Enrichment analysis applied to disease prognosis

Catia M Machado[1,2], Ana T Freitas[2] and Francisco M Couto[1]*

[1]*LaSIGE, Departmento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal*

[2]*Instituto de Engenharia de Sistemas e Computadores/Instituto Superior Técnico, Lisboa, Portugal*

ABSTRACT

Enrichment analysis is normally used to identify relevant biological features that can be used to describe a set of genes under analysis that, for example, share a common expression profile.

In this article we propose the exploitation of enrichment analysis for a different purpose: the evaluation of a disease prognosis. With this application of enrichment analysis we expect to identify clinical and biological features that best differentiate between patients that suffered a specific disease event from those that did not. The features thus identified will be used to create patient profiles, which will in turn be evaluated through similarity and supervised classification approaches to predict the occurrence of the event.

This article presents the enrichment analysis methodology proposed for a prognosis study, in which we use the disease hypertrophic cardiomyopathy and its most severe manifestation, sudden cardiac death, as a case study.

1 INTRODUCTION

Enrichment analysis is normally 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 the gene set under analysis more than would be expected by chance. Such biological features are said to be enriched, or overrepresented, and are then used to formulate a biological interpretation of the gene set.

The ontology most commonly used in these analyses is the Gene Ontology (Ashburner et al. 2000, Robinson and Bauer 2011, Zhang et al. 2010), although other resources such as MeSH and KEGG are also explored (Leong and Kipling 2009). Strategies based on multiple vocabularies have also been developed, namely in pharmacogenomics, including the Human Disease Ontology and the Pharmacogenomics Knowledge Base (Hoehndorf et al. 2012). LePendu et al. propose a method to generate annotations when using vocabularies other that the Gene Ontology, testing its feasibility with the Disease Ontology (LePendu et al. 2011).

In terms of statistical methods, the most commonly used is the Fisher's exact test (Robinson and Bauer 2011, Huang et al. 2009), with more recent implementations also using Bayesian techniques (Bauer et al. 2010).

Enrichment analyses are normally divided in three categories: Singular Enrichment Analysis (SEA), Gene Set Enrichment Analysis (GSEA) and Modular Enrichment Analysis (MEA). SEA works with a user-selected gene set and iteratively tests the enrichment of each individual ontology concept in a linear mode. GSEA also evaluates the enrichment of ontology concepts individually, but considering all the genes in the experiment and not just a user-selected gene set. MEA works with a user-selected gene set, but incorporates into the analysis the relationships between concepts represented in the ontologies, thus evolving from a term-centric approach to a biological module-centric approach (Huang et al. 2009).

Several tools have been developed that implement one or more of these approaches. Examples of these tools are Onto-express (Khatri et al. 2002), GSEA (Subramanian et al. 2005), and GOToolBox (Martin et al. 2004) (a detailed list of tools was collected by Huang et al. 2009).

In this work we propose to adapt the enrichment analysis to develop a disease prognosis methodology, with the goal of predicting if specific events may or may not occur in a given patient. The enrichment analysis will be applied to identify the set of clinical and genetic features that might assist us in the differentiation of the patients for whom the event occurred from the patients for whom it did not. The identified features will then be used to create profiles for the individual patients. In order to differentiate between the two sets of patients, the profiles will be subjected to an evaluation step, in which we will explore a similarity and a classification approach. In the similarity approach, different semantic similarity measures (Pesquita et al. 2009) and a relatedness measure (Ferreira and Couto 2011) will be tested to compare the profiles, followed by machine learning algorithms such as clustering and nearest neighbors. In the classification approach, the patient profiles will be analyzed with supervised classification algorithms such as random forests (Breiman 2001) and Bayesian networks (Berner 2007) (see Fig. 1).

* To whom correspondence should be addressed.

The datasets to be used in the implementation of this methodology were collected by biomedical experts in the context of medical practice, and are thus characterized by a small number of clinical features and a high number of missing values, among other aspects. With this work, our purpose is to evaluate if the application of an enrichment analysis to this type of dataset can result in the extraction of relevant knowledge from controlled vocabularies to improve the quality of the dataset and, consequently, the quality of the predictions made from it.

As a case study we will consider the disease hypertrophic cardiomyopathy (HCM). This is a genetic disease that is the most frequent cause of sudden cardiac death (SCD) among apparently healthy young people and athletes (Maron et al. 2009, Alcalai et al. 2008). It is characterized by a variable clinical presentation and onset, and there are approximately 900 mutations in more than 30 genes currently known to be associated with it (Harvard Sarcomere Mutation Database). Due to these characteristics, HCM is very difficult to diagnose. The prognosis is by no means easier, since the severity of the disease varies even between direct relatives. It has been observed that the presence of a given mutation can correspond to a benign manifestation in one individual and result in SCD in another (Maron et al. 2009, Alcalai et al. 2008).

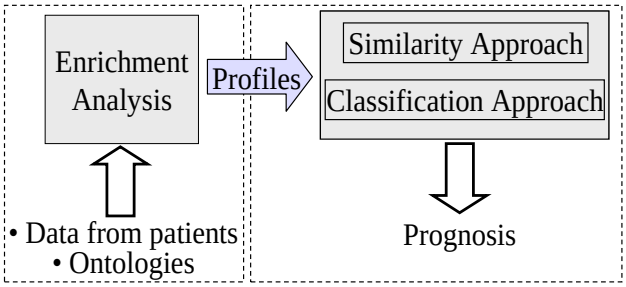

Fig. 1. Schematic representation of the prognosis methodology. The methodology is composed of two units: the first (left-side) receives as input data from patients mapped to biomedical ontologies (or controlled vocabularies in general). It will apply an enrichment analysis to identify a list of ontology terms considered to be enriched, which will be used to create profiles for the patients. These profiles will then be subjected to an evaluation step (the second unit, on the right-side) that will result in the evaluation of the prognosis for individual patients. For the implementation of the second unit, we will explore a similarity and a classification approach.

Due to the importance of the prognosis of HCM in terms of SCD, this will be the event analyzed in our present study.

This work is currently under development, and in the rest of the article the focus will be on our proposed application of enrichment analysis to disease prognosis. In the following sections we present the dataset and the methodology. In the methodology section we begin by drawing a parallel between the application of this analysis in the context of gene expression data analysis and in the context of the prognosis methodology. Finally, we present how the enrichment analysis will be conducted with data from HCM patients, and how the patient profiles will be created from the results obtained.

2 DATASET

The data necessary for the diagnosis and the prognosis of HCM has been represented in a semantic data model, with mappings established between the concepts in the model and four controlled vocabularies: the National Cancer Institute Thesaurus (NCIt) (version 10.03) (Sioutos et al. 2007), the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) (version 2010_01_31) (SNOMED), 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). A total of 85.8% of the clinical concepts represented in the model was mapped either to NCIt or SNOMED-CT, in identical proportion (42.9%).

Table 1 contains all the clinical features to be used in the present work. With the exception of two of these features, *Sporadic* and *Hypertrophy morphology*, they are represented in the semantic model and have an established mapping with NCIt or SNOMED CT.

Table 1. Clinical features considered in the enrichment analysis and their possible values.

Feature	Possible values
Cardioverter defibrillator Non-obstructive HCM Obstructive HCM Resuscitated sudden death Sudden death Non-sudden death Sudden death family history Familial Sporadic	-1;1*
Blood pressure	normal; hypertension; hypotension
Gender	male; female
Age	1; 2; 3; 4
Hypertrophy morphology	apical; centric; concentric

* The values -1 and 1 correspond respectively to the absence or the presence of the feature in the patient.

Familial and sporadic indicate if the patient has either the familial (hereditary) or the sporadic form of HCM.

The age values correspond to the following intervals, in years: (1) [0,20]; (2)]20,40]; (3)]40,60]; (4) >60.

The genetic features are the mutations associated with the disease, with possible values {-1,1}, i.e. absence or presence of the mutation in the genome of the patient. The genes in

which each mutation occurs are currently being mapped to the Gene Ontology.

Both clinical and genetic features have been previously collected for 80 patients from Portuguese hospitals and molecular biology research laboratories, for the evaluation of associations between genetic and clinical factors. The clinical features presented in Table 1 are considered by the medical experts as the most relevant for the diagnostic and the prognosis of HCM, and were thus the only ones provided for our present study. Table 2 shows the percentage of patients that have a known value for each of the clinical features and the total of 569 mutations tested.

Table 2. Percentage of patients that have a known value for each of the clinical features and for the total number of mutations.

Feature	Patients (%)
Cardioverter defibrillator	96
Non-obstructive HCM	36
Obstructive HCM	36
Resuscitated sudden death	96
Sudden death	100
Non-sudden death	100
Sudden death family history	37
Familial	96
Sporadic	96
Blood pressure	39
Gender	96
Age	60
Hypertrophy morphology	96
Mutations	76

3 METHODOLOGY

The first step in any enrichment analysis is the definition of the list of entities to be analyzed.

Considering the case of gene expression analysis, the complete list of genes under analysis is called the *population set*. As referred in the Introduction, the GSEA receives this list as input. However, both SEA and MEA require two sets of genes as input: a user-selected gene set, which is called the *study set* and is a sub-set of the population set; and the population set. The criterion used to select the study set can be (and normally is) the level of expression of the genes in the biological setting under analysis, meaning that the study set will be the set of genes that are considered to be over- and/or under-expressed. The evaluation of the existence of enriched ontology terms is then made for the study set in respect to the entire population set. This means that for an annotated term to be considered as enriched its annotation rate has to be higher in the study set than in the population set.

Considering the application of the enrichment analysis to the prognosis of disease-related events, the population set is the complete list of patients with the disease. Since we are interested in obtaining a list of enriched ontology terms for the set of patients for whom the event occurred and also for the set of patients for whom it did not, each set will be in turn considered the study set. Fig. 2 shows an exemplificative representation of the population and study sets in a gene expression experiment, and their counterparts in the prognosis analysis.

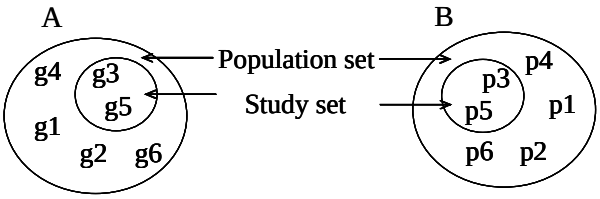

Fig. 2. Population set and study set in (A) a gene expression analysis and (B) the prognosis of disease-related events. In this example, the population set in A is composed of 6 genes, g1 to g6, and the study set of 2 genes, g3 and g5. In accordance, the population and study sets in B are composed of 6 and 2 patients, respectively.

3.1 Definition of patient profiles

Our aim is to define the patient profiles based on the result of individual enrichment analyses performed with different controlled vocabularies.

In order to assess the feasibility of this methodology, we will begin by performing analyses with the Gene Ontology and the NCIt.

Considering our case study of SCD occurrence in HCM patients, we intend to evaluate the existence of ontology terms that can assist us in separating patients with SCD from patients without SCD.

When performing the analysis with the Gene Ontology, the terms which enrichment will be evaluated depend on the mutations the patients have. Firstly, the list of mutations that all the patients in the study set have (e.g. patients with SCD, with mutation value =1) is compiled; secondly, the list of non-redundant mutated genes is retrieved from the list of mutations; finally, the list of Gene Ontology terms used to annotate the mutated genes is retrieved. The terms annotated to the patients in the rest of the population are retrieved in the same manner. The frequency of occurrence of the annotations is then calculated based on the patients, i.e., how many patients in the study set and the population set are annotated with the term. For each term, a patient can only be counted once, even if he/she has more than one mutation through which the term can be identified.

When performing the analysis with NCIt, the terms which enrichment will be evaluated depend on the values of the clinical features. For the features with possible values {-1,1}, they will be considered if having value equal to 1 (thus being present in the patient); for the categorical features, all will be considered except when there are no known values for any of the patients in the set. The terms annotated to the

features are retrieved based on the mappings already defined between them and the NCIt. The following two features exemplify the procedure for boolean and categorical variables, respectively:

- *Non sudden death*: when value =1, retrieve and use the term *Non_Sudden_Cardiac_Death* (and its parent terms).
- *Blood pressure*: when value =*hypertension*, retrieve and use the term *Hypertension* (and its parent terms).

The frequency of occurrence of the annotations is calculated as before, i.e., how many patients in the study set and the population set are annotated with the term.

We will test both SEA and MEA approaches. Since GSEA produces a list of enriched terms for the entire set of entities, it is not as interesting for our study as the other two.

The lists of enriched terms that result from the analysis with each controlled vocabulary will be compiled and used as a template-profile for the respective set of patients (e.g. with SCD). The individual profiles will be defined as follows: for each patient and each ontology term, it is checked if the patient is annotated with the term; if true, a pair *variable/term* is created for that patient. The complete set of pairs *variable/term* thus obtained is the profile for that specific patient.

The pairs *variable/term* will substitute the original variables in the second unit of the prognosis methodology (Fig. 1).
In this study we include in the group of patients with SCD both patients that died due to a sudden cardiac arrest and patients that suffered at least one resuscitated sudden cardiac arrest (which can be either alive or dead). The group of patients without SCD includes all the other patients.

4 DISCUSSION AND CONCLUSIONS

In this article we present a novel prognosis prediction methodology based on an enrichment analysis. This type of analysis is normally used in contexts such as gene expression analysis for the identification of functional annotations that might be used to explain the differences in expression. Here we propose to use enrichment analysis for the identification of ontology terms that might be used to explain the differences between the group of patients for whom a given disease event occurred and the group of patients for whom it did not occur. The ontology terms considered to be enriched will assist in the creation of profiles for individual patients. These profiles will then be used to evaluate for new patients if the event might occur or not.
An important aspect of the present analysis is the dataset: it contains data from patients, and was collected in the context of their medical evaluation. As such, it reflects two important aspects of the nature of clinical records: only the information deemed relevant by the medical experts is present; not all of the information is available for all of the patients.
Our interest is precisely in evaluating if it is feasible to extract relevant knowledge from controlled vocabularies that can enrich the dataset, and thus allow its exploitation with data mining algorithms.
In a first approach, we will test only two vocabularies: the Gene Ontology and the NCIt. Although this means that some of the features will not be considered due to the inexistence of annotations, we expect to be able to evaluate the applicability of the methodology.
The data under analysis in this study has been provided by several Portuguese institutions, including hospitals and molecular biology research laboratories.

ACKNOWLEDGEMENTS

This work was supported by the FCT through the Multi-annual Funding Program, the doctoral grant SFRH/BD/65257/2009 and the SOMER project (PTDC/EIA-EIA/119119/2010).
The authors would like to thank to Alexandra R. Fernandes, Susana Santos and Dr. Nuno Cardim for collecting and providing the dataset.

REFERENCES

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. and Sherlock,G. (2000) Gene Ontology: tool for the unification of biology. *Nat. Gene.t*, **25,** 25–29.

Robinson,P.N. and Bauer,S. (2011) *Introduction to Bio-Ontologies*. (Chapter 8) CRC Press Taylor & Francis Group.

Zhang,S., Cao,J., Kong,Y.M. and Scheuermann,R.H. (2010) GO-Bayes: Gene Ontology-based overrepresentation analysis using a Bayesian approach. *Bioinformatics*, **26**(7), 905-911.

Leong,H.S. and Kipling,D. (2009) Text-based over-representation analysis of microarray gene lists with annotation bias. *Nucleic Acids Res.* **37**(11), e79.

Hoehndorf,R., Dumontier,M., Gkoutos,G.V. (2012) Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. *Bioinformatics*, **28**(16), 2169-75.

LePendu,P., Musen,M.A., Shah, N.H. (2011) Enabling enrichment analysis with the Human Disease Ontology. *J. Biomed. Inform.* **44**(Suppl 1), S31-8.

Huang,D.W., Sherman,B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**(1), 1-13.

Bauer,B., Gagneur,J. and Robinson,P.N. (2010) GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Res.* **38**(11), 3523-3532.

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

Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. and Mesirov,J.P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550.

Martin,D., Brun,C., Remy,E., Mouren,P., Thieffry,D. and Jacq,B. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol.* **5**, R101.

Pesquita,C., Faria,D., Falcão,A.O., Lord,P. and Couto,F.M. (2009) Semantic Similarity in Biomedical Ontologies. *PLoS Comput. Biol.* **5**(7): e1000443.

Ferreira,J. and Couto,F. (2011) Generic semantic relatedness measure for biomedical ontologies. *International Conference on Biomedical Ontologies (ICBO)*, 2011.

Breiman,L. (2001). Random forests. *Mach. Learn.* **45**, 5–32.

Berner,E.S. (Editor) (2007) *Clinical decision support systems: theory and practice*. (Chapter 3) Health Informatics Series, 2nd Edition.

Maron,B.J., Maron,M.S., Wigle,E.D. and 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. *J. Am. Coll. Cardiol.* **54**, 191-200.

Alcalai,R., Seidman,J.G. and Seidman,C.E. (2008) Genetic Basis of Hypertrophic Cardiomyopathy: from Bench to the Clinics. *J. Cardiovasc. Electrophysiol.* **19**, 104-110.

Harvard Sarcomere Mutation Database: http://genepath.med.harvard.edu/~seidman/cg3/

Sioutos,N., Coronado,S., Haber,M.W., Hartel,F.W., Shaiu,W.L. and Wright,L.W. (2007) NCI Thesaurus: a Semantic Model Integrating Cancer-Related Clinical and Molecular Information. *J. Biomed. Inform.* **40**, 30-43.

Systematized Nomenclature of Medicine-Clinical Terms (SNOMED), http://www.ihtsdo.org/snomed-ct/

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. *Stud. Health Technol. Inform.* **136**, 9–14.

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 Biol*. **6**, R44

Machado,C.M., Couto,F.M., Fernandes,A.R., Santos,S. and Freitas,A.T. (2012) Toward a translational medicine approach for hypertrophic cardiomyopathy. *International Conference on Information Technology in Bio- and Medical Informatics* (ITBAM), 2012.