Knowledge Discovery and Risk Prediction for Chronic Diseases: an Integrated Approach

Anju Verma[1], Maurizio Fiasché[1,2], Maria Cuzzola[1], Francesco C. Morabito[2] and Giuseppe Irrera[1]

[1] CTMO - Transplant Regional Center of Stem Cells and Cellular Therapy, "A. Neri", Hospital "Morelli" of Reggio Calabria, Italy
[2] DIMET, University "Mediterranea" of Reggio Calabria, Italy
maurizio.fiasche@unirc.it

Abstract. A novel ontology based type 2 diabetes risk analysis system framework is described, which allows the creation of global knowledge representation (ontology) and personalized modeling for a decision support system. A computerized model focusing on organizing knowledge related to three chronic diseases and genes has been developed in an ontological representation that is able to identify interrelationships for the ontology-based personalized risk evaluation for chronic diseases. The personalized modeling is a process of model creation for a single person, based on their personal data and the information available in the ontology. A transductive neuro-fuzzy inference system with weighted data normalization is used to evaluate personalized risk for chronic disease. This approach aims to provide support for further discovery through the integration of the ontological representation to build an expert system in order to pinpoint genes of interest and relevant diet components.

Keywords: Knowledge discovery, knowledge representation, chronic disease ontology, personalized risk evaluation system.

1 Introduction

Populations are aging and the prevalence of chronic diseases which persists for many years is increasing. The chronic diseases such as cardiovascular disease, type 2 diabetes and obesity have high global prevalence, have multifactorial etiology. These diseases are mainly caused by interactions of a number of common factors including genes, nutrition and life-style. For ontology based personalized risk evaluation for type 2 diabetes, a Protégé-based ontology has been developed for entering data for type 2 diabetes and linking and building relationships among concepts. The ontological representation provides the framework into which information on individual patients for disease symptoms, gene maps, diet and life history details can be inputted, and risks, profiles, and recommendations derived.

A personalized risk evaluation system has been used for building the personalized modeling. Global models capture trends in data that are valid for the whole problem space, and local models capture local patterns which are valid for clusters of data. Both models contain useful information and knowledge. Local models are also adaptive to new data as new clusters and new functions that capture patterns of data in these clusters. A local model can be incrementally created. Usually, both global and local modeling approaches assume a fixed set of variables and if new variables, along with new data, are introduced with time, the models are very difficult to modify in order to accommodate these new variables. However new variables can be accommodated in only personalized models, as they are created ‘‘on the fly’’ provided that there is relevant data for them [1]. The personalized risk evaluation using ontological based data is the main approach of the paper.

2 Chronic Disease Ontology

Ontology is a systematic account of being or existence. Ontology in terms of bioinformatics can be interpreted as the representation of the existing domain of the knowledge of life. Ontology is used to reason and make inferences about the objects within the domain [2]. Ontology is concerned with making information and knowledge explicit; it includes descriptions of concepts and their relationships. Ontology describes a hierarchical structure of concepts and the relationships built in order to extract new knowledge.

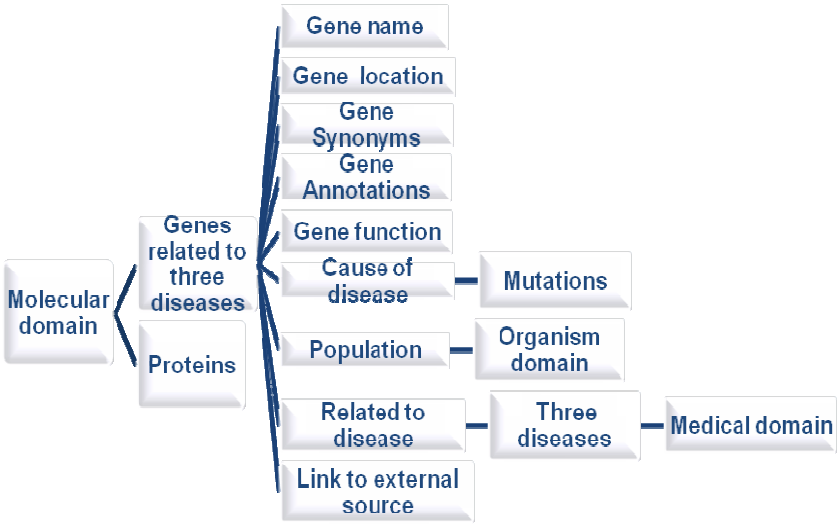

Fig.1. General structure of molecular domain in the chronic disease ontology.

Ontology is generally written as a set of definitions of the formal vocabulary of objects and relationships in the given domain. It supports the sharing and reuse of formally represented knowledge among systems [3, 4]. As a database technology, ontologies are commonly coded as triple stores (subject, relationship, object), where a network of objects is formed by relationship linkages, as a way of storing semantic information [5, 6].A standardized ontology framework makes data easily available for advanced methods of analysis, including artificial intelligence algorithms, that can tackle the multitude of large and complex datasets by clustering, classification, and rule inference for biomedical and bioinformatics applications. The main advantages of building ontology are to extract and collect knowledge; share knowledge; manage terminology; store, retrieve and analyze; find relationships between the concepts; discover new knowledge and reuse knowledge for decision support system.

Chronic disease ontology consists of five major domains namely; organism domain, molecular domain, medical domain, nutritional domain and a biomedical informatics map domain. These domains or classifications contain further subclasses and instances. Each subclass has a set of slots which provide information about each instance and have relationships among other slots, instances and concepts. Each gene instance has different information associated with the gene and also has relationships with other domains (Figure1). The chronic disease ontology can be updated manually and regularly with new knowledge and information providing a framework to keep an individual's specific information (medical, genetic, clinical and nutritional), to discover new knowledge and to adapt as required for personalized risk prediction and advice.

3 Type 2 Diabetes Personalized Risk Evaluation System

Type 2 diabetes mellitus is one of the most common chronic "lifestyle" diseases with a high prevalence throughout the world [7]. There are two main types of diabetes mellitus; type-1 and type-2. Type 2 diabetes is the most common type of diabetes and globally about 90% of all cases of diabetes are type 2 diabetes [8]. There have been several models, namely, 'The global diabetes model' [9, 10], 'The diabetes risk score' [11], the 'Archimedes diabetes model' [12, 13], the Diabetes risk score in Oman [14], the 'Genetic Risk Score' [15]. All these models predict risk of future complications associated with type 2 diabetes in people with already diagnosed type 2 diabetes.

The global diabetes model (GDM) is a continuous, stochastic micro simulation (individual by individual approach) model of type 2 diabetes. The GDM is a computer program and predicts longevity, quality of life, medical events and expenditures for groups and individuals with type 2 diabetes. The GDM calculates rates and probabilities of the medical events in diabetic individuals [9, 10]. It has been reported that from the existing methods for predicting risk of type 2 diabetes, The Archimedes Model predicts the risk with better sensitivity and specificity than other models [16]. Recently, the 'Genetic Risk Score' has been developed which uses multiple genetic as well as conventional risk factors [15]. Because these methods calculate risk of type 2 diabetes globally and they are not the same as the proposed methodology in this thesis, which involves calculations of personalized risk. The aim of the current

research is to create a personalized model for predicting risk of type 2 diabetes. Genetic variables have been used along with clinical variables to create a personalized model to predict risk of type 2 diabetes. The next section of this paper describes the methods used for creating a diabetes risk model using genetic markers along with clinical variables.

3.1 Step 1: Selection of Features for building the personalized risk evaluation system for type-2 diabetes:

The first step to build the model was feature selection which has been done by using different methods including signal to noise ratio and t-test. This analysis was done using NeuCom and Siftware. NeuCom is a computer environment based on connectionist (Neuro-computing) modules. NeuCom is self-programmable, learning and reasoning tool. NeuCom environment can be used for data analysis, modeling and knowledge discovery. Siftware is an environment for analysis, modeling and profiling of gene expression data. NeuCom and Siftware have been developed at Knowledge Engineering and Discovery Research Institute (KEDRI, http://www.kedri.info).

Results achieved from signal to noise ratio are exactly similar to student's t-test. According to signal to noise ratio and t-test for the combined male and female subjects, genes ANGPTL3, ANGPT4, TNF, FLT1, MMP2 and CHGA are ranked highest. Interestingly, gene CHGA has not been ranked at same high position for male and female subjects separately. The first six genes of highest importance for male and female subjects were selected for further analysis and to build personalized sex-specific risk prediction model. As genes are ranked differently as per signal to noise ratio for male and female subjects, different genes have been selected for personalized modeling for male and female subjects. Different methods were then used for type 2 diabetes risk prediction methods such as multiple linear regression using NeuCom (global, inductive method), WWKNN and TWNFI (personalized methods) [17].

3.2 Step2: Building Personalized risk evaluation model:

As every person has a different genetic admixture, therefore personalized prediction and treatment is required for each person. In personalized modeling, a model is created for a single point (subject record) of the problem space only using transductive reasoning. A personalized model is created "on the fly" for every new input vector and this individual model is based on the closest data samples to the new samples taken from a data set. The K-nearest neighbors (K-NN) method is one example of the personalized modeling technique. In the K-NN method, for every new sample, the nearest K samples are derived from a data set using a distance measure, usually Euclidean distance, and a voting scheme is applied to define the class label for the new sample [18, 19]. In the K-NN method, the output value y for a new vector x is calculated as the average of the output values of the k nearest samples from the data set D. In the weighted K-NN method (WKNN), the output y is calculated based not only on the output values (e.g. class label) y of the K, NN samples, but also on a weight w , that depends on the distance of them to x . In Weighted-weighted K nearest neighbor algorithm for transductive reasoning (WWKNN) the distance

between a new input vector and the neighboring ones is weighted, and also variables are ranked according to their importance in the neighborhood area.

Transductive neuro-fuzzy inference system with weighted data normalization (TWNFI) is an improved, advanced and more complex transductive and dynamic neural-fuzzy inference system with local generalization, in which, either the Zadeh-Mamdani type fuzzy inference engine [20,21] or the Takagi-Sugeno fuzzy inference engine [22] can be used. The local generalization means that in a sub-space (local area) of the whole problem space, a model is created and this model performs generalization in this area.

Table 1. Examples of TWNFI personalized models for two different male subjects; high risk and low risk; with weight of variables and genes with global weights representing importance of the variables.

	Subject 1 (High risk male)	Subject 2 (Low risk male)	
Input Variables	Weights of input variables	Weights of input variables	Global weights/ importance (male)
Age (years)	0.7729	0.9625	0.8393
Haemoglobin (g/L)	0.8521	0.7847	0.8429
Fasting blood glucose (mmol/L)	0.7507	0.9352	0.8769
Cholesterol (mmol/L)	0.7478	0.752	0.8104
Triglycerides (mmol/L)	0.6961	0.7413	0.8327
ANGPTL3	0.7617	0.9269	0.9254
FGF1	0.7295	0.641	0.8228
FLT1	0.651	0.7059	0.8096
MMP2	0.6797	0.8802	0.9009
TNF	1	0.8495	0.8699
ANGPT4	0.6705	1	0.904
Actual output			
Predicted output with Multiple linear regression	0.7963	0.1378	
Predicted output with WWKNN	1.127	0	
Predicted output with TWNFI	1.002	0	

In the TWNFI model, Gaussian fuzzy membership functions are used in each fuzzy rule for both antecedent and consequent parts. In TWNFI data is first normalized and then it looks for nearest samples. TWNFI performs a better local generalization over new data as it develops an individual model for each data vector that takes into account the new input vector location in the space. Table 1 shows results from example of personalized model built for two male subjects. Subject 1 belongs to class 1(with type 2 diabetes) and subject 2 belongs to class0 (without type 2 diabetes). It was found that highest accuracy was achieved with the TWNFI method. TWNFI not

only gives highest accuracy, also gives weights of variables as per their importance for risk of disease.

For each subject in present example, separate weight of each variable has been presented and compared with global weights of variables for male subjects. It is very interesting that male subject 1 and 2 both have higher values of fasting blood glucose, cholesterol and triglycerides, the genes were more important factors to predict the risk of type 2 diabetes for male subject 2. By comparing weights for each variable of each subject, it was found that for male subject 1, gene TNF was found to be the most important gene associated with type 2 diabetes while for male subject 2, ANGPT4 gene has been weighted the highest, while for all the male subjects the ANGPTL3 gene has been found most important factor for type 2 diabetes. TWNFI along with high accuracy and importance of variables also provides set of rules based on the clusters formed based on nearest neighbors. Each rule contains a lot of information for each variable. Rules or profiles for male subjects were generated on the basis of nearest samples.

4 Integration Framework for Chronic Disease Ontology and Personalized Modeling

This section explains the framework for integrating the chronic disease ontology and a personalized risk evaluation system. The challenge is to create computational platforms that dynamically integrate the ontology and a set of efficient machine learning methods, including new methods for personalized modeling that would manifest better accuracy at a personal level and facilitate new discoveries in the field of bioinformatics. The chronic disease ontology that was described in section 2 will be used to integrate personalized modeling and ontology. The current chronic disease ontology contains most of genes which are common for three chronic interrelated diseases (cardiovascular disease, type-2 diabetes and obesity).

Data for personalized modeling was collected during a Government Research Program with the title: “Biologia e Impiego dei Progenitori Endoteliali nell’ Arteriopatia Obliterante Periferica” sponsored from Italian Ministry of the Health. The collected dataset which was used for predicting risk of type 2 diabetes included clinical and genetic variables and it was found that for male and female subjects different combinations of genes are more predictive of type 2 diabetes. So these genes were updated in the chronic disease ontology and the missing genes and information related to these genes was also added in to the chronic disease ontology. Similarly, any other information derived from personalized model can be added to the chronic disease ontology and the new relationships and discoveries within the chronic disease ontology can be used to improve personalized risk evaluation system (Figure2).

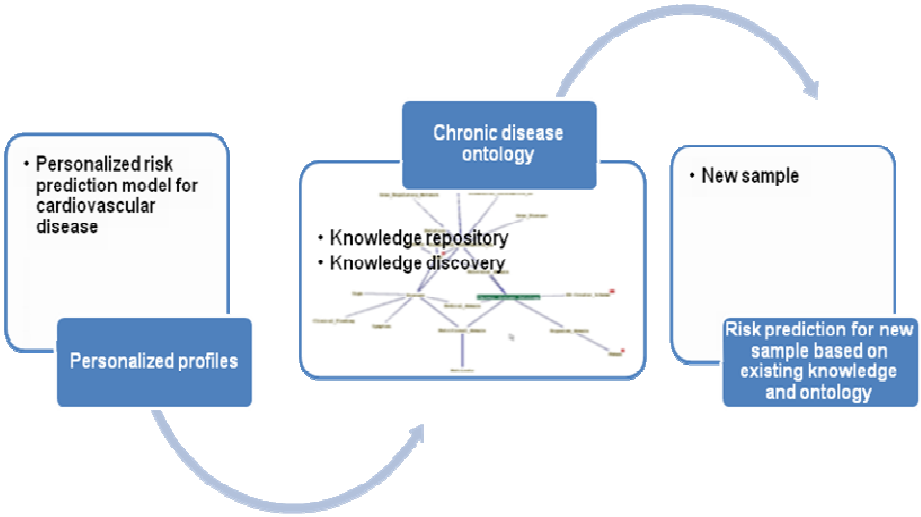

Fig.2. Example of framework for use of knowledge from the chronic disease ontology (CDO) to personalized model.

The framework uses the chronic disease ontology based data and knowledge embedded in the ontology. It also allows the adaptation of new knowledge by entering the results of the machine learning system to ontology. The main modules (Figure 3) are: an ontology module, a machine learning module (TWNFI) and an interface to import and export knowledge from and to ontology. The ontology and machine learning module evolve through continuous learning from new data. Results from the machine learning procedures can be entered back into the ontology thus enriching its knowledge base and facilitating new discoveries. Integration of the chronic disease ontology and personalized model can be done for diabetes.

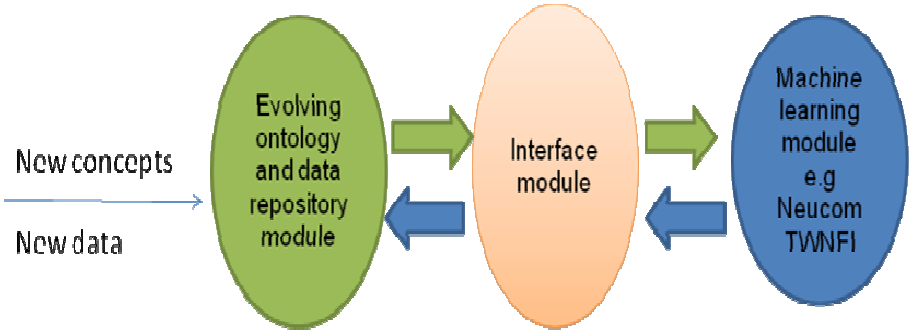

Fig.3. The ontology-based personalized decision support (OBPDS) framework consisting of three interconnected parts: (1) An ontology/database module; (2) Interface module; (3) A machine learning module.

It can be explained with the help of an example as the information obtained from personalized model for type 2 diabetes in the Italian dataset, such as the gene matrix metalloproteinase (MMP2), responsible for protein binding in normal person and mutated form is responsible for high risk of type 2 diabetes in the Italian male population can be added to the ontology and if a new male subject comes which is from Italian population, the same information can be used next time.

Similarly, it has been found that the gene hypoxia inducible factor 1 (HIF1A) acts as a normal transcription binding factor but mutation in gene is related to type 2 diabetes in females in Italian population. This information can be added to ontology and can be applied to the analysis for the next new subject from a similar population and with similar clinical features for risk prediction. Similar process can be applied for predicting risk of obesity.

Recently, it was found that gene FTO in its inactivated state protects from risk of obesity [23]. Polymorphism in FTO gene is strongly and positively correlated to body mass index which is common measure of obesity. This knowledge has been updated in the chronic disease ontology and the system is able to use this knowledge if a similar subject with high body mass index comes, it can identify that FTO gene is active and the person may have a predisposition to obesity if dietary intake exceeds physical activity.

5 Conclusions and Future Plans

It has been found that the male subjects have high values of cholesterol and triglycerides and are more prone to type 2 diabetes; For male and female subjects different combinations of genes have association with type 2 diabetes; for male subjects, genes ANGPTL3, MMP2, ANGPT4, TNF, FGF1 and FLT1 appear to be the most important genes associated with risk of type 2 diabetes; for female subjects, genes ANGPTL3, ANGPTL4. HIF1A, TNSF15, FLT1 and TNF appear to be the most important factors for determining risk of type 2 diabetes [24].

The explained framework for the integration of the ontology and the personalized modeling techniques illustrates the integration of personalized method and ontology database for better recommendations and advice and explains how existing knowledge and new knowledge can be used together for better life style, risk evaluation and recommendations[25,26]. As, Diabetes has global prevalence and none of the methods so far published have combined clinical and genetic variables together. The section 3 we have described how a model can be built using clinical and genetic variables. For personalized modeling, different methods such as WWKNN and TWNFI were used and compared [27, 28, 29, 30]. It has been found that TWNFI gives highest accuracy along with importance of each gene and variable by optimizing each variable and weight which can be used for better prediction and recommendations.

Our future plan is to extend the personalized risk evaluation system explained in this paper with more genes and more set of clinical and general variables. Still a better prediction system can be developed, if nutritional information and other

environmental variables are known (e.g. exposure to sun for vitamin D) along with clinical and genetic variables are available.

We also plan to extend the chronic disease ontology with the new knowledge and information in terms of; New data, genes and also in terms of medical information such as anatomical and physiological information about the organs involved in the type 2 diabetes. The chronic disease ontology is evolving and vast project and can be carried out for years. The only limitation of evolving the chronic disease ontology is that the new information has to be added manually, at present there is no such tool which can automatically update the existing information without duplicating or removing the existing knowledge in ontology.

Acknowledgments: This study has been supported by Foundation of Research and Technology by TIF scholarship through an affiliate of Pacific Channel Limited and Knowledge Engineering and Discovery Research Institute, AUT. Many Thanks to Prof. Nik Kasabov for his support.

References

1. Kasabov, N.: Global, local and personalized modeling and profile discovery in Bioinformatics: An integrated approach, Pattern Recognition Letters, 28(6), 673-685 (2007).
2. Gruber T. R.: A translation approach to portable ontologies. Knowledge Acquisition 5, 199-220 (1993).
3. Fensel D.: Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. 2 ed. Springer, Heidelberg (2004).
4. Chandrasekaran B., Josephson J. R., Benjamins V. R.: What are ontologies, and why do we need them? Intelligent Systems and Their Applications 14: 20-26 (1999).
5. Owens A.: Semantic Storage: Overview and Assessment. Technical Report IRP Report 2005, Electronics and Computer Science, U of Southampton (2005).
6. Berners-Lee T., Hendler J., Lassila O.: The Semantic Web. Scientific American (May 17) (2001).
7. The FIELD Study Investigators. (2004). The need for a large-scale trial of fibrate therapy in diabetes: the rationale and design of the Fenofibrate Intervention and Event Lowering in Diabetes (FIELD) study. ISRCTN64783481. Cardiovascular Diabetology. 2004; 3:9.
8. New Zealand Guidelines Group (2003(a)). Management of diabetes. New Zealand Guidelines Group, Wellington. Retrieved from : http://www.nzgg.org.nz/guidelines/dsp_guideline_popup.cfm?guidelineID=36
9. Brown, J. B., A. J. Palmer, et al. (2000(a)). "The Mt. Hood challenge: cross-testing two diabetes simulation models." Diabetes Research and Clinical Practice 50(3): S57-S64.
10. Brown, J. B., A. Russell, et al. (2000(b)). "The global diabetes model: user friendly version 3.0." Diabetes Research and Clinical Practice 50(3): S15-S46.
11. Lindstrom, J. and J. Tuomilehto (2003). "The diabetes risk score. A practical tool to predict type-2 diabetes risk." Diabetes Care 26(3): 725-731.
12. Eddy, D. M. and L. Schlessinger (2003(a)). "Archimedes. A trial-validated model of diabetes." Diabetes Care 26(11): 3093-3101.
13. Eddy, D. M. and L. Schlessinger (2003(b)). "Validation of the Archimedes diabetes model." Diabetes Care 26(11): 3102-3110.

14. Al-Lawati, J. A. and J. Tuomilehto (2007). "Diabetes risk score in Oman: A tool to identify prevalent type-2 diabetes among Arabs of the Middle East." Diabetes Research and Clinical Practice 77: 438-444.
15. Cornelis, M., L. Qi, et al. (2009). "Joint effects of common genetic variants on the risk of type-2 diabetes in U. S. men and women of European ancestry." Annals of Internal Medicine 150: 541-550.
16. Stern, M., K. Williams, et al. (2008). "Validation of prediction of diabetes by the Archimedes Model and comparison with other predictiong models." Diabetes Care 31(8): 1670-1671.
17. Song, Q. and Kasabov, N.: TWNFI - a transductive neuro-fuzzy inference system with weighted data normalization for personalized modeling, Neural Networks, 19(10), 1591-1596 (2006).
18. Vapnik, V. N. (1998). Statistical Learning Theory: Wiley Inter-Science.
19. Mitchell, M. T., Keller, R., et al (1997). Explanation-based generalization: A unified view. Machine Learning, 1(1), 47-80.
20. Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8, 338-353.
21. Zadeh, L. A. (1988). Fuzzy logic. IEEE Computer, 21, 83-93.
22. Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. IEEE Transactions on Systems, Man, and Cybernetics, 15, 116-132.
23. Fischer, J., L. Koch, et al. (2009). "Inactivation of the Fto gene protects from obesity." Nature 458: 894-899.
24. Verma, A. (2010). An Integrated Approach for Ontology Based Personalized Modeling: Chronic Disease Ontology, Risk Evaluation and Knowledge Discovery. LAP LAMBERT Academic Publishing.
25. Kasabov,N., Song, Q., Benuskova, L., Gottgtroy, P., Jain, V., Verma, A., Havukkala, I., Rush, E., Pears, R., Tjahjana, A., Hu, Y., MacDonel, S.,(2008). Integrating Local and Personalised Modelling with Global Ontology Knowledge Bases for Biomedical and Bioinformatics Decision Support, Chapter 4, In: Smolin et al (eds) Computational Intelligence in Bioinformatics, Springer.
26. Kasabov, N. and Y. Hu (2010) Integrated optimisation method for personalised modelling and case study applications, Int. Journal of Functional Informatics and Personalised Medicine, vol.3,No.3,236-256.
27. Fiasché, M., Verma, A., Cuzzola, M., Iacopino, P., Kasabov, N. and Morabito, F. C. (2009). Discovering Diagnostic Gene Targets and Early Diagnosis of Acute GVHD Using Methods of Computational Intelligence over Gene Expression Data. In: Artificial Neural Networks – ICANN 2009. Part II, LNCS 5769/2009, pp 10-19. Springer Berlin / Heidelberg, ISBN/ISSN: 978-3-642-04276-8.
28. Fiasché, M., Cuzzola, M., Fedele, R., Iacopino, P., Morabito, F.C. (2010). Machine Learning and Personalized Modeling based Gene Selection for acute GVHD Gene Expression Data Analysis. In: Artificial Neural Networks – proceedings of ICANN 2010 Part I, LNCS 6352.
29. Fiasché, M., Cuzzola, M., Irrera, G., Iacopino, P., Morabito, F.C. Advances in Medical Decision Support Systems for Diagnosis of Acute Graft-versus-Host Disease: Molecular and Computational Intelligence Joint Approaches. Frontiers in Biology. Higher Education Press and Springer -Verlag GmbH, doi: 10.1007/s11515-011-1124-8.
30. M. Fiasché, M. Cuzzola, P. Iacopino, N. Kasabov , F. C. Morabito. *Personalized Modeling based Gene Selection for acute GvHD Gene Expression Data Analysis: a Computational Framework Proposed*. Australian Journal of Intelligent Information Processing Systems, Vol 12, No 4 (2010): Machine Learning Applications (Part II).