Analysing Microarray Data using the Multi-functional Immune Ontologiser

Sabah Khalid[1, 2], Karl Fraser[2], Mohsin Khan[1], Ping Wang[3], Xiaohui Liu[2], Suling Li[1, *]

[1]Molecular Immunology Group, Microarray Facility, Division of BioSciences, Brunel University, Uxbridge, UB8 3PH, UK
[2]Intelligent Data Analysis Group, Department of Information Systems and Computing, Brunel University, Uxbridge, UB8 3PH, UK
[3]Immunology Group, Institute of Cell and Molecular Sciences, Barts and London School of Medicine, London, UK

Abstract

Gene expression microarrays are a prominent experimental tool in functional genomics allowing researchers to gain a deeper understanding of biological processes. To date, no such tool has been developed to allow researchers with a specialised biological research interest to distinctively identify those genes and gene functionalities associated more strongly with the research area. Based on this functional analysis capability we present a specialised multi-functional Immune Ontologiser – a software, specialised for immunologists to annotate multiple genes from microarray datasets within two new ontologies: a newly structured Immune Ontology focussed at immunology and haematology and a uniquely curated ImmunoArray-PubOntology. The Immune Ontology functionally annotates genes identifying immunology related functions enriched with upregulated or downregulated genes of interest. The ImmunoArray-PubOntology compares and contrasts gene functionality of microarray datasets, comparing genes of interest with the differential gene expression matrices published amongst immunology-related microarray literature. This aspect facilitates literature mining by extracting publications containing gene sets of interest in a well-structured immunological context where the literature has been categorised according to disease types. The software consists of a query-optimised database of two parts – the ImmunoGene-database and a unique Database of Immunological Microarray Publications (DIMP) to provide the user with a more detailed insight into other studies involving their genes and research groups investigating similar research areas. Using our Immune Ontologiser software to analyse tolerance array data we identify 70 interesting up-regulated genes in terms of their functionality within tolerance. Furthermore, from these 70 genes we identify 15 genes to have immunology-related functions. More interestingly, the remaining 55 genes were not previously known to be directly involved within the immunology related condition and hence we have identified target genes for future investigation. Among the 70 genes, 21 have been identified by our software to be studied within various immunology-related diseases via microarray experiments performed by other laboratories.

The software and database schema is freely available at ftp://ftp.brunel.ac.uk/cspgssk. Additional material is available online at http://www.brunel.ac.uk/about/acad/health/healthres/researchareas/mi/publications/supplementary. A detailed microarray protocol is available at http://www.ebi.ac.uk/arrayexpress under the Accession Number: E-MEXP-283.

Key words: ontology, immunology, literature mining, microarrays, immune tolerance.

* Email: Dr Suling Li - Su-ling.li@brunel.ac.uk

1 Introduction

The advent of microarray technology has opened up a gateway for researchers to concurrently measure gene expression levels for thousands of genes (even entire genomes) in a single experiment [1]. Yet, even the simplest microarray experiments generate vast amounts of data, making it difficult to comprehend biologically. Consequently it becomes a necessity for researchers to mine the generated data in order to extract biological meaning, which is apt for the specific research area being investigated. Furthermore it is important to determine the genes, which have been differentially expressed within the data. This process of data mining ensures that the initial microarray-generated dataset is narrowed down to include only those genes that are of most interest to the researcher. Having identified such interesting genes, the next challenge is to relate the genes of interest to their respective biological categories in order to make their biological functionalities known to the researcher.

The extraction of such functional information can be obtained using ontologies, which provide a structured description of biological information that is extremely useful for computational management [2]. One of the most widely used ontologies is the Gene Ontology (GO) developed by the Gene Ontology Consortium [3,4], which categorises genes according to their biological process, molecular function and cellular localisation. Taking the importance of functional genomics for microarray data into consideration, there exist several software systems that exploit the GO, designed with the purpose of functionally annotating any given set of differentially expressed genes into their respective functional categories. Examples of such software include OntoExpress [5,6], MAPPFinder [7,8], GoMiner [9,10] and DAVID-Database for Annotation, Visualisation and Integrated Discovery [11,12].

Although the functional data generated for individual genes from the aforementioned software use the extremely comprehensive GO database, further mining and selection of genes is imperative before embarking upon more detailed analysis of the interesting genes. This is essential as it is not feasible to explore every gene with an associated biological function. In the same way that a microarray experiment identifies valuable genes giving crucial direction for further research, functional annotation should govern an even more detailed analysis of the interesting genes. This is important as researchers are generally extensively involved with investigating a particular area of biology and hence ultimately interested in genes and gene functionalities that are highly related to their research, leading to more in-depth work.

In light of this, software such as those mentioned above pose two specific problems. First, it is illogical and laborious to select such interesting genes manually by sifting through the entire functional data generated, especially since tools such as GoMiner can generate up to 20,000 gene functionalities. Second, the scope of these tools is general, as they do not functionally characterise genes in a specific context (i.e. by autonomously categorising genes that fall within a specific branch of biology such as immunology or oncology). The latter situation becomes problematic when the immunologist for example, wishes to functionally categorise genes that are known to be involved in immunology only. If current tools were employed one would be compelled to search through several hundred functional categories by eye amongst a complex GO hierarchy, in order to determine which are immunology-related. This is no doubt a time consuming process. Furthermore, examining the GO for specific gene functionalities becomes extremely tedious due to its polyhierarchical structure in which GO functions and their respective genes are repeated several times. For these reasons, specialised software are required, which would reduce such burden from researchers interested in functionally categorising genes that belong to specific research disciplines.

Following the completion of a microarray experiment the resulting biological discoveries are stored in descriptive full text, which has resulted in the literature becoming the most useful

source of knowledge. Many literature-mining tools have been developed in order to allow researchers to retrieve other literature containing the same genes i.e. place genes of interest into context relative to published medical literature. Currently, access to scientific abstracts within the main biological literature database, PubMed consisting of 15 million abstracts is via keyword search [13]. However, due to the wealth of biological information available and the continuous increase in size of scientific literature databases, the number of entries retrieved from keyword searches, may be huge. Although the user may find abstracts relating to scientific research that are useful, they may be buried amongst hundreds that are irrelevant, through which the user would have to sift through and filter [14]. Similarly, with respect to molecular experiments using multiplex strategies such as gene expression microarrays [15], such keyword searches are of limited use when the user is interested in identifying relevant literature for perhaps hundreds of genes (as a result of their own microarray experiments). Some current applications, such as PubMed or MedMiner [13] are only able to display abstracts based on one gene, whereas others such as PubMatrix [15] accept no more than 100 genes. Taking into consideration the amount of information that can be acquired from literature mining coupled with the increasing use of gene expression microarrays, there is currently no tool to provide a structured literature output based on a multiple gene process that is only associated with a researchers area of expertise.

To this end, we have developed the multi-functional Immune Ontologiser software, which automatically categorises and annotates microarray gene expression clusters of interest at the general biological level called the Bio-Ontology, as well as identifying biological processes more related to immunology/haematology within an Immune-Ontology that possesses specificity, all at the click of a button. The Immune Ontologiser has a unique function, which focuses on literature mining. This aspect is designed to simultaneously compare an unlimited number of genes of interest from a microarray experiment with interesting gene expression datasets published in microarray-based literature related to immunology. The software not only retrieves the relevant literature, it also presents all of the interesting differentially expressed genes listed within the publication and more specifically highlights the users genes of interest. Each study identified is supported by a hyperlink to the publication in PubMed allowing further analysis of the genes involved. Thus this functionality is aimed at providing the user with a detailed insight into the other immunological studies involving their genes whilst also providing information about gene functionality amongst other diseases. Lastly, the Immune Ontologiser comprises an advantageous feature called differential gene expression analysis. Selecting informative genes from microarray experiments is one of the most important data analysis steps for deciphering biological information embedded within the experiment. The differential gene expression analysis feature aids this selection process allowing the import of microarray gene expression data and in turn the automated identification of differentially expressed genes based on a user-defined threshold. This saves the biologist from performing this task in a traditional spreadsheet program. Accepting genes from any species, this feature of the software can be used as a stand-alone function.

2 Methods

The Immune Ontologiser uses a MySQL [16] relational database that is used to generate the hierarchical trees and the parent-child relationships within them. This section describes this database schema together with the data collection techniques.

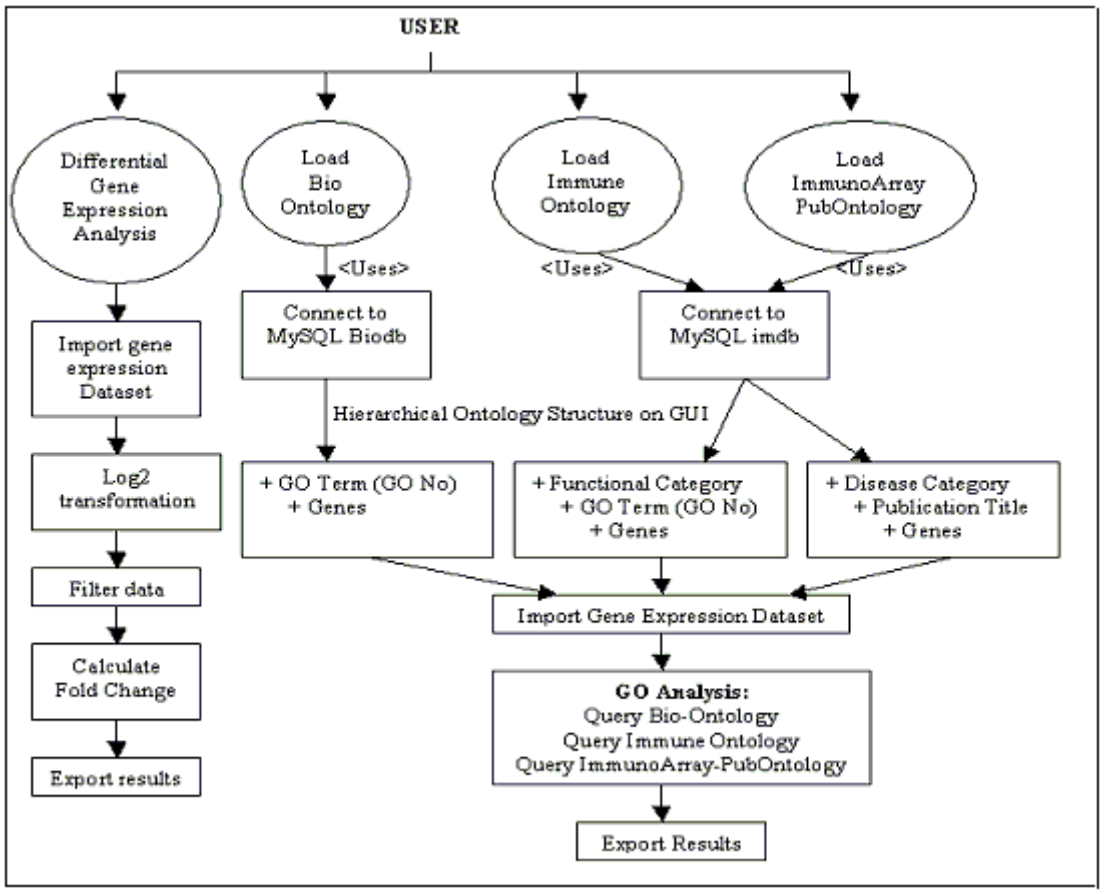


Figure 1. The structure of the Immune Ontologiser modelling its functionality. The Immune Ontologiser consists of four panels. The "Differential Gene Expression Analysis" panel requires the import of gene expression data, log2 transformation, and data filtering and finally the identification of differentially expressed genes. The "Bio-Ontology", "Immune-Ontology" and ImmunoArray-PubOntology" connect to the underlying MySQL database representing the data as various hierarchical trees on the interface. This is followed by the querying of the hierarchical tree type of interest with the users microarray gene list.

2.1 Implementation

Processed data from the Immune Ontologiser is presented on a multiple panel graphical user interface (GUI) in a format that provides the user with a flexible and intuitive view of their queried data. The GUI is user-friendly interacting with users via menus, mouse clicks and user-input dialogs and consists of two panels, each with a set functionality procedure and output (Figure 1). The first panel called "Differential Gene Expression Analysis" is used to automatically identify differentially expressed genes according to user-defined criteria following a microarray experiment. This has traditionally involved a spreadsheet package with the user manipulating the gene expression values. To calculate differential gene expression, the microarray data must be imported as a tabbed delimited text file where the gene expression ratios have been normalised and filtered. When the data is imported it is simultaneously log transformed to base2 (log2). This transformation is a requirement for concluding valid differential gene expression results for microarray data. The user has the advantage to define a significant log threshold and hence determine the genes for which the fold change is to be calculated upon. Any genes for which the fold change is not calculated because the log threshold criterion is out of range display "NaN". Alternatively the Immune Ontologiser can display the fold changes for the entire dataset. Sorting the fold change column allows the analyst to select the fold change above which the gene expression between the two microarrays is considered differential. Such genes are considered significantly

differentially expressed and hence are the basis for further biological investigation. The first four columns of the text file must represent the accession number of the gene, a short gene description, and the gene expression ratios from the first and second microarray experiment, respectively.

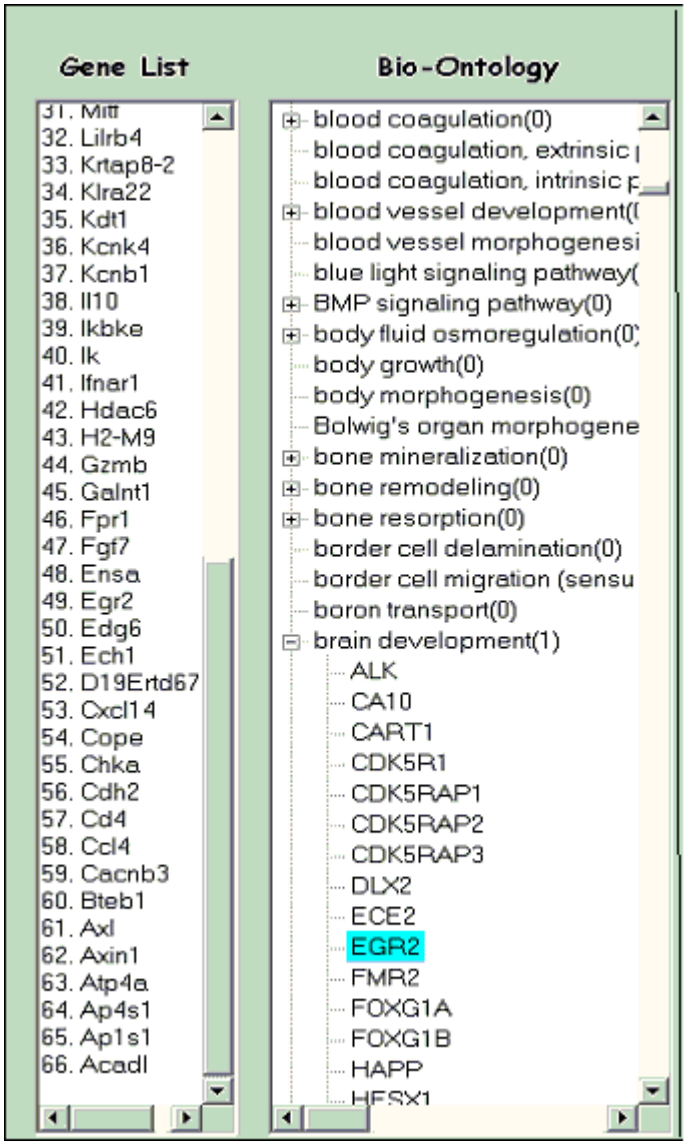


Figure 2. The Bio-Ontology hierarchy. The genes highlighted within the Bio-Ontology are those from the imported microarray gene list. The gene, *EGR2* has functionality in brain development. The un-highlighted genes are also known to be involved in brain development, but are not found within the imported microarray gene list.

The second panel is used to generate three types of trees depending on the nature of the queries in the form of a hierarchical treeview structure when the underlying MySQL database of the Immune Ontologiser is queried. The hierarchies are generated individually and presented on the GUI each time a new gene list is to be imported which involves placing genes at the nodes describing the gene function. The treeview structures are individually searched to find and highlight the genes of interest imported by the user. They are labelled: Bio-Ontology, Immune-Ontology and ImmunoArray-PubOntology for functional gene analysis and literature searching (Figure 1). The Bio-Ontology created using our MySQL Bio-database is aimed at giving an insight into the underlying biological functions of genes from microarray gene expression datasets for *Homo Sapiens* and *Mus Musculus* (Figure 2). Each biological function within our Bio-Ontology hierarchy is a parent node presented as a

Gene Ontology term (GO term) together with associated genes being presented as child nodes. Whereas the Gene Ontology consists of GO term's that are repeated amongst a complex polyhierarchical structure, our Bio-Ontology is structured not to contain such heavily nested, multi-level GO-term-GO-term relationships that are normally found embedded within the GO hierarchy. Using this simplified approach we create a user-friendly representation of the GO data in which the repetitive visualisation of GO terms and their corresponding genes is avoided.

The Immune Ontology for functional annotation is derived from our ImmunoGene database and displays immunology/haematology related biological categories for genes of interest (Figure 3). This aspect is advantageous for researchers who are interested in the specific immunology or haematology related biological functions of their interesting genes or those who are only interested in those genes from an interesting gene set that are more functionally involved in immunology or haematology. Thus, helping them to extract precise biological meaning from their data and discover novel functions for their interesting genes, e.g. identifying gene functionalities that are not previously known to be involved within the disease under investigation or to guide further in-depth research within the specialised research interest. The ultimate goal of such analyses is to enable researchers to delve deeper into the underlying molecular mechanisms and understand the intricate details regulating a particular biological process of interest. The current version of the ImmunoGene-database consists of 1630 human and 2113 mouse immunology/haematology related biological processes clustered into 27 distinct sub-groups. The ImmunoGene database contains a total of 1926 human genes and a total of 2043 mouse genes known related to immunology/haematology (Table 1). The parent nodes represent the 27 immunology related biological functional categories, whereas the children nodes represent functional sub-groups of more specific biological processes belonging to each parent category. The query returns a count beside each of the 27 functional categories, representing the total number of gene appearances from a user's gene list, thereby allowing the user to identify the immunological category most enriched with their genes of interest.

Table 1. 27 Parent biological categories and corresponding children sub-groups contained within the Immune Ontology

Functional Category	No. Of Sub Groups Human/ Mouse	No. Of Genes* Human/ Mouse	Functional Category	No. Of Sub Groups Human/ Mouse	No. Of Genes* Human/ Mouse
Behaviour	3/3	38/157	Death	55/60	1981/2697
Biological Process Unknown	1/0	20/0	Detection and Response to Stimulus	92/101	4076/4402
Blood and Circulation	53/72	352/452	Development	60/63	1746/2006
Bone	20/20	84/312	Enzyme Activity	15/26	207/661
Cell Activation	100/107	684/1659	Homeostasis	28/39	263/655
Cell Adhesion	15/17	347/348	Hormone	25/22	78/86
Cell Cycle	73/103	1220/1682	Immune Response	72/104	1315/1694
Cell differentiation	30/28	275/709	Localisation	121/118	1127/2295
Cell growth and/or maintenance	52/46	1584/1193	Metabolism	473/428	6876/9963
Cell motility	8/11	204/260	Signal Transduction	118/151	2996/3646
Cell organisation and biogenesis	35/48	179/293	Transcription	27/31	1431/1817
Cell-cell signalling	4/3	195/242	Translation	7/15	51/65
Central nervous system	45/65	292/555	Viral Life cycle	25/9	80/51
Cytokine Production	73/78	373/635			

*** The number of genes involved within each functional category including duplicate gene entries arising from one gene belonging to more than one sub-group. The total number of genes including duplicates within all 27 categories is 28074 for human and 38535 for mouse.**

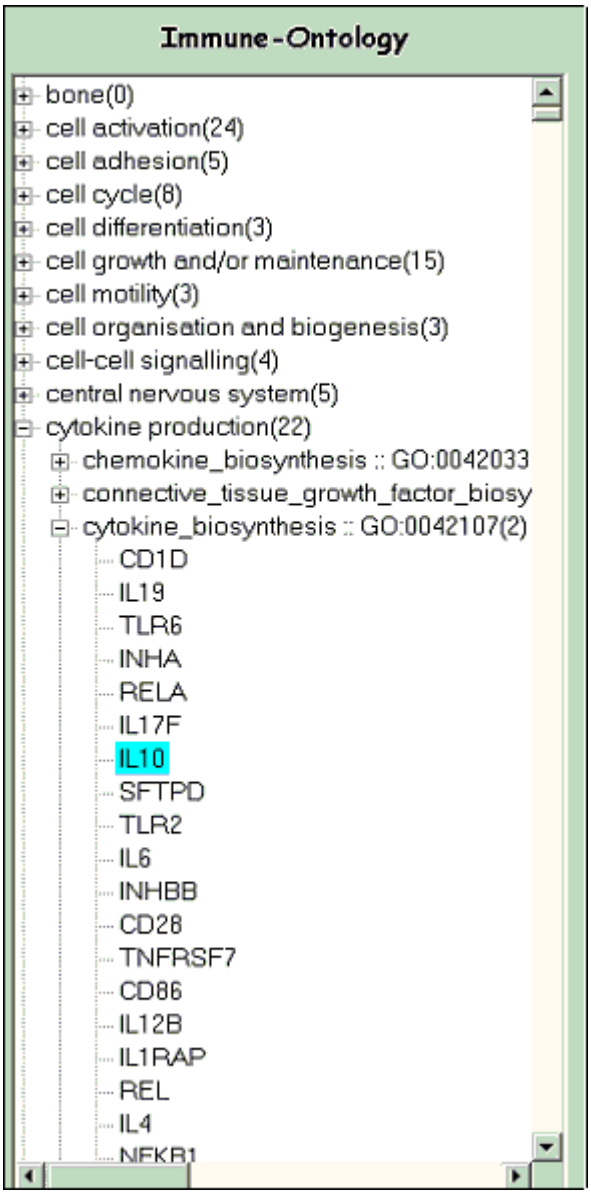


Figure 3. The Immune-Ontology to identify immunology/haematology related processes. The *IL10* gene from the microarray gene list has functionality in cytokine production. More specifically, it plays a role in cytokine biosynthesis (GO:0042107).

Our unique ImmunoArray-PubOntology enables the extraction of multiple immunology related published literature containing interesting genes identified by a researcher's microarray experiment allowing a gene expression comparison. More specifically, genes considered to be significant are simultaneously compared and identified within the interesting differential gene expression matrices published within the literature. Hence, giving insights into the involvement of genes within various immunology related disease(s) or within diseases affecting genes involved in the immune response. The underlying Database of Immunological Microarray-related Publications (DIMP) contains articles derived from the main biological literature database, PubMed [17]. It is designed to contain literature specifically related to immunology diseases in which the genes involved have been identified using human or mouse samples through microarray technology. The literature is organised by grouping publications according to immunological disease types. The parent nodes represent the immunological disease category and the child nodes represent the publications pertaining to the parent category, presented as hyper-links to the article itself for further information. Taking the mouse over a publication title brings up an information window displaying the

complete reference of the paper leading to the abstract. Expanding the child nodes reveal the genes as HUGO identifiers that have been identified in the study under investigation (Figure 4). All results obtained from querying the ontologies can be exported in Excel format.

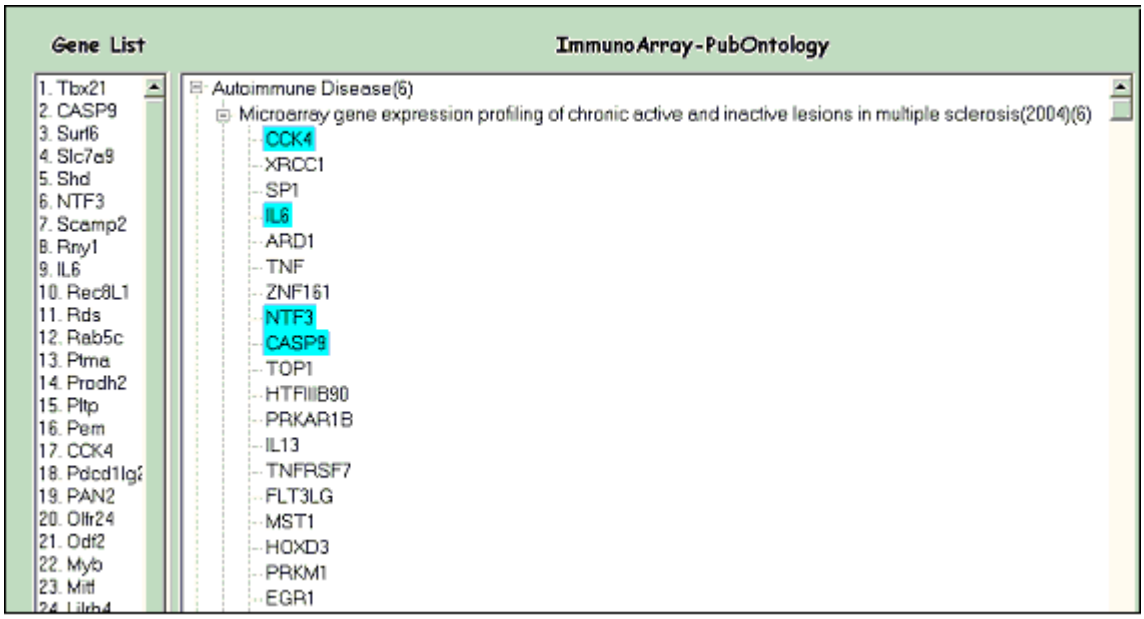


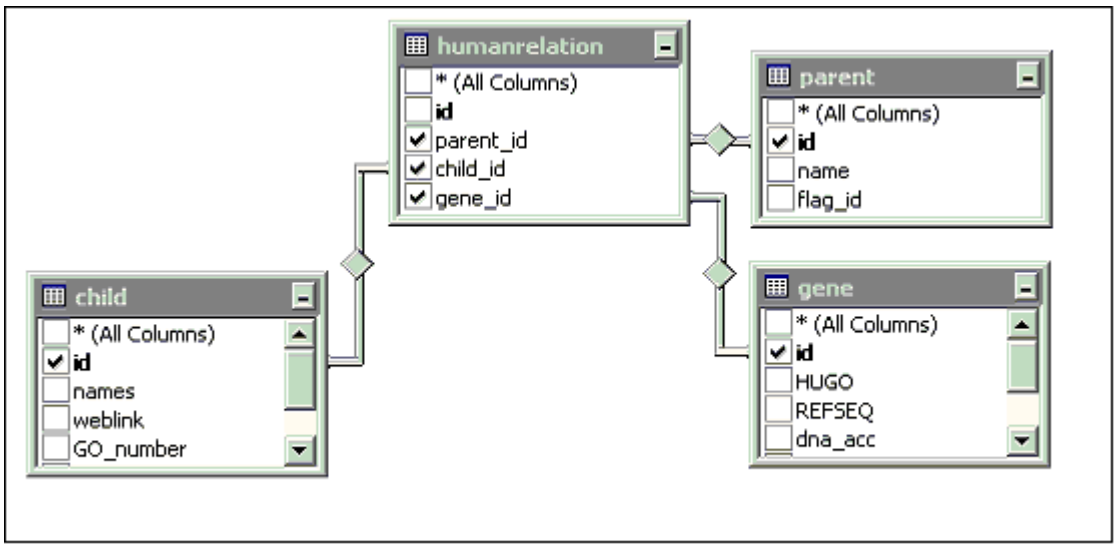
Figure 4. Hierarchical structure of the ImmunoArray-PubOntology

From the imported microarray gene list, *CCK4*, *IL6*, *NTF3* and *CASP9* have also been identified in the study investigating microarray gene expression profiling of chronic active and inactive lesion in multiple sclerosis, belonging to the category autoimmune disease. In addition, we have designed a flexible database allowing the user to insert, update or delete old records. Such changes are possible once the user has downloaded a local copy of the database.

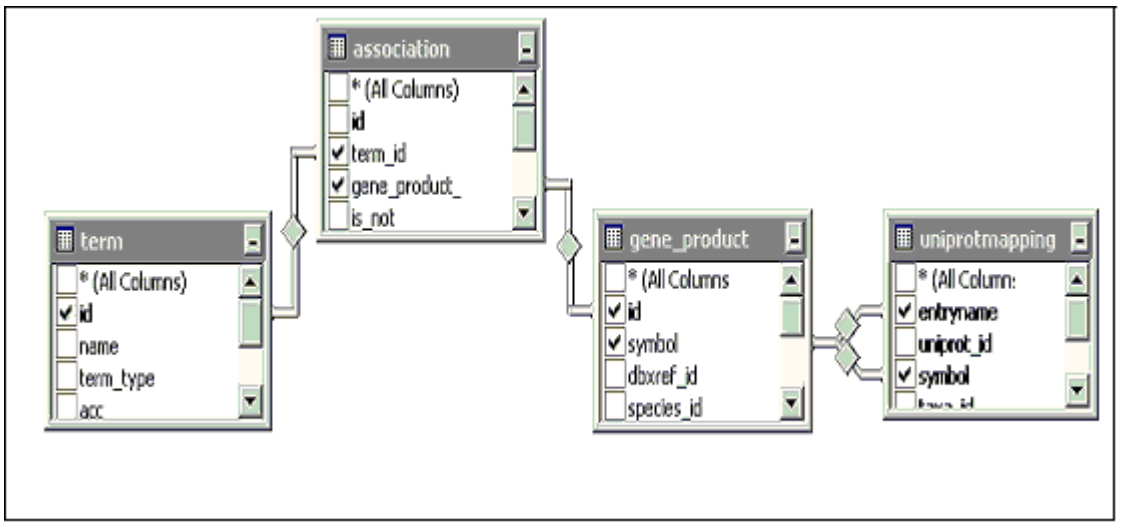
2.2 System Architecture

The Immune Ontologiser has been implemented in Visual Basic.Net and uses a relational database comprising tables to store all gene ontology data implemented in MySQL. This database is searched when the user imports a microarray gene expression dataset and queries an ontology type. The database schema representing the ImmunoGene and Bio-database is highly optimised for rapid querying and retrieval of information for large numbers of genes and compatible with a MySQL database server no later than 3.23.58 (Figure 5a and 5b). The ImmunoGene database tables are used to store the data required to build the Immune Ontology and ImmunoArray-PubOntology and are named: parent, child, gene and human/mouse relation. The parent, child and gene table represent the main functional category, sub-category and genes with respect to the Immune Ontology and immunological disease category, publications titles and genes with respect to the ImmunoArray-PubOntology. The relationship between the parent and child is one-to-many and between the child and gene table is many-to-many. For this reason, a relation table is used to connect the parent, child and gene tables to their respective nodes, and whose primary keys are foreign keys in the relation table. The bio-database comprises the uniprotmapping, gene_product, term and association tables used to store the gene ontology data required to build the Bio-Ontology. The term and gene_product tables represent the biological functions displayed within the ontology and the genes associated with each term, respectively. The Visual Basic.Net application communicates with the MySQL database through an ODBC MySQL

connector available from http://www.mysql.com displaying the data in an appropriate format on the GUI providing the user with a flexible and intuitive view of biological relationships (Figure 2 – 4).



a) The parent, child and gene tables are linked via the humanrelation table within the MySQL database. The column "id" is a primary key in the parent, child and gene table and foreign keys in the relation2 table.



b) The term, gene_product and uniprotmapping files are linked via the association table. The "id" column is the primary key in the gene_product and term table and foreign keys within the association table.

Figure 5. The database schema used for the Immune Ontologiser

2.3 Constructing the ImmunoGene-database and Bio-database

The ImmunoGene-database consists of immunology/haematology related genes from several commercial microarrays collected from BD Biosciences Clontech [18] and SuperArray Bioscience Corporation [19] [see additional file 1: Table 1]. All arrays were commercial except one, which was an immunology/haematology gene list obtained from a gene chip designed for functional immunological studies [20]. The rational behind selecting these arrays was to collect all known genes involved in immunology/haematology related processes that are used in microarray research focussing on human and mouse genes in order to cover

species-specific differences. The mouse and human gene lists were merged into two independent master lists and the gene names converted to HUGO id's using the program, MatchMiner [21,22]. The final gene list consisted of 2049 human and 2043 mouse immunology/haematology genes represented by HUGO ids. We identified the biological processes for these gene lists using the Gene Ontology [4] to include basic bodily processes concentrating mainly on functions related to immunology/haematology resulting in 1630 human functions and 2113 mouse functions together with their corresponding genes. These biological processes were grouped into 27 broad functional parent categories according to functions that were related, resulting in the final contents for the ImmunoGene-database. The biological data representing the Bio-database is derived from the Gene Ontology Consortium [3,4]. It has been modified to consist of the biological processes hierarchy and ensures that each biological function represented by a GO term within the hierarchy is not nested amongst several other GO terms or repeated. Thus, each GO term is displayed once only, together with the corresponding genes in a much more user-friendly structure, avoiding GO's polyhierarchical structure.

2.4 Developing the Database of Immunological Microarray Publications (DIMP)

We conducted a massive published literature search to accumulate all human and mouse genes that have been identified from microarray experiments relating to immunological diseases or conditions to date using the scientific literature database PubMed [17]. We manually queried PubMed adopting several search criteria, to ensure the capture of all relevant publications as comprehensively as possible relating to the microarray experiments carried out using samples from humans and mouse. For thorough identification of articles, our search criteria used text words that included several sets of keywords including immune disease and microarray, immunity and microarray and immunology and microarray. These keywords were identified amongst the abstracts and titles of all the available literature on PubMed and hence helped filter out unrelated articles allowing us to retrieve relevant abstracts. We categorised all the identified published literature into major categories according to the specific areas of immunological diseases such as autoimmune disease, immunological cancers or experimental conditions. We analysed for each relevant abstract the corresponding full-text article and extracted the gene identifiers that the authors published as being differentially expressed within the gene expression matrices following the microarray experiments they conducted. For the differentially expressed genes extracted from each article we converted the given gene identifier to HUGO identifiers using the program MatchMiner [21,22]. The database is updated monthly, thereby continuously expanding the existing database with the ever-increasing literature arising from microarray methods and data analysis.

3 Results

To demonstrate the functionality of the Immune Ontologiser we exploited the software to analyse the results of our own microarray study investigating the molecular mechanisms underlying immune tolerance. The aim of our microarray experiment was to identify differential gene expression in tolerant versus activated CD4+ T cells from mice, in order to understand the underlying mechanisms regulating tolerance, using cDNA microarrays. The gene expression profiles for activated T cells (A2) and tolerant T cells (T2) were obtained using an oligonucleotide array of 10,000 known mouse genes. Microarray production, sample preparation, hybridisation, scanning and data analysis were performed by according to the MIAME (Minimum Information About a Microarray Experiment) guidelines. For exact

experimental protocol, array layout and complete data; see the public repository, Array Express [23] accession number E-MEXP-283.

3.1 Identifying Differentially Expressed Genes

Our microarray dataset comprised 7287 mouse genes whose expression values had been normalised and filtered to exclude any expression ratios considered to be unreliable using Acuity 3.1 [24] [see additional file 2: Table 2]. We imported this dataset into our Immune Ontologiser software with the aim of identifying significantly differentially expressed genes between the activated and tolerant gene expression profiles. For clarity, GE1 represented the raw gene expression ratios from the activated CD4+ T cells (A2) and GE2 represented the raw gene expression ratios from tolerant CD4+ T cells (T2).

Table 2. Significant differentially expressed genes

Microarray Comparison	Log2 Threshold	No. Of genes fulfilling log2 Threshold	No. Of genes significantly differentially expressed
A2 vs. T2	=>+1.5 in A2	3440	430*
T2 vs. A2	=>+1.5 in T2	2169	70*
A2 and T2	=>+1.5 in A2+T2	111	111*
A2 vs. T2	=<-1.5 in A2	337	208^
T2 vs. A2	=<-1.5 in T2	146	47^
A2 and T2	=<-1.5 in A2+T2	74	74^

The second column represents the log2 threshold chosen in one microarray or both. +1.5 and –1.5 is selected as the standard upregulated or downregulated threshold, respectively. Expression levels above or below this threshold were considered to be of interest. The third column represents the number of genes satisfying the log2 threshold selected from the array consisting of 7287 genes. *The number of genes significantly up regulated =>1.5 fold from the number of genes above the log2 threshold. ^ The number of genes significantly down regulated =>1.5 fold from the number of genes fulfilling the log2 threshold.

We first exploited the Immune Ontologiser's differential gene expression analysis feature for identifying genes specifically upregulated in activated CD4+ T cells compared with tolerant CD4+ T cells to gain an understanding of the genes involved in T cell activation. Hence, after log2 transformation of the dataset we defined a threshold of >=+1.5 compared with the normal control within the A2 microarray to select significantly expressed genes. From the entire dataset of 7287 genes, 3440 genes fulfilled the criteria and were subsequently used to calculate the differential gene expression. The expression of these genes is considered biologically significant within the A2 cDNA microarray representing activated CD4+ T cells compared with control CD4+ T cells. Further filtering the dataset to select genes with a >=1.5 fold higher level of expression in activated CD4+ T cells than in tolerant CD4+ T cells, identified 430 genes. We then reversed to compare the gene expression profiles from tolerant CD4+ T cells with activated CD4+ T cells, to identify the genes significantly upregulated in tolerance, with the aim of investigating such genes further and researching their roles within tolerance. Again, we chose a threshold of >=+1.5 compared with the normal control, but this time within the T2 dataset, which identified 2169 genes. From these, 70 genes showed a >=1.5 fold higher level of expression in tolerant CD4+ T cells than in activated CD4+ T cells, i.e. were differentially expressed >=1.5 fold. In order to identify the genes simultaneously

expressed in both activated and tolerant CD4+ T cells, a >=+1.5 threshold was defined for both A2 and T2 datasets which identified 111 genes (Table 2, Figure 6).

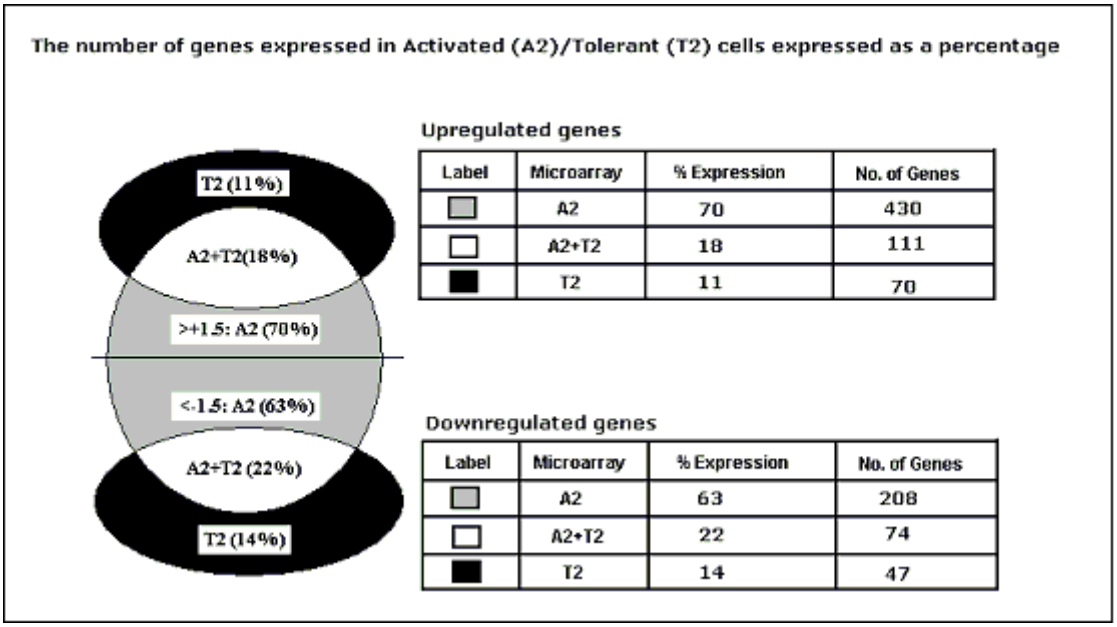


Figure 6. The number of genes expressed in Activated (A2)/Tolerant (T2) cells expressed as a percentage. A pictorial representation of Table 2, displaying the percentage of gene expression in up regulated and down regulated A2 and T2 cells. The percentage is calculated as the number of genes in each microarray divided by the total number of genes up regulated or down regulated.

Using the Immune Ontologiser we analysed our gene expression dataset further, finding specific genes that are down regulated in activated and tolerant conditions. We set a threshold criterion of <=-1.5 (i.e. –1.5, –1.6, -1.7) representing down regulated genes compared with the normal control in the A2 dataset, which identified 337 genes. From these, 208 genes were a further down regulated more than 1.5 fold in comparison with the T2 dataset. In contrast, 146 genes fulfilled the threshold criterion of <=-1.5 in T2 compared with the normal control from which 47 were a further down regulated more than 1.5 fold in comparison with A2. Identifying genes down regulated more than 1.5 fold in both A2 and T2 revealed 74 genes (Table 2, Figure 6). Details of the genes within these differentially expressed datasets are provided in additional file 3: Table 3a-3f. Having identified informative and meaningful genes from our microarray data we can now extract biological meaning for each of the differential gene expression datasets.

3.2 Functional Analysis of Tolerant Genes using the Bio-Ontology and Immune Ontology

The Immune Ontologiser places genes of interest from a microarray experiment within the context of biological processes in our user-friendly Bio-Ontology, highlighting the genes within the GO functions they are associated with and at the same time displays the number of genes from the user's gene list that are identified within each GO function. The Bio-Ontology gives the overall view and functionality of interesting genes amongst all of the biological processes currently in the Gene Ontology [3,4] while the Immune-Ontology extracts biological functions that are more related to immunology and haematology for the genes of interest and allows the user to identify biological process(es) that may be more important by seeing which biological category is most enriched with genes from the users interesting genes. We queried the Bio-Ontology and Immune-Ontology with genes significantly differentially expressed in tolerance.

Table 3. Immunology/haematology related functions for differentially expressed genes in Tolerance

Functional Category	Sub group*		2.2.1.1.1.1.1 Functional Category	Sub group*	
	A	B		A	B
Behaviour	2	0	Death	31	7
Biological Process Unknown	0	0	Detection and Response to Stimulus	77	21
Blood and Circulation	0	5	Development	25	8
Bone	0	0	Enzyme Activity	0	4
Cell Activation	41	0	Homeostasis	4	1
Cell Adhesion	7	0	Hormone	0	0
Cell Cycle	9	5	Immune Response	33	5
Cell differentiation	4	1	Localisation	25	2
Cell growth and/or maintenance	31	5	Metabolism	81	32
Cell motility	3	0	Signal Transduction	32	21
Cell organisation and biogenesis	5	0	Transcription	8	3
Cell-cell signalling	4	1	Translation	0	0
Central nervous system	6	3	Viral Life cycle	3	0
Cytokine Production	22	0			

***Sub-groups A represent the number of subgroups within the functional category containing upregulated genes from the T2 dataset. Sub-groups B represent the number of subgroups within the functional category containing downregulated genes from the T2 dataset.**

Before proceeding we converted our gene Accession Number identifiers to HUGO identifiers, to make them compatible with the Immune Ontologiser using the program MatchMiner [21,22]. Querying the Bio-Ontology with our upregulated tolerant T cell dataset showed genes to be widely spread amongst various biological processes. The gene, *EGR-2* for example is shown in signal transduction (GO:0007165), which relates to immunological processes but also more widely, within brain development (GO:0007420), peripheral nervous system development (GO:0007422) and mechanosensory behaviour (GO:0007638). Details of the results together with a complete listing of the genes identified within each biological function are provided in additional file 4: Table 4.

We then subjected the Immune-Ontology to the same dataset used to query the Bio-Ontology. Table 3 shows the number of genes identified within each immunology/haematology related function for the 70 genes involved in tolerance upregulated and 47 genes down regulated. It is evident from Table 3 that the functional group metabolism consists of the majority of the genes from the 70 that we identified to be upregulated in tolerant CD4+ T cells. It is important to remember that genes can occur within more than one functional process within a

main biological category. Hence, genes from our T2 dataset have been identified within metabolism a total number of 81 times. Metabolism is a basic bodily process one would expect to find enriched with the involvement of many genes. More specific to immunology, one can see that the more immunology related functions enriched with the genes of interest involved in tolerance are detection and response to stimulus, cell activation, immune response, death, signal transduction and cytokine production (Table 3). Details of the genes involved within each biological function are provided in additional file 5: Table 5. The specific research area to be further investigated from the aforementioned microarray experiment is the molecular mechanisms underlying T cell tolerance. Identifying genes from the genes of interest that have biological functions highly involved within immunology are of extreme value to the researcher. Extracting such genes and their respective functionalities using the Immune Ontology govern further analysis enabling researchers to move towards the ultimate goal of unwinding the molecular intricacies underlying in this case, T cell tolerance.

Amongst the 70 genes upregulated in tolerance compared with activated T cells we confirmed the expression of 15 genes via RT-PCR (manuscript in preparation). Amongst these 15 confirmed genes were the genes *T-bet*, also known as *T-box21* (*Tbx21*), the early growth response transcription factor 2 gene *EGR-2* (*Krox-20*), the lymphokine interleukin 10 (*IL10*), Granzyme B (GZMB), the chemokine CCL4 and the interferon regulatory factor, IRF1. These genes were highly expressed in the tolerant T cells with mRNA expression levels ranging from 24.000 for IRF1 to 38.202 for CCL4. Over-expression of the *EGR-2* gene has been associated with the inhibition of T cell activation and promotion of tolerance also known as anergy through many different mechanisms [25,26]. As an example, searching the functions of the *EGR-2* gene via the general multi-level GO hierarchy from the Gene Ontology Consortium identified a wide range of biological functions, many of which are not immunology related, such as morphogenesis or organismal physiological processes. However, buried amongst the complex hierarchical GO terms are many functions specifically related to immunology and using the Immune Ontology aspect of our Immune Ontologiser software, we have identified these specific biological functions and in turn identified *EGR-2* to be involved in cell differentiation, the central nervous system, development, metabolism and transcription. We have presented a much more user-friendly structure, where the user does not have to search through the entire GO hierarchy by eye to extract immunology related biological functions.

3.3 Retrieving Multiple Immunological Studies Based on our Tolerant Gene Dataset using DIMP

The Immune Ontologiser enables the extraction of immunology related published literature containing the genes of interest identified from a researcher's microarray experiment allowing a gene expression comparison. It does so by comparing interesting genes from a microarray dataset with genes identified through microarrays known to be involved in immunology related diseases or conditions. Thus providing the user with a further insight into other immunological research areas their genes may be involved in, providing access to the original article. The articles are organised within a structured ontology grouping publications according to the immunological disease type or condition.

We subjected our publications database consisting of literature where the microarray experiments used human and mouse samples to the same set of genes we used to query the Immune Ontology and Bio-Ontology. We aimed to identify areas of immunological research where these genes were also considered biologically significant. For detailed results see additional file 6: Table 6. From our upregulated genes of interest involved in tolerance our software identified 5 genes: *CD4*, *CACNB3*, *PTMA*, *IL10* and *RAB5* involved in the autoimmune disease multiple sclerosis using cDNA microarrays [27]. Comparative analysis

by Mycko *et al*., 2004, highlighted different sets of genes, including genes of inflammatory characteristics, apoptosis related and stress-induced, indicating their potential role in multiple sclerosis pathogenesis. Other genes from our dataset of interest, *GZMB* and *PLTP* were also identified through microarray analysis of peripheral blood mononuclear cells (PMBCs) from multiple sclerosis [28] detecting a profile of immune cell activation, autoantigen upregulation, and enhanced E2F pathway transcription. The transcription factor *EGR-2* together with many other genes from our dataset were also identified by a study analysing the expression profiles in transgenic mice with cardiac-specific over-expression of the tumour necrosis factor (*TNF1.6*) resulting in autoimmune myocarditis [29]. The study suggested that TNF-alpha over-expression activates not only the inflammatory response, but also humoral immune responses within the transgenic hearts.

Hence, using this aspect of our software enables researchers to accumulate the immunological related microarray based literature available on PubMed for all the genes simultaneously from a microarray dataset that an analyst may define as being biologically significant. Importantly, researchers do not need to manually search through the articles to find which genes from their microarray dataset are present within the retrieved articles, as our software is designed to immediately highlight all the matched genes to the user. Researchers can thereby gain a wider understanding of the involvement of their genes in different diseases together with a comparison of the gene expression values and at the same time perhaps investigate the microarray techniques used by other researchers and perhaps gain ideas and direction for further development into their own research, without the pain-staking task of searching through all the available literature via keyword searches.

4 Discussion

The motivation for the Immune Ontologiser is simple. Since the development of the GO consortium and the assignment of GO terms to genes, scientists have been able to functionally annotate interesting genes using computational tools that take advantage of the GO. However, researchers are currently unable to concisely identify only those biological categories involved in a specific biological system of the human body, e.g. the reproductive system or specific branch of biology such as immunology, for interesting gene expression datasets. Using GoMiner for example, the user would have to search through the summary data of hundreds of categories one at a time to find all processes related to the immune or any other system the researcher is interested in. Simply knowing every single biological, molecular and cellular function for gene expression datasets, amounting to thousands of gene functionalities is not biologically meaningful to the biologist until it is further mined. Obtaining such specific information is of significance to biological experts within specialised research areas to carry out more exhaustive research with respect to understanding any underlying biological pathway or mechanism. Furthermore, since the development of an immunology-related microarray chip containing primarily known genes for functional immunological studies [20], there has been no specific immunology related ontology to allow users of microarrays to identify gene functions that relate more to immunology. Here, we concentrate on the immunology/haematology aspect of biology to help biologists rapidly identify gene functionality within an immunology ontology. The second motivation arises from the lack of this informative aspect in related programs that allows the comparative analysis of an interesting gene list from a microarray experiment with genes identified via microarray methods in immunology-related published studies. Current applications are unable to identify multiple publications for numerous genes of interest at once, nor do they possess the ability to accentuate such genes within the interesting gene expression datasets in a well-structured manner. The aim of this unique aspect is to facilitate the comparison of

results between scientific groups investigating similar research areas whilst gaining further knowledge. The final advantage of this powerful tool is its flexibility, with its unique ability to automate the identification of differentially expressed genes between microarrays, again a very advantageous application to integrate, yet ignored by other available related programs. Typically known as a mundane and time consuming task carried out in Excel, Immune Ontologiser presents such genes via quick and simple steps, allowing the user to define a threshold in a single microarray on which the fold change should be calculated.

In relation to the Immune Ontologiser several tools have been developed to provide functional insights into the results obtained from high-throughput gene expression profiling (Table 4). FatiGO is a web-accessible application that extracts relevant GO terms for differentially expressed genes from a microarray experiment [2,30]. However, the output from FatiGO is restricted to only one level of the GO hierarchy per query. In contrast, GeneLynx [31,32] provides information about gene functionality at all three levels of the GO via AmiGO in a tabular format. However, a hierarchical tree representation can only be viewed for one level of the GO hierarchy. Unfortunately this tool is limited to the analysis of only one gene at any time. MAPPFinder, in contrast provides the fundamental tree representation of the GO hierarchy, with summary and statistical data in line with each category, but the genes themselves are presented in an auxiliary table. This limitation is overcome by both GoMiner and our Immune Ontologiser that integrate both GO categories and genes in one hierarchical view. The functionality of GoMiner is similar to the GO chart module of a related program, DAVID. The gene classification data however, is displayed differently. DAVID displays a bar chart where the length of the bar represents the number of gene identities in each category. Selecting an individual bar opens a new HTML table displaying the individual genes. GoMiner presents the individual genes within the hierarchical tree structure together with the total number of genes involved in the category. An enhanced feature of GoMiner compared with DAVID and MAPPFinder is that it provides intuitive tree and DAG (directed acyclic graph) views of genes embedded within the GO hierarchy, which provides a qualitative and quantitative picture of the complex, multiple parent-child relationships of some GO categories. However, DAVID can only display such views through hyperlinks of GO terms to QuickGO developed at the European Bioinformatics Institute [33,34].

EnsMart is a web accessible program that provides a vast amount of functional annotation by extracting data held in the Ensembl database [35,36]. Its gene ontology section also provides a web-link to QuickGO like DAVID. It is very flexible accepting several accession types including Affymetrix probe sets, GenBank and LocusLink Id's and annotating genes from several species including human, mouse, rat and fly. However, limitations in comparison to DAVID and MAPPFinder include EnsMart's inability to provide graphic summaries of GO categories, protein domains or biochemical pathways, producing only tabular output. DAVID and MAPPFinder both allow the viewing of genes within the context of biochemical pathways via KEGG charts, as does GoMiner via BioCarta and KEGG biological pathways. Lastly Resourcerer addresses the issue of being able to make comparisons of gene expression patterns between species, linking genes surveyed in one species to the corresponding genes in other species [37,38]. Analysis of gene expression in model organisms, particularly the mouse and rat has become a fundamental tool for the study of human development and disease. Resourcerer aims to link such genes to the corresponding human genes as well as compare the genes across other species, including dog, cattle and pig. However, it does not provide graphic summaries or annotations from GO or KEGG, thus limiting its utility as a tool for functional annotation. All the above tools for functional annotation of genes categorise genes within the entire GO, whether it be the biological, molecular and cellular aspect simultaneously or just one part. The functions identified represent all the functions for the gene(s) of interest, lacking specificity and order. If researchers are more specifically

interested only in those genes from their microarray data that are involved in the circulatory system, digestive system, nervous system, muscular and skeletal system or reproductive system for example, none of the software above allows them to view these specific genes and their associated specific functions at a glance.

With respect to the functionality of our ImmunoArray-PubOntology, many other literature-mining tools have also been developed. Such tools are of tremendous use to the scientific community allowing the sharing of knowledge and facilitating connections between researchers around the world, especially since it is not feasible or practical for one laboratory to perform microarray experiments of every nature. Current literature mining tools have their own defined types of output, advantages and limitations and either use only one or a combination of methodologies, namely information retrieval (IR), entity recognition (ER) and information extraction (IE) [39]. The most commonly used literature mining tools are PubMed and MedMiner that use a process called information-retrieval enabling researchers to identify relevant papers. MedMiner allows the querying of multiple databases, currently GeneCards [40] and PubMed [17], via the combination of keywords or by using word frequencies to display relevant abstracts based upon the frequency of which a keyword appears within an abstract. Upon the identification of any given keyword both PubMed and MedMiner can present the user with thousands of articles. Only after reading each article may the researcher reach a conclusion as to whether the article is of interest. In an attempt to make this process more efficient, an application called iHOP (Information Hyperlinked over Proteins) was developed [41]. Based on the more advanced methodology called ER, which identifies biological entities such as genes or proteins mentioned in the text, iHOP presents sentences taken directly from the abstracts of relevant literature, thereby giving researchers the control to decide whether the information offered by a sentence is of any interest. Furthermore, the sentences are utilised to generate a gene model represented as a dynamic graph with the aim of displaying all associations between the genes in collected sentences. However, iHOP, PubMed and MedMiner can only be used to extract literature for one gene at any time and hence are not applicable for microarray data where the emphasis is on extracting relevant information for hundreds of gene simultaneously. In contrast, the online tool PubGene can be applied for interpreting microarray gene expression data. Using a more graphical approach PubGene [42] generates a functional association model based on the co-occurrence of genes within a common MEDLINE record. PubGene employs the more superior IE process enabling specific facts to be extracted from the literature, on the assumption that if two genes are co-mentioned within the title or the abstract of a MEDLINE record there is an underlying biological relationship. Although this can result in complex literature based gene networks, the retrieved articles range from a variety of research domains and are not structured or presented in any meaningful fashion. Often, researchers are involved in a specific field of biology and hence may only be interested in literature related to their area of expertise. To fulfil this requirement, the IR-based tool, XplorMed [14] summarises MEDLINE search results according to eight subjects via keywords, allowing the user to obtain papers in a particular research area. However, XplorMed is limited to allow the further investigation and identification of one gene within articles at a time. For the functional analysis of multiple genes, is the software PubMatrix [15]. PubMatrix, also an IR application requires the user to input a gene list together with a second list of pre-defined functionalities, e.g. diabetes, aging or infection and submit the data to PubMatrix for processing. In turn, the software searches PubMed and identifies the number of articles in which each gene is associated with functionality from the second search list. This is advantageous if the user only wants to search for articles in which genes have been associated with their pre-defined functionalities. Unfortunately, the gene list accepted by PubMatrix is limited to a maximum of 100 genes and the second list can only consist of 10 categories. Lastly, is the online tool GoPubMed [43]. Although not applicable for microarray data this

application has been designed in an attempt to present the results from literature mining in a more organised manner using the Gene Ontology. GoPubMed is an IR system in which keywords are submitted to the PubMed database, to obtain corresponding literature. GO terms are simply extracted from the retrieved abstracts and presented within the Gene Ontology structure, hence allowing abstracts to be categorised according to an ontology.

Table 4. Comparison of the Immune Ontologiser with similar software

Functionality	Immune Ontologiser	FatiGO	MAPP Finder	Gene Lynx	EnsMart	GoMiner	DAVID
Differential gene expression Analysis	YES	NO	NO	NO	NO	NO	NO
Functional annotation of genes within GO biological processes	YES	YES	YES	YES	YES	YES	YES
Functional annotation within a distinct haematology/ immunology GO	YES	NO	NO	NO	NO	NO	NO
Showing the most important biological process from microarray data	YES	NO	NO	NO	NO	NO	NO
Comparison of microarray datasets with immunology related published microarray literature	YES	NO	NO	NO	NO	NO	NO
DAG Tree	NO	NO	NO	NO	NO	YES	YES
Species	Human Mouse	Human Mouse Fly	Human Mouse	Human Mouse	Human Mouse Rat Fly	All from Gene Ontology Consortium	Human Mouse Rat Fly
View GO terms in a hierarchy	YES	YES	YES	NO	NO	YES	YES
Hyperlinked cross references	YES	NO	NO	NO	NO	YES	YES

Our Immune Ontologiser software accepts an unlimited list of interesting genes from microarray datasets and based on the IR technique, searches our publication database to extract immunology related literature in which such genes of interest are also believed to be biologically meaningful, i.e. significantly differentially expressed as a result of gene expression analysis. We have structured our immunology related literature by grouping together publications investigating similar diseases or conditions. Using this strategy we have

been able to provide researchers with a more meaningful view of data stored within the relevant articles. The user is not compelled to identify important articles within their research field on a gene by gene basis and nor do they have to search the content of each article on an individual basis to identify which genes from their microarray datasets are involved within the publication.

Limitations are also present with the Immune Ontologiser as the software is currently limited to identifying gene names that are presented as a result of microarray experiments and accepts only HUGO identities for *Homo Sapiens* and *Mus Musculus* species. We aim to enhance this functionality allowing researchers to biologically annotate data via other identifier types as well as allow the software to be applied to other species. The current version of the Immune Ontologiser is specific to biological process annotation of microarray data. We plan to incorporate molecular function and cellular localisation within the ontology for the genes within our ImmunoGene-database as well as incorporate ontologies specific for other fields, e.g. oncology. In the future we intend to retrieve literature related to a particular research area direct from the biomedical resource, PubMed. Moreover, we would like to investigate computational methods that will potentially allow us to read the full text of relevant articles and using entity recognition, extract and present the gene names that are significant in microarray experiments, together with their gene expression values and functionalities as described by the paper. This capability could prove to be powerful for researchers in the context of comparing more in-depth information embedded within the literature. However recognising and retrieving gene names from the full text article could pose further challenges as on many occasions the results comprising such information are presented within bitmap images. This methodology would not only be valuable for the literature mining aspect of our software but also for the enhancement of our Immune Ontology with genes and their respective functions identified from pertinent immunological literature.

5 Conclusions

Overall, we believe that the multifunctional Immune Ontologiser is an attractive package for the scientific community working with microarrays. The Immune Ontology and unique microarray immunological publication ontology are extremely beneficial for immunologists analysing microarray datasets. The first aspect of the software is aimed at more specifically categorising and identifying genes of interest involved in immunology/haematology related biological processes. The latter functionality focuses on literature-mining allowing further exploitation of biological data and comparison of microarray gene expression data. As there is no such tool currently available to allow researchers to distinctly identify those genes and gene functionalities that are more highly involved within specialised research areas (such as immunology) we have created ontologies focussed more towards immunology. In similar ways, this concept can be applied to develop ontologies in other fields. In addition, the Immune Ontologiser's current design elements provide automated solutions for differential gene expression and identification of important genes that are significantly differentiated in one biological condition compared with another, which are then used to enable researchers to rapidly discover biological meaning in large datasets consisting of lists of genes.

Availability and requirements

- Project name: Immune Ontologiser
- Project Home Page: Databases including software executable can be accessed from ftp://ftp.brunel.ac.uk/cspgssk
- Operating system: Tested on Windows 2000 and Windows XP

- Programming language: Visual Basic.Net and MySQL
- Other requirements: Microsoft .NET Framework version 2.0 Software Development Kit (SDK), MySQL database server no later than 3.23.58 and the MySQL Connector/ODBC 3.51.

Acknowledgements

This study was partly supported by grants from the UK Medical Research Council (MRC) (Grant number: G0300520) and the Brunel University Studentship.

6 References

[1] M. Schena, D. Shalon, R.W. Davis and P.O. Brown. 2005. Quantitative monitoring of gene expression patterns with complementary DNA microarray. *Science,* 270, 467-470.

[2] F. Al-Shahrour, R. Diaz-Uriarte and R. Dopazo. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20, 578-580, (2004).

[3] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25, 25-9 (2000).

[4] Gene Ontology [http://www.geneontology.org]

[5] P. Khatri, S. Draghici, G.C. Ostermeier and S.A. Krawetz. Profiling gene expression using onto-express. *Genomics* 79, 266-70 (2002).

[6] OntoExpress [http://vortex.cs.wayne.edu:8080]

[7] S.W. Doniger, N. Salomonis, K.D. Dahlquist, K. Vranizan, S.C. Lawlor and B.R. Conklin. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biology* 4, R7 (2003).

[8] GENMAPP [http://www.genmapp.org/]

[9] B.R. Zeeberg, W. Feng, G. Wang, M.D. Wang, A.T. Fojo, M. Sunshine, S. Narasimhan, D.W. Kane, W.C. Reinhold, S. Lababidi, K.J. Bussey, J. Riss, J.C. Barrett and J.N. Weinstein, J.N. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology* 4, R28 (2003).

[10] GoMiner [http://discover.nci.nih.gov/gominer]

[11] G. Dennis, B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane and R.A. Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* 4, P3 (2003).

[12] DAVID [http://www.david.niaid.nih.gov/]

[13] L. Tanabe, U. Scherf, L.H. Smith, J.K. Lee, L. Hunter and J.N. Weinstein. MedMiner: an Internet text-mining tool for biomedical interpretation, with application to gene expression profiling. *Biotechniques* 27, 1210-1217 (1999).

[14] C. Perez-Iratxeta, P. Bork and M.A. Andrade. XplorMed: a tool for exploring MEDLINE abstracts. *Trends in Biochemical Sciences* 26, 573-575 (2001).

[15] K.G. Becker, D.A. Hosack, G. Dennis, R.A. Lempicki, T.J. Bright, C. Cheadle and J. Engel. PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics* 4, 61 (2003).

[16] MySQL [http://www.mysql.com]

[17] PubMed at NCBI [http://www.ncbi.nlm.nih.gov/]

[18] Clontech [http://www.clontech.com/clontech]

[19] Bioscience Corporation [http://www.superarray.com/home.php] SuperArray]

[20] J. Waukau, P. Jailwala, Y. Wang, H.J. Khoo, S. Ghosh, X. Wang and M.J. Hessner. The design of a gene chip for functional immunological studies on a high-quality control platform. *Annals of the New York Academy of Sciences* 1005, 284-287 (2003).

[21] K.J. Bussey, D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W.C. Reinhold, B. Zeeberg, W. Ajay and J.N. Weinstein. MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biology* 4, R27 (2003).

[22] MatchMiner [http://discover.nci.nih.gov/matchminer]

[23] ArrayExpress [http://www.ebi.ac.uk/arrayexpress]

[24] Axon Instruments [http://www.axon.com]

[25] M. Safford, S. Collins, M.A. Lutz, A. Allen, C. Huang, J. Kowalski, A. Blackford, M.R. Horton, C. Drake, R.H. Schwartz and J.D. Powell. Egr-2 and Egr-3 are negative regulators of T cell activation. *Nature Immunology* 6, 472-480 (2005).

[26] D.B. Parkinson, A. Bhaskaran, A. Droggiti, S. Dickinson, M. D'Antonio, R. Mirsky and K.R. Krox-20 inhibits Jun-NH2-terminal kinase/c-jun to control Schwann cell proliferation and death. *Journal of Cell Biology* 164, 385-394 (2004).

[27] M.P. Mycko, R. Papoian, U. Boschert, C.S., Raine and K.W. Selmaj. Microarray gene expression profiling of chronic active and inactive lesions in multiple sclerosis. *Clin Neurol Neurosurg.* 106, 223-229 (2004).

[28] A.H. Iglesias, S. Camelo, D. Hwang, R. Villanueva, G. Stephanopoulos and F. Dangond. Microarray detection of E2F pathway activation and other targets in multiple sclerosis peripheral blood mononuclear cells. *J Neuroimmunol.* 150, 163-77 (2004).

[29] Z. Tang, B.S. McGowan, S.A. Huber, C.F. McTiernan, S. Addya, S. Surrey, T. Kubota, P. Fortina, Y. Higuchi, M.A. Diamond, D.S. Wyre and A.M. Feldman, A.M. Gene expression profiling during the transition to failure in TNF-alpha over-expressing mice demonstrates the development of autoimmune myocarditis. *J Mol Cell Cardiol.* 36, 515-530 (2004).

[30] FatiGO [http://fatigo.bioinfo.cnio.es]

[31] B. Lenhard, W.S. Hayes and W.W Wasserman. GeneLynx: a gene-centric portal to the human genome. *Genome Research* 11, 2151-2157 (2001).

[32] Genelynx [http://www.genelynx.org/]

[33] E. Camon, M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox and R. Apweiler. The Gene Ontology Annotation (GOA) project: implementation of GO in Swiss-Prot, TrEMBL and InterPro. *Genome Research* 13, 662-672 (2003).

[34] QuickGO [http://www.ebi.ac.uk/ego]

[35] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox and E. Birney. EnsMart: a generic system for fast and flexible access to biological data. *Genome Research* 14, 160-169 (2004).

[36] EnsMart [http://www.ensembl.org/multi/martview]

[37] J. Tsai, R. Sultana, Y. Lee, G. Pertea, S. Karamycheva, V. Antonescu, J. Cho, B. Parvizi, F. Cheung and J. Quackenbush. RESOURCERER: a database for annotating and linking microarray resources within and across species. *Genome Biology* 2, 2.1-2.4 (2002).

[38] TIGR [http://pga.tigr.org/tigr-scripts/magic/r1.pl]

[39] L.J. Jensen, J. Saric and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. Nature Reviews Genetics 7, 119-129 (2006).

[40] M. Rebhan, V. Chalifa-Caspi, J. Prilusky and D. Lancet. GeneCards: Integrating information about genes, proteins and diseases. *Trends Genet.* 13, 163 (1997).

[41] R. Hoffman and A. Valencia. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics* 21, ii252-ii258 (2005).

[42] T.K. Jenssen, A. Laegreid, J. Komorowski and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics* 28, 21-28 (2001).

[43] R. Delfs, A. Doms, A. Kozlenkov and M. Schroeder. GoPubMed: ontology-based literature search applied to Gene Ontology and PubMed. *Nucleic Acids Research* 33, W783-6 (2004).