Front. Biol. 2011, 6(4): 263–273
DOI 10.1007/s11515-011-1124-8

REVIEW

Advances in medical decision support systems for diagnosis of acute graft-versus-host disease: molecular and computational intelligence joint approaches

Maurizio FIASCHÉ (✉)[1,2], Maria CUZZOLA[2], Giuseppe IRRERA[2], Pasquale IACOPINO[3], Francesco Carlo MORABITO[1]

[1] *DIMET, University "Mediterranea" of Reggio Calabria, Italy*
[2] *Transplant Regional Center of Stem Cells and Cellular Therapy, "A. Neri" , Reggio Calabria, Italy*
[3] *School of Hematology, University of Messina, Italy*

Abstract **Acute graft-versus-host disease (aGVHD) is a serious systemic complication of allogeneic hematopoietic stem cell transplantation (HSCT) causing considerable morbidity and mortality. Acute GVHD occurs when alloreactive donor-derived T cells recognize host-recipient antigens as foreign. These trigger a complex multiphase process that ultimately results in apoptotic injury in target organs. The early events leading to GVHD seem to occur very soon, presumably within hours from the graft infusion. Therefore, when the first signs of aGVHD clinically manifest, the disease has been ongoing for several days at the cellular level, and the inflammatory cytokine cascade is fully activated. So, it comes as no surprise that progress in treatment based on clinical diagnosis of aGVHD has been limited in the past 30 years. It is likely that a pre-emptive strategy using systemic high-dose corticosteroids as early as possible could improve the outcome of aGVHD. Due to the deleterious effects of such treatment particularly in terms of infection risk posed by systemic steroid administration in a population that is already immune-suppressed, it is critical to identify biomarker signatures for approaching this very complex task. Some research groups have begun addressing this issue through molecular and proteomic analyses, combining these approaches with computational intelligence techniques, with the specific aim of facilitating the identification of diagnostic biomarkers in aGVHD. In this review, we focus on the aGVHD scenario and on the more recent state-of-the-art. We also attempt to give an overview of the classical and novel techniques proposed as medical decision support system for the diagnosis of GVHD.**

Keywords computational intelligence, gene selection, GVHD, machine learning, personalized modelling, wrapper

Introduction

Recently, there have been major advances in our knowledge of basic immunology. In parallel, although much of information has been obtained from preclinical models and far less from correlations with clinical observations or treatment, our awareness of the complexity of the pathophysiology of acute graft-versus-host disease (aGVHD) is significantly increased (Socié and Blazar, 2009).

At the same time, the interplay with bioinformatics, defined as the branch of information sciences for the analysis, modeling, simulation and knowledge discovery of biological phenomena, such as genetic processes, has stimulated synergistic research in many cross-disciplinary areas.

The potential applications of microarray technology are numerous and include identifying markers for classification, diagnosis, disease outcome prediction, target identification and therapeutic responsiveness. However, microarray analysis might not identify unique markers (e.g. a single gene) clinically useful for same diseases. Indeed, diagnosis and prediction of the biological state/disease is likely to be more accurate by identifying clusters of gene expression profiles (GEPs) performed by macroarray analysis. Based on profile, it is possible to set a diagnostic test: a sample can be taken from a patient, the data related to the sample processed, and,

Received November 30, 2010; accepted March 21, 2011
Correspondence: Maurizio FIASCHÉ
E-mail: maurizio.fiasche@unirc.it

eventually, a profile related to the sample can be obtained (Kasabov, 2007a).

This profile can be matched against existing gene profiles and based on similarity, it can be used with certain probability to confirm a diagnosis of disease or if the patient is at risk of developing it in the future. This approach could be used to detect aGVHD as well as to discover curative therapy for several malignant and non malignant disorders. Until now, aGVHD diagnosis is merely based on clinical criteria eventually confirmed by biopsy of one of the three target organs (skin, gastrointestinal tract, or liver) (Ferrara and Reddy, 2006).

Unfortunately, there is no definitive diagnostic blood test for aGVHD, although a lot of circulating serum proteins have been described as potential biomarkers in recent limited studies (Socié, 2009).

As aforementioned, most of the current knowledge of the pathophysiology of aGVHD is derived from animal models. It was first noted that irradiated mice infused with allogeneic marrow or spleen cells died as result of syndrome recognized as aGVHD (van Bekkum and Vries De, 1967). This post-transplantation complication is a complex multistep reaction described by Ferrara et al. (2009) as a three-phase phenomenon that occurs in a step-wise and sequential manner (Fig. 1). The starting point is the tissue damage to the recipient by the radiation/chemotherapy pre-transplant conditioning regimen. The first step triggers several effects: danger signals, activation of host antigen presenting cells (APCs), storm of inflammatory cytokines, cell mediators and co-stimulatory molecules recruitment in the damage sites. It is useful to remember that conditioning regimen not only affects immune cells but also simultaneously endothelial and epithelial cells. The second step involves the activation, proliferation, differentiation and migration of alloreactive donor T cells, which expand and differentiate into effector cells. Host APCs alone are sufficient to activate donor T cells, although, both host- and donor-derived APCs migrate in secondary lymphoid organs after allogeneic HSC transplant (Beilhack et al., 2005). Finally, in the third step, the effector phase, the activated donor T cells mediate cytotoxicity against target host cells by Fas-Fas ligand interactions, perforin-granzyme B, and the additional production of cytokines, such as TNF-α. This additional promotion of inflammation ultimately may cause extensive destruction of target tissues in the transplant recipient (Jacobsohn and Vogelsang, 2007).

This paper is organized as follows: in "GVHD: prediction and diagnosis" section, we focus on recent advances in molecular pathophysiology and on how these data could be translated into the development of innovative diagnostic strategy for patients with aGVHD. In "The computational intelligence approach" section, the problem under study is defined and the benefits and limitations of classical computational intelligence approach are commented. In "The personalized modeling" section, we give an overview of gene selection wrapper methods and, in particular, we discuss a personalized modeling gene selection algorithm (Hu et al., 2009). Some experimental data allowing identification of a compact set of transcript genes from GEP data are provided in "Experiment" section; in the sub-sections, the joint use of computational and molecular evidence, proposed for the first time by Fiasché et al. in 2009 and 2010, to confirm the early statement of aGVHD in real cases is described. Finally, in "Biomedical conclusions and future work" section some comments with respect to possible future developments of the present research are given.

GVHD: prediction and diagnosis

Despite the introduction of new immunosuppressant drugs in the prophylaxis regimens, neither new agent nor new strategy (i.e. T-depletion) has been associated with a clear reduction in incidence of aGVHD without compromising engraftment or relapse rate of malignancies. Thus, current standard regimens commonly employ a calcineurin inhibitor plus methotrexate or mycophenolate mofetil. Moreover, antithymocyte globulin (ATG) is administered if the donor is unrelated or mismatched (Ayuk et al., 2008). The immunosuppressive agents used to

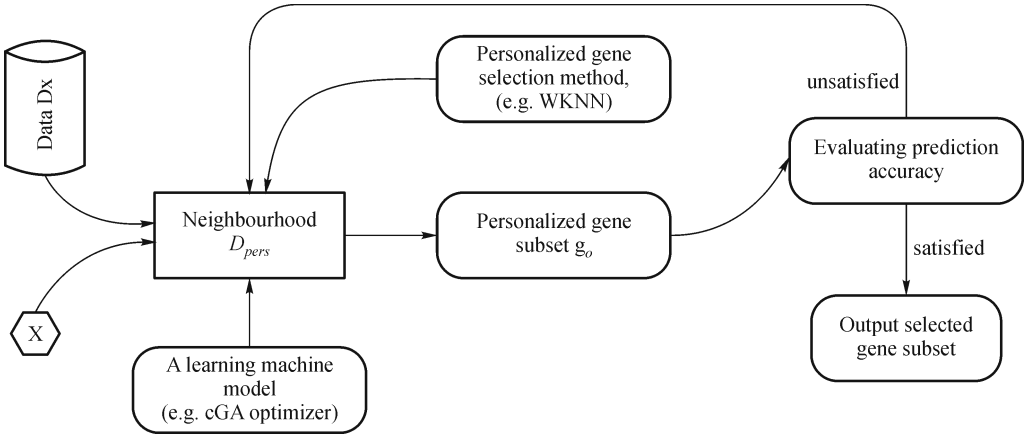

Figure 1 A diagram of personalized modeling-based gene selection method (PMGS).

prevent aGVHD and methylprednisolone can be used to treat established aGVHD, achieving response rates of 30%–70%. No standard therapy exists once a patient has developed aGVHD refractory to corticosteroids, although a number of agents have been tried with limited success (Ho and Cutler, 2008).

Nowadays, standardized criteria for the aGVHD diagnosis are based on clinical signs and can be confirmed by biopsy, which is an invasive and labor-intensive approach. However, symptoms of aGVHD are frequently non-specific, especially when the disease is isolated to gastrointestinal tract or liver. In such scenario, a purely clinical diagnosis is challenging and it becomes even more difficult when aGVHD has late onset (Brodoefel et al., 2010).

Since the outcome of aGVHD is associated with an early start of immunosuppressive therapy, it is essential to achieve rapid diagnosis and accurate differentiation from other conditions. On the other hand, undesirable consequences of unnecessary over treatment are the increase of both opportunistic infections and relapse rate.

In addition, one of the main problems facing the management of aGVHD following hematopoietic stem cell transplantation (HSCT) is that early events leading to disease occur very soon, presumably within hours from the graft infusion. Therefore, when aGVHD is clinically manifest, the disease has been ongoing for several days at the cellular level, and the inflammatory cytokine cascade is fully activated. So, it comes as no surprise that progress in treatment based on clinical diagnosis of aGVHD has been limited in the past 30 years (Bacigalupo et al., 2010).

As noted above, initial control of aGVHD is critical, but no current laboratory test can be used to predict or to diagnose the disease. Even biopsy may give false negative and false-positive results (Firoz et al., 2006; Ferrara, 2008). Thus, there is a clear need for adjunctive diagnostic tool to aid in early management decision. The use of a biomarker-based diagnostic approach has proven extremely useful in other serious medical conditions, and this strategy has been advocated also in aGVHD.

Recently, some research groups have proposed novel approaches to facilitate aGVHD diagnosis based on molecular methods and proteomic analysis. In a pilot study, Buzzeo et al. (2008) have investigated the genome-wide transcriptional response occurring in allogeneic HSCT patients during aGVHD onset. This study, based on microarray analysis, has identified a large number of transcriptional immune markers that significantly were up- or down-expressed during the development of aGVHD. Thus, monitoring changes in immune gene response could predict the GVHD risk in allogeneic-HSCT patients. Further studies designed to define risk of GVHD through proteomic approaches, have demonstrated the power and relevance of an individualized approach for GVHD management (Weissinger et al., 2007; Paczesny et al., 2009a, 2009b). In particular, Paczesny et al. (Paczesny et al., 2010a) have demonstrated that elafin protein could provide prognostic and diagnostic information specific for skin aGVHD but no for other target tissues. A different approach has been proposed by some of the authors (Fiasché et al., 2009) in a preliminary study, using computational intelligence analysis on gene expression data. This study has provided evidence that a joint computational-biological approach can be successfully applied to assess early aGVHD and to discover biomarker transcript genes useful for aGVHD diagnosis.

Other likely candidate biomarkers for diagnosis and prediction of aGVHD are based on the changes observed in the many cellular players that are implicated in the different pathophysiology phases of aGVHD (Paczesny et al., 2010b). For instance, it is known that regulatory T cells (Treg) frequency at GVHD onset are significantly reduced in patients who develop severe aGVHD compared to patients without GVHD (Magenau et al., 2010). The point is that, as the authors themselves stated Treg frequency at aGVHD onset has only modest diagnostic value. This concept is strengthened by evidence that there is no correlation of Treg frequency with the histological or clinical severity of gastrointestinal aGVHD (Lord et al., 2011).

From the methodological point of view, the above examples illustrate how the armamentarium to assess molecular and cellular immune profiling is increasing rapidly. In our study, we choose to investigate the immune interactions occurring during aGVHD by transcriptomic quantifications. We attempted to obtain a transcriptomic profile in the blood measuring RNA quantity in circulating nucleated cells. Since transcriptional activity is largely dependent on environmental factors, RNA levels change dynamically over time in blood (Chaussabel et al., 2010). Therefore, we believe that the status and dynamic changes in the cellular make up of the immune system can be better monitored by profiling transcripts when the alloreactive T cells attack the recipient tissues during aGVHD (Payne et al., 1996).

The computational intelligence approach

Bioinformatics brings together several disciplines: molecular biology, genetics, microbiology, mathematics, chemistry and biochemistry, physics, and, of course, informatics. Many processes in biology, are dynamically evolving and their modeling requires evolving methods and systems. In bioinformatics new data are being made available with a tremendous speed that would require the models to be continuously adaptive. Knowledge-based modeling, that includes rule and knowledge discovery, is a crucial requirement.

In what follows, we sum up the main phases of information processing and problem solving in most of the bioinformatics systems (Kasabov, 2007a):

1) Data collection: Collecting biological samples and

processing them.

2) Feature analysis and feature extraction: Defining which features are more relevant and therefore should be used when creating a model for a particular problem (e.g. classification, prediction, decision making).

3) Modeling the problem: Defining inputs, outputs, and type of the model (e.g. probabilistic, rule-based, connectionist), training the model, and statistical verification.

4) Knowledge discovery *in silico*: New knowledge is gained through the analysis of the modeling results and the model itself.

5) Verifying the discovered knowledge *in vitro* and *in vivo*: Biological experiments both in the laboratory and in real life to confirm the discovered knowledge.

Some tasks in bioinformatics are characterized by:

1) Small data sets, few samples.

2) Static data sets, i.e. data do not change in time once they are used to create a model.

3) No need for online adaptation and training on new data.

For these tasks the traditional statistical and artificial intelligence (AI) techniques are well suited. The traditional, off-line modeling methods assume that data are static and no new data are going to be added to the model. Before the model creation, data are analyzed and relevant features are selected, again in an off-line mode. The offline mode usually requires many iterations of data propagation for estimating the model parameters.

Unfortunately, most of the tasks for data analysis and modeling in bioinformatics are characterized by:

1) Large dimensional data sets that are updated regularly.

2) A need for incremental learning and adaptation of the models from input data streams that may change their dynamics in time.

3) Knowledge adaptation based on a continuous stream of new data.

Data collection of GVHD poses these 3 problems, and during our study there was evidence that classical AI methods had some problems when new data were added to the model (Fiasché et al., 2010a). Personalized modeling and the ECOS-like integrated approach (Kasabov, 2007a) for new data and existing modeling could be the solution for this very complex task.

The personalized modeling

The main idea of personalized modeling is to create a model for each objective sample, which is able to discover the most important information specifically for this sample. Since personalized modeling focuses on the individual sample rather than simply on the global problem space, it can be more appropriate to build clinical decision support systems for new patients. A previous work has reported that using personalized modeling can achieve better classification results than the results from global modeling (Song and Kasabov, 2006). The framework proposed by Fiasché et al. (2010a) used a personalized modeling based gene selection method (PMGS) (Hu et al., 2009) for macroarray data analysis integrating new data with the existing models (Kasabov, 2007a).

Gene selection methods and personalized modeling

The advent of microarray technology emphasized the problem to identify which genes are most important for diagnosing different diseases (e.g. cancer diagnosis) and prognosis task. Generally, most developed gene selection methods can be categorized into two groups: filter and wrapper methods. Filters and wrappers differ in the way the feature subsets are evaluated. Filter approaches remove irrelevant features (genes) according to general characteristics of the data, measuring the intrinsic characteristic of genes. Wrapper gene selection methods, by contrast, apply machine learning models (usually classification models) to feature subsets and use cross-validation to evaluate the score and to estimate the usefulness of feature subsets. In theory, wrappers should provide more accurate classification results than filters (Langley, 1994). The use of "tailor-made" feature subsets should provide better classification accuracy for the corresponding classifiers, since the features are selected according to their contribution to the classification accuracy of the classifiers. A standard wrapper gene selection method can be summarized as follows:

For a given training data set $D = \{x_{ij}, y_i\} | x \in X, y \in Y, i = 1,...,n, j = 1,...,m$, pertaining to a pattern recognition task. x_{ij} and y_i denote the *jth* gene's value of sample i and the class label of sample i, respectively. The optimized gene selection method including a classifier and a subset of genes are able to maximize the prediction accuracy, i.e. obtain the maximum correctness of the mapping from input set X to output set Y. Thus, a typical wrapper gene selection method is formulated in the following way to minimize the expected risk:

$$\Lambda(f_\sigma) = \int \lambda(y, f_\sigma(\sigma x)) \mathrm{d}P(x, y)$$

where Λ is the expected risk, λ is a loss function, f_σ is a family of learning functions (e.g. a classifier or regression model), P is an evaluating function on the training data x, and σ denotes a vector indicating whether the gene i ($i = 1,...,n$) is selected or not. Wrapper gene selection methods can generally yield high classification accuracy using a particular classifier with an expensive computational cost. In wrapper method, the gene selection process is heavily dependent on a search engine, a search area (data), and an evaluation criterion for optimizing the learning model (Guyon and Elisseeff, 2003). We have found in our previous macroarray data experiments (Verma et al., 2009) that the global modeling cannot provide precise and sufficient information for a new coming data vector under different circumstances, and also the selected subset of genes are not promising to be the biomarker genes. More importantly, it is difficult to incorporate previous developed

models or existed knowledge into global modeling. For a new data sample, the whole (global) problem space usually contains much noise information that prevents the learning algorithm working properly on it, though the information may be useful for the global modeling. The noise information in the global problem space should be excluded to obtain a satisfactory result from the analysis. As discussed by Song and Kasabov (2006) and Verma et al. (2009), personalized modeling focuses on the individual sample rather than simply on the global problem space, so that creating a personalized problem space specifically for the new data can be a more appropriate solution to analyze new coming data sample in medical area. Personalized modeling is a relative new method in bioinformatics research, which is less found in literature. A representative work was published (Kasabov, 2007b). One main difficulty in gene selection procedure is the learning function optimization issue for evaluating the candidate genes during the training process. Genetic algorithm (GA) is a powerful method that is useful for exploring the combination of features and principally is able to converge to the best solution. However, classical GA is often criticized for its huge computational cost and the difficulty of parameter setting. Thus, we have employed a compact genetic algorithm (cGA) (Harik et al., 1999) to search for the optimal solution in the binary encoded problem space (gene selection), owing to its ability to converge toward the optimum significantly faster comparing to conventional GAs.

Algorithm of PMGS method

In our study (Fiasché et al., 2010a), we planned to compare with classical wrapper (global model), a new gene selection method based on personalized modeling for GEP data obtained by macroarray analysis, especially for GVHD diagnosis and prognosis. In the proposed PMGS method, we employ wrappers to search candidate gene sets, then use the selected most significant genes to profile individual data sample. This gene selection method can incorporate any classifier models for optimizing the learning function during the training process. In this study, we have investigated two classification algorithms, including Weighted distance K-nearest neighbor (WKNN) (Kasabov, 2007b) and Naïve Bayes to make a comparison. Figure 2 illustrates the process of the new proposed PMGS method using wrapper for searching candidate genes. PMGS method starts with the creation of a personalized problem space (Dpers) for an objective data sample. A statistical model (SNR) selects a subset of genes based on their ranking scores from Dpers space. Then, these candidate genes are evaluated by a classifier, which is incorporated into a cGA-based optimizer. This training process is iterated until the stopping criterion is met. The final selected genes from cGA optimizer are then used to construct the classifier for validating the testing data sample. The algorithm that outlines the process of PMGS method is formally described by Fiasché et al. (2010b).

Experiment

The goal of this study is to design a model to select a compact set of genes that can profile the pattern of objective macroarray data.

Data

Fifty-nine HSCT patients were enrolled between March 2006 and July 2008 in Transplants Regional Center of Stem Cells and Cellular Therapy "A. Neri" Reggio Calabria, Italy, during the development of a Research Program of Minister of the Health with the title: "Project of Integrated Program: Allogeneic Hemopoietic Stem Cells Transplantation in Malignant Hemopathy and Solid Neoplasia Therapy–Predictive and prognostic value for graft vs. host disease of chimerism and gene expression." Because experimental design plays a crucial role in a successful biomarker search, the first step in our design was to choose the most informative specimens in order to achieve adequate matching between positive cases GVHD (Yes) and negative controls GVHD (No) to avoid bias. This objective is best achieved through a database containing high-quality samples linked to quality controlled clinical information. Patients with clinical signs of GVHD (Yes) were enrolled, and in more than 95% of them, the disease GVHD was confirmed by biopsy including those with grade I. We used 26 samples from GVHD (Yes) patients that were taken at the time of diagnosis and we selected 33 samples from patients that did not experience GVHD (No). All together Yes/No patient groups comprised a validation set. As suggested by theoretical and practical reasons, we have chosen to assess changes of transcript genes in whole blood. Herein we report some of these reasons:

- In HSCT patients, the recovery of immune cells is very slow during the first month following stem cell transplantation (Mohty et al., 2002). Because circulating lymphocyte count is very low, it is impossible to select cell subsets by using magnetic methods.
- Using gene expression assay, it needs six hours to obtain the laboratory results starting from clinical suspicion. It is not possible and useful to add another laboratory-step in work flow-chart, indeed our aim in this project was to identify diagnostic support models with specific characteristics: easy, fast and feasible.
- Acute GVHD is triggered by immunological events, as well described in this review. As referred by Chaussabel et al. (2010), the status of the immune system can be best monitored by profiling transcripts in blood providing a "snap shot" of the complex immune networks.

Therefore, total RNA was extracted from peripheral whole-blood samples using a RNA easy Mini Kit (Qiagen) according to the manufacturer's instructions. Reverse transcription of the purified RNA was performed using Superscript III Reverse Transcriptase (Invitrogen). A multigene expression assay to test occurrence of GVHD was carried out

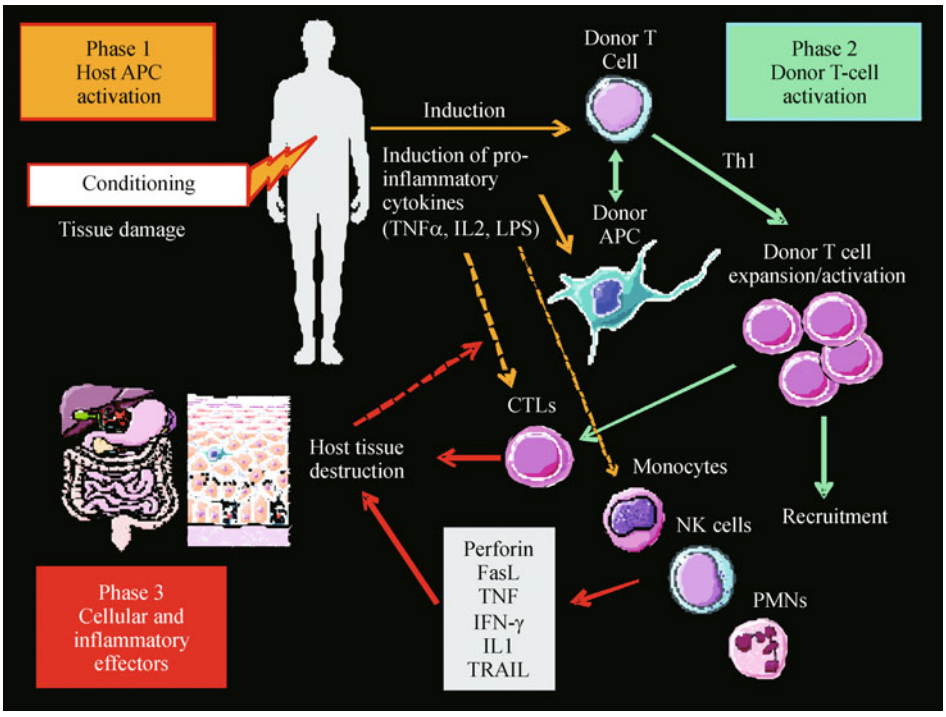

Figure 2 Three phases of GVHD immuno-pathogenesis.

with TaqMan® Low Density Array Fluidic (LDA-macroarray card) based on Applied Biosystems 7900HT comparative ddCT method, according to manufacturer's instructions. Expression of each gene was measured in triplicate, and then normalized to the reference gene 18S mRNA, that was included in macroarray card. As for the design of the macroarray card, we selected 47 candidate genes from the published literature, genomic databases, pathway analysis. The 47 candidate genes were involved in immune network and inflammation pathogenesis. Finally, a group of 7 new patients was enrolled for testing the new approach reported in the present paper. We will define this new data set where three cases have been diagnosed as aGVHD cases.

Experimental setup—wrapper approach

Results of applications of wrapping and filtering methods using classical machine learning techniques as classifiers published in previous works are shown in Table 1 (Fiasché et al., 2009, 2010a). In these applications the procedure consists of a pre-processing task (with gene selection techniques) and of classification/prediction task. Therefore the starting point has been a suitable division of the global data set in training and testing data set. The performance of the pre-processing task has been measured on the training data set while the classification/prediction performance has been evaluated on the testing data set. Herein the training data set consists of 29 patients' samples [13 GVHD (Yes) and 16 GVHD (No)] and the test data set consists of 30 patients' samples [13 GVHD (Yes) and 17 GVHD (No)]. Best results with the minimum number of genes have been obtained using SVM as classifiers. SVMs use a kernel function to implicitly map data to a high dimensional space. Then, SVMs construct the maximum margin hyperplane by solving an optimization problem on the training data. Sequential minimal optimization (SMO) (Platt, 1998) has been used in this paper to train a SVM. However, due to the high computational cost, it is not very practical to use the wrapper method to select genes for SVMs. In all methods utilized in our analyses (Fiasché et al., 2009, 2010a), the search algorithm was the best-first with forward selection, starting with the empty set of genes. The search for the best subset was based on the training data only. Once the best subset has been determined, and a classifier has been built from the training data (reduced to the best features found), the performance of that classifier was evaluated on the test data. The genes selected using wrapper methods are shown in Table 1 in comparison with gene selected from the previous analysis (Fiasché et al., 2009). A leave-one-out cross validation (LOOCV) procedure was performed to investigate the robustness of the method over the training set: in 29 runs, the subset of 5 genes was selected 29 times (100%) by the SMO. "Experimental results" section will show the performance of this technique estimated on the testing data.

Experimental setup—personalized modeling based gene selection method

The PMGS approach described in the section "Algorithm of PMGS method" is here applied for comparison with the techniques used in "Experimental setup – wrapper approach"

section. Differently from previous analysis, the final selected most important genes for each data sample may be different and selected frequency of some genes is significantly high. This means that they can be recognized highly representative of the data pattern. For example, *CASP1, FOXP3, ICOS, CD52* are the most important genes for 20 samples and *CASP1* is often present in the best subgroups (Fiasché et al., 2010a). As previously discussed, the main goal of PMGS method is to discover the personalized information for each sample (can be a patient clinical sample), rather than simply to compare the classification accuracy with published results in literature. For this purpose, PMGS is designed to be able to give a detailed profile for the new testing data sample (a new patient sample), which can contribute to clinical decision support system. Here we give an example to demonstrate how PMGS can visualize the analysis result from a data sample (a patient). For the sample 7, PMGS method selects 3 genes (*CASP1, FOXP3, ICOS*) and the classifier successfully predicts sample 7 diseased. To help visualize the result, we plot 36 nearest neighbors of sample 7 of the data set in a 3-D space (3 most important genes) (see Fig. 3). It is easy to elucidate that the sample 7 is more likely to be in the diseased group, since most of its nearest neighbors belong to diseased group. (Note: the value on x and y axis is normalized to [0, 1]).

However, for this section, a LOOCV procedure was performed to investigate the robustness of the method over the training set: in 29 runs, the personalized best subset was selected 29 times (100%). In the next section "Experimental setup—integrated approach for new data and existing models" is remarked the performance of this technique estimated on the testing data.

Experimental setup—integrated approach for new data and existing models

Herein we utilize the PMGS approach described in "Algorithm of PMGS method" section, to compare it with the techniques used in "Experimental setup— wrapper approach" section and to integrate global existing models (obtained in "Experimental setup — wrapper approach" section and that we will call M) (Fiasché et al., 2010a) with personalized models and all these models with new data (Kasabov, 2007a; Fiasché, 2010).

With this method, we assume that an existing model M performs well in part of the problem space, but there is also a new data set D that does not fit into the model. The existing model M is first used to generate a data set D0 of input–output data samples through generating input vectors (the input variables take values in their respective range of the problem space where the model performs well) and calculating their corresponding output values using M (Fiasché, 2010). The data set D0 is then used to evolve an initial model M0 with the use of an evolving system and to extract rules from it where each rule represents a local model (a prototype). We have tested the new data set D, defined in "Data" section by using this approach.

Here we have 7 patients' samples with a new situation for the clinical symptoms and for the general clinical situation. Furthermore a different modality of infusion of the cell is occurred. So, particularly in these situations the integration of the existing models is a good method. The Model M seen above, does not perform very well on the new data D as shown in Table 3, while the model is used to train on a data set D0 in a subspace of the problem space where it performs well.

Table 1 The 13 genes selected from CFS (correlation-based feature selection) with their names and meaning; the 7 genes selected through the wrapper-Naïve Bayes method (Fiasché et al., 2009) are marked with °; the 5 genes selected with SVM are marked with *

Gene name	Official full name	Immune function
BCL2A1	BCL2-related protein A1	Anti- and pro-apoptotic regulator.
CASP1°*	Caspase 1, apoptosis-related cysteine peptidase	Central role in the execution-phase of cell apoptosis.
CCL7	Chemokine (C-C motif) ligand 7	Substrate of matrix metalloproteinase 2
CD83	CD83 molecule	Dendritic cells regulation.
CXCL10°	Chemokine (C-X-C motif) ligand 10	Pleiotropic effects, including stimulation of monocytes, natural killer and T cell migration, and modulation of adhesion molecule expression.
EGR2°	Early growth response 2	Transcription factor with three tandem C2H2-type zinc fingers.
FAS	TNF receptor superfamily, member 6	Central role in the physiologic regulation of programmed cell death.
ICOS°*	Inducible T cell co-stimulator	Plays an important role in cell-cell signaling, immune responses, and regulation of cell proliferation.
IL4	Interleukin 4	Immune regulation.
IL10°*	Interleukin 10	Immune regulation.
SELP	Selectin P	Correlation with endothelial cells.
SLPI°	Stomatin (EPB72)-like 1	Elemental activities such as catalysis.
STAT6	Transducer and activator of transcription 6, interleukin-4 induced	Regulation of IL4- mediated biological responses.
*Foxp-3**	Forkhead box P3	Regulatory T cells play important roles in the maintenance control of transplantation tolerance.
CD52°*	CD52 antigen	B cell activation.

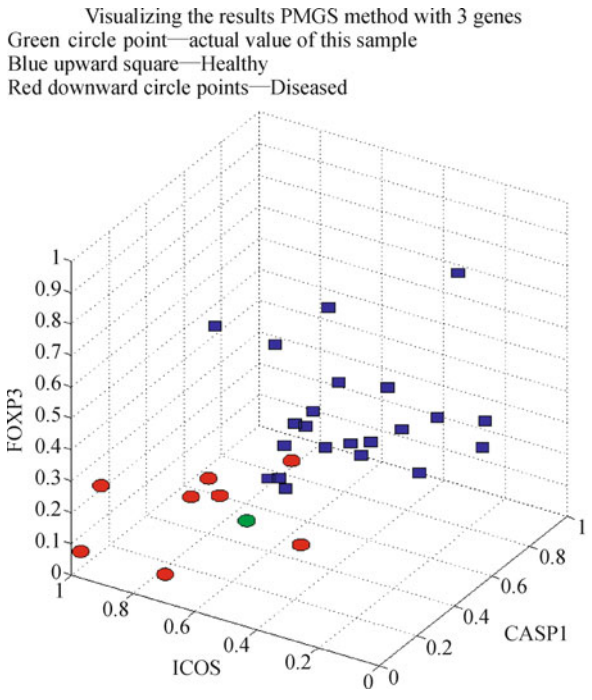

Figure 3 The profile of sample 7 in GVHD data set.

The new data set D is in another subspace of the problem space. Data D0tr extracted from D0 is first used to evolve one of the models seen above M0, and with the rule extracted the model M is transformed into equivalent local models. The model M0 is further evolved on Dtr into a new model Mnew, the first representing data D0tr and the last two, data Dtr. Although on the test data D0tst both models performed equally well, Mnew generalizes better on Dtst. Building alternative models of the same problem could help to better understand the problem and to choose the most appropriate model for the task. The new model created for each step is trained (D0tr) with the personalized model for 50 runs (9 cases have been removed since not clinically good for the training) and a LOOCV (92% for the integrated method) has been calculated in comparison with single standard methods on the new data set D of 7 patients (Table 3).

Experimental results

In Table 2 is shown that classical classifiers previously used (Fiasché et al., 2009; 2010a) obtain similar results with the PMGS method on the training data set of 29 samples and on testing data set of 30 patients. The results confirm that it is possible to classify the GVHD using a selected number of variables. Only one case escaped all our classification models, which achieved 97% accuracy in a LOOCV on the testing data set. Instead, the classical Computational Intelligence methods give results of inferior quality than new integrated approach on the new testing data D, indeed there is an evident improvement of prediction results with this integrated and online technique, with a LOOCV of 92%. In conclusion, following points 4 and 5 of bioinformatics paradigm described in "The Computational computational intelligence approach" section, it is evident that with this method, it has been possible to give support predictive results for diagnosis for each of new case, independently from medical diagnosis, simply applying this last technique on the data collected from medical team. Comparing results with classical diagnostic methods, it has been shown that accuracy of integrated approach has been very good, while with classical computational intelligence methods, there are some errors in the predicted results, technically explained from algorithm working especially from partitioning of the problem space. With the integration of existing model and new data, this problem seems solved.

Biomedical conclusions and future work

Our interest was to select fewer number of molecular biomarkers from an initial gene panel and exploiting this to develop a fast, easy and non-invasive diagnostic test.

During the study, joining biological and computational analysis it has been possible using an online tool to identify the best personalized cluster of informative genes for each patient building an important medical decision support system for diagnosis of aGVHD. The proposed method provides a good overall accuracy to confirm disease development in HSCT patients.

The subset of gene transcripts selected by computational analysis included genes responsible for cross-talk between Th2 immune effectors cells and T/B cell co-stimulatory molecules such as CD52 and ICOS.

At onset of aGVHD, we detected down-expression of IL-10, IL-4 and STAT6 (signal transducer and activator of transcription 6, regulator of transcriptional activation in response to interleukin-4). This observation may indicate a

Table 2 Experimental results of a CFS with ANN classifier and a wrapper method combined with SVM and of the PMGS with Naïve Bayes and with WKNN

Method	Training set	Test set
CFS-ANN	28(29)	29(30)
Wrapper-SVM	29(29)	29(30)
PMGS-naïve Bayes	27(29)	29(30)
PMGS-WKNN	29(29)	29(30)

The starting set has been divided into training set and test set, a leave one-out cross-validation has been calculated for the two subsets.

Table 3 Experimental results of a CFS with ANN classifier and a wrapper method combined with SVM with PMGS

Method	Test set (Dtst)
CFS-ANN	3(7)
Wrapper-SVM	4(7)
PMGS-Naïve Bayes	3(7)
PMGS-WKNN	4(7)
(i)PMGS- Naïve Bayes	6(7)
(i)PMGS- WKNN	6(7)

In the last two rows the integrated approach was marked with (i).

possible suppression of these immune pathways associated to bad immune actions. This idea is in accordence with previous study (Foley et al., 2008) that reported a GVHD-protective role mediated by IL-4 and IL-10.

In similar way, we detected a down-expression of CD52 and ICOS transcripts. These findings are noteworthy given that co-stimulatory molecules expressed on T cells critically regulate donor T cell activation and play a critical role in the development of aGVHD (Fujimura et al., 2010). Most importantly, as previously discussed, they have considerable therapeutic potential for GVHD. Indeed, low dose of alemtuzumab (Campath-1H), an anti-CD52 monoclonal antibody, can successfully treat steroid-refractory aGVHD (Varadi et al., 1996; Schub et al., 2011). Furthermore, we highlight the fact that defective expression of ICOS can impair the immune-protective effectors during GVHD. In support of this possibility, a previous and a recent report indicated ICOS as a regulatory molecule for T cell responses during GVHD. In particular, it has been shown that ICOS signal inhibits GVHD development mediated by CD8 positive effector cells in HSCT (Yu et al., 2006).

According to previous reports, mediators of apoptosis cells and dendritic cell activators were also involved. Recently, two specific molecules, namely CXCL10 and CCL7, were implicated in network of immune cell interactions in aGVHD (Paczesny et al., 2010b).

These chemokines mediate the T cell trafficking to target organ during aGVHD development. Our study draws attention to increased expression of CXCL10 and CCL7 as informative biomarker of alloreactive disease. This result confirms early Piper's observation about CXCL10 role in the pathogenesis of GVHD in the skin (Piper et al., 2007).

Altogether, our results strongly outlined the importance and utility of non-invasive tool for aGVHD diagnosis based on GEP. We believe that to achieve benefit from GEP analysis, it is very important to determine:

1) the transcript levels of immune effector cells in early time post-engraftment in order to better recognize polarization of Th2 cells;

2) the transcripts correlated with the CD8 positive cell action. In a recent report, Kato et al. (2010) outlined the role of CD8 T effector cells upon chronic exposure to alloantigens demonstrating that alloantigenic stimuli rather than homeostatic factors are critical to sustaining continuous proliferation of alloreactive CD8 T cells during GVHD;

3) the expression level of Foxp-3 transcript gene, surrogate marker of Treg cells as indicator of immune-suppressive control action during aGVHD. Miura et al. (2004) demonstrated the association of Foxp3 regulatory gene expression with GVHD. Wolf et al. (2007) demonstrated that Treg frequencies decreased linearly with increasing grades of GVHD at onset, and correlated with eventual maximum grade of GVHD.

In conclusion, our assessment model may prevent the need for an invasive procedure for diagnosis of GVHD. Indeed, this study demonstrates that the proposed integrated methodology for the personalized selection of gene diagnostic targets and their use for support in diagnosis of GVHD results in a satisfactory 92% accuracy over independent test data set of HSCT population. We plan to extend the system as a personalized model including all clinical and genetic variables, testing with new data samples this method and for a larger group of patients to capture their peculiarity. As a classifier, a spiking neural network can be explored (Kasabov, 2010). The authors are engaged in this direction.

References

Ayuk F, Diyachenko G, Zabelina T, Panse J, Wolschke C, Eiermann T, Binder T, Fehse B, Erttmann R, Kabisch H, Bacher U, Kröger N, Zander A R (2008). Anti-thymocyte globulin overcomes the negative impact of HLA mismatching in transplantation from unrelated donors. Exp Hematol, 36(8): 1047–1054

Bacigalupo A, Lamparelli T, Milone G, Sormani M P, Ciceri F, Peccatori J, Locasciulli A, Majolino I, Di Bartolomeo P, Mazza F, Sacchi N, Pollicheni S, Pinto V, Van Lint M T, the Gruppo Italiano Trapianto Midollo Osseo (GITMO) (2010). Pre-emptive treatment of acute GVHD: a randomized multicenter trial of rabbit anti-thymocyte globulin, given on day + 7 after alternative donor transplants. Bone Marrow Transplant, 45(2): 385–391

Beilhack A, Schulz S, Baker J, Beilhack G F, Wieland C B, Herman E I, Baker E M, Cao Y A, Contag C H, Negrin R S (2005). *In vivo* analyses of early events in acute graft-versus-host disease reveal sequential infiltration of T-cell subsets. Blood, 106(3): 1113–1122

Brodoefel H, Bethge W, Vogel M, Fenchel M, Faul C, Wehrmann M, Claussen C, Horger M (2010). Early and late-onset acute GVHD following hematopoietic cell transplantation: CT features of gastrointestinal involvement with clinical and pathological correlation. Eur J Radiol, 73(3): 594–600

Buzzeo M P, Yang J, Casella G, Reddy V (2008). A preliminary gene expression profile of acute graft-versus-host disease. Cell Transplant, 17(5): 489–494

Chaussabel D, Pascual V, Banchereau J (2010). Assessing the human immune system through blood transcriptomics. BMC Biol, 8(1): 84–98

Ferrara J L (2008). Advances in the clinical management of GVHD. Best Pract Res Clin Haematol, 21(4): 677–682

Ferrara J L, Levine J E, Reddy P, Holler E (2009). Graft-versus-host

disease. Lancet, 373(9674): 1550–1561

Ferrara J L, Reddy P (2006). Pathophysiology of graft-versus-host disease. Semin Hematol, 43(1): 3–10

Fiasché M (2010). Implementations of Evolving Integrated Multimodel Systems, Algorithms and Applications in Biomedical Field. DissertationTip. DIMET, University "Mediterranea" of Reggio Calabria

Fiasché M, Cuzzola M, Fedele R, Iacopino P, Morabito F C (2010a). Machine Learning and Personalized Modelling based Gene Selection for acute GvHD Gene Expression Data Analysis. In: Artificial Neural Networks — proceedings of ICANN 2010 Part I, LNCS6352

Fiasché M, Cuzzola M, Iacopino P, Kasabov N, Morabito F C (2010b). Personalized Modelling based Gene Selection for acute GvHD Gene Expression Data Analysis: a Computational Framework Proposed. Australian Journal of Intelligent Information Processing Systems, 12 (4): Machine Learning Applications (Part II)

Fiasché M, Verma A, Cuzzola M, Iacopino P, Kasabov N, 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, Berlin: Springer Heidelberg, pp. 10–19

Firoz B F, Lee S J, Nghiem P, Qureshi A A (2006). Role of skin biopsy to confirm suspected acute graft-vs-host disease: results of decision analysis. Arch Dermatol, 142(2): 175–182

Foley J E, Mariotti J, Ryan K, Eckhaus M, Fowler D H (2008). Th2 cell therapy of established acute graft-versus-host disease requires IL-4 and IL-10 and is abrogated by IL-2 or host-type antigen-presenting cells. Biol Blood Marrow Transplant, 14(9): 959–972

Fujimura J, Takeda K, Kaduka Y, Saito M, Akiba H, Yagita H, Yamashiro Y, Shimizu T, Okumura K (2010). Contribution of B7RP-1/ICOS co-stimulation to lethal acute GVHD. Pediatr Transplant, 14 (4): 540–548

Guyon I, Elisseeff A (2003). An introduction to variable and feature selection. J Mach Learn Res, 3(7–8): 1157–1182

Harik G R, Lobo F G, Goldberg D E (1999). The compact genetic algorithm. IEEE Trans Evol Comput, 3(4): 287–297

Ho V T, Cutler C (2008). Current and novel therapies in acute GVHD. Best Pract Res Clin Haematol, 21(2): 223–237

Hu Y, Song Q, Kasabov N (2009). Personalized modelling based gene selection for microarray data analysis. In: Köppen M, Kasabov N, Coghill G (eds.). ICONIP 2008, Part I. LNCS. Springer, Heidelberg, vol. 5506, pp. 1221–1228

Jacobsohn D A, Vogelsang G B (2007). Acute graft versus host disease. Orphanet J Rare Dis, 2(1): 35–44

Lord J D, Hackman R C, Gooley T A, Wood B L, Moklebust A C, Hockenbery D M, Steinbach G, Ziegler S F, McDonald G B (2011). Blood and gastric FOXP3$^+$ T cells are not decreased in human gastric graft versus host disease. Biol Blood Marrow Transplant, 17(4): 486–496

Kasabov N (2007a). Evolving Connectionist Systems: The Knowledge Engineering Approach, 2nd ed. Springer, London

Kasabov N (2007b). Global, local and personalized modelling and profile discovery in bioinformatics: An integrated approach. Pattern Recognit Lett, 28(6): 673–685

Kasabov N (2010). To spike or not to spike: A probabilistic spiking neural model. Neural Networks, 23(1): 16–19

Kato K, Cui S, Kuick R, Mineishi S, Hexner E, Ferrara J L M, Emerson S G, Zhang Y (2010). Identification of stem cell transcriptional programs normally expressed in embryonic and neural stem cells in alloreactive CD8$^+$ T cells mediating graft-versus-host disease. Biol Blood Marrow Transplant, 16(6): 751–771

Langley P (1994). Selection of relevant features in machine learning, In: Proceedings of AAAI Fall Symposium on Relevance, 140–144

Magenau J M, Qin X, Tawara I, Rogers C E, Kitko C, Schlough M, Bickley D, Braun T M, Jang P S, Lowler K P, Jones D M, Choi S W, Reddy P, Mineishi S, Levine J E, Ferrara J L, Paczesny S (2010). Frequency of CD4$^+$CD25(hi)FOXP3$^+$ regulatory T cells has diagnostic and prognostic value as a biomarker for acute graft-versus-host-disease. Biol Blood Marrow Transplant, 16(7): 907–914

Miura Y, Thoburn C J, Bright E C, Phelps M L, Shin T, Matsui E C, Matsui W H, Arai S, Fuchs E J, Vogelsang G B, Jones R J, Hess A D (2004). Association of Foxp3 regulatory gene expression with graft-versus-host disease. Blood, 104(7): 2187–2193

Mohty M, Gaugler B, Faucher C, Sainty D, Lafage-Pochitaloff M, Vey N, Bouabdallah R, Arnoulet C, Gastaut JA, Viret F, Wolfers J, Maraninchi D, Blaise D, Olive D (2002). Recovery of lymphocyte and dendritic cell subsets following reduced intensity allogeneic bone marrow transplantation. Hematology, 7(3): 157–64

Paczesny S, Braun T M, Levine J E, Hogan J, Crawford J, Coffing B, Olsen S, Choi S W, Wang H, Faca V, Pitteri S, Zhang Q, Chin A, Kitko C, Mineishi S, Yanik G, Peres E, Hanauer D, Wang Y, Reddy P, Hanash S, Ferrara J L M (2010a). Elafin is a biomarker of graft-versus-host disease of the skin. Sci Transl Med, 2(13): ra2

Paczesny S, Hanauer D, Sun Y, Reddy P (2010b). New perspectives on the biology of acute GVHD. Bone Marrow Transplant, 45(1): 1–11

Paczesny S, Krijanovski O I, Braun T M, Choi S W, Clouthier S G, Kuick R, Misek D E, Cooke K R, Kitko C L, Weyand A, Bickley D, Jones D, Whitfield J, Reddy P, Levine J E, Hanash S M, Ferrara J L M (2009a). A biomarker panel for acute graft-versus-host disease. Blood, 113(2): 273–278

Paczesny S, Levine J E, Braun T M, Ferrara J L M (2009b). Plasma biomarkers in graft-versus-host disease: a new era? Biol Blood Marrow Transplant, 15(Suppl): 33–38

Payne D K, Sullivan M D, Massie M J (1996). Women's psychological reactions to breast cancer. Semin Oncol, 23(Suppl 2): 89–97

Piper K P, Horlock C, Curnow S J, Arrazi J, Nicholls S, Mahendra P, Craddock C, Moss P A H (2007). CXCL10-CXCR3 interactions play an important role in the pathogenesis of acute graft-versus-host disease in the skin following allogeneic stem-cell transplantation. Blood, 110(12): 3827–3832

Platt J (1998). Fast training of support vector machines using sequential minimal optimization. Advances in Kernel Methods–Support Vector Learning. MIT Press

Schub N, Günther A, Schrauder A, Claviez A , Ehlert C, Gramatzki M, Repp R (2011). Therapy of steroid-refractory acute GVHD with CD52 antibody alemtuzumab is effective. Bone Marrow Tranplant, 46: 143–147

Socié G (2009). Graft-versus-host disease: proteomics comes of age. Blood, 113(2): 271–272

Socié G, Blazar B R (2009). Acute graft-versus-host disease: from the bench to the bedside. Blood, 114(20): 4327–4336

Song Q, Kasabov N (2006). TWNFI — Transductive weighted neuro-fuzzy inference system and applications for personalized modelling. Neural Netw, 19(10): 159–596

van Bekkum D W, Vries De M J (1967). Radiation Chimaeras, London: Logos Press

Varadi G, Or R, Slavin S, Nagler A (1996). *In vivo* CAMPATH-1 monoclonal antibodies: a novel mode of therapy for acute graft-versus-host disease. Am J Hematol, 52(3): 236–237

Verma A, Fiasché M, Cuzzola M, Iacopino P, Morabito F C, Kasabov N (2009). Ontology Based Personalized Modelling for Type 2 Diabetes Risk Analysis: An Integrated Approach. In: Leung C S, Lee M, Chan J H (Eds.): ICONIP 2009, Part II, LNCS 5864, Springer, Heidelberg, pp. 360–366

Weissinger E M, Schiffer E, Hertenstein B, Ferrara J L, Holler E, Stadler M, Kolb H J, Zander A, Zürbig P, Kellmann M, Ganser A (2007). Proteomic patterns predict acute graft-versus-host disease after allogeneic hematopoietic stem cell transplantation. Blood, 109(12): 5511–5519

Wolf D, Wolf A M, Fong D, Rumpold H, Strasak A, Clausen J, Nachbaur D (2007). Regulatory T-cells in the graft and the risk of acute graft-versus-host disease after allogeneic stem cell transplantation. Transplantation, 83(8): 1107–1113

Yu X Z, Liang Y, Nurieva R I, Guo F, Anasetti C, Dong C (2006). Opposing effects of ICOS on graft-versus-host disease mediated by CD4 and CD8 T cells. J Immunol, 176(12): 7394–7401