Integration of Genetic and Medical Information Through a Web Crawler System

Gaspar Dias[1], José Luís Oliveira[1],
Francisco-Javier Vicente[2], and Fernando Martín-Sánchez[2]

[1] Universidade de Aveiro, IEETA/DET, 3810 193 Aveiro, Portugal
[2] Instituto de Salud Carlos III (ISCIII), Madrid, Spain

Abstract. The huge amount of information coming from genomics and proteomics research is expected to give rise to a new clinical practice, where diagnosis and treatments will be supported by information at the molecular level. However, navigating through bioinformatics databases can be a too complex and unproductive task.
In this paper we present an information retrieval engine that is being used to gather and join information about rare diseases, from the phenotype to the genotype, in a public web portal – diseasecard.org.

1 Introduction

The decoding of the human genome and of other human beings has been promoting a better understanding of the evolution process of the species and the relation between diseases and genes. The integration of these massive amounts of genetic information in the clinical environment is expected to give rise to a new clinical practice, where diagnosis and treatments will be supported by information at molecular level, i.e., molecular medicine [1-3].

The generalization of molecular medicine requires an increased exchange of knowledge between clinical and biological domains [4]. Several databases already exist, covering part of this phenotype-to-genotype connection. A major hindrance to the seamless use of these sources, besides the quantity, is the use of ad-hoc structures, providing different access modes and using different terminologies for the same entities. The specificity of these resources and the knowledge that is required to navigate across them leave its plain usage just to a small group of skilled researchers.

This problem is even more important if we deal with rare disease where the relation between phenotype and genotype is typically strong (around 80% have genetic origins). Rare diseases are those affecting a limited number of people out of the whole population, defined as less than one in 2,000. Despite this insignificant rate, it is estimated that between 5,000 and 8,000 distinct rare diseases exist today, affecting between 6% and 8% of the population in total. In the European Community this represents between 24 and 36 million of citizens.

Many rare diseases can be difficult to diagnose because symptoms may be absent or masked in their early stages. Moreover, misdiagnosis, caused by an inadequate knowledge of these diseases, is also common. In this scenario, it is a major goal to maximize the availability of curate information for physicians, geneticists, patients, and families.

Consider a patient with Fabry disease, a rare disease with a prevalence of less than 5 per 10,000. So far, the main sources for finding biomedical information about this disease are basically two: bibliography and the Internet.

J.L. Oliveira et al. (Eds.): ISBMDA 2005, LNBI 3745, pp. 78–88, 2005.

The Web is a valuable source of information provided one knows where and how to look for the information. The easiest and most direct manner of making a search is to use the traditional search engines. To this date, a search engine such as Google indexes more then 8 billion web pages turning the search for a rare disease into an adventure. Just to cite an example, Fabry disease produces about 105,000 entries. To be able to deal with such a number of resources would be impossible if they are not previously filtered.

In this paper we present the DiseaseCard portal, a public information system that integrates information from distributed and heterogeneous public medical and genomic databases. In [5] we have present a first version of DiseaseCard, a collaborative system for collecting empirical knowledge disseminated along research centers, organization and professional experts. In this paper we present an automatic information retrieval engine that is now the computational support for DiseaseCard. Using a predefined navigation protocol, the engine gathers the information in real-time and present it to the user through a familiar graphic paradigm.

2 Information Resources Selection

The selection of the sources for biomedical information is crucial for DiseaseCard. *Nucleic Acids Research* (NAR) publishes annually "*The Molecular Biology Database Collection*", a list of the most important databases hierarchically classified by areas in this field of interest. In the 2005 update [6], there were 719 databases registered in the journal, 171 more than the previous year, a number, although appreciably lower than the one obtained in Google, still unmanageable by a primary care physician. Each database has its own domain, its own architecture, its own interface, and its own way of making queries, i.e., resources are heterogeneous and distributed.

There are resources specialized in rare diseases accessible via the Internet and that can be queried for all the information available about a pathology. However, they are oriented towards clinical aspects of the disease, relegating genetic aspects to the background. For instance, IIER (iier.isciii.es/er/) or ORPHANET (www.orphanet.net) are websites that in Spain and France, respectively, are points of reference.

2.1 Public Databases on Biomedicine

Back to the practical case from which we started, the goal of DiseaseCard is the integration of heterogeneous and distributed genetic and clinical databases, under a common appearance and navigation method. Once Diseasecard does not generate information by itself it is crucial to look upon reliable, curate and frequently updated data sources. To achieve this, several databases registered in NAR have been selected. Guaranteed scientific reliability, exact and frequently updated data and public and free access are common characteristics shared by these databases.

Based on this list we can map several pathways along the web which we use to build a navigation protocol and obtain Diseasecard information.

1. **Orphanet** is a database dedicated to information on rare diseases and orphan drugs. It offers services adapted to the needs of patients and their families, health professionals and researchers, support groups and industrials.

2. **IIER** is the Spanish database of rare diseases, a 'meeting place' for professionals, patients, family, organizations, industry, media and society in general.
3. **ClinicalTrials** provides information about federally and privately supported clinical research in human volunteers. It gives information about a trial's purpose, who may participate, locations, and phone numbers for more details.
4. **OMIM** is the database of human genes and genetic disorders, compiled to support research and education on human genomics and the practice of clinical genetics. The OMIM Morbid Map, a catalog of genetic diseases and their cytogenetic map locations, is now available.
5. **GenBank** is a database that contains publicly available DNA sequences for more than 170000 organisms.
6. The NCBI **dbSNP** is database of genome variation that complements GenBank by providing the resources to build catalogs of common genome variations in humans and other organisms.
7. The **EMBL** Nucleotide Sequence Database is maintained at the European Bioinformatics Institute (EBI) in an international collaboration with the DNA Data Bank of Japan (DDBJ) and GenBank at the NCBI (USA). Data is exchanged amongst the collaborating databases on a daily basis.
8. **Entrez-Gene** provides a single query interface to curated sequence and descriptive information about genetic loci. It presents information on official nomenclature, aliases, sequence accessions, phenotypes, EC numbers, MIM numbers, UniGene clusters, homology, map locations, and related web sites.
9. **Swiss-Prot** is a protein sequence database. It provides a high level of annotation such as description of protein function, domains structure etc.
10. The **ProDom** database contains protein domain families automatically generated from the SWISS-PROT and TrEMBL databases by sequence comparison.
11. **KEGG** (Kyoto Encyclopedia of Genes and Genomes) is the primary database resource of the Japanese GenomeNet service for understanding higher order functional meanings and utilities of the cell or the organism from its genome information.
12. **GeneCards**, is an automated, integrated database of human genes, genomic maps, proteins, and diseases, with software that retrieves, consolidates, searches, and displays human genome information.
13. The **PubMed** database includes over 14 million citations for biomedical articles back to the 1950's. These citations are from MEDLINE and additional life science journals. PubMed includes links to many sites providing full text articles and other related resources.
14. **PharmGKB** is an integrated resource about how variation in human genes leads to variation in our response to drugs. Genomic data, molecular and cellular phenotype data, and clinical phenotype data are accepted from the scientific community at large. These data are then organized and the relationships between genes and drugs are then categorized into different categories.
15. **HGNC** (HUGO), the Human Gene Nomenclature Database is the only resource that provides data for all human genes which have approved symbols. It contains over 16,000 records, 80% of which are represented on the Web by searchable text files.

16. The goal of the Gene Ontology Consortium (**GO**) is to produce a dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing.
17. **EDDNAL** is the European Directory of DNA diagnostic Laboratories. It aims to disseminate information among medical genetics health-care professionals concerning the availability of DNA-based diagnostic services for rare genetic conditions in Europe. EDDNAL also seeks to promote the highest standards of genetic testing as well as to facilitate research into the development of new diagnostic tests.

3 Navigation Workflow

With all the databases filtered, we have developed a conceptual protocol that follows logical rules combining the two perspectives: genetic and clinical [5]. This protocol allows users to follow a specific pathway from the disease to the gene (Figure 1).

For a given disease, the protocol start search for its symptoms, for general information and related centers of reference (Pathology). This is the entry point of the protocol. A disease can be related to a disorder or a mutation in the genetic pattern of the patient (Disorder & Mutation), like a polymorphism (Polymorphism) for instance. This is due to a change in the sequence of nucleotides (Nucleotide sequence) causing a change in the corresponding protein (Protein: sequence) with a domain and belonging to a family (Domain & Families). As a result the 3D structure of the protein is altered (Protein: 3D structure) and therefore its biological function (Protein: function). The protein takes part in a metabolic pathway (metabolic pathway), carrying out enzymatic reactions (Enzyme). It is at this level where the drugs act, if available (Drug). Clinical trials are carried out with the drugs in Europe and the USA and also pharmacogenetics and pharmacogenomics research (Pharmacogenomics & Pharmacogenetics research). There is bibliography about the disease and there are centers where the genetic diagnosis is made (Genetic Test/Lab). There is also genetic information relevant for R&D such as the exact location in the chromosome, official name of the affected gene, genetic data integrated in an individual card for each gene and information relative to the "Molecular function, Biologic process and Cellular component" in which the gene(s) is (are) involved, included in Gene Ontology.

From the clinical perspective, we have divided the protocol into user profiles fitting the different types of users, according to their information needs. In the top part of the protocol, area 1 (in white), there are the resources that provide information useful to primary care physicians such as information about bibliography, centers for genetic diagnosis and available treatments. The patient asks the primary care physician for the information about the disease and the doctor gets the information by querying DiseaseCard from his own office PC connected to the Internet.

Generally, the hospital specialist, to whom the patient was referred by the primary care physician, does the follow up in this type of pathologies. The specialist looks for more specific and detailed information than the primary care doctor, the areas 1 and 4. Next to the specialist is the hospital geneticist, which uses the information available in the areas 2 and 3 of the protocol.

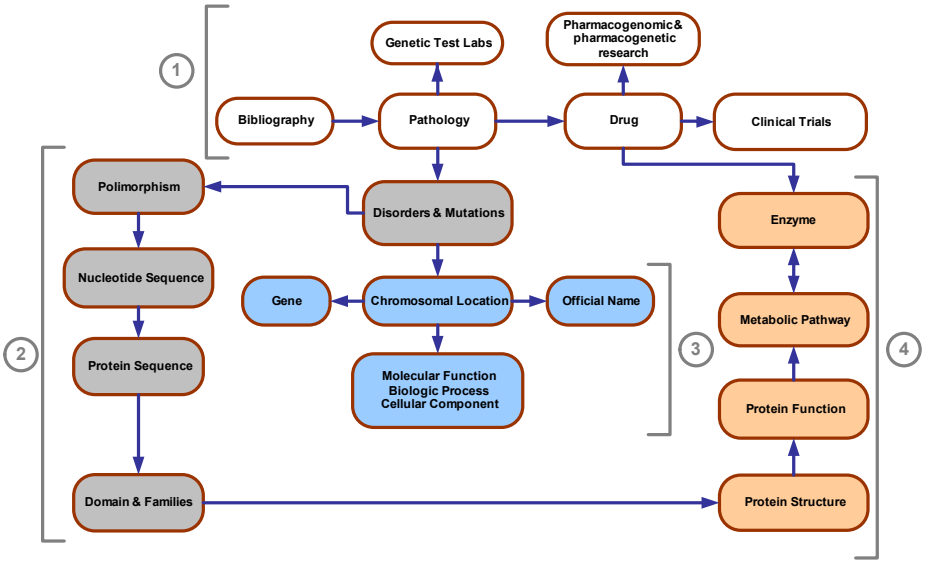

Fig. 1. Map of the concepts/databases used in Diseasecard. This illustration shows a navigation path, showing different areas associated with different clinical users: Area 1 – Primary care physician; Areas 2 and 3 – Geneticist; Areas 1 and 4 – Specialist, Pharmacologist

Due to their highly specialized training these professionals require exact genetic information, a need addressed by DiseaseCard by offering the possibility of designing ad hoc individualized cards for each pathology and user.

4 An Information Retrieval Engine

In a human-based navigation scheme each user would have to follow a predefined procedure in order to build one disease card manually. He has to navigate, in a given pathway, through several web pages and get web links (URLs) to fill diseasecard's concepts. Figure 2 describes the resource pathways that have to be followed in order to collect all diseasecard information. The "core" databases in this protocol are Orphanet, OMIM, Entrez-Gene, HGNC and Expasy from which most of the navigation nodes can be reached.

Since the cards are built based on a predefined template [5], the data sources to be explored are always the same so the respective queries/URLs are equal except their query ids. Exploring this commonality we can map the building task into a single protocol and construct an automatic retrieval engine.

With this aim, we develop a module (Cardmaker) integrated in Diseasecard System which automatically builds cards based on this single protocol. The protocol is described previously in XML using a specific schema (XML Protocol Descriptor, or XPD). With the protocol descriptor approach, instead of having a hard coded search engine we achieve a dynamic and more comprehensive system to assemble each card information.

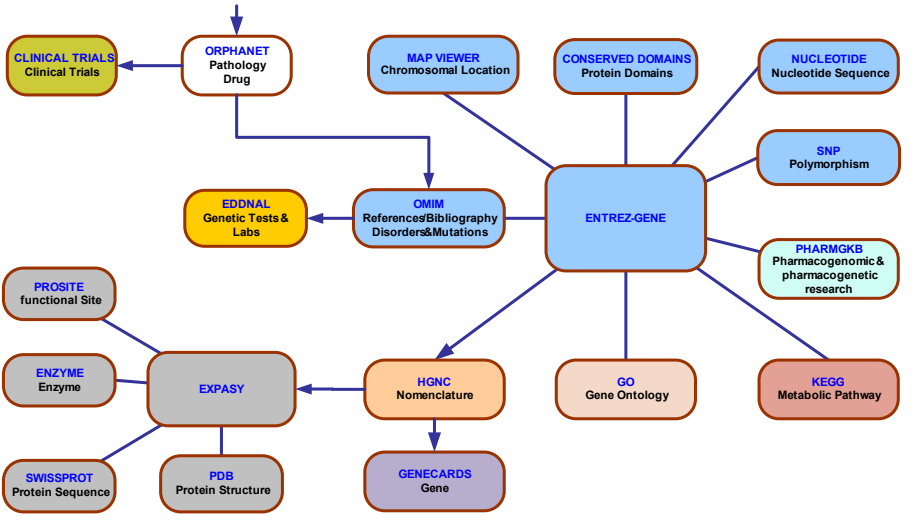

Fig. 2. Data sources network. Each box represents a data source containing respective retrievable concepts. Some of these concepts lead to other data sources and other concepts

Through this protocol, the expert user chooses the sources to be consulting during the querying process. This protocol is then used in general queries and interpreted by a parser which converts the XML syntax into search/extract actions (dynamic wrappers) used to explore and retrieve information from web resources. Once this file is defined, it is "plugged" into the system and it is then ready to generate cards.

Our goal is to focus on the development of a more flexible integration system that helps diseasecard's users to gather information from heterogeneous databases in an automated and transparent way.

A main characteristic of the diseasecard portal is that it only manages URL links to the relevant information. If a particular resource is mapped in a XPD file, to extract these links the system has to execute the following steps.

1. Download the web page specified in the protocol entry;
2. Search for data associated to diseasecard concepts;
3. Store extracted URL in a concept item;
4. Step into the next protocol entry;

This procedure is repeated until the end of the protocol.

Figure 3 illustrates an example of part of a protocol. Each block is a protocol entry that represents a data source with respective retrievable card concepts. Some of these concepts are also key parameters to navigate into other data sources. As an example, in the *Orphanet* data source the system can retrieve *pathology*, *Drug* and *Omim* concepts. The first two items are stored in a diseasecard instance and the third is used to step into OMIM site. Now, in OMIM, the system searches for *Reference*, *Disorders and mutations* and *Entrez Gene*'s links and repeats the process until the end of protocol.

The protocol is stored in an "xpd" file that is divided in *wrapper* elements, each associated to a single database. Inside each element it is specified the name of the

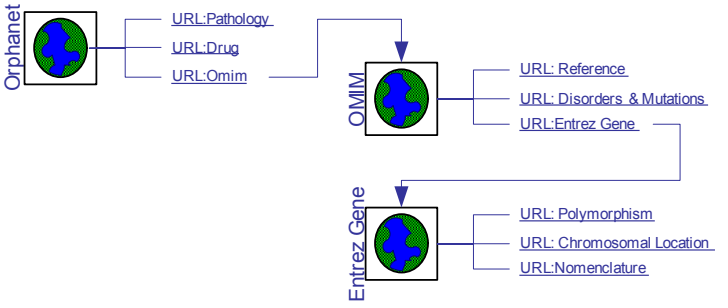

Fig. 3. A protocol preview with three data sources and respective retrievable card concepts. Each URL link is stored in a concept item and in some cases it is the entry point to access other sites

resource, respective URL and filtering terms to search and extract the relevant information. These filtering or matching terms are based on *Regular Expressions*[1] [7]. Each *wrapper* element will be interpreted by the *parser* that generates a run-time retrieval component. The following text is a piece of the protocol showed in Figure 3.

```
<wrapper>
   <resource-name>Orphanet</res   ource-name>
   <resource-url key-origin="_ext"><![CDATA[http://orphanet.url?KEY]]>
     </resource-url>
   <search-for>
      <regex for pathology><search-for>
      <regex><![CDATA[<regex for omim code>]]></regex>
      <put-into>_omim</put-into>
   </search-for>
</wrapper>
```

Once this process is mapped in a XML protocol descriptor, we need an engine to interpret the protocol language and to execute and control the automated card construction. To do this work we have developed and integrated into the Diseasecard System the XPD Engine, which is illustrated in Figure 4. Note that this engine works inside the *Cardmaker* tool.

This engine generates diseasecard instances based on the user request and on an XPD file. A *Parser* converts the XML language into a set of tasks called the *Task Plan*. These tasks will be used afterward to create web wrappers dynamically when requested by the *Controller*. Based on the *Task Plan*, the *Controller* manages these wrappers which generate card concepts that are finally assembled into a new card and stored in the Diseasecard's Database.

[1] A regular expression (abbreviated as regexp, regex or regxp) is a string that describes or matches a set of strings, according to certain syntax rules. Regular expressions are used by many text editors and utilities to search and manipulate bodies of text based on certain patterns

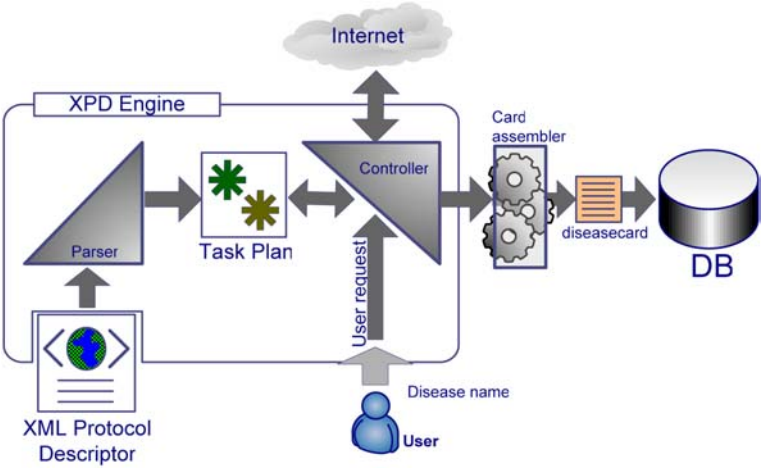

Fig. 4. The XPD Engine architecture

While this XPD engine is enough generic to be applied in any field of knowledge where information retrieval from public database must be performed, we show here its functionality inside the framework of diseasecard portal.

5 DiseaseCard Portal

DiseaseCard is a web portal publicly available that provides an integrated view over medical and genetic information with respect to genetic diseases (www.diseasecard.org). The end-user will typically start the investigation by searching for a disease or providing its name. Upon the identification of the disease, the portal will present a structured report, containing the details of the pathology and providing entry points to further details/resources either on the clinical and genetic domains.

Several aspects of older version of Diseasecard have been changed. We have decided to integrate the *Cardmaker/XPD Engine* tool into the querying process and this has implied a complete different usage paradigm – from a collaborative user-based annotation system (previous version) into an automatic information retrieval application (actual version). Although instead of having two different user operations, card querying and card creation, the system merges both tasks into a single one, providing a more useful and intuitive interface. *Cardmaker* works in the background if database doesn't have any card associated to the user request. Otherwise the system returns a card or a set of cards that matches with the user request (Figure 5).

Table 1 shows the main steps of Diseasecard evolution since it was created. The first goal that was behind the conception of Diseasecard was the development of a web tool where a group of experts could collaborate on-line in order to share, store and spread their knowledge about rare diseases (DC1.0). Since we have concluded that the majority of the diseasecard protocol could be automated we have developed an alternative way of creating cards through the *Cardmaker* tool. This utility was available only for authenticated users (DC2.0). After some improvements on *Cardmaker* we realized that this tool has a good time performance and it can create reliable

diseasecard contents. So we decided to delegate this task totally to the system. With this approach, card creation is transparent to any user because the tool works in run-time when he asks for some card that isn't defined in database (DC3.0).

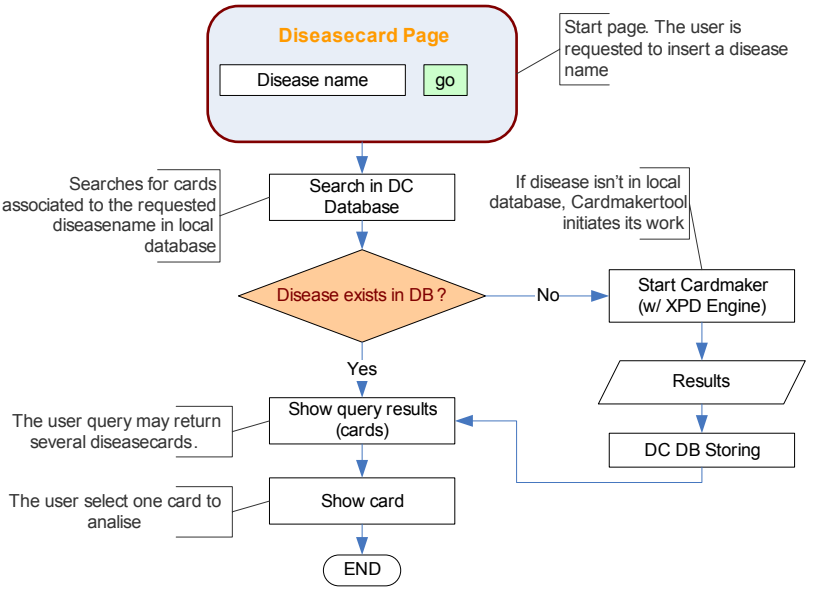

Fig. 5. Behavior of the Diseasecard querying operation. Based on the disease name requested by user, the Diseasecard checks for cards related to requested disease name in the database. If exists, the system shows the card contents. If not, the system launches the Cardmaker/XPD Engine. If this operation is succeeded, then the results are showed in the card view

Table 1. Diseasecard's versions

Diseasecard Version	Main Features
Diseasecard 1.0	Collaborative web tool
Diseasecard 2.0	Collaborative web tool + Automated card creation
Diseasecard 3.0	Automated card creation

The next illustrations show two perspectives of the Diseasecard System. Our concern in the main page (Figure 6) is to provide a simple and intuitive interface in order to facilitate the user's search. The main page provides two different query modes which are "*search for disease*" through a single word or expression related to any disease or search from the "*list of available diseases*" where the user selects a disease from an alphabetic list.

Figure 7 appears after selecting a disease from the returned list originated by a user query. It shows on the left side the diseasecard structure which contains all diseasecard's concepts and respective links to associated web resources. This example shows a rare disease called *Fanconi Anemia*. The selected concept in this case is *pathology* and it links to an *Orphanet* page which contains a description of this disease showed on the right side of the page.

Fig. 6. Diseasecard.org – The main page

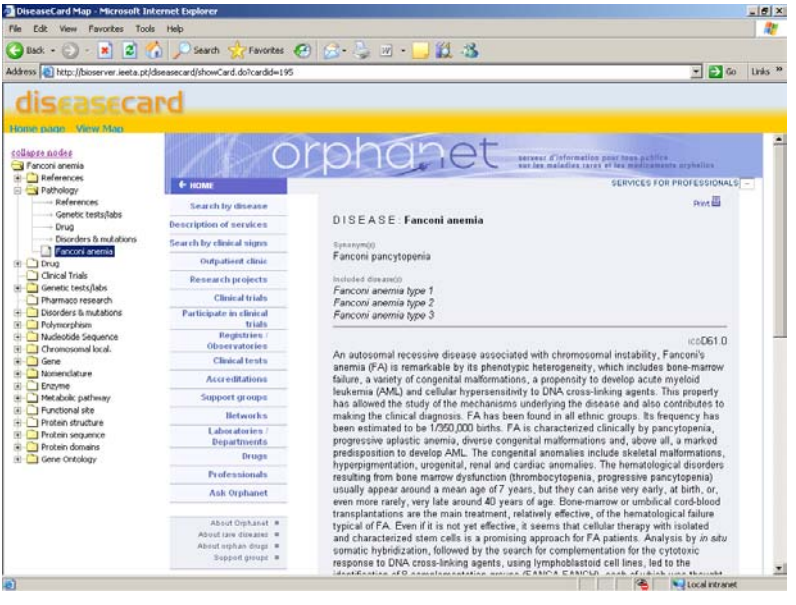

Fig. 7. Card detail. After selecting a disease, the details are displayed in this page

6 Conclusions

The recent advances on genomics and proteomics research bring up a significant grow on the information that is publicly available. However, navigating through genetic and

bioinformatics databases can be a too complex and unproductive task for a primary care physician. Moreover, considering the rare genetic diseases field, we verify that the knowledge about a specific disease is commonly disseminated over a small group of experts.

The diseasecard is a web portal for rare diseases, based on an automatic information retrieval engine, that provides transparently to the user a virtually integration of distributed and heterogeneous information – using a pathway that goes from the symptom to the gene. With this system medical doctors can access genetic knowledge without the need to master biological databases, teachers can illustrate the network of resources that build the modern biomedical information landscape and general citizen can learn and benefit from the available navigation model.

Acknowledgement

This portal was developed under the endorsement of the INFOGENMED project and of the NoE INFOBIOMED, both funded by the European Community, under the Information Society Technologies (IST) program.

References

1. R. B. Altman, "Bioinformatics in Support of Molecular Medicine" Em AMIA Annual Symposium, Orlando.," presented at Proc AMIA Symp., 1998.
2. B. D. Sarachan, M. K. Simmons, P. Subramanian, and J. M. Temkin, "Combining Medical Informatics and Bioinformatics toward Tools for Personalized Medicine," *Methods Inf Med*, vol. 42, pp. 111-5, 2003.
3. D. B. Searls, "Data Integration: Challenges for Drug Discovery," *Nature Reviews*, vol. 4, pp. 45-58, 2005.
4. Martin-Sanchez, I. Iakovidis, S. Norager, V. Maojo, P. de Groen, J. Van der Lei, T. Jones, K. Abraham-Fuchs, R. Apweiler, A. Babic, R. Baud, V. Breton, P. Cinquin, P. Doupi, M. Dugas, R. Eils, R. Engelbrecht, P. Ghazal, P. Jehenson, C. Kulikowski, K. Lampe, G. De Moor, S. Orphanoudakis, N. Rossing, B. Sarachan, A. Sousa, G. Spekowius, G. Thireos, G. Zahlmann, J. Zvarova, I. Hermosilla, and F. J. Vicente, "Synergy between medical informatics and bioinformatics: facilitating," *J Biomed Inform*, vol. 37, pp. 30-42., 2004.
5. J. L. Oliveira, G. Dias, I. Oliveira, P. Rocha, I. Hermosilla, J. Vicente, I. Spiteri, F. Martin-Sánchez, and A. S. Pereira, "DiseaseCard: A Web-Based Tool for the Collaborative Integration of Genetic and Medical Information," presented at Biological and Medical Data Analysis: 5th International Symposium (ISBMDA'2004), Barcelona, Spain, 2004.
6. M. Y. Galperin, "The Molecular Biology Database Collection: 2005 update," *Nucleic Acids Research*, vol. 33, 2005.
7. Regular Expressions in Java. (n.d.) Retrieved June 17, 2005, from http://www.regular-expressions.info/java.html