Exploring ontological knowledge for translational medicine

Catia M. Machado
LASIGE, Dep. de Informática
Instituto de Engenharia de Sistemas e Computadores
- Investigação e Desenvolvimento
Universidade de Lisboa
Portugal
cmachado@xldb.di.fc.ul.pt

Francisco M. Couto
LASIGE, Dep. de Informática
Universidade de Lisboa
Portugal
fcouto@di.fc.ul.pt

Ana T. Freitas
Instituto de Engenharia de Sistemas e Computadores
- Investigação e Desenvolvimento/Instituto Superior Técnico
Universidade de Lisboa
Portugal
atf@inesc-id.pt

Ontologies provide a means to formally describe a domain knowledge in a structured manner that can be shared between people and computers alike.

In translational medicine approaches (i.e., approaches focused on the improvement of human health by bridging the gap between basic science research and clinical practice) ontologies play an invaluable role in knowledge representation and in data integration [1]-[2].

Ontologies are also a fundamental tool in the implementation of the semantic web proposed by Tim Berners-Lee *et al.* [3], where the Web of documents is replaced by the Web of data, thus allowing the manipulation of data over disparate domains and solving most of the problems previously reported for data integration.

One of the possible results of translational medicine is the concretization of personalized medicine, where individual genotypes (i.e., the genetic information of a person) can be associated with the phenotype (i.e., the observable characteristics) [4]. The identification of such associations can then 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).

Disease prognosis framework

The goal of our work is the development of a disease prognosis framework that leverages the representation of knowledge in the form of ontologies. This framework comprises two components: a data representation and integration component; and a data enrichment and analysis component. At the core of the integration component is a semantic model developed in the semantic web technology OWL [5]. On the one hand, this model allows the conceptual representation of data from heterogeneous domains of knowledge, which facilitates its integration. On the other hand, it promotes the discovery of new disease knowledge through the query of explicitly stated information and through inference. The second component incorporates knowledge from existing ontologies in the form of overrepresented ontology terms into the dataset under analysis. The use of these terms as features is expected to improve the quality of the predictions made with the dataset, which will be done through the exploration of data mining models reflecting the genotype-phenotype associations of the disease (Fig. 1).

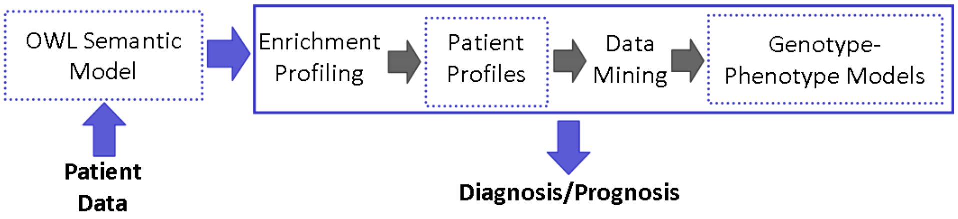

Figure 1: Disease framework. This framework comprises a data representation and integration component, and a data enrichment and analysis component. The core of the first component is an OWL semantic model that provides the conceptual representation of disease knowledge, which assists in the integration 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.

Data representation and integration

In the first component of the framework, we used as case study the disease hypertrophic cardiomyopathy (HCM). This is an autosomal dominant genetic disease with approximately 1000 mutations in more than 50 genes currently known to be associated with it [6], and is the most frequent cause of sudden cardiac death (SCD) in apparently healthy young people and athletes [7]-[8].

The current version of the semantic model is a result of several iterations, which included reusing existing vocabularies through the incorporation of their knowledge and the creation of mappings at concept-level [9]. The following four vocabularies were considered for this purpose: Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [10], National Cancer Institute Thesaurus (NCIt) [11], Sequence Ontology [12], and Gene Regulation Ontology [13].

The model was developed modularly, mainly due to the need to integrate data about fundamentally different real-world entities. The resulting three modules are: *Clinical Evaluation*, containing administrative concepts and clinical data elements necessary for the diagnosis and the prognosis of HCM patients; *Genotype Analysis*, containing concepts associated with the genetic testing of biological samples; and *Medical Classifications*, an auxiliary module containing medical standards used in the characterization of clinical elements such as patient symptoms.

In total, the modules contain 66 concepts and 102 properties. The module *Clinical Evaluation* (ce:) imports the other two modules through non-hierarchical relationships defined between them (e.g., ce:*Subject* ce:*hasBiologicalSample* ga:*Biological Sample*, where ga: indicates a class from the module *Genotype Analysis*).

Data enrichment and analysis

The main goal of the second component of the framework is the identification of genotype-phenotype associations that can improve the diagnosis/prognosis of the disease under analysis. Since the datasets are collected in the context of medical practice, they are frequently characterized by a small number of clinical features and a high number of missing values, which impairs their use for knowledge extraction purposes. To deal with these limitations, we have proposed to explore enrichment analysis with this type of dataset to extract relevant knowledge from controlled vocabularies to improve the quality of the dataset and, therefore, improve the quality of the predictions made with it [14]-[15].

Enrichment analyses (also known as overrepresentation analyses) are extensively used for the functional analysis of large lists of genes identified with high-throughput technologies, such as expression microarrays. They exploit 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. The ontology most commonly used in these analyses is the Gene Ontology [16]-[17].

In our implementation of an enrichment analysis we explored the Singular Enrichment Analysis (SEA) [18]. This enrichment approach works with a gene set (the *study set*) selected by the researcher from a reference set of genes (the *population set*) and iteratively tests the enrichment of each individual ontology term in a linear mode.

The first implementation of this methodology was done with genetic data and the Gene Ontology [15]. This ontology allows the annotation of biological products with terms describing the molecular functions (MF) they perform, the biological processes (BP) in which they are involved, and the cellular components (CC) where they are located or of which they are a component.

Our adaptation of the Singular Enrichment Analysis to a dataset of patients resulted in two different analyses: an enrichment profiling, and a differential enrichment. Both analyses were applied with the end purpose of identifying a set of ontological terms that can be used to profile the patients, terms that will be incorporated into the original dataset as features. Each of the analyses was built to test slightly different forms of obtaining the set of profiling terms. In the enrichment profiling, the aim was 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 the patients 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 genes mutated in the patients with the complete set of genes in the Human genome. In the differential enrichment analysis, the aim was to identify differentiating features between a group of patients with a particular characteristic (e.g., being positive for a disease-related event) and all the patients with the disease. This analysis is also based on the set of mutations the patients have, considering the mutations in the study group vs. the mutations in all the patients.

As in the first component of the framework, HCM was used as case study. The analyzed disease-related event was the occurrence of SCD, and so two different study sets were considered: patients with SCD and patients without SCD (no-SCD). In the enrichment profiling, SCD and no-SCD were tested in turn against the Human genome, whereas in the differential enrichment SCD and no-SCD were tested in turn against the entire population of HCM patients.

The enrichment profiling analysis identified several 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. Examples of the enriched terms are *regulation of heart rate* and *adult heart development* (both BP), *myosin heavy chain binding* (MF), and *striated muscle myosin thick filament* (CC).

The differential enrichment analysis identified one term for SCD and five terms for no-SCD (p-value < 0.1, not considering multiple-testing correction). The SCD term is the MF *structural constituent of muscle* (p-value = 0.08), and the no-SCD terms are: *negative regulation of ATPase activity* (p-value = 0.08) and *regulation of ATPase activity* (p-value = 0.09; both BP); *striated muscle thin filament* (p-value = 0.08) and *troponin complex* (p-value = 0.08; both CC); and *troponin C binding* (p-value = 0.08; MF).

The results obtained with these analyses indicate that the enrichment profiling analysis is useful for the characterization of patients, as it allowed the identification of meaningful terms associated with HCM. Notwithstanding, the preliminary results here referred where obtained only for the genetic data and thus we are currently performing the analysis of the clinical data with the vocabularies SNOMED-CT and NCIt. Additionally, we are testing the enrichment methodology with a second dataset, of patients with chronic obstructive pulmonary disease, considering the same vocabularies as for HCM.

The evaluation of the enrichment methodology will be done by applying data mining algorithms to the datasets: without enriched terms; and adding the enriched terms as features to the dataset. We expect that the results obtained with the algorithms will be improved by the inclusion of the ontology terms.

The disease framework was designed do that it can be used for any genetic disease. On the data representation and integration component, the modular development of the semantic model facilitates its extension and reutilization. Similarly, the enrichment and mining methodology in the second component was developed to explore the most of the case-study dataset but avoiding overfitting.

Once finalized, the disease framework will play an important role in the concretization of translational medicine by assisting medical doctors in the definition of the appropriate treatments and preventive actions for individual patients.

Acknowledgments

This work was supported by national funds through the Fundação para a Ciência e a Tecnologia under the Multi-annual Funding Program, the doctoral grant SFRH/BD/65257/2009, the projects SOMER (PTDC/EIA-EIA/119119/2010), PEst-OE/EEI/LA0021/2013, and PTDC/EIA-EIA/112283/2009.

References

[1] Albani S. & Prakken B. (2009) The advancement of translational medicine-from regional challenges to global solutions. *Nature Medicine* **15**(9):1006-1009.

[2] Woolf S.H. (2008 The meaning of translational research and why it matters. *JAMA* **299**(2):211-213.

[3] Berners-Lee T., Hendler J. & Lassila O. (2001) The Semantic Web. *Scientific American* **284**(5):34-43.

[4] 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**:241251.

[5] OWL Web Ontology Language Current Status: `http://www.w3.org/standards/techs/owl#w3call`.

[6] Heart Genetics - Cardiovascular genetic panel: `http://www.heartgenetics.com/#tests`.

[7] Maron,B., Maron,M., Wigle,E. & 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**:191200.

[8] Alcalai,R., Seidman,J. & Seidman C (2008) Genetic Basis of Hypertrophic Cardiomyopathy: from Bench to the Clinics. *Journal of Cardiovascular Electrophysiology* **19**:104110.

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

[10] SNOMED-CT: `http://www.ihtsdo.org/snomed-ct/`

[11] Sioutos,N., Coronado,S., Haber,M.W., Hartel,F.W., Shaiu,W.L. & Wright,L.W. (2007) NCI thesaurus: A semantic model integrating cancerrelated clinical and molecular information. *Journal of Biomedical Informatics* **40**:3043.

[12] 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.

[13] 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**:914.

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

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

[16] Robinson,P. & Bauer,S. (2011) Overrepresentation analysis. In N. Britton, X. Lin, H.M. Safer, M. Singh and, A. Trammontano (eds.), *Introduction to Bio-Ontologies, Mathematical and Computational Biology Series*,pp.181218. CRC Press, Taylor and Francis Group.

[17] 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.

[18] Huang,D., Sherman,B. & Lempicki,R. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**:113.