., Vandersteen, J.G. & Pryor, R.J. (2003). High-resolution genotyping by amplicon melting analysis using lcgreen. *Clinical Chemistry*, **49**, 853–860. 69

Woolf, S.H. (2008). The meaning of translational research and why it matters. *JAMA*, **299**, 211–213. 1, 8

Zhang, S., Cao, J., Kong, Y. & Scheuermann, R. (2010). GO-Bayes: Gene Ontology-based overrepresentation analysis using a Bayesian approach. *Bioinformatics*, **26**, 905–911. 26