D51 ANALYSIS MODULES IDENTIFICATION (M12)

This deliverable provides a first classification of the modules which are common to the various aspects of data analysis and therefore should be considered as the data analysis building blocks of GenData.

The deliverable is organized in four sections: the first section concerns the analysis of genomic data. It presents the main contributions of the GenData project to the analysis and annotation of the raw sequences produced by Next-Generation Sequencing experiments. The second section presents a framework for managing and analyzing genomic data by combining OLAP analysis and Data Mining. The third section addresses the analysis of genomic metadata, which consist of sequence annotations and related information coming from external sources. Finally, the last section addresses the analysis of data types other than genomic data (e.g., clinical data) which are anyhow related to the GenData model.

Each section is organized as follows. Subsection "Introduction" presents the main issues in the field and it clearly states the aim of the research. Subsection "State-of-the-art" gives a big picture of the context under analysis and it overviews the most recent research advances in the field. Finally, Subsection "Activity description" thoroughly describes the activities undertaken in the context of GenData to address the specific issues.

INDEX

SECTION 1

ANALYSIS OF GENOMIC DATA

1.1 Peak shape analysis

Anna Maria Paganoni

Dipartimento di Matematica "Francesco Brioschi"

Politecnico di Milano

Introduction

Peak shape project idea is to take into consideration the shape of CHIP-Seq peaks, detecting statistically significant shape differences and associating the shape to a functional role and a biological meaning. For this purpose we use different techniques for clustering peaks, from multivariate analysis on shape indices, to functional data analysis treating peaks as curves.

State-of-the-art

At present the analysis of ChIP-Seq data is mainly restricted to the detection and the investigation of enriched areas of the genome, namely the peaks, and almost only signal intensity is considered in the analysis.

Motivated by the fact that these peaks show very different shapes, we propose to take into account also the shape. The idea is that statistically significant shape differences are associated with a functional role and a biological meaning.

Activity description

In the first part of the work, we study ChIP-Seq peaks through multivariate statistical techniques, by selecting five indices that summarize the shape. We use clustering algorithms on these indices in order to assess whether the peaks can be divided into groups according to their shapes. The resulting clusters are then validated, by checking for artifacts through available open chromatin regions experiments. Clusters are also characterized in terms of motif analysis and in terms of co-occurrences with other transcription factors. In this step bioinformatics tools are used, as well as random forests and multiple correspondence analysis. The application of this analysis pipeline to publicly available ChIP-Seq for the erythroid transcription factor GATA-1 in K562 cells leads to the identification of three clusters, suggesting the existence of statistically significant differences in peak shape inside a single ChIP-Seq. Such differences are associated with motif occurrences and with the presence of a novel protein complex.

In the second part of the project, we are going to investigate more deeply the connections between peak shapes and the nature of the bindings. For this purpose we are going to study the correlation between peak shapes and expression data, by using publicly available RNA-Seq experiments. An important step forward will be to consider the peaks as curves, embedding the problem in the functional data analysis framework. Data registration and functional clustering techniques, such as the k-mean alignment (or k-medoid alignment), will be modified and adapted for the analysis of ChIP-Seq peak shape. In this way peak shape will be studied in a more natural and direct way.

1.2 Discovering New Gene Functionalities From Random Perturbations Of Known Gene Ontological Annotations

Paolo Ciaccia, Giacomo Domeniconi, Gianluca Moro, Roberto Pasolini, Claudio Sartori

Dipartimento di Informatica, Scienza e Ingegneria

Università di Bologna

Introduction

Prediction of associations between items and features characterizing them is a common machine learning task which is often performed in several application domains. In bioinformatics, several terminologies and ontologies are available to describe structural and functional features of biomolecular entities. Among them, the most developed and relevant is the well-known Gene Ontology (GO) [9] The association of its terms to biomolecular entities, mainly genes and proteins, is widely used to annotate, and thus characterize, them. This task, for an organism, can be seen as a binary matrix in which each row corresponds to a gene and each column to a term, if there is an association between a gene and a term, the corresponding cell will be 1, otherwise 0.
The goal of our research is to discover new GO term annotations of different organism genes based on available GO annotations, we apply different supervised algorithms, differently from the common methods in literature, which involve the use of unsupervised techniques, such as clustering or SVD. To apply supervised algorithms to the prediction problem, we assign labels to the originally unlabeled GO annotations based on a random perturbation of the annotation matrix. In so doing, we create a training matrix with missing annotations; thus, we can train the model to recognize new annotations.
This methodology, as said, other than being useful and of interest for the bioinformatics community, could be directly useful also for the GenData project, since the *genome space* can be reduced to a similar matrix, where each row corresponds to a region (gene) and each column to a dataset (experiment). We plan to develop this aspect in the continuous of our research.

State-of-the-art

Different methods have been proposed to predict biomolecular annotations. In [16], decision trees and Bayesian networks were suggested to learn patterns from available annotation profiles and predict new ones. Along this line, Tao and colleagues [26] improved by using a k-nearest neighbour (k-NN) classifier to make a gene inherit the annotations that are common among its nearest neighbour genes in a gene network.
In [15] and [11], Khatri and colleagues suggested a prediction algorithm based on the Singular Value Decomposition (SVD) method of the gene-to-term annotation matrix, which is implicitly derived from the count of co-occurrences between pairs of terms in the available annotation dataset.
This prediction method based on basic linear algebra was then extended in [8], by incorporating gene clustering based on gene functional similarity computed on Gene Ontology annotations. It was further enhanced by automatically choosing its main parameters, including the SVD truncation level, based on the evaluated data [7]. The SVD has also been used with annotation co-occurrence weights based on gene-term frequencies [10, 21].
Other methods based on linear algebraic evaluation of co-occurrences exist; in particular the ones related to Latent Semantic Indexing (LSI) [12], which have been originally proposed in Natural Language Processing. Among them, the probabilistic Latent Semantic Analysis (pLSA) [14] gives a

well defined distribution of sets of terms as an approximation of the co-occurrence matrix. It uses the *latent* model of a set of terms to increase robustness of annotation prediction results. In [17] and [20], pLSA proved to provide general improvements with respect to the truncated SVD method of Khatri and colleagues [15].
In bioinformatics, *topic modeling* has been leveraged also by using the Latent Dirichlet Allocation (LDA) algorithm [3]. In [2] and [19], LDA was used to subdivide expression microarray data into clusters. Besides, they defined a new model able to consider a given dependence between genes; this dependence is introduced in the model through a variable that represents a categorization of the genes and that can be inferred from a priori knowledge on the evaluated genes. Very recently, Pinoli et al. [22] took advantage of the LDA algorithm, toghether with the Gibbs sampling [5, 23], to predict gene annotations to GO terms.

Activity description

All direct and indirect annotations of a set of genes can be represented by using binary matrices. Let $\boldsymbol{G}$ be the set of genes of a certain organism and $\boldsymbol{T}$ a set of feature terms. We define the annotation matrix $\boldsymbol{A^{TxG}}$ as the matrix whose columns (rows) correspond to terms (genes). For each gene and for each term, the value of the entry of the annotation matrix is set according to the following rule:

- 1, if the gene is annotated either to term or to any of its descendants
- 0, otherwise

As previously mentioned, unsupervised methods can be easily applied to an annotation matrix $\boldsymbol{A}$. Our proposal consists of creating a data representation that allows using supervised algorithms, which are based on data pre-labeling, to discover unknown associations between genes and feature terms.

Given the element $\boldsymbol{A(g, t)}$ of the annotation matrix, we want to predict if the gene *g* is likely to be annotated to the term *t* or not. This can be represented as a supervised problem, in which the label can be 0 or 1 according to the presence or absence of annotation between the gene and the term, while all other annotations of the gene represent the features of the record.

The problem of labeling each record arises. Given an annotation matrix, our proposal is to use as input a version of the matrix with less annotations (referred as outdated matrix since it may resemble an outdated annotation dataset version), derive from such input matrix the features of the data model and consider as label of each record the presence or absence of an annotation in a more complete matrix (referred as updated matrix since it may resemble a newer annotation dataset version). Given the term considered for the prediction, called *class-term*, the representation of the data is created by taking as features, for each gene, all the annotations to all the other terms in an outdated version of the matrix $\boldsymbol{A_{v0}}$, while the label is given by the value of the class-term in the updated version of the matrix $\boldsymbol{Av1}$**.**

This data representation is exactly the same as that of a supervised classification problem represented in a Vector Space Model. The representation of the data described above, however, requires two versions of the annotation matrix to create the training model, while the purpose of our research is to provide a method actually usable by biologists, that works using only a single version of the gene-term annotations matrix.
To overcome the problem just mentioned, we start from the observation that the input matrix already contains errors, namely missing annotation, therefore firstly we can consider using only this matrix to obtain the training data representation, assuming $\boldsymbol{A_{v0}=A_{v1}}$.

This task, as mentioned, is the same of an unbalanced binary classification, the number of gene-term annotated couples is much lower than those without. In the literature [6] it is known that the unbalancing of classes leads to classification models highly skilled in the recognition of the highest frequency classes with respect to those with lower, which are the ones of interest. Considering this, it may be helpful to use a representation model that contains a greater number of errors, to better train the prediction model.
If we consider that the annotations of genes and features are discovered by teams of biologists that work independently from of each other, a reasonable hypothesis is that the new annotations discovered by the entire scientific community, on the whole, do not have any kind of bond or rule. This should be equivalent to a random process of discovery of the errors.
This has lead to our thesis according to which new gene annotations can be discovered by artificially increasing in the input matrix A_{v1} the instances belonging to the least represented class, namely by randomly increasing the number of missing annotations. This is achieved by randomly deleting known annotations in the matrix A_{v0}.
Thus, to get the data to train the classification model, we propose to create a new matrix by randomly perturbing A_{v1}, in which some annotations are eliminated with a probability p.

Once created the training matrix, we can use any supervised algorithm to train the prediction model and then validate it with starting from A_{v1} to discover the missing annotations with respect to an updated version A_{v2}.
The predicition model provides a probability distribution *pd(g,t)*, called *likelihood*, concerning the presence of an annotation of the term *t* with the gene *g*. To provide predictions of only new annotations, only those annotations that were not present in the outdated version of the matrix are taken into account. The supervised process described above, is repeated for all the terms in T, giving as final output a list of predictions of new annotation profiles ordered according to their likelihood.

The workflow of the proposed methodology for the discovery of new annotations can be summarized as follows:

- We extracted an outdated version of the annotation matrix.
- We randomly perturbed the annotation matrix to get a modified version of it, with some missing annotations.
- By running the prediction algorithm, we got a list of predicted annotations ordered by their confidence value (i.e. their corresponding likelihood value).
- We selected the top *P* predictions and we counted how many of these *P* predictions were found confirmed in the updated version of the annotation matrix.
- For each experiment, steps 2, 3, 4 were repeated 10 times by varying the random seed.

Bibliography

[1] Barutcuoglu, Z., Schapire, R. E., and Troyanskaya, O. G. (2006). Hierarchical multi-label prediction of gene function. Bioinformatics, 22(7):830–836.

[2] Bicego, M., Lovato, P., Oliboni, B., and Perina, A. (2010). Expression microarray classification using topic models. In Proceedings of the 2010 ACM Symposium on Applied Computing, pages 1516–1520. ACM.

[3] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022.

[4] Canakoglu, A., Ghisalberti, G., and Masseroli, M. (2012). Integration of biomolecular interaction data in a genomic and proteomic data warehouse to support biomedical knowledge discovery. In Computational Intelligence Methods for Bioinformatics and Biostatistics, pages 112–126. Springer.

[5] Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. The American Statistician, 46(3):167–174.

[6] Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter, 6(1):1–6.

[7] Chicco, D. and Masseroli, M. (2013). A discrete optimization approach for svd best truncation choice based on roc curves. In Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on, pages 1–4. IEEE.

[8] Chicco, D., Tagliasacchi, M., and Masseroli, M. (2012). Genomic annotation prediction based on integrated information. In Computational Intelligence Methods for Bioinformatics and Biostatistics, pages 238–252. Springer.

[9] Consortium, G. O. et al. (2001). Creating the gene ontology resource: design and implementation. Genome research, 11(8):1425–1433.

[10] Done, B., Khatri, P., Done, A., and Draghici, S. (2007). Semantic analysis of genome annotations using weighting schemes. In Computational Intelligence and Bioinformatics and Computational Biology, 2007. CIBCB'07. IEEE Symposium on, pages 212–218. IET.

[11] Done, B., Khatri, P., Done, A., and Draghici, S. (2010). Predicting novel human gene ontology annotations using semantic analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 7(1):91–99.

[12] Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 281–285. ACM.

[13] G. Pandey, V. K. and Steinbach, M. (2006). Computational approaches for protein function prediction: A survey. Technical report, Minneapolis, MN, USA.

[14] Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57. ACM.

[15] Khatri, P., Done, B., Rao, A., Done, A., and Draghici, S. (2005). A semantic analysis of the annotations of the human genome. Bioinformatics, 21(16):3416–3421.

[16] King, O. D., Foulger, R. E., Dwight, S. S., White, J. V., and Roth, F. P. (2003). Predicting gene function from patterns of annotation. Genome research, 13(5):896–904.

[17] Masseroli, M., Chicco, D., and Pinoli, P. (2012). Probabilistic latent semantic analysis for prediction of gene ontology annotations. In Neural Networks (IJCNN), The 2012 International Joint

Conference on, pages 1–8. IEEE.

[18] Perez, A. J., Perez-Iratxeta, C., Bork, P., Thode, G., and Andrade, M. A. (2004). Gene annotation from scientific literature using mappings between keyword systems. Bioinformatics, 20(13):2084–2091.

[19] Perina, A., Lovato, P., Murino, V., and Bicego, M. (2010). Biologically-aware latent dirichlet allocation (balda) for the classification of expression microarray. In Pattern Recognition in Bioinformatics, pages 230–241. Springer.

[20] Pinoli, P., Chicco, D., and Masseroli, M. (2013). Enhanced probabilistic latent semantic analysis with weighting schemes to predict genomic annotations. In Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on, pages 1–4. IEEE.

[21] Pinoli, P., Chicco, D., and Masseroli, M. (2013). Weighting scheme methods for enhanced genome annotation prediction. In Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB), 2013 10th International Meeting on.

[22] Pinoli, P., Chicco, D., and Masseroli, M. (2014). Latent dirichlet allocation based on gibbs sampling for gene function prediction. In Proceedings of the International Conference on Computational Intelligence in Bioinformatics and Computational Biology, pages 1–7. IEEE Computer Society.

[23] Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 569–577. ACM.

[24] Raychaudhuri, S., Chang, J. T., Sutphin, P. D., and Altman, R. B. (2002). Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. Genome Research, 12(1):203–214.

[25] Tanoue, J., Yoshikawa, M., and Uemura, S. (2002). The genearound go viewer. Bioinformatics, 18(12):1705–1706.

[26] Tao, Y., Sam, L., Li, J., Friedman, C., and Lussier, Y. A. (2007). Information theory applied to the sparse gene ontology annotation network to predict novel gene function. Bioinformatics, 23(13):i529–i538.

SECTION 2

GOLAM: A FRAMEWORK FOR ANALYZING GENOMIC DATA

Lorenzo Baldacci, Matteo Golfarelli, Simone Graziani, Stefano Rizzi

Dipartimento di Informatica, Scienza e Ingegneria

Università di Bologna

This section presents GOLAM, a framework for OLAP analysis and mining of matches between genomic regions extracted using the GenData Language. The goal of GOLAM is to overcome the current limitations of genome analysis methods, that are normally based on browsing. This is done by partially automating and speeding-up the analysis process on the one hand, by making it more flexible and introducing a multi-resolution view of data on the other. In this paper we focus on conveying its potential and on describing its functional architecture.

Introduction

One of the analysis services envisioned in GenData 2020 is related to multi-resolution analysis of matches between genomic regions. To understand the importance of this service, consider the real scenario in which the user (typically, a biologist or a geneticist) knows the regulatory mechanisms of a gene and wants to discover other genes with similar regulation mechanisms. To this end she uses a genome browser [1] to inspect the functional feature she is interested in for the reference gene, which gives her a visual overview for the regulatory function. With this picture in mind, the biologist browses genome data looking for similar areas, aimed at discovering if the same regulatory mechanism holds for other genes too. This process is now completely manual; the user totally relies on her own experience to detect similarities, and has no support for browsing and analyzing the matching areas of the genome. To bridge this gap, GenData proposes to automate the detection of similar areas using similarity search techniques, and to analyze the resulting matches using OLAP and mining techniques. The underlying functional flow we envision is sketched in Figure 1 and can be summarized as follows:

1. The biologist uses a genome browser to enter a selection of regions of interest (probe area) and to set a *filter* that delimits the portion of shared genomic data against which the probe will be matched (genome space).

2. A similarity search is performed to identify those areas (target areas) in the genome space that (approximately) match with the probe area.

3. The resulting matches (together with the related meta-data) are loaded on-the- onto a multidimensional cube, made available to biologists for OLAP analyses and mining.

4. Biologists can iteratively change either the probe area, or the search space, or both and trigger a new analysis session.

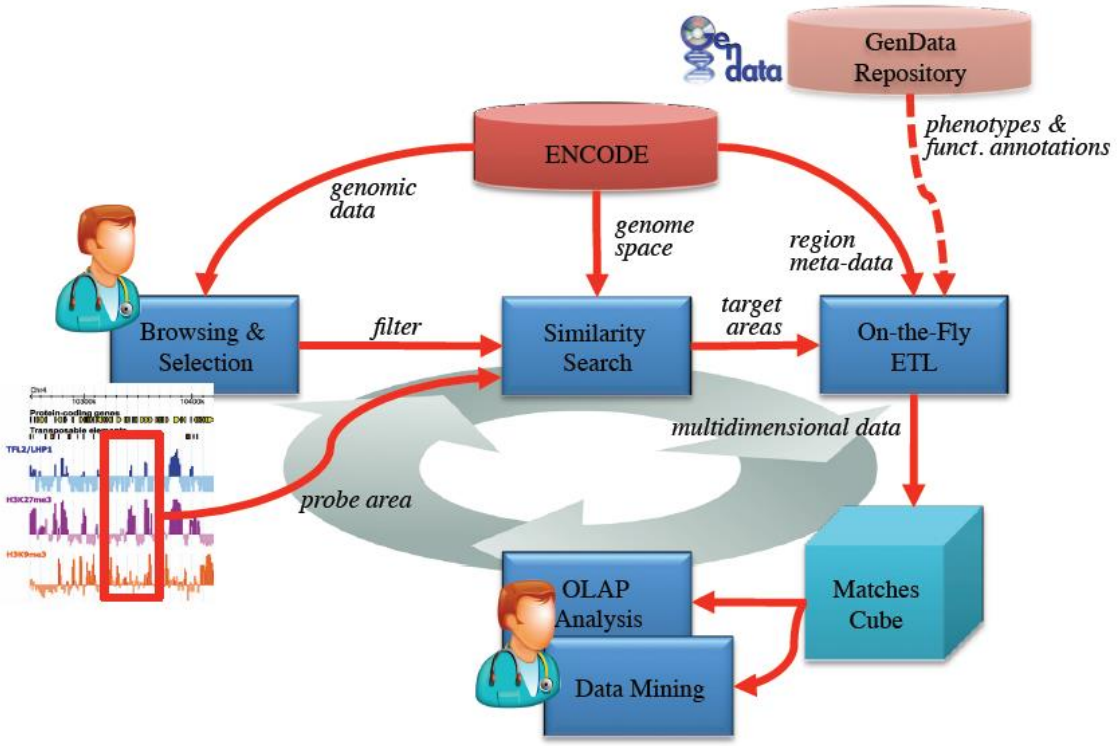


Figure 1. The GOLAM approach

Since OLAP and mining features are mixed, this approach can be properly classified as OLAM [2], so we call it Genome OLAM or, briefly, GOLAM.

First we present the GOLAM framework focusing on its functional architecture and then we provide an OLAP session as an example of multidimensional analysis enabled by the proposed framework. The GOLAM framework will have to be flexible enough to consider two sources of information: on the one hand it will have to be integrated with GenData model as it will provide the similarity search as an analysis service and, on the other, it will have to enrich the similarity search results with ENCODE data and meta-data in order to provide the user a more comprehensive analysis.

ENCODE, the Encyclopedia of DNA Elements, is a public research kicked off and funded by the US National Human Genome Research Institute in 2003 [3]. It is a joint effort of a worldwide consortium comprising research groups which have agreed on standards and open data policies regulating their information. Its goal is to identify and describe the regions of 3 billion base-pair human genome that are important for different kinds of functions.

The section is organized as follows. Section “Related work” provides an overview of related literature regarding OLAP technologies applied to genomic contexts and data mining applications as well. In section “Activity description” we describe in detail the GOLAM framework.

State-of-the-art

Data warehousing and OLAP technologies are quite mature and widely used in business contexts, however not much work exists about their application to genomics. In the literature, a few data warehouse applications for biological data have been proposed. Some of them (e.g., [4, 5, 6, 7]) tackle the challenge of integrating heterogeneous data sources to produce a reconciled database (i.e., an Operational Data Store in the classical data warehousing terminology) but do not provide a subject-oriented multidimensional schema, so we will not further discuss them. In this section we will focus on those applications dealing with OLAP analysis, data warehouse modeling, and data mining techniques to genomics.

Although some publications advise the usage of OLAP as a tool for information discovering (as in [8]), at the best of our knowledge, only in a few cases it has actually been employed. In [9] the authors used OLAP technologies to analyze gene expression data, gathered from tests done on soybean roots, with the purpose of discovering information to develop soybean cultivars resistant to the SCN pest. In [10], important correlations between deletion patterns in the chromosome Y and patient populations have been found through the application of OLAP and data mining techniques. In both studies, the fact of interest differs from ours; while we focus on approximate genome area matches, [9] uses gene expression data |the result of DNA microarray experiments and [10] applies OLAP analysis to the result of clustering performed on deletion patterns in the genome.

Conceptual and logical modeling of data warehouses with respect to genomic and more generally biomedical data is another not well studied topic with only a few existent publications. The authors of [11] propose models for three gene expression data spaces (*sample, annotation,* and *gene expression*) based on star and snowflake schemata. A more exhaustive study of the challenges and requirements of a biomedical multidimensional model is presented in [12], along with a new schema called BioStar. The authors claim that through this schema they can address typical challenges of the genomics domain, such as fast evolving structures, imprecise and incomplete data, etc. This last work in particular differs from ours because we propose a true multidimensional data model based on the GenData data model, while BioStar is a general framework to be used as a way to model various types of biomedical data.

In the context of genomic data, data mining approaches have been successfully applied and the room for genomic data analysis is due to increase in the future. Genomic data mining has been applied to tackle several challenges deriving from the complexities and volumes of genomic data: high-dimensional biological data, problems showing a larger number of candidate predictions compared to the number of observations, noisy data, volume of data to analyze, partially unknown and complex domains. However, over the past years a great number of applications mined interesting information out of genomic databases. In [13], data mining techniques have been successfully employed for inferring prognosis and chemotherapeutic patients' information from their genomic expression signature. Profiling techniques have been tuned for RNA expression in [14], so that expression patterns could be measured on patient samples. Finally, and perhaps most importantly, mining techniques have been applied despite the technical differences of data sets coming from different laboratories enabling therefore investigations hardly achievable with traditional methods [15].

Activity description

In this section we present the GOLAM framework for OLAP analyses and mining of matches between genome regions resulting from various types of biological experiments (such as ChIP-seq). With reference to Figure 1, the main phases and component of the framework can be described as follows:

- *Browsing & Selection*. In this phase, the biologist uses a genome browser [1] to explore the ENCODE data. On the one hand she defines a probe area, i.e., a set of regions belonging to the same cell line but possibly to different experiments on which her interest is focused. On the other, she writes filters to determine the genome space, i.e., a portion of ENCODE data against which the probe area will be compared.
- *Similarity Search.* During this phase, the probe area is compared against the genome space to determine a set of target areas. A target area is a set of regions whose similarity with the probe area is above a given threshold, and is determined as an extension of an alignment between the probe area and the genome space; slightly extending the target area beyond the alignment is useful to enable biologists to give a wider context to their analyses. Like normally done in string matching and string alignment [16, 17], the similarity between two areas is computed based on the pairwise similarities between the regions belonging to the two areas (see [18, 19] for instance). A match is then defined by a pair of matching regions, belonging to the probe and to the target areas respectively, and by their similarity.
- *On-The-Fly ETL*. This phase aims at loading the matches resulting from similarity search, together with the related meta-data, on a multidimensional cube. Each region participating in a match is associated to a wide set of information describing, among the others, its experiment and cell line. This information is currently retrieved from ENCODE; in the near future, it will be possible to extract further information describing the related phenotypes and functional annotations from the GenData repository so as to integrate it with the ENCODE data. Noticeably, the ETL process must be done on-the-fly since both the probe and the genome space are chosen by the biologist at analysis time. Besides, to cope with the potentially huge size of the genome space, we adopt a twofold approach. If (depending on the specific probe area and genome space) the number of target areas resulting from similarity search is tractable, all regions belonging to all target areas are loaded into the cube; otherwise, only the regions of each target area that match with at least one region of the probe area are loaded.
- *OLAP Analysis.* Analysis sessions in biology are inherently exploratory and dynamic, so an OLAP approach is well suited since it provides powerful and flexible tools to explore data from different perspectives and levels of detail. So, at this stage, biologists can analyze region matches using classical OLAP operators such as roll-up, slice-and-dice, etc.
- *Data Mining*. Data about region matches can be also used as a starting point for a number of data mining analyses that we know to be of some interest for biologists. For example, they could be used to determine the pattern of probe features (e.g., regions and chromosomes) that determines the highest number of matches or the one that characterizes at best a specific feature in the genome space (e.g., a pathology listed in the clinical data).

As suggested by Figure 1, the whole workflow is iterative, so as to give biologists the possibility of dynamically changing or refining both the probe area and the genome space, and compare the new results with those obtained at previous iterations.

As mentioned above, the size of the data being processed may be significant. To give an idea of the data volumes, we remark that a typical probe includes up to 100 experiments, each featuring a maximum of 10 regions. However, in the average, biologists work with probes that include about 50-100 regions overall. The expected number of matches obviously depends on several factors, especially on the size of the selected genome space; in a realistic setting, biologists expect to find a number of target areas ranging from 103 to 105, each featuring approximately the same number of regions as the probe area. Figure 2 shows an example of an OLAP session supported by our framework.

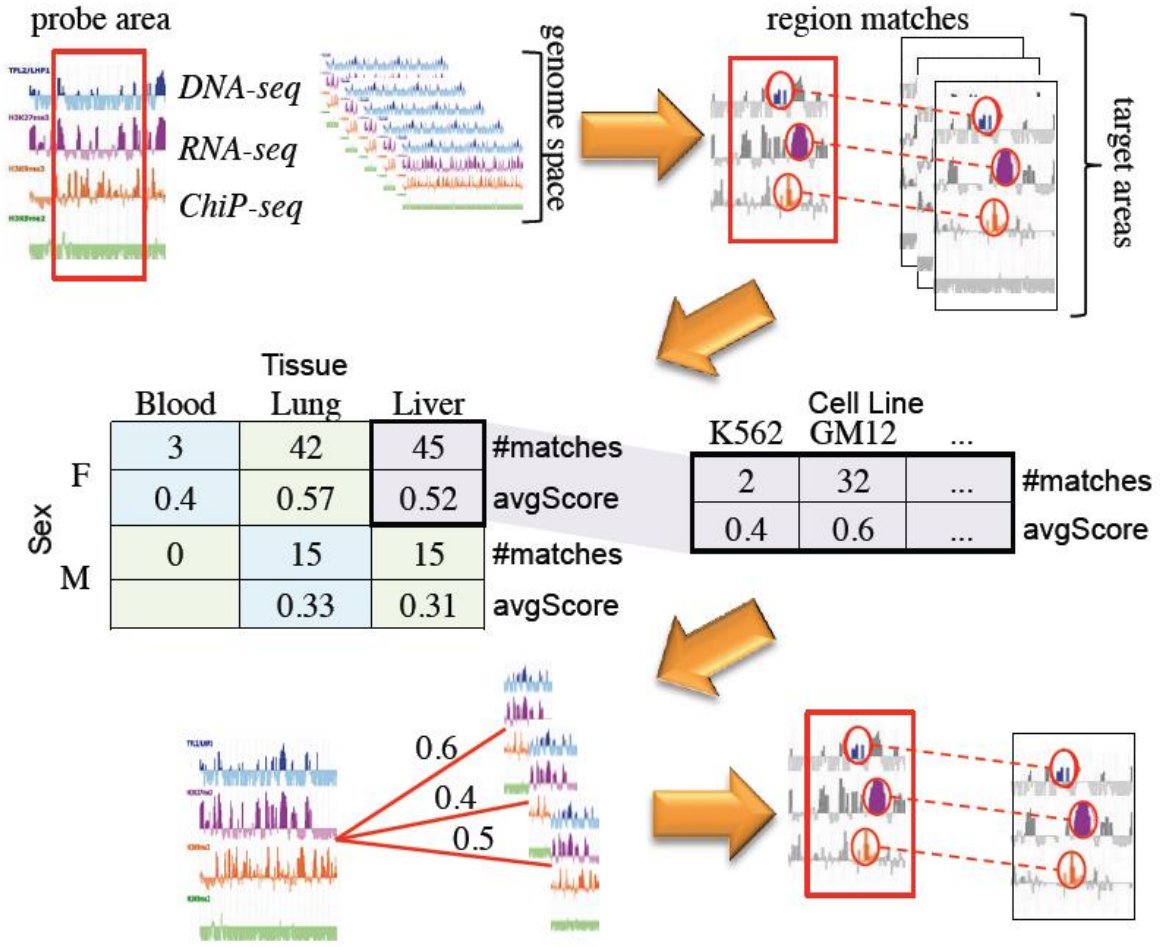


Figure 2: Example of OLAP analysis with GOLAM

The biologist would like to understand if a given set of regions, found in a set of experiments using a common genome browser, could be a characterizing pattern for some pathology. First she selects the area of interest as the probe area, delimits the genome space, and starts a similarity search (Figure 2,

top-left). The probe is compared with the genome space and a set of target areas are detected (Figure 2, top-right). The resulting region matches are then loaded into a multidimensional cube and explored using OLAP techniques. For instance, a pivot table that aggregates matches by Sex and Tissue is created (Figure 2, mid-left); then a drill-down to Cell Line is performed to focus on regions belonging to livers of female patients (Figure 2, mid-right). Finally, a drill-through operation enables the biologist to visualize the matching areas and, eventually, focus her investigation on one of them (Figure 2, bottom-left and bottom-right).

Bibliography

［1] Kent, J., Sugnet, C., Furey, T., Roskin, K., Pringle, T., Zahler, A., Haussler, D.: The human genome browser at UCSC. Genome Res. (12) (2002) 996-1006

[2] Han, J.: OLAP mining: Integration of OLAP with data mining. In: Proc. Working Conf. on Database Semantics, Leysin, Switzerland (1997) 3-20

[3] Raney, B., Cline, M., Rosenbloom, K., Dreszer, T., Learned, K., Barber, G., Meyer, L., Sloan, C., Malladi, V., Roskin, K., Suh, B., Hinrichs, A., Clawson, H., Zweig, A., Kirkup, V., Fujita, P., Rhead, B., Smith, K., Pohl, A., Kuhn, R., Karolchik, D., Haussler, D., Kent, J.: ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res. (39) (2011) D871-D875

[4] Shah, S., Huang, Y., Xu, T., Yuen, M., Ling, J., Ouellette, F.: Atlas{a data warehouse for integrative bioinformatics. BMC Bioinformatics 6(1) (2005) 34-49

[5] Cornell, M., Paton, N., Hedeler, C., Kirby, P., Delneri, D., Hayes, A., Oliver, S.: GIMS: an integrated data storage and analysis environment for genomic and functional data. Yeast 20(15) (2003) 1291-1306

[6] Fischer, M., Thai, Q., Grieb, M., Pleiss, J.: DWARF: a data warehouse system for analyzing protein families. BMC Bioinformatics 7(1) (2006) 495-504

[7] Fellenberg, K., Hauser, N., Brors, B., Hoheisel, J., Vingron, M.: Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis. Bioinformatics 18(3) (2002) 423-433

[8] Stelzer, G., Dalah, I., Stein, T.I., Satanower, Y., Rosen, N., Nativ, N., Oz-Levi, D., Olender, T., Belinky, F., Bahir, I., et al.: In-silico human genomics with GeneCards. Human genomics 5(6) (2011)

[9] Alkharouf, N., Jamison, C., Matthews, B.: Online analytical processing (OLAP): a fast and e ective data mining tool for gene expression databases. BioMed Research International 2005(2) (2005) 181-188

[10] Dzeroski, S., Hristovski, D., Peterlin, B.: Using data mining and OLAP to discover patterns in a database of patients with Y-chromosome deletions. In: Proc. AMIA. (2000) 215-219

[11] Markowitz, V., Topaloglou, T.: Applying data warehouse concepts to gene expression data management. In: Proc. BIBE, Bethesda, Maryland (2001) 65-72

[12] Wang, L., Zhang, A., Ramanathan, M.: BioStar models of clinical and genomic data for biomedical data warehouse design. International Journal of Bioinformatics Research and Applications 1(1) (2005) 63-80

[13] Havaleshko, D., Cho, H., Conaway, M., Owens, C., Hampton, G., Lee, J., Theodorescu, D.: Prediction of drug combination chemosensitivity in human bladder cancer. Molecular Cancer Therapeutics 6(2) (2007) 578-586

[14] Ma, X.J., Patel, R., Wang, X., Salunga, R., Murage, J., Desai, R., Tuggle, T., Wang, W., Chu, S., Stecker, K., Raja, R., Robin, H., Moore, M., Baunoch, D., Sgroi, D., Erlander, M.: Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. Archives of Pathology and Laboratory Medicine 130(4) (2006) 465-473

[15] Lee, J.K., Williams, P.D., Cheon, S.: Data mining in genomics. Clinics in Laboratory Medicine 28(1) (2008) 145-166

[16] Monge, A.E., Elkan, C.: An efficient domain-independent algorithm for detecting approximately duplicate database records. In: Proceedings Workshop on Research Issues on Data Mining and Knowledge Discovery. (1997)

[17] Brown, P.F., Pietra, V.J.D., de Souza, P.V., Lai, J.C., Mercer, R.L.: Class-based n-gram models of natural language. Computational Linguistics 18(4) (1992) 467-479

[18] Chikina, M.D., Troyanskaya, O.G.: An effective statistical evaluation of ChIPseq dataset similarity. Bioinformatics 28(5) (2012) 607-613

[19] Smith, T., Waterman, M.: Identification of common molecular subsequences. Journal of Molecular Biology 147 (1981) 195-197

SECTION 3

GENOMIC METADATA MINING

3.1Discovering frequent correlations from genomic metadata

Elena Baralis, Luca Cagliero, Tania Cerquitelli, Silvia Chiusano, Paolo Garza

Dipartimento di Automatica e Informatica

Politecnico di Torino

Introduction

In the last years an increasingly large amount of genomic data acquired through Next Generation Sequencing experiments has been generated and annotated either automatically or manually by domain experts. Genomic metadata consist of annotations associated with sequences generated by Next Generation Sequencing experiments as well as with data acquired from external sources.
Analyzing the correlations hidden in these huge metadata collections is an interesting yet challenging task. For example, genomic regions are commonly annotated with the name of the matching gene. Discovering significant correlations between genes is appealing for fighting genetic disorders [3,11]. However, the manual inspection of sequence annotations become very time consuming or even unfeasible when the number of analyzed sequences becomes very large. Hence, with the advent of NGS technologies, there is an increasing need for automatic data mining tools able to automatically discover interesting patterns from genomic metadata.

In the context of GenData, our activities focus on applying unsupervised data mining techniques to large datasets consisting of genomic metadata sets with the aim of supporting domain experts in the analysis of large NGS experimental data. More specifically, we applied an established weighted itemset mining algorithm [20] to different types of genomic metadata with the twofold aim at

1. Discovering a worthy subset of patterns representing valuable correlations among multiple genes, and
2. Characterizing genetic disorders by discovering significant correlations between disorders, genes, and authors of scientific papers.

Task (1) entails extracting from gene count datasets combinations of two or more genes for which the corresponding counts are above a given (user-specified) threshold. These patterns are then used by domain experts to support further analyses. The significance of the extracted patterns has been validated using the Gene functional classification provided by DAVID Bioinformatics Database (http://david.abcc.ncifcrf.gov/).

Task (2) entails characterizing genetic disorders/genes with a subset of most authoritative authors and their corresponding publication records. The goal is to simplify the task of literature review on a specific genetic disorder/gene by recommending to domain experts articles written by outstanding scientists according to objective quality indexes on author reputation (e.g., number of author citations, H-index, etc.).

State-of-the-art

A large body of work has been devoted to studying effective and efficient data mining algorithms to transform large amounts of genetic data into useful knowledge [2]. For example, clustering algorithms have been proposed to identify groups of genes that are strongly correlated with each other, but uncorrelated with those of other groups [3,5]. In [6] a step further towards the generation of 3D gene clusters has been made. The authors proposed ParTriCluster, an algorithm that discovers groups of genes behaving similarly across samples and time stamps. The research community also proposed effective classification techniques, i.e. supervised data analysis methods, to correlate gene expression patterns with given classification labels [7,9]. In the context of Gene Expression Data analysis, frequent itemset and association rule mining [10] have been exploited to (i) extract biologically relevant co-expressions among multiple genes [11]; (ii) discover correlations between environmental effects and gene expressions [12]; (iii) profile gene expressions according to a worthwhile subset of gene correlations [13]; (iv) determine biological data duplicates [14].

A parallel effort has also been devoted to developing novel itemset mining algorithms that are able to effectively cope with high-dimensional biological data (e.g. Gene Expression Datasets or Count Datasets containing thousands of genes and tens of samples) [15], [16]. However, to perform itemset and rule mining, gene expression or count values are commonly discretized into a predefined number of bins prior to executing the mining process. Specifically, experts are asked to partition gene expression values into three discrete subsets (e.g., lowly expressed, highly expressed for GEDs, high-count and low-count of Count Datasets). A preliminary attempt to overcome this issue has been made in [28]. The authors proposed to apply a weighted itemset mining algorithm [28] to discover significant correlations among genes from Gene Expression Data. Similar to [20] in GenData we also exploited an established weighted itemset mining algorithm [20] to analyze genomic data. Unlike [28] the activities related to GenData address the analysis of different types of data, i.e., they target the analysis of genomic metadata (e.g., the RNA-Seq/Chip-Seq Count Datasets). Furthermore, to evaluate itemset significance ad hoc quality measures computed on item weights (e.g., the number of author citations) have also been proposed.

Activity description

Our activities entail the following steps: (i) Genomic metadata retrieval and preprocessing, (ii) Discovery of correlations among multiple genes from gene count datasets, and (iii) Characterization of genetic disorders.

Genomic metadata retrieval and preprocessing

Genomic metadata consist of annotations associated with Next Generation Sequencing experiments as well as metadata acquired from external sources. Examples of genomic metadata sets are:

- *Experiment profile datasets*, which collect profile features related to NGS experiments (see

Figure 1),

- *Count datasets*, which consist of sets of records, where each record corresponds to a different gene and it contains the gene counts corresponding to each experiment sample (see Figure 2),
- *Phenotype datasets*, which collect phenotype features related to each sample (see Figure 3),
- *Online Mendelian Inheritance in Man Catalog (http://www.omim.org)*, which contains for each genetic disorder, the list of related genes/proteins, as well as the list of scientific articles ranging over the disorder (see Figure 4).

<**fileURL**, http://www.ieo.pj1/myProcessedfile>
<**assembly_REF**,
http://www.ensamble.org/Homo_sapiens/hs19>
<**expType**, 'ChIP-seq'>,
<**expName**, 'Transcription Factors Binding Sites'>
<**machine**, 'Illumina 2000'>
<**anAlgo**, 'MACS'>
<**pValue**, '20'>
<**disease** = 'myelodysplastic neoplasm'>
<**sex** = 'female'>
…

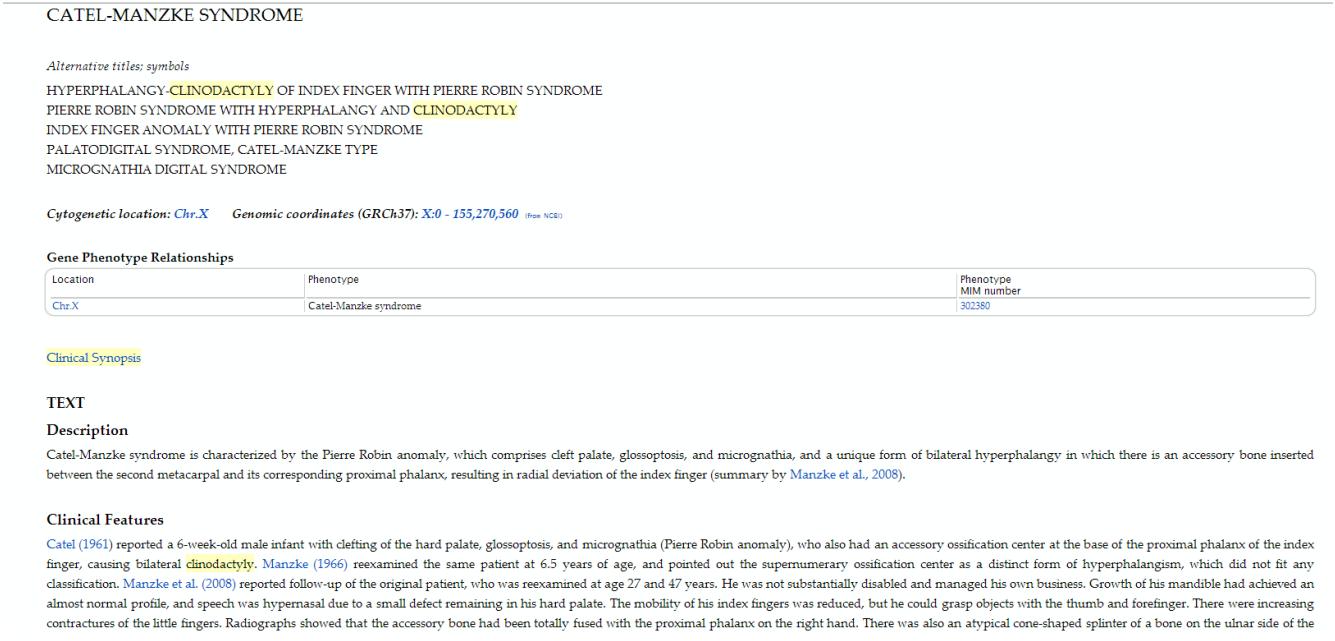
Figure 1. Example of experiment profile dataset

Gene Sample_1_count Sample_2_count Sample_3_count
ENSG00000000003 1354 216 215
ENSG00000000005 712 134 4
ENSG00000000419 450 547 516
…

Figure 2. Example of RNA-Seq count dataset

sample.id num.tech.reps tissue.type gender age race
ERS025098 2 adipose F 73 caucasian
ERS025092 2 adrenal M 60 caucasian
ERS025085 2 brain F 77 caucasian
…

Figure 3. Example of phenotype dataset



CATEL-MANZKE SYNDROME

Alternative titles; symbols

HYPERPHALANGY-CLINODACTYLY OF INDEX FINGER WITH PIERRE ROBIN SYNDROME
PIERRE ROBIN SYNDROME WITH HYPERPHALANGY AND CLINODACTYLY
INDEX FINGER ANOMALY WITH PIERRE ROBIN SYNDROME
PALATODIGITAL SYNDROME, CATEL-MANZKE TYPE
MICROGNATHIA DIGITAL SYNDROME

Cytogenetic location: *Chr.X* ***Genomic coordinates (GRCh37):*** *X:0 - 155,270,560* (from NCBI)

Gene Phenotype Relationships

Location	Phenotype	Phenotype MIM number
Chr.X	Catel-Manzke syndrome	302380

Clinical Synopsis

TEXT

Description

Catel-Manzke syndrome is characterized by the Pierre Robin anomaly, which comprises cleft palate, glossoptosis, and micrognathia, and a unique form of bilateral hyperphalangy in which there is an accessory bone inserted between the second metacarpal and its corresponding proximal phalanx, resulting in radial deviation of the index finger (summary by Manzke et al., 2008).

Clinical Features

Catel (1961) reported a 6-week-old male infant with clefting of the hard palate, glossoptosis, and micrognathia (Pierre Robin anomaly), who also had an accessory ossification center at the base of the proximal phalanx of the index finger, causing bilateral clinodactyly. Manzke (1966) reexamined the same patient at 6.5 years of age, and pointed out the supernumerary ossification center as a distinct form of hyperphalangism, which did not fit any classification. Manzke et al. (2008) reported follow-up of the original patient, who was reexamined at age 27 and 47 years. He was not substantially disabled and managed his own business. Growth of his mandible had achieved an almost normal profile, and speech was hypernasal due to a small defect remaining in his hard palate. The mobility of his index fingers was reduced, but he could grasp objects with the thumb and forefinger. There were increasing contractures of the little fingers. Radiographs showed that the accessory bone had been totally fused with the proximal phalanx on the right hand. There was also an atypical cone-shaped splinter of a bone on the ulnar side of the

Figure 4. Online Mendelian Inheritance in Man Catalog

We retrieved Chip-seq Count Datasets from the Genomic Data Model (GDM) using the following query written in the Genomic Query Language (GQL):

```
G = SELECT (PROVIDER=='UCSC' OR PROVIDER=='RefSeq') ANNOTATIONS;

S = SELECT (EXPTYPE=='Chip-Seq' AND CELL='K562') ENCODE_BROAD;

M = MAP(AVG(Signal)) G S;
```

where in the first statement the annotations related to two specific gene types are selected, in the second statement the Chip-Seq experiments related to a specific cell (e.g., K562) are considered, while in the last statement the selected experiments are mapped to the genes of interest.

Furthermore, from the Online Mendelian Inheritance in Man Catalog (http://www.omim.org) we retrieved the list of currently known genetic disorders, the corresponding genes/proteins, as well as the state-of-the-art scientific articles ranging over each genetic disorder together with the names of the corresponding authors .

Metadata sets are preprocessed to make them suitable for the subsequent weighted itemset mining process. Notably, gene count values require no discretization, because, as discussed in the following paragraph, gene count values are considered as item weights.

The analysis of the genomic metadata entails two specific tasks, whose descriptions are given below.

Discovery of correlations among multiple genes from gene count datasets

This task entails extracting from gene count datasets frequent weighted itemsets, which represent combinations of two or more genes for which the corresponding counts are above a given (user-specified) threshold. These patterns are particularly interesting in genomics because they represent extensions of regulatory network connections [11,13]. More specifically, state-of-the-art approaches focus on discovering pairwise correlations among genes, whereas itemset mining approaches allow us to discover and analyze underlying high-order correlations among genes [28].

Traditional itemset mining algorithms (e.g., Apriori, FP-Growth [2]) show some limitations while coping with continuous values, because they are unlikely to occur many times in the source data. Hence, a data discretization process is commonly applied to continuous item values prior to itemset mining. However, the discretization step applied to gene count datasets could bias the quality of the mining result because experts have to assume a reliable gene count distribution. Consequently, biologists and physicians often analyze and compare the results produced by different discretization methods [13], [17]. To overcome this issue, we adopted a more effective approach [28] to discovering itemsets from genomic metadata while avoiding the discretization step. Rather than discretizing gene count values before executing the itemset mining process, we represent per-sample gene count values as item weights. In other words, we consider genomic metadata as weighted datasets [18] for which expression values are mapped to item (gene) occurrences within each sample. Then, weighted itemsets are extracted from weighted data. Since item weights can be continuous, discovering weighted itemsets instead of traditional (not weighted) ones prevents experts from discretizing GEDs before analyzing them. This approach improves the effectiveness of the knowledge discovery process.

Several weighted itemset mining algorithms (e.g., [18], [19]) have been proposed to consider item weights during the itemset extraction process. In this context, we adopted the weighted itemset mining strategy that has recently been proposed in [20]. To demonstrate the significance of the extracted patterns, we validated the mining result using the Gene functional classification provided by DAVID Bioinformatics Database (http://david.abcc.ncifcrf.gov/). Specifically, we compared the functional groups identified by DAVID with those produced by the itemset miner.

Characterization of genetic disorders

The Online Mendelian Inheritance in Man Catalog collects, for each genetic disorder, the list of mostly related genes, as well as the most authoritative articles ranging over the disease.

This task entails characterizing genetic disorders/genes with a subset of most authoritative authors. The goal is to simplify the exploration of the state-of-the-art publications by suggesting to analysts the names of the most outstanding experts on a specific disorder/gene as well as their corresponding publication records.

To address this task, we used the same weighted itemset mining strategy adopted to accomplish Task (1). Unlike the former case, each transaction corresponds to a distinct scientific paper, while items represent either scientific paper authors or genetic disorders or genes. The relevance weights assigned to each author correspond to an objective quality index related to author reputation (e.g., number of author citations, H-index, etc.). In such a way, itemsets containing combinations of outstanding authors have, on average, an higher support value.

We focused on itemsets containing:

(i) a disorder and a set of authors, or
(ii) a gene and a set of authors.

By ranking itemsets of type (i) or (ii) in order of decreasing support we selected top interesting itemsets. These itemsets can be used to recommend to experts the reading of the articles written by most outstanding scientists and ranging over a specific disorder or gene, respectively.

Bibliography

[1] D. Clark and N. Pazdernik, Molecular Biology: Understanding the Genetic Revolution. Elsevier Science, 2012.

[2] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Addison Wesley, 2005.

[3] A. Ben-Dor and Z. Yakhini, "Clustering gene expression patterns, in Proceedings of the third annual international conference on Computational molecular biology, ser. RECOMB '99, 1999, pp. 33-42.

[4] Y. Cheng and G. M. Church, Biclustering of expression data, in ISMB, 2000, pp. 93-103.

[5] W.-H. Au, K. C. C. Chan, A. K. C. Wong, and Y. Wang, "Attribute clustering for grouping, selection, and classification of gene expression data," IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 2, no. 2, pp. 83-101, Apr. 2005.

[6] R. B. Arafujo, G. H. T. Ferreira, G. H. Orair, W. Meira, R. A. C. Ferreira, D. O. G. Neto, and M. J. Zaki, "The partricluster algorithm for gene expression analysis," Int. J. Parallel Program., vol. 36, no. 2, pp. 226-249, Apr. 2008.

[7] Y. Lu and J. Han, Cancer classification using gene expression data, Inf. Syst., vol. 28, no. 4, pp. 243-268, Jun. 2003.

[8] M. Khashei, A. Zeinal Hamadani, and M. Bijari, A fuzzy intelligent approach to the classification problem in gene expression data analysis, Know.-Based Syst., vol. 27, pp. 465-474, Mar. 2012.

[9] M. A. Iwen, W. Lang, and J. M. Patel, "Scalable rule-based gene expression data classification," in ICDE, 2008, pp. 1062-1071.

[10] R. Agrawal, T. Imielinski, and Swami, Mining association rules between sets of items in large databases, in ACM SIGMOD 1993, 1993, pp. 207-216.

[11] C. Creighton and S. Hanash, Mining gene expression databases for association rules." Bioinformatics, vol. 19, no. 1, pp. 79-86, 2003.

[12] R. Martinez, N. Pasquier, and C. Pasquier, Computational intelligence methods for bioinformatics and biostatistics, F. Masulli, R. Tagliaferri, and G. M. Verkhivker, Eds., 2009, ch. Mining Association Rule Bases from Integrated Genomic Data and Annotations, pp. 78-90.

[13] P. C. Saez, M. Chagoyen, A. Rodriguez, O. Trelles, J. Carazo, and A. P. Montano, Integrated analysis of gene expression by association rules discovery, BMC Bioinformatics, vol. 7, no. 1, pp. 54, 2006.

[14] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," IEEE Trans. Knowl. Data Eng., vol. 19, no. 1, pp. 1-16, 2007.

[15] F. Pan, G. Cong, A. K. H. Tung, J. Yang, and M. J. Zaki, "Carpenter: finding closed patterns in long biological datasets,in Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, ser. KDD '03. ACM, 2003, pp. 637–642.

[16] G. Cong, A. K. H. Tung, X. Xu, F. Pan, and J. Yang, "Farmer: finding interesting rule groups in microarray datasets, in Proceedings of the 2004 ACM SIGMOD international conference on management of data, ser. SIGMOD '04. ACM, 2004, pp. 143-154.

[17] V. Belcastro, V. Siciliano, F. Gregoretti, P. Mithbaokar, G. Dharmalingam, S. Berlingieri, F. Iorio, G. Oliva, R. Polishchuck, N. Brunetti-Pierri, and D. di Bernardo, Transcriptional gene network conference from a massive dataset elucidates transcriptome organization and gene function." Nucleic acids research, vol. 39, no. 20, pp. 8677-8688, Nov. 2011.

[18] W. Wang, J. Yang, and P. S. Yu, "Efficient mining of weighted association rules (WAR)," in Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'00, 2000, pp. 270-274.

[19] K. Sun and F. Bai, "Mining weighted association rules without preassigned weights," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 4, pp. 489-495, 2008.

[20] L. Cagliero and P. Garza, Infrequent weighted itemset mining using frequent pattern growth," IEEE Transactions on Knowledge and Data Engineering, vol. 99, no. PrePrints, p. 1, 2013.

[21] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases, in VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, J. B. Bocca, M. Jarke, and C. Zaniolo, Eds. Morgan Kaufmann, 1994, pp. 487-499.

[22] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, Discovering frequent closed itemsets for association rules," in Proceedings of the 7th International Conference on Database Theory, ser. ICDT '99. London, UK, UK: Springer-Verlag, 1999, pp. 398–416. [Online]. Available: http://dl.acm.org/citation.cfm?id=645503.656256

[23] J. Roberto and J. Bayardo, "Efficiently mining long patterns from databases, in SIGMOD 1998, L. M. Haas and A. Tiwary, Eds., 1998, pp. 85-93.

[24] M. Mampaey, N. Tatti, and J. Vreeken, “Tell me what I need to know: Succinctly summarizing data with itemsets,” in Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2011.

[25] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 2000, pp. 1-12.

[26] E. J. Yeoh, M. E. Ross, S. A. Shurtleff, K. W. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, and Cheng, “Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, Cancer Cell, vol. 1, no. 2, pp. 133-143, 2002.

[27] K. Ishii, T. Washio, T. Uechi, M. Yoshihama, N. Kenmochi, and M. Tomita, Characteristics and clustering of human ribosomal protein genes, BMC Genomics, vol. 7, no. 1, pp. 1-16, 2006.

[28] E. Baralis, L. Cagliero, T. Cerquitelli, S. Chiusano, P. Garza, Frequent weighted itemset mining from gene expression data, Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on , vol., no., pp.1,4, 10-13 Nov. 2013 doi: 10.1109/BIBE.2013.6701681

3.2Extracting salient features from breast cancer patients with the GenData Genome Query Language and the GELA software

Paolo Atzeni, Emanuel Weitschek

Dipartimento di Ingegneria - Sezione di Informatica e Automazione

Università Roma Tre

Introduction

Breast cancer is the most commonly diagnosed cancer in women and is their second leading cause of death. In this work we extract salient patient features from The Cancer Genome Atlas, the largest repository of both clinical and genomic cancer data, by using the GenData Genome Query Language. By means of the analysis of these features we are able to automatically classify the patient sample into tumoral and control cases with the aid of the Gene Expression Logic Analyzer (GELA) software, which identifies classification formulas ("if then rules") for each class present in the data set. The analysis is possible thanks to the integration of Clinical, DNA Sequencing, Mutations, RNA Sequencing, and DNA Methylation data performed with the GenData model.

State-of-the-art

The Cancer Genome Atlas (TCGA) database [1] contains a comprehensive genomic characterization and analysis of more than 20 cancer type tissues to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing. The project began in 2006 and is a coordinated joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) both of the National Institute of Health (NIH). The main aim of TCGA is the improvement of the ability to diagnose, treat and prevent cancer. The cancer data is available through the TCGA Data Portal, a free platform to search, download, and analyze data sets containing clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes. A comprehensive documentation is available at https://wiki.nci.nih.gov/display/TCGA.

Following experimental data types are released:

- Tissue pathology data
- Images
- Pathology reports
- Copy-number alterations for non-genetic platforms
- Epigenetic data
- Data summaries, such as genotype frequencies
- Clinical data (biotab and xml)
- DNA Sequencing (whole genome, whole exome, mutations)
- DNA Methylation
- Gene expression data (miRNA, mRNA, Total RNA, Micro- and Protein-arrays)
- Copy numbers

Activity description

For the GenData project we focus on Clinical data, Mutations, RNA-Seq, and DNA-Methylation. Following processing and analysis steps are performed:

1. Data extraction from The Cancer Genome Atlas (TCGA) with focus on the breast cancer pathology (900 patients and controls)

2. Transformation to the GenData format with the tcga2gendata Java software (under release)

3. Patient features data matrix extraction through the Genome Query Language (GQL)

Patient	$Feature_1$	$\cdots$	$Feature_m$	Class
$sample_1$	$value_{(1,1)}$	$\cdots$	$value_{(1,m)}$	BRC
$sample_2$	$value_{(2,1)}$	$\cdots$	$value_{(2,m)}$	BRC
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$sample_n$	$value_{(n,1)}$	$\cdots$	$value_{(n,m)}$	Control

Figure 1. Matrix example

4. Knowledge extraction with Gene Expression Logic Analyzer (GELA) and supervised/unsupervised machine learning techniques: the aim is to extract salient features (clinical variables, genes, mutations, etc.) that enable to automatically classify the samples in their belonging class (breast cancer or control).

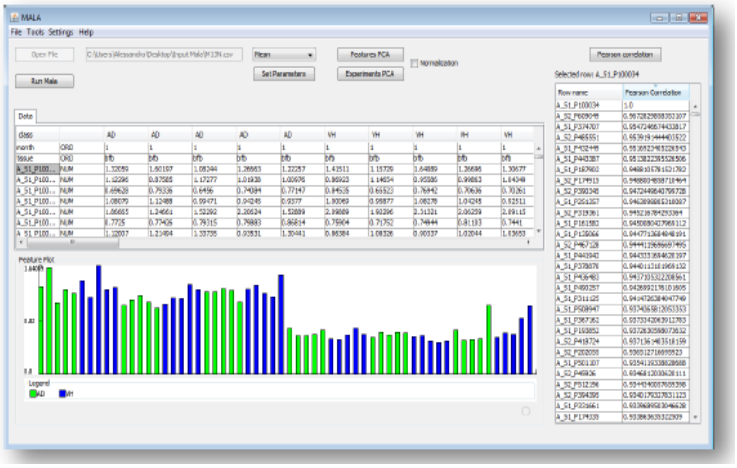


Figure 2. The GELA software

Gene Expression Logic Analyzer (GELA) is a clustering and rule based supervised classification software, particularly engineered for gene expression analysis. The aims of GELA are to cluster the gene expression profiles in order to discover similar genes and to classify the experimental samples. GELA converts the numeric gene expression profiles to discrete, and implements a method named Discrete Cluster Analysis (DCA) to cluster the genes based on the discretization step. Moreover, it uses an integer programming method for selecting the characteristic genes for each class, and adopts the lsquare method for computing the logic classification formulas (”if then rules”). GELA also integrates standard statistical methods for gene expression profiles data analysis: Principal Component Analysis (PCA) to group similar genes and experiments, and Pearson Correlation Analysis to find a list of correlated genes to a selected gene. It is an evolution of the MALA tool [2] available at dmb.iasi.cnr.it/mala.php. GELA is going to support the classification and clustering of RNA-Seq, DNA Methylation, and Mutations data.

Bibliography

[1] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. **N**ature 455, 1061–1068 (2008)

[2] Emanuel Weitschek, Giovanni Felici, Paola Bertolazzi. MALA: A Microarray clustering and classification software. DEXA - Database and Expert Systems Applications, 201-205 (2012)

SECTION 4

4. ANALYSIS OF OTHER DATA TYPES

4.1 Pattern mining from clinical data

Pierangelo Veltri

Dipartimento di Scienze Mediche e Chirurgiche

Università degli Studi "Magna Grecia" di Catanzaro

Sergio Greco, Giuseppe Tradigo

Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica

Università della Calabria

Introduction

Genomic medicine focus on understanding of the molecular basis of disease [1] and on translating this knowledge into clinical practice. The identification of pathologies related to genetic alterations in fact, has potential diagnostic, prognostic and therapeutic benefits.[2].
Applications of genomic medicine are different: individualized therapy base on own genetics [3] which improves quality of patient care, pharmacogenomics, genetic markers identification to use in screening analysis, linkage studies which correlate phenotype with cromosomic markers, drug discovery and development. In the genomic era, transformation of clinical practice is best exemplified in cancer care: oncologists include BCRA1/BCRA2 test in breast and ovarian cancer investigation [4].
Complexity of diagnostics and genome sequencing analysis, is due to generation and management of vast amounts of data of different types and complexities.
Moreover, the sharing of genomic data and then pattern and information extraction, is very difficult because submission of data in databases of interest comes from a large number of scientists from different communities. As result, there are often errors in annotations, redundancies or use of different nomenclature [5]. This limit has improved with ontologies and standardized medical languages development. Ontologies define controlled vocabulary for the sharing of information in a domain [6] by hierarchical definitions of important concepts.
The integration of these highly heterogeneous elements, allows the mining of pattern for the creation of new knowledge to apply in clinical practice. Extraction of relevant clinical information, can support the understanding of wrong gene expression and related disease. An integration system model of both clinical and genomic data is studied. It has an ontologic base linking wrong gene expression with related disease.

State-of-the-art

Genomic medicine is a data intensive field that attempts to explain the molecular basis of disease and to translate this knowledge into clinical practice by its integration with other diagnostic data. It uses information from genomes and their derivatives (e.g. RNA and proteins) to guide medical decision making in personalized medicine [7].

Different genomic and clinical database are available, e.g. Genbank, EntrezGene and dbSNP, the NCBI database on gene variations. Furthermore, a database of risk genes, OMIM (Online Mendelian Inheritance in Man), reveals an increasing number of genomic factors that should be useful in risk assessment in prediction models [8].
Standardized vocabularies are required to ensure a semantic integration of heterogeneous data and interoperability between sources. UMLS (Unified Medical Language System), SNOMED and HL7 are example of existing standards; ontologies resolves instead semantic inconsistencies. Many ontologies each relating to specific biomedical domains, have been developed [9]. An ontology contains a representation of all the concepts present in a domain and all the relationships between them. Besides, because ontologies can be grown over time as new data are available, new links are created and new knowledge assimilated. Clinical Bioinformatics Ontology (CBO) is an ontology example which provides to identify concepts related to clinically significant findings. Gene Ontology (GO), is a taxonomy used to describe normal molecular function of proteins, cellular component in which they operate and biological processes they are involved [10]. GO dataset are available in flat files, XML and MySQL format. Moreover, several methods to adapt a general ontology for specific application are present in literature, such statistical methods to adapt OBO Foundry Disease Ontology (DO) for the identification of gene-disease association in [11] and its use as a controlled vocabulary to annotate genome in terms of disease [12]. Disease Ontology has been also incorporated into open source tools (e.g. Geneanswer) to this purpose.
Different database integration methods are evaluated and compared in [1]. Data warehouse (e.g. UCSC Genome Browser, EnsEMBL Database Project) or database federation approaches allows architecture integration. While data warehousing is the consolidation of all data into a single database, in a database federation, data origins remains autonomy and distributed along the network: data models are preserved and linked by mapping. Wrappers are used to interface with data sources. Other integration processes are database federations with mediated schemas [13] and peer management system [14]. If data warehouse are appropriate in cases where performance, local control and privacy are key issues, mediated schema approach facilitates general queries and integration of different types of data. The genotype to phenotype correlation, is an example of these vague questions.

Activity description

A clinical-genomic database definition, is desirable to increase knowledge base for physicians. To extract useful information, an integration step is required. Integration is fundamental about querying across different data sources and integrated clinical–genomic architecture is essential to allows a wide information availability (open data type) to physicians. Possible applications of this model, is forecasting cancer risk by data analysis or linkage studies which allow to localize susceptibility genes involved in phenotype etiology.

The aim of the activity, is to realized an integrated system linking medical and genomic data to support clinical practice according to Italian electronic health record guide lines. Centralized database also contains biological databank references and genetic data obtained by counseling activity. Different types of data are involved: data from diagnostics, imaging, anamnesis and more from functional genomic and proteomic investigation. Web based information is also used. In this wide scenery of sources, it must to operate both with structured data (e.g. laboratory analysis) and unstructured data (e.g. clinical records annotation). Furthermore, information extraction and data exchange methods are studies and following procedural steps are involved in model realization:

- Clinical data study;

- Genomic data study;
- Databanks and ontologies investigation;
- Data exchange and integration system survey.

For useful feature extraction, complex data as biomedical signals or medical images, must be processed before. Each source extracts genomic and/or clinical information from own clinical records generating metadata, so creation of Meta Health Records (MHR) database (called MetaInfo), represents the first realization step. In MetaInfo, a few aggregate useful information will be gathered to allow a macroscopic research. Useful information are: (i) Personal (Surname, Name, BirthDate, Sex) and (ii) Clinical (Diagnosis, DiagnosisDate, Stage, Therapy, Results, DeathDate). Generation of MHR is from medical records of each source and share a common identification pattern. In order to overcome the problem of different terminology in sources, extraction process is sustained by controlled medical languages and ontologies (e.g. UMLS or Clinical Bioinformatics Ontology) which support integration. Ontologies include elements such as concepts, relationships between concepts, and their properties. They may also provide the capability to generate logical inferences by defining rules and axioms. System collects extracted data and integrate them in a new clinical/genetic queryable database.

Architecture's intents are: (i) information space wider, (ii) information increase for clinicians and insiders and (iii) possibility to enter information in exchange for information. Identified actors for the system are pharmacologists, biologists, laboratory technicians and clinicians to populate database. Moreover, patients and generic technical–administrative users can access database with different views.

Possible queries that could benefit from a ClinicalGenomic database are also identified, such as identification of genes pattern with a better survival rate in a specific cancer or stratification of a patient group with the same disease and molecular profile, on the basis of their response to specific drug treatement. To realized model, the identification of textual usable data from electronic record is a key point: these information must be few but fundamentals at the same time. Final dataset (clinical and genomic) includes all features highly related with interest domain (e.g. cardiovascular diseases). Furthermore, this knowledge extraction process from large health dataset, involves a significant risk for privacy breach, thus data anonymization is necessary. In fact, as technologies evolve and data proliferate, it's more difficult to protect genetic information.

Prediction models could be also applied, such as machine learning feature selection and pattern classification algorithms for molecular cancer classification or phenotype prediction. These methods have statistical and clinical relevance in cancer detection for a variety of tumor types by high dimensional data analysis [15],because they allow to derive unknown features of patients by known features of similar cases. To this purpose, an improvement of cancer risk assessment models based on genetic tree predictive algorithms is integrated.

A possible application of the presented integration system, is represented by control of genetic counseling and data collection. Genetic counseling is process whereby individuals are informed about genetic issues related to themselves and/or their family and are counseled about the potential consequences of the information [16]. It provides different settings such as predictive testing to an asymptomatic person at risk of cancer disorder, counseling in relation to prenatal screening and testing, diagnosis of a rare genetic syndrome. Hence, its main advantage resides in improvement of prevention protocols.

In our activity, we started to recruit data from counseling (in anonymous way) performed in an Italian south region, to correlate clinical and genomic data and to create an instance of a clinical-genomic database. Part of the preliminary results of this study has been reported in [17].

Bibliography

[1] B. Louie et al, Data integration and genomic medicine, Journal of Biomedical Informatics 40 (2007) 5–16

[2] Johanne Tremblay, Pavel Hamet, Role of genomics on the path to personalized medicine, Metabolism clinical and experimental 62 (2013) s2 – s 5

[3] Sander C, Genomic medicine and the future of health care. Science 2000;287(5460):1977–8.

[4] Louise Bouchard, I. Blancquaert, F. Eisinger, W.D. Foulkes, G. Evans,H. Sobol, C. Julian-Reynier, Prevention and genetic testing for breast cancer: variations in medical decisions, Social Science & Medicine 58 (2004) 1085–1096

[5] James D. Cavalcoli, Genomic and Proteomic Databases: Large-Scale Analysis and Integration of Data, TCM Vol. 11, No. 2, 2001

[6] Furkh Zeshan, Radziah Mohamada , Medical Ontology in the Dynamic Healthcare Environment , Procedia Computer Science 10 (2012) 340 – 348

[7] Geoffrey S. Ginsburgemail, Huntington F. Willard, Genomic and personalized medicine: foundations and application, Translational Research Volume 154, Issue 6 , Pages 277-287, December 2009

[8] Bianco A.M., Marcuzzi A., Zanin V., Girardelli M.,Vuch J., Database tools in genetic diseases research , Genomics 101 (2013) 75–85

[9] Anita Burgun, Desiderata for domain reference ontologies in biomedicine, Journal of Biomedical Informatics 39 (2006) 307–313

[10] Ashburner et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium , Nat Genet. 2000 May ; 25(1): 25–29. doi:10.1038/75556.

[11] Pan Du et al. , From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations Vol. 25 ISMB 2009, pages i63–i68 doi:10.1093/bioinformatics/btp193

[12] Osborne et al., Annotating the human genome with Disease Ontology, BMC Genomics 2009, 10(Suppl 1):S6

[13] Chawathe S, The TSIMMIS Project: integration of heterogeneous information sources. In: Proceedings of IPSJ conference. 1994. p.7–18

[14] Gribble S, et al., What can databases do for peer-to-peer? WebDB Workshop on Databases and the Web, June 2001

[15] Habtom W. Ressom et al. Front Biosci, Classification algorithms for phenotype prediction in genomics and proteomics, Jan 1, 2008;13:691-708

[16] M Delatycki, Genetic Counseling , Brenner's Encyclopedia of Genetics, 2nd edition, Volume 3 doi:10.1016/B978-0-12-374984-0.00612-4

[17] An Architecture for integrating Genetic and Clinical data Giuseppe Tradigo, Claudia Veneziano, Sergio Greco, Pierangelo Veltri. IEEE ICCS 2014

4.2 Intensional Query Answers for clinical data

Mirjana Mazuran, Elisa Quintarelli, Letizia Tanca

Dipartimento di Elettronica, Informazione e Bioingegneria

Politecnico di Milano

Introduction

The application of data mining techniques to extract useful knowledge from datasets has been widely applied in the literature. By mining frequent patterns from repositories, the users can be provided with synthetic, albeit approximate, information on dataset contents. Indeed, as a consequence of the richness of information we have at our disposal, when a user faces the (frequently enormous amount of) available data she often does not know its features and needs to figure out "where to start from" to make sense of it, in order to distinguish between useful information and the "noise" generated by all the uninteresting data around it.

Giving a summarized view of a vast integrated biological or clinical dataset can thus increase the user comprehension, cutting the costs of the lengthy interactions needed to find the most interesting features of a given (large) dataset. The idea here is to collect frequent patterns, in the form of *association rules*, to be stored and maintained as condensed knowledge about the dataset, and used to support query sense-making and refinement. Henceforth this condensed knowledge will be called *intensional, approximate knowledge.*

We imagine two possibilities of using intensional approximate knowledge:

1. the user queries the initial clinical repository, but she can instruct the system to provide also an approximate answer, extracted from the intensional repository;
2. the user directly queries the approximate, intensional base containing the association rules that have been previously mined.

State-of-the-art

In Frege and Russell's studies on the foundations of Mathematics, the term *intension* [1-5] suggests the idea of denoting sets by means of the properties shared by their objects rather than by exhibiting them. A logical formula has both an extension and an intension, its intension being the formula itself, while its extension is the set of objects that satisfy it. Thus, the intensional definition of a set of (mathematical) objects specifies all the properties of those objects, that is, the necessary and sufficient conditions for an element to belong to that set. On the other hand, the extensional definition of a set lists all (and only) the objects of the set. As acknowledged by Aristotle himself in his work about "categories", the real problem is that very seldom an intensional definition is possible, since finding a minimal and complete set of properties (features) that precisely characterize a collection of data is easier in mathematics than in real life. A 100% accurate description of the world is impossible: however an approximate description – i.e. with an accuracy lower than 100% - of a collection of data is still possible.

The problem of providing (exact) intensional answers by means of integrity constraints has been initially addressed in [3] in the relational databases context. Some works [6,7] have proposed to extract

approximate intensional knowledge by means of data mining techniques, and to use it to intensionally describe sets in an approximate way with the objective of storing the intensional descriptions (instance patterns) for later querying. These summarization techniques are proposed for relational or XML data and address also the objective of performance, achieved by querying the stored patterns instead of using classical indexing techniques.

Activity description

Inexperienced users need the support of knowledge-retrieval system able to search, retrieve and "highlight" quick, though approximate information starting from simple inputs. We have defined an appropriate template for writing queries over intensional, approximate knowledge, so that the GUI guides the user in composing queries to be applied to the previously mined association rules. The templates are very specific and allow the user to compose such queries as "Find the most relevant information about tumors in a given Italian region". This kind of query is helpful especially in wide-range analyses involving a huge amount of data and to discover frequent correlations of values, which can be used later to write more specific queries. Also, as another example, we could think about a regional health director interested in "Retrieving the most wards in the hospitals of a given region", in order to be able to provide these wards with proper equipment; yet another example could involve a doctor looking for "General information about births in Italy".

In the first scenario, the GenData user composes an SQL query but (if requested) she will receive first the intensional, approximate answer. The IQ4GenData module will parse the input SQL query and rewrite it in order to apply it to the mined association-rule base. Then, if requested, the extensional answer – that is, the answer containing the actual data, will also be provided.

Example 1: "Retrieve all information about births in Italy"

Input (classical) SQL query:

```
SELECT b.*
FROM birth AS b
WHERE b.state=" Italy"
```

While in the textual document, for readability reasons, the result of a query will be shown as a set of implications, the mined association rules are currently stored by IQ4GenData in a relational database having the following schema:

```
Nodes (node_id, table, attribute, value)
Rules (rule_id, support, confidence, num_antec, num_cons)
Antec_cons (ant_cons_id, node_id, rule_id, ant_cons)
```

The following piece of SQL code is the automatically rewritten query, to be applied to the association-Rule database presented above and retrieves every rule containing a reference to birth in Italy.

```
SELECT rules.rule_id, support, confidence
FROM rules
INNER JOIN (SELECT rule_id
          FROM antec_cons AS ac, nodes AS n
          WHERE ac.node_id=n.node_id AND n.table=" birth"  AND
                        n.attribute=" state"  AND
          n.value=" Italy"  AND ac.ant_cons=' A' ) AS a1 ON
                        rules.rule_id=a1.rule_id
INNER JOIN (SELECT rule_id
          FROM antec_cons AS ac, nodes AS n
          WHERE ac.node_id=n.node_id AND n.table=" birth"  AND ac.ant_cons=' C' )
                        AS a2 ON
          rules.rule_id=a2.rule_id
```

The intensional output of the IQ4GenData module will be a set of association rules that can be represented either in a relational format, or as an XML document, or in a more intuitive/graphical format. An example of possible result is:

```
1. birth.sex= Male birth.state=italy  => birth.pregnancy_term=full term
2. birth.state=italy  => birth.ward=maternity ward
3. birth.pregnancy_term=full term birth.state=italy  =>
   birth.weight=[2600,4200]
```

These three rules allow the user to discover that:

1. In Italy, births of male babies are mostly related to full-term pregnancies
2. Births in Italy mostly take place in maternity wards.
3. The weight of babies born in Italy is mostly within the interval [2600,4200].

In general, the classes of SQL queries that can be managed by the IQ4GenData module are:

1. Simple or complex restrictions queries: used to impose a simple, or complex (containing AND and OR operators), condition on the value of an attribute
2. Count queries: used to count the number of tuples having a specific content
3. Top-k queries: used to select the best k answers satisfying a counting and grouping condition.

The second querying possibility, i.e. the case when the user directly queries the approximate, intensional base containing the association rules, is supported by the WHAT ABOUT statement:

```
WHAT ABOUT Ward
WHERE Region = Campania AND N_patients_per_month>= 1000
WITH CONFIDENCE 0.9
```

This will trigger the intensional knowledge system to return every association rule containing:

- attributes from the relations in the WHAT ABOUT list (in this example,Ward)
- the elements that satisfy the conditions (in our case, Region = Campania AND N_patients_per_month>= 1000)

- having confidence greater or equal to the stated value (for example 0.9)

The next step for intensional query answering in the GenData project will be the application of IQ4GenData to handle genomic annotations and to support a full process of data exploration.

Bibliography

[1] Chalmers, D.J. (2002). On Sense and Intension, in Tomberlin, ed., Philosophical Perspectives 16: Language and Mind, Blackwell, pp. 135-82

[2] Pirotte, A., Roelants, D., Zimányi, E. (1991). Controlled Generation of Intensional Answers. IEEE Trans. Knowl. Data Eng. 3(2): 221-236

[3] Motro, A.. Using integrity constraints to provide intensional answers to relational queries. In Proc. 15th International Conference on Very Large Databases, pages 237–246, Amsterdam, 1989

[4] Gal, A. and Minker, J. Informative and Cooperative Answers in Databases Using Integrity Constraints. Technical Report CS-TR-1911, University of Maryland, September 1987

[5] Kanellakis, P.C., Kuper, G., Revesz, P. (1995). Constraint Query Languages. Journal of Computer and System Sciences, 51(1)

[6] Baralis E., Garza P., Quintarelli E., Tanca L.: Answering XML queries by means of data summaries. ACM Trans. Inf. Syst. 25(3) (2007)

[7] Mazuran, M., Quintarelli, E., Tanca, L. (2012). Data Mining for XML Query-Answering Support. IEEE Trans. Knowl. Data Eng. Engine 24(8): 1393-1407