Supporting Translational Research on Inherited Cardiomyopathies through Information Technology

C. Larizza1; M. Gabetta1; G. Milani1; M. Bucalo1; F. Mulas2; A. Nuzzo2,3; V. Favalli4; E. Arbustini4; R. Bellazzi1

1Dipartimento di Ingegneria Industriale e dell’Informazione, Università di Pavia, Pavia, Italy; 2Bioinformatics Unit, Centre for Tissue Engineering, University of Pavia, Pavia, Italy; 3Department of Biotechnology, BU Oncology, Nerviano Medical Sciences, Viagrande (CT), Italy; 4IRCCS Fondazione Policlinico S. Matteo, Pavia, Italy

Keywords
Translational research, biomedical research, dilated cardiomyopathy

Summary
Objectives: The INHERITANCE project, funded by the European Commission, is aimed at studying genetic or inherited Dilated cardiomyopathies (DCM) and at understanding the impact and management of the disease within families that suffer from heart conditions that are caused by DCMs. The biomedical informatics research activity of the project aims at implementing information technology solutions to support the project team in the different phases of their research, in particular in genes screening prioritization and new gene-disease association discovery.

Methods: In order to manage the huge quantity of scientific, clinical and patient data generated by the project several advanced biomedical informatics tools have been developed. The paper describes a layer of software instruments to support transcription of the results of the project in clinical practice as well as to support the scientific discovery process. This layer includes data warehousing, intelligent querying of the phenotype data, integrated search of biological data and knowledge repositories, text mining of the relevant literature, and case based reasoning.

Results: At the moment, a set of 1,394 patients and 9,784 observations has been stored into the INHERITANCE data warehouse. The literature database contains more than 1,100,000 articles retrieved from the Pubmed and generically related to cardiac diseases, already analyzed for extracting medical concepts and genes.

Conclusions: After two years of project the data warehouse has been completely set up and the text mining tools for automatic literature analysis have been implemented and tested. A first prototype of the decision support tool for knowledge discovery and gene prioritization is available, but a more complete release is still under development.

1. Introduction
Cardiomyopathies are defined as primary myocardial disorders of unknown cause and are classified into four main subtypes, based on ventricular morphology and physiology: hypertrophic (HCM), dilated (DCM), restrictive (RCM), and arrhythmogenic right ventricular cardiomyopathy (ARVC) [1]. Overlapping phenotypes (for example, dilatation of a hypertrophied ventricle) can occur in advanced phases of the disease. DCM is defined as a myocardial disorder characterized by the presence of left ventricular dilatation and systolic impairment, in the absence of abnormal loading conditions (e.g. hypertension, valve disease) or coronary artery disease sufficient to cause global systolic dysfunction [1]. Right ventricular dilatation and dysfunction may also be present. Family screening and genetic studies have identified more than 20 disease-causing genes to date [2].

Currently, patients with DCM are treated in accordance with international guidelines for the management of heart failure with little consideration of the possible influence of the underlying aetiology on the response to treatment. Recent studies suggest that this might result in sub-optimal or inappropriate therapy in some patients [3]. For example, knowledge that a DCM patient is carrier of a LMNA gene mutation might be of major importance when deciding on device therapy.

The INHERITANCE project (Integrated Heart Research In Translational Genetics of Cardiomyopathies in Europe), a multi-disciplinary, multi-centre research project funded by the European Commission, seeks to study the genetics of inherited DCM and to understand the impact and management of the condition within families that suffer from DCMs.

The INHERITANCE translational strategy is based on a clinical algorithm that seeks to determine disease-specific features (red flags) that associate with different types of DCM or suggest specific genetic or metabolic pathways of disease. A reverse
translational strategy will be run in parallel to establish or confirm the association of DCM phenotypes with clinical markers that occur in patients with DCM caused by the different genes.

The INHERITANCE project is structured into six research areas that study different facets of the DCM condition, including clinical cardiogenetics, -omics, i.e. genetic testing, transcriptomics, proteomics and metabolomics, animal studies, structural studies, treatments, and biomedical informatics, which aims to implement information technology solutions to support the project team in managing the huge quantity of scientific, clinical and patient data generated by the project. This paper focuses on the biomedical informatics methods and tools that have been made available to the INHERITANCE researchers.

The rest of the paper is organized as follows:

- Section 4 shows the results of the IT solutions obtained so far.

2. The General Architecture of the INHERITANCE Knowledge Management System

INHERITANCE, on top of a database application collecting the clinical data of DCM patients, implements a layer of software instruments to support translation of the results of the project in clinical practice as well as to support the scientific discovery process. This layer includes data warehousing (the i2b2 block shown in Figure 1), intelligent querying of the phenotype data, integrated search on biological data and knowledge repositories, text mining of the relevant literature, and case-based reasoning. We refer to these components as the knowledge-management system of INHERITANCE.

The overall design of the knowledge management architecture is described in Figure 1. The data warehouse is populated through a set of automated queries that extract patients' data from the INHERITANCE database. The data are then made available to the researchers through a data mining and exploration tool. Literature is analyzed through a text mining strategy based on Natural Language Processing (NLP). Finally a decision support tool, exploiting the patient database, the text-mining tool and software solutions to automatically access biomedical databases, provides support in refining patient’s diagnosis. The Decision Support System (DSS) consists of a CBR module that finds in the available Case Base (CB) the most similar cases to the current one on the basis of selected features. The literature analysis has a double use in the DSS: new cases extraction and feature prioritization. In the first case it can be used to extend the CB (made by the INHERITANCE patient set) with new "virtual" cases described in the scientific literature. The features of such patients are extracted from the scientific articles using NLP techniques. The second application of the literature analysis tool is the selection among a large set of features of the most relevant ones for the case comparison. The features relevance is derived from their occurrence in the DCM articles.

In the following we will describe each of the components included in the final architecture and the state-of-art of the project.
3. INHERITANCE Data Warehouse

The INHERITANCE project collects patients’ data in a Electronic Medical Record (EMR), called Cardioregister (https://cardioregister.com/Pages/Main.aspx), a web-based system designed to store, exploit and download data of patients and families with DCM, offering the ability to produce customized reports. The data collected in Cardioregister by the clinical centers involved in the project are periodically uploaded, with an automatic procedure, into a data warehouse for data exploration and dynamic querying.

The data warehouse exploited in the INHERITANCE project is based on the i2b2 (Informatics for Integrating Biology and the Bedside) software system [4] (http://www.i2b2.org/software). The goal of the i2b2 project is to provide clinical investigators with a software infrastructure able to integrate in a data warehouse clinical records and research data. Moreover, it provides a powerful and easy to use query generation tool useful for extracting patients’ sets meeting specified criteria. The choice of using the open-source i2b2 infrastructure has been, therefore, suggested by the availability of many tools ready to use for data exploration and stratification, the possibility of easily adding new modules for specific purposes and, finally, the easy integration of data coming from different sources (e.g. data extracted from clinical reports using NLP based tools). The i2b2 project has been successfully adopted in more than fifty hospitals and universities in the United States and more than fifteen in the rest of the world. It is also the starting point for the implementation of open-source extensions, such as tranSMART [5]. An interesting example of application of the i2b2 infrastructure is given by the ONCO-i2b2 project described in [6].

The i2b2 data model is based on the star schema [7] for which central table is the fact table where each row represents a single observation (fact) about a patient. The basic attributes of an observation are: the patient and provider codes, the code for the concept observed, a start and end date and the value. Besides the fact table, the i2b2 data repository contains four dimension tables containing further descriptive and analytical information that characterize the facts. Facts are associated to concept codes and the hierarchical structure of these codes, together with their descriptive terms and some other information, form the i2b2 ontology. The four dimensions in the i2b2 data warehouse are: Patient, Visit, Concept and Provider dimensions. In particular, the Concept dimension table maps the terms vocabulary to the codes of the patient observations.

3.1 Running the i2b2 system

As mentioned above, the infrastructure created by the i2b2 software allows investigators to perform queries on the data warehouse. This infrastructure consists of a collection of interoperable services (called cells) taking part of the overall i2b2 infrastructure (called hive). In the INHERITANCE implementation we exploit the Project Management cell, necessary to manage the users and their authorizations, the Data Repository cell, which provides the i2b2 Clinical Research Chart (CRC) access functionalities and the Ontology cell, which allows every term describing a patient’s fact in the CRC database to be used in a query. The Ontology Cell manages the vocabulary of concepts and medical terms, and contains information about the relationships between the elements for the entire structure. Only facts mapped into the ontology can be queried by the system to select sets of patients.

The i2b2 web client query interface allows ad-hoc queries by research clinicians. Such queries extract specific patient sets by returning their data or the aggregate numbers. The terms used in a query are those in the data repository associated with patient observations.

Within the INHERITANCE project we have also developed a new plugin to cope with the specific need of viewing and exporting the overall set of observations related to a patients’ set.
3.2 ETL

To import the data into the INHERITANCE data warehouse an ETL (Extract, Transform and Load) tool has been used. ETL tools are powerful instruments designed to:

- collect data from a variety of sources (more types of files and databases in the same process);
- move and transform the data (applying cleaning procedures and integrity checks);
- create aggregates or disaggregates;
- store data with a certain frequency, typically daily, in a large dimensional database, optimized for OLAP activities.

The ETL process in the INHERITANCE project has been implemented using Spoon-Kettle (kettle.pentaho.com), a graphic and powerful tool of Pentaho Data Integration (www.pentaho.com).

Figure 2 shows how Kettle has been exploited to import into the INHERITANCE warehouse the data coming from Cardio-register.

As shown in Figure 2, Kettle has been used to handle these steps:

I. Extraction of patient information;
II. Creation of new ontologies;
III. Loading of the observations into the data warehouse.

Every step of the process is performed by means of a Kettle transformation pipeline, as described in the following.

3.2.1 Extraction of the Patients’ Information

The information about the patients of interest is retrieved by executing an SQL script and transferred into a staging table where few checks and data format conversions are performed in order to obtain a dataset compliant with the Oracle database. During the extraction the patients' set is de-identified by removing any identifier and an i2b2 ID code is assigned to each patient. A table mapping the i2b2 IDs to the original patients' IDs is available, but the framework doesn't provide any module for accessing this table. Only through an ad hoc module, made accessible exclusively with given privileges, it will be possible to re-identify patients for specific purposes. At the end of this step, the patients' list is inserted into the Patient dimension table.

The i2b2 framework exploits different user roles by allowing only privileged users to access the whole patient set.

3.2.2 Creation of New Ontologies

Once the patients' list has been imported, it is necessary to create the ontologies related to the relevant clinical concepts within the project domain. In addition to Demographic information (age, age at diagnosis, age at genetic study, age at follow up, consanguineous parents and sex), we have considered the Physical Examinations (chest pain, diastolic and systolic blood pressure, dyspnea, height, syncope and weight) and the most important clinical extra cardiac concepts not directly related to the disease, but involved in its diagnosis (red flags).

In particular, we included: auditory problems, central nervous system problems, cutaneous problems, endocrine problems, mental retardation, neuromuscular problems, ocular problems, other non-cardiac traits, skeletal problems, syndromic traits and dysmorphic.

Each set of concepts (Demographic, Red Flags, Physical Examination) is associated to a table in the INHERITANCE metadata database, which models the overall ontology hierarchical structure, while each concept is represented as an instance of the Concept dimension table. Figure 3 shows, as an example, the red flags ontology structure. In this example the red flags ontology root contains multiple folders that represent the comorbidities that can occur with Dilated Cardiomyopathy. Each folder has 3 mutually exclusive leaves (yes/no/unknown) that can be used (as any other concept defined in the ontology) to define queries to filter out patients’ sets.

3.2.3 Loading of the Observations

In this final step, after a check on the birth date, the list of observations, one for each concept, is created for each patient and then transferred into the Observation fact table of INHERITANCE. Despite in this project we are studying inherited DCMs, the current version of the CRC doesn’t provide the possibility of explicitly managing the families members, so at the moment patients’ relatives are treated separately from the probands.
4. Reasoning and Decision Support Tools

A crucial aspect of the INHERITANCE knowledge management system is the definition of a tool that can guide clinicians in properly ranking the DCM causative genes, so that their screening can be effectively performed in the clinic, as well as in supporting the scientific discovery process. Our efforts have been directed towards i) providing a literature mining tool to extract biomedical information from the literature and to exploit this information for the discovery of new knowledge and ii) supplying a tool for analogical reasoning that takes advantage of the knowledge coming from previous solved DCM cases to suggest more effective diagnostic actions.

The main goal is to prioritize around 20 genes for screening on the basis of the patients' signs/symptoms. The large variability of the patients' data and the limited amount of formalized knowledge available requires the design of a decision support tool able to provide instruments for analogical reasoning to clinicians, including case similarity, information retrieval and text mining. Each clinical case is usually described by hundreds of features, including anamnesis and family information, lifestyle, lab tests and exams, ECGs, echo-cardiography data. Among the collected data, red flags are the most relevant ones as they could suggest specific genetic disease association.

To cope with this problem, we have implemented a combined reasoning process useful to exploit both the knowledge coming from the scientific literature and the previous experience on similar cases. In particular, when already diagnosed cases similar to the case under study are available, a CBR-based strategy is exploited to suggest the set of genes to test for possible mutations; otherwise, when the available case base (CB) is not sufficient, or to boost the suggestion supplied by the CB system, an automatic literature analysis tool can be used to process the scientific literature to drive towards new knowledge discovery and/or obtain a ranked list of candidate genes. In the following, the two approaches are described with the addition of the third one, which couples the literature mining and the case-based strategy.

4.1 Tools for Automated Literature Analysis

Automated literature analysis is becoming an essential need in current biomedical research. Text Mining (TM) and Natural Language Processing (NLP) provide algorithms and techniques for automated elaboration of textual content. This task is particularly important in the early stage of any study, in which resuming the available knowledge is crucial to formulate initial hypotheses and plan next tasks. The challenge is to broaden the search of potentially useful information to generate new hypotheses [8]. For instance, an added value could be to suggest that a candidate gene is often related to another gene, which has not been previously considered.

Within INHERITANCE we focused on genetic studies, in which a set of initial hypotheses of gene-disease association is made on some candidate genes, so that the first step is to explore the recent literature to confirm their possible role in the disease mechanism. We developed a tool able to extract the concepts of interest (genes and medical terms, like pathologies) using a structured knowledge base, the Unified Medical Language System (UMLS) [9], by which we can derive genes/disease annotation. Moreover, we also implemented similarity metrics, based on a relevance measure of the terms for each gene, to identify which terms each gene shares with the others. In this way we can derive a graph, in which the nodes connection reflects how tightly related those terms are, in accordance to the available literature.

The analysis method we propose aims to derive a literature-based gene/medical term association by extracting both UMLS terms and genes from the abstracts of the scientific literature. The overall analysis consists of three main steps:

- querying PubMed via Web Services to retrieve the most recent literature about specific genes/diseases
- automatically extracting concepts (genes/disease) from PubMed abstracts based on NLP techniques
- constructing annotation/co-citation networks to interpret available knowledge and suggest new hypotheses that can be tested.

The details of the literature analysis system and the medical concepts extraction are summarized below and described in [10]. This approach, that combines a keyword representation of the gene with a weighting procedure of the keywords, has proven to be effective in order to associate the genes in accordance with their expression profile [11].

The strategy is based on a search engine, which exploits the NCBI Web Service implementation of the Entrez Programming Utilities [12] to access to Entrez data via the Simple Object Access Protocol (SOAP). The module queries PubMed and retrieves a set of XML-formatted scientific papers dealing with the genes of interest, which can be automatically analyzed. In particular, through a language processing environment called GATE [13], we handled different text mining and analysis steps, aimed at identifying medical concepts and genes from the abstracts.

First, the Text Tokenization phase manages the identification of parts of text separated by blank spaces and the language-dependent exceptions to basic rules. Sentences are then separated using the Sentence Splitting component, and each previously identified token is assigned to a grammatical class. In a separate step, the lemma belonging to every token and the noun phrases within the text are identified. Finally, we developed two modules that were specifically designed for our purposes. The UMLS Concept Extractor module starts the analysis by generating a set of all possible substrings of each noun phrase in the document, considering the token itself and its lemma. A query is sent to the UMLS Metathesaurus database [14], which contains health-related concepts coming from many different sources. When a positive matching arises, the system identifies the official name of the found concept and assigns a new annotation to the document.

The second module, called Gene Extractor, is aimed at finding genes names cited in the text, this time relying on the Entrez Gene NCBI’s database [15]. Each token
found with the Text Tokenization step, which fits with a standard gene representation (e.g., with one capitalized letter and not composed only by digits), is sent as a first query to the database. In case that a match with an entry in NCBI Gene is found, a second query retrieves the official name of the previously identified gene.

In our specific case this pipeline has been used to retrieve the articles related to an initial set of candidate genes; the set of abstracts is then further analyzed by means of the Gene Extractor in order to build a larger set of genes which are cited together with the first candidate ones. Each gene in this augmented set is used to query PubMed for publications, that are again processed with the text mining pipeline, now ending with the UMLS Concept Extractor. Finally, the system is able to associate a set of UMLS terms (called “annotation profile”) to both the candidate genes and their co-cited genes. The annotation profiles of each gene are then used to build graphical and quantitative representations of the gene network as will be described in the Results section.

4.2 CBR-based System

Case Based Reasoning (CBR) techniques [16] provide tools to apply the analogical reasoning paradigm on clinical and biological data associated with real patients’ cases.

Each case is described by a set of features and a suitable distance measure between two cases is defined in order to find the cases most similar to the current case and exploit the knowledge about their diagnosis/treatment. The CBR system we have purposely developed for this project [17] is designed to deal with flexible case description and allows the comparison of two patients coming from different databases where different terms are used to describe their features set. Such flexibility is needed since: i) cases may come from two different databases, so that they are described by features representing the same concept which are codified according to different terminologies (e.g., ICD9-CM and SNOMED-CT); ii) the features describing the cases may be different as they don’t represent exactly the same concepts, but rather similar concepts; for example, because a concept may be more general (e.g., “Headache”) than another (e.g., “Episodic cluster headache”). Once this approach has been implemented the system is able to compare (to produce a distance measure between two cases) also when their data come from heterogeneous databases.

The main advantage coming from this approach is given by the capability to consider the features describing the patient at different levels of abstractions; to accomplish this task we have chosen to find our data representation on biomedical ontologies and terminologies, with particular reference to UMLS, and to develop a semantic distance algorithm able to exploit this data representation. In the current implementation we manage only binary features, while in the future we will also consider quantitative findings when computing the “case distance”. In our case, it was possible to consider data included in Cardioregister and data coming form other sources (such as excel files) with slightly different data definitions.

After the mapping phase, the data are represented as an array of concepts belonging to the UMLS Metathesaurus; each concept is represented by its Concept Unique Identifier (CUI) and by a Boolean modifier, which indicates if the finding referring to the concept is asserted or negated. One of the main advantages of our approach stands in the fact that the achieved distance scores exploit the same semantic environment used to map the data.

The system calculates the distance score $Dist()$ between $P_A$ (the array containing the data of the new patient) and $P_B$ (the generic patient of the CB) as follows:

$$Dist(P_A, P_B) = w^+ \cdot \text{semDist}(P^+, P^+) + w^- \cdot \text{semDist}(P^-, P^-),$$

where $P^+$ and $P^-$ are sub-arrays of $P_A$ and $P_B$ containing the CUIs of the concepts asserted (i.e., findings that are asserted in the case description), while $P^-_N$ and $P^+_S$ are the sub-arrays for the negated CUIs. $w^+$ and $w^-$ are normalization factors. This function returns 0 when two identical cases are compared. The function $\text{semDist()}$ is the semantic distance between two arrays of CUIs; it is calculated on the basis of one of the metrics presented in [18]:

$$\text{semDist}(P_A, P_B) = \sum \min (\text{clinDist}(P_{A,x}, P_{B,y})) / \text{size}(P_A)$$

where $P_A$ and $P_B$ are arrays of CUIs, $P_{A,x}$ is the i-th component of $P_A$ (i.e. a CUI) and $P_{B,y}$ is the j-th component of $P_B$, and the clinical distance $\text{clinDist}()$ between two CUIs is the size of the shortest path joining them in UMLS ontology considered [18].

Despite several types of relations exist inside the UMLS Metathesaurus (e.g., “may_be_a”, “has_associated_finding”), we have considered exclusively the “is_a” relation, as discussed in [19]. When any path doesn’t connect two concepts, the system assigns an arbitrary high clinical distance. Given a new case, its distance from every case of the CB is computed, so the user can visualize and explore the data about the most similar cases, in order to acquire potentially useful information for the diagnosis/treatment of the incoming patient.

4.3 Literature Based Prioritization

The tool for automated literature analysis described in Section 3.1 can be exploited also for candidate gene prioritization. In this case the process to obtain a ranking for a set of candidate genes follows these steps:

- all Pubmed abstracts are retrieved and included into an abstracts database;
- all Pubmed abstracts are analyzed in order to extract the concepts of interest (of genes and medical terms); this process takes advantage of the synonyms coming from UMLS metathesaurus and Gene database. The results of this analysis (annotation of genes and medical terms cited in the Pubmed abstracts) are stored in the literature database so that the association between each article and the extracted concepts is made available for the next step.

After these steps (which create the annotated literature database and, therefore, are performed once), the tool can complete its analysis in two different ways: the first one is driven by a specific real patient case, while the second one is aimed at updating the knowledge about the main disease (DCM) aspects.
The first analysis is patient-specific and, therefore, it is repeated for every investigation:

- for each finding asserted in the current case, Pubmed is queried in order to obtain the reference to the directly related articles. Since all the data associated with the patients have been mapped onto the UMLS metathesaurus, the search can be optionally extended also to the articles related to concepts that are semantically similar to the one present in the patient description. Following this analysis, the system generates a list of the associated genes along with the co-occurrence frequencies. These frequencies are used to obtain the ranked list of candidate genes for the specific patient.

The second type of analysis includes the following steps:

- by considering the MeSH thesaurus, 17 disease categories are taken into account. These MeSH terms are direct descendant of the “Diseases” main term and represent the macro-categories into which the diseases can be split (e.g. “Musculoskeletal Diseases”, “Nervous System Diseases”). The goal is to identify a set of articles where the “Dilated Cardiomyopathy” main disease is co-cited with diseases belonging to these macro categories (that represent the comorbidities).

The Pubmed queries resulting from this approach are therefore composed by “Dilated Cardiomyopathy” (searched within the whole abstract) and one of the 17 disease categories (searched, with all their descendant in the MeSH tree, among the MeSH terms describing the article).

- once this set of articles is identified, it is possible to exploit the already performed literature analysis in order to weight the importance of each candidate gene in the context. In this case the context is represented by the main disease associated to a particular family of other diseases that represent the comorbidities of the real patients.

The result of the first analysis (i.e. the patient-specific one) is an augmented ranked list of genes related to the patient case, where not all the suggested genes might belong to the list of the ones previously known as associated to the specific disease (DCM in our case). This could, therefore, lead to the discovery of new gene/disease associations.

The second type of analysis, instead, provides a general overview of the relationship between the main disease and the macro-categories of possible comorbidities; the result of this process leads to 17 prioritization lists (one for each comorbidity subgroup) of the candidate genes, where each gene is scored with the percentage of articles citing it on the total number of articles belonging to the set.

5. Results

The overall INHERITANCE software infrastructure has been implemented following a series of steps necessary for tailoring the application to the specific medical domain. The basic phase has been conducted in strict collaboration with the medical experts bringing to the identification of the key medical concepts to be included in the system ontology. This step is preliminary to the development of the ETL procedures to import the data from Cardioregister into the data warehouse. The following step has been devoted to the design and implementation/customization of a set of software decision support tools. In particular, the CBR system has been completely designed and developed for the project even though it is general and could be easily adapted to other medical applications. The literature analysis tool has manifold applications, depending on the concepts extracted from the articles and the citation frequency computation. In the previous section, we provided two examples of application of the tool, one for gene list prioritization and another for “virtual” patients extraction, but it can be similarly applied to knowledge discovery and hypotheses verification.

5.1 The INHERITANCE Data Set

At the moment, a set of 1,394 patients has been imported from Cardioregister into the INHERITANCE data warehouse and 9,784 observations have been inserted into the Observation fact table. The Concept dimension table contains 467 rows, each row corresponding to one of the different values assumed by each concept. Figure 3 shows the three possible values of the Skelatal Problem finding (“yes”, “no” and “unknown”) that represent leaves in the INHERITANCE ontology. Currently, the abstract database contains 1,105,550 abstracts retrieved from Pubmed and generically related to cardiac diseases. The search within the abstract database of the most relevant UMLS semantic types (e.g. Clinical Drugs, Diseases, Genes, etc.) related to the project clinical domain produced about 12 million annotations related to 51,977 distinct UMLS concepts. Unfortunately, the complete INHERITANCE patients’ set, coming from the different clinical centers involved in the project, has not been still imported, so the full validation still needs to be performed.

5.2 Literature Based Knowledge Discovery

Starting from 20 candidate genes known to be involved in DCM, we applied this strategy with the final aim of identifying a set of DCM-related genes to be further investigated.

The 50 most recent abstracts, obtained by querying PubMed for the candidate genes were used to extract other 2414 co-cited genes and their most frequently associated terms. This procedure allowed to represent each gene by an annotation profile, composed of UMLS terms indicating diseases or symptoms and the counts of their occurrences in PubMed entries. The annotation profiles were composed, on average, of a set of 60 terms and their corresponding frequencies. Thanks to a text mining weighting scheme, known as $\text{TF-IDF}$, the counts were transformed into values that reflect the importance of each term for each gene. Within these methods, a term is rated as important for a gene if it occurs frequently in its articles and if it is specifically assigned to the publications related to that gene.

In order to have a graphical visualization of the groups of similarly annotated genes, we created an association network,
where the genes are linked on the basis of the cosine similarity of their annotation profile.

▶ Figure 4 shows a portion of the network obtained by connecting the genes if the cosine metric exceeded the 99th percentile of the distribution of pair-wise similarities. Dark nodes in the network represent genes that may be taken into consideration from researchers interested in DCM. In particular, we have selected the 5 most important terms for each gene, according to TF-IDF, and used them to highlight additional genes tightly associated with DCM in the literature (triangular-shaped) and other genes related to the more general UMLS term “Cardiomyopathies” (square-shaped).

To quantitatively characterize the importance of the nodes in our network, we have also computed the betweenness centrality [20]. This topological measure is defined as the number of shortest paths that go through a considered node, and represents the influence of that node in the flow of information in the network. Nodes with a high betweenness typically make possible the communications in the network among clusters of nodes characterized by high internal connectivity. We observed these properties also in the DCM network, where the size of the nodes was adjusted proportionally to the value of the node’s betweenness. For instance, by analyzing the important associated terms, we noticed that the known DCM-related gene LDB3, characterized by a high betweenness centrality, links a group of nodes that are strongly associated with Cardiomyopathies (in the upper right side of the image) to a cluster (in the left side), which is more frequently related to Myopathy and Muscular dystrophy.

5.3 CBR System

The preliminary testing process of the CBR-based system has been performed on a simulated set of cases. We chose to adopt this strategy in order to be able to measure the CBR system performances also in the early stages of the INHERITANCE project, when the data warehouse was still in development and the data weren’t already stored into the project database. However, the CBR system has been already arranged to extract directly from the i2b2 CRC the most similar cases.

Working side by side with the physicians involved in the project, we created a simulated benchmark of DCM patients in order to test the capability of the CBR system to associate a single case with the most similar cases in terms of actual diagnosis.

The simulated patient set has been designed considering the main known relationships between the different types of DCMs and their typical phenotype markers [21]. In particular we derived from literature, for each class of interest, the event rates of all the features suitable to describe the patient; in this step we have also exploited the experiences of the physicians involved in the project. Afterward, we randomly generated each simulated case considering the probabilities of each feature to be true according to the class it belongs to.
We defined four classes of patients:

- **not affected** by cardiomyopathy (no_CM)
- **affected** by a not-specified cardiomyopathy (CM)
- **affected** by DCM with Dystrophin (DCM_dys) mutation
- **affected** by DCM with Lamin (DCM_lmna) mutation

Each patient is described by sixteen phenotype features, related to diagnoses and symptoms (e.g. "Chest Pain", "Dyspnea", "Muscular Dystrophy"); each phenotype feature has the asserted/negated modifier associated. For each class we generated 25 cases.

The UMLS version used for testing the system is UMLS-2010AB, while the vocabularies selected for the internal data representation are: i) Medical Subject Headings (MeSH), ii) UMLS Metathesaurus (MTH) and iii) SNOMED Clinical Terms (SNOMED CT). For the purposes of this test, it was enough to structure the patient description into a single section containing all the features.

The test has been performed as follows: 3 patients per class (12 patients in the aggregate) were randomly extracted from the original set of 100 patients; in turn, these patients were considered as the new case and the CBR system evaluated its distance score from the remaining 99 cases. The results are shown in Table 1.

From the data shown in this table, it is clear that the semantic distance algorithm works properly, in fact 11 of the considered 12 cases are correctly classified and, in general, all the 4 classes achieve a better distance score with themselves than with other classes.

### 5.4 Literature Based Gene Prioritization

In this section we will show the results of both literature-based gene prioritization strategies: the patient-driven one and, on the other side, the one aimed at investigating the genetic implications between the main disease (DCM) and its possible comorbidities.

Both prioritization strategies exploit the literature database, described in Section 3.3, that actually contains more than 1,100,000 articles, already analyzed for extracting UMLS concepts and genes; in particular, for the scope of this analysis, we considered the UMLS concepts belonging to the semantic class "Genes & Molecular Sequences" and the Gene annotations.

We will show two patient-specific gene prioritization examples, in order to point out both potentialities and current limits of this kind of analysis. The first phenotype, called *Case 1*, presents, besides DCM, "Creatine Phosphokinase Serum Increased" (sCPK Incr.), while the second one, *Case 2*, has DCM and "Left Ventricular Noncompaction" (LVNC). The cases' description is reported in Table 2.

The data of these cases, already mapped onto the UMLS Metathesaurus (as described in section 3.3), were used to generate a query to Pubmed; the query gener-

---

**Table 1**

Distance scores achieved in the test on the simulated patients' set. For each class of the compared cases, the scores are computed as the mean values of the similarities between the case under examination and all the compared cases belonging to the same class. Scores represented vertically are the mean values for the class they belong to. Scores in bold are the best for each new case or class.

<table>
<thead>
<tr>
<th>cases under examination</th>
<th>no CM</th>
<th>CM</th>
<th>DCM dys</th>
<th>DCM lmna</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#1</td>
<td>#2</td>
<td>#3</td>
<td>#3</td>
</tr>
<tr>
<td>no CM</td>
<td>2.28</td>
<td>3.26</td>
<td>7.95</td>
<td>8.04</td>
</tr>
<tr>
<td>CM</td>
<td>9.95</td>
<td>9.07</td>
<td>3.51</td>
<td>4.5</td>
</tr>
<tr>
<td>DCM dys</td>
<td>11.63</td>
<td>10.97</td>
<td>6.3</td>
<td>7.58</td>
</tr>
<tr>
<td>DCM lmna</td>
<td>13.26</td>
<td>14.03</td>
<td>5.21</td>
<td>5.43</td>
</tr>
</tbody>
</table>

---

**Table 2**

Description of the two phenotypes used for testing the literature based gene prioritization. The percentages associated to the mutations are related to our real patient set and represent the frequency of each mutation among the real cases that are defined by the phenotypes.

<table>
<thead>
<tr>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutation</td>
<td>Dystrophin (DMD Gene) 50% Lamin (LMNA Gene) 29% Other genes 21%</td>
</tr>
<tr>
<td>Features (UMLs codes)</td>
<td>Cardiomyopathy, Dilated (C0007193) Creatine Phosphokinase Serum Increased (C0241005)</td>
</tr>
</tbody>
</table>
The resulting article sets were made up, respectively, by 30 articles for Case 1 and 40 articles for Case 2; for each article we considered the genes cited in its abstract and then we built a weighted list of genes for the two cases in exam. The obtained gene lists are represented in Table 3.

For Case 1, the results show that the sCPK Incr. feature performs well as a predictor of the possible mutations observed in our patient set; in fact the frequency of cases with LMNA or DMD gene mutation (that are the most cited gene in the relative literature), among the patients with DCM and sCPK Increased, is 79%. This good performance is due, besides the very nature of the considered feature as a mutation predictor, also to current status of the scientific literature, where the association between DCM, sCPK Incr. and these mutations is strong.

The results obtained with Case 2, instead, show the actual limitations of this approach. Therefore, this approach can be used as a starting point for a more advanced analysis.

### Table 3
List of candidate genes for the two phenotypes obtained by the literature analysis. The actual mutated genes are reported in bold.

<table>
<thead>
<tr>
<th>Gene</th>
<th># Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMD</td>
<td>9</td>
</tr>
<tr>
<td>LMNA</td>
<td>3</td>
</tr>
<tr>
<td>DES</td>
<td>1</td>
</tr>
<tr>
<td>FKRP</td>
<td>1</td>
</tr>
<tr>
<td>LAMA2</td>
<td>1</td>
</tr>
<tr>
<td>TAZ</td>
<td>2</td>
</tr>
<tr>
<td>MYH7</td>
<td>2</td>
</tr>
<tr>
<td>ACTC1</td>
<td>1</td>
</tr>
<tr>
<td>LMNA</td>
<td>1</td>
</tr>
<tr>
<td>TNNT2</td>
<td>1</td>
</tr>
</tbody>
</table>

### Table 4
List of 17 disease categories used to prioritize genes to be tested for a mutation.

<table>
<thead>
<tr>
<th>Categories of Comorbidities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathological Conditions, Signs and Symptoms</td>
</tr>
<tr>
<td>Congenital, Hereditary, and Neonatal Diseases and Abnormalities</td>
</tr>
<tr>
<td>Nervous System Diseases</td>
</tr>
<tr>
<td>Musculoskeletal Diseases</td>
</tr>
<tr>
<td>Animal Diseases</td>
</tr>
<tr>
<td>Nutritional and Metabolic Diseases</td>
</tr>
<tr>
<td>Skin and Connective Tissue Diseases</td>
</tr>
<tr>
<td>Hemic and Lymphatic Diseases</td>
</tr>
<tr>
<td>Virus Diseases</td>
</tr>
<tr>
<td>Immune System Diseases</td>
</tr>
<tr>
<td>Endocrine System Diseases</td>
</tr>
<tr>
<td>Female Urogenital Diseases and Pregnancy Complications</td>
</tr>
<tr>
<td>Respiratory Tract Diseases</td>
</tr>
<tr>
<td>Neoplasms</td>
</tr>
<tr>
<td>Nervous System Diseases</td>
</tr>
<tr>
<td>Nutritional and Metabolic Diseases</td>
</tr>
<tr>
<td>Pathological Conditions, Signs and Symptoms</td>
</tr>
<tr>
<td>Respiratory Tract Diseases</td>
</tr>
<tr>
<td>Skin and Connective Tissue Diseases</td>
</tr>
<tr>
<td>Stomatognathic Diseases</td>
</tr>
<tr>
<td>Substance-Related Disorders</td>
</tr>
<tr>
<td>Virus Diseases</td>
</tr>
</tbody>
</table>

Figure 5   Graphical representation of the gene/comorbidity association. This can be used to derive the gene prioritization lists on the basis of the co-citation frequency of genes and comorbidity categories.
approach. In fact only 10% of the real patients’ mutations (TAZ and LMNA gene) were identified; the reasons of this performance are mainly due to the inadequacy of the feature to be a good mutation predictor and also to the existing scientific literature, where most of the mutated genes of our patients’ set (91%) have not yet associated with DCM and LVNC.

In order to test the second approach, we choose, among the 26 available, the 17 disease categories which could better cover the spectrum of the different types of comorbidities that can occur with DCM. These categories (Table 4), with DCM, were used to compose the queries to be sent to Pubmed. The 17 prioritization lists so far obtained can be graphically represented as bar charts showing the gene/comorbidity association frequencies (Figure 5).

The results show that, for different phenotypes (DCM + comorbidities), it is possible to identify different prioritization lists among the candidate genes selected by the physicians. This clearly shows that, as soon as comorbidities are observed on the single patient, it is possible to update/rank the gene list.

6. Conclusions

The main task of the INHERITANCE project is to investigate the molecular basis of inherited DCM. To this end we have developed and implemented a set of software tools to support data management and decision support, including:

I. A data warehouse for fast phenotype data exploration based on the i2b2 system.

II. A tool for automatic literature analysis and literature-based discovery

III. A system for supporting reasoning and decision on a single case, with the aim to prioritize gene screening and to discover new gene-disease associations.

The project, after two years, has already collected more than one thousand patients coming from four medical centers. The data warehouse has been completely set up and the text mining tools for automatic literature analysis have been implemented and tested. The decision support tool for knowledge discovery and gene prioritization is currently still under development. In particular, the CBR system will be expanded with a distance metric taking into account also continuous features. Moreover, the literature annotation process is continuously run in order to increase the base of analyzed articles and making so more efficient the overall reasoning process. When the literature analysis will be completed, we plan to run periodically an update process, which will add the mining results of the most recent articles and keep, therefore, the literature base up to date.

Besides the IT tools described in this paper, we are working also at the implementation of a collaborative wiki environment based on the semantic web technologies, that will represent a unique tool of dissemination of help for clinician and researchers involved in DCM programs.

Acknowledgments

This work was supported by the INHERITANCE project (grant 241924) funded by the European Commission (FP7). We thank Lorenzo Monserrat, HealthEncode and the Cardioregister team for their effective collaboration.

References