Article

Selection of Proteins for Human MHC Class II Presentation

Li Jiang[1, 2], Ole Lund[2, 3] and Jinquan Tan[1, 3]

We investigated the predicted function of proteins eluded from human MHC class II molecules. Peptides that are presented by MHC class II were obtained from the SYFPEITHI database and the corresponding proteins were found in the SWISSPROT database. The functions of these proteins were predicted using the protfun server. Our analysis showed that human proteins presented by MHC class II molecules are likely to be in the cell envelope, be a receptor or involved in immune responses. Presented proteins from bacteria and virus, on the other hand, are more likely to be involved in regulatory functions, translation, transcription as well as replication. These results can lead to better understanding the autoimmunity and the response to infections. ***Cellular & Molecular Immunology*****. 2005; 2(1):49-56.**

Key Words: bioinformatics, protein function, immunology, MHC class II

Introduction

Immune cells do not interact with, or recognize, an entire immunogenic molecule. They recognize discrete sites on the macromolecule called epitopes, or antigenic determinants (1). The part of a processed antigen that binds to the major histocompatibility complex (MHC) molecules is called an epitope, which was derived from antigen restriction element. MHC molecules bind short peptides and display them on the cell surface for recognition by the T-cell receptor (TCR) of T cell (2).

The MHC is referred to as the HLA complex in humans. The classical MHC molecules are encoded by loci within three subregions: MHC class I, MHC class II and MHC class III. They are associated with intercellular recognition and with self/nonself discrimination. The loci constituting the MHCs genes are highly polymorphic; that is, many alternate forms of genes, or alleles, exist at each locus. Because MHC molecules function as antigen-presenting structures, the particular set of MHC molecules expressed by an individual influences the repertoire of antigens to which that individual's helper T (Th) cells and cytotoxic T (Tc) cells can response to. For this reason, the MHC has been implicated in the susceptibility to disease and in the development of autoimmunity. MHC class II genes encode proteins expressed primely on antigen-presenting cells (macrophages, dendritic cells, and B cells), where they present processed antigenic peptides to Th cells (1).

The MHC class II molecules can bind a variety of peptides (in general these peptides are derived from exogenous proteins, either self or nonself, which are degraded with the endocytic processing pathway) and present these peptides to $CD4^+$ Th cell. Various hydrolytic enzymes within the acidic endocytic compartments degrade exogenous antigens, internalized by phagocytosis or endocytosis. Within rough endoplasmic reticulum (RER), newly formed class II MHC molecules with peptides are transported to the plasma membrane. The isolated peptides generally contain 13-18 amino acid residues (a core binding sequence comprising 7-10 amino acids), and those binding to a particular MHC class II molecule often have internal conserved "motifs" (1).

Sequences from pathogens provide a huge amount of potential vaccine candidates, as the specific peptides bind to MHC class II molecules. MHC-binding peptides are also potential tools for diagnosis and treatment of cancer (3). Binding of a peptide to an MHC molecule is prerequisite for recognition by the T cells, but only certain peptides can bind to any given MHC molecule. It is estimated that only one in 100 to 200 peptides actually binds to a particular MHC molecule (4).

Therefore, a good computational prediction of what kind of protein (what function), degraded into MHC-binding peptides and presented by MHC class II molecules could be useful in vaccine research. The work presented here aims at predicting the function of proteins and determining which functional class of proteins, are more likely to be presented

[1]Department of Immunology, Wuhan University School of Medicine, Wuhan 430071, Hubei, China.

[2]Center of Biological Sequence Analysis, Biocentrum-DTU, the Technical University of Denmark, DK-2800 Lyngby, Denmark.

[3]Corresponding to: Dr. Jinquan Tan, Department of Immunology, Wuhan University School of Medicine, Wuhan 430071, Hubei, China. E-mail: jinquan_tan@hotmail.com. Or Dr. Ole Lund, Center of Biological Sequence Analysis, Biocentrum-DTU, the Technical University of Denmark, DK-2800 Lyngby, Denmark. E-mail: lund@cbs.dtu.dk.

Received Jan 21, 2005. Accepted Feb 2, 2005.

by MHC class II.

In the last few decades, advances in molecular biology and equipment available for research in this field have allowed the increasingly rapid sequencing of large portions of the genomes of several species. Information science has been applied to biology to produce the field called Bioinformatics. The most pressing task in bioinformatics involves the analysis of sequence information. Computational Biology is the name given to this process, and it involves developing methods to predict the structure and/or function of newly discovered proteins and structural RNA sequences. The human genome project has led to the discovery of many human protein coding genes which were previously unknown. As a large fraction of these are functionally uncharacterized, it is of interest to develop methods for predicting their molecular functions from sequences. To predict the function of unknown protein sequence needs available data and bioinformatics tools-software.

Traditionally, protein function has been related directly to the three-dimensional structure of the polypeptide chain, which currently, for most sequences are hard to compute. One way of predicting the function of a protein is to show that it aligns well with a protein of known function. This will only work when a homologue with known function has been found. An alternative method operates in the "feature" space of all sequences, and is therefore complementary to methods that are based on alignment and the inherent, position-by-position quantification of similarity between two sequences. One available method is ProtFun, which is developed as an entirely sequence-based method that identifies and integrates relevant features that can be used to assign proteins of unknown functional classes, and enzyme categories for enzymes. As one may expect that proteins performing similar functions would share some attributes even though they are not at all related at the global level of primary structure. The essential types of post-translational modification (PTMs) include N- and O-glycosylation, (S/T/Y) phosphorylation, and cleavage of N-terminal signal peptides controlling the entry to the secretary pathway. ProtFun can thus be used to predict the function of a protein even if no homologue with known function is known (5).

The strategies for the elucidation of protein function may benefit from a number of functional attributes that are more directly related to the linear sequence of amino acid, and hence easier to predict than protein structures. The attributes include features associated with post-translational modifications and protein sorting, but also much simpler aspects such as the length, isoelectric point and composition of the polypeptide chain (5).

As long as the sequences of the proteins, which produce the peptides and bind to MHC class II molecules, are available, one can predict the functions of these proteins by the ProtFun server which may use mainly the PTMs characters of those sequences. As the reaction between peptides, lymphocyte cell receptors and MHC molecules are the principle concept for developing adaptive immune. Eventually the function prediction will attribute to better understanding of both autoimmunity and infect immunity.

Materials and Methods

Generating a dataset with peptides binding to MHC class II molecules

The peptides were obtained from SYFPEITHI, a database of MHC binding peptides, which is freely available on the web at (http://syfpeithi.bmi-heidelberg.com/scripts/MHCServer.dll/home.htm). This public website is a database for MHC ligands and peptide motifs. It contains a collection of MHC class I and class II ligands and peptide motifs from humans and other species, such as apes, cattle, chicken and mice. All motifs currently available are accessible as individual entries. Searches for MHC alleles, MHC motifs, natural ligands, T-cell epitopes, source proteins/organisms and references are possible (6, 7).

We selected "FIND YOUR LIGAND, MOTIF OR EPITOPE" on the home page of SYFPEITHI. In humans MHC HLA-DP, HLA-DQ and HLA-DR regions encode class II molecules. The 69 alleles whose name starts with HLA-DP, HLA-DQ or HLA-DR were available in "select MHC type" in the SYFPEITHI database.

The 759 motifs were classified into four categories: 1) T-cell epitope (T); 2) Unknown motifs: T-cell epitopes (UT); 3) Example for ligand (L); 4) Unknown motifs: Ligand (UL). We used the letter T, L, UT and UL to represent above four categories. A FASTA file was generated for each motif sequence. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. 1) The description line starts with a greater than symbol (">"); 2) The word following the greater than symbol (">") immediately is the "ID" (name) of the sequence, the rest of the line is the description; 3) The description is optional; 4) All lines of text usually should be shorter than 60 characters, 5) The sequence ends if there is another greater than symbol (">") symbol at the beginning of a line and another sequence begins. The following is an example:

> Example ID envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHC
TNLMNTTVTTGLLLNGSYSENRTQIWQKHRTSNDSALI
LLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHS

Finding the original proteins

A short peptide sequence can be shared among several proteins. This means that some short sequences can be found in more than one protein. The website Swiss-Prot [http://us.expasy.org/sprot/] contains annotated sequences of proteins. We used the sequences of MHC class II binding peptide as entry to trace the original proteins. We only kept one protein matching to each peptide. The following rules were used to select the proteins: 1) Only choose the protein original from human, viruses, bacteria and parasites; 2) Only choose the protein for the peptides labeled with (T, L, UT or UL); 3) If there are multiple choices matching the same peptide, we checked SYFPEITHI website manually to ensure that the protein matched the exact Swiss-Prot name. A random protein on the list was selected if the Swiss-Prot name unavailable; 4) if we were still not able to choose a suitable protein for a peptide sequence then it was discarded.

Table 1. The most significant functions found in the overlapping dataset (Set 1)

	Total	Human	Nonhuman	Viruses	Bacteria	Parasites
Positive	Transport and binding (13.26)	Transport and binding (14.70)	Nonenzyme (3.10)	Regulatory function (3.25)	Transcription regulation (2.43)	Transport and binding (5.45)
Negative	Replication and transcription (-12.56)	Replication and transcription (-14.29)	Fatty acid metabolism (-5.65)	Fatty acid metabolism (-5.08)	Fatty acid metabolism (-3.25)	Translation (-3.45)

The human and parasite groups are similar both in over-represented proteins function such as transport and binding, and in under-represented functions like replication, transcription and translation. Virus and bacterium groups behave in a reverse way: the over-represented ones are regulatory function and transcription regulation; the under-represented one is fatty acid metabolism.

Table 2. The most significant functions found in the nonoverlapping dataset (Set 2)

	Total	Human	Nonhuman	Viruses	Bacteria	Parasites
Positive	Immune response (5.73)	Immune response (6.07)	Transcription (2.00)	Transcription (2.50)	Transcription regulation (2.23)	Transport and binding (3.37)
Negative	Replication and transcription (-4.55)	Replication and transcription (-5.49)	Fatty acid metabolism (-3.18)	Fatty acid metabolism (-2.62)	Fatty acid metabolism (-2.38)	Replication and transcription (-3.15)

In this dataset, the human and parasite groups are similar both in over-represented functions such as transport and binding as well as immune response which mostly the membrane proteins are involved in, and in under-represented functions like replication and transcription. Virus and bacterium groups are identical in both predicted functions.

For each peptide we extracted from SYFPEITHI, we tried to find a suitable protein matching it. We divided these proteins into four groups according to their original species: human, viruses, bacteria and parasites. We also added another two groups: nonhuman (sum of the three types of pathogens) and total (all four groups). The numbers of entries in our database were 562, and included 467 human, 56 virus, 33 bacterium, 6 parasite, and 95 nonhuman sequences.

For each protein we predicted the function using the ProtFun server (http://www.cbs.dtu.dk/services/ProtFun/) at CBS. Some protein predictions were not available because certain protein sequences are not allowed (for example, sequence length exceeding 4,000 amino acids) as well as some, which caused errors during the job. Eventually the output dataset (Table1, Set 1) was reduced to 539 which consisted of 453 human, 47 virus, 33 bacterium, 6 parasite and 86 nonhuman sequences.

Generating a concise dataset without overlapping proteins

Even though each peptide was only assigned to one protein there were still a large number of the repeated proteins due to one protein can contain more than one MHC class II binding peptide. After deleting redundant, we had a dataset of 209 sequences consisting of 162 human, 24 viruses, 19 bacteria, 4 parasites and 47 non-human.

After predicting the protein functions with the ProtFun server, the remaining dataset (Table 2, Set 2) consisted of: 159 human, 21 virus, 19 bacterium, 4 parasite and 44 nonhuman sequences (203 in total).

Generating a random protein dataset

To find out if the MHC class II-binding proteins have any significant function we generated two datasets of proteins randomly chosen from Swiss-Prot with the same size as the overlapping and nonoverlapping datasets described above.

Processing the protein sequences on ProtFun 2.1 server

ProtFun is a method for prediction of protein function for a subset of classes from the Gene Ontology classification scheme. This subset including several pharmaceutically interesting categories: transcription factors, receptors, ion channels, stress and immune response proteins, hormones and growth factors can also be predicted. The method relies on protein sequences as the sole input, it does not rely on sequence similarity, but instead on sequence derived protein features such as predicted post translational modifications (PTMs), protein sorting signals and physical/chemical properties calculated from the amino acid composition. This allows for prediction of the function for orphan proteins which no homology can be found (8).

For each input sequence, the ProtFun server (http://www.cbs.dtu.dk/services/ProtFun/) predicts cellular role, enzyme class and Gene Ontology category. The output scores consist of two numbers: the first number is the estimated probability that the entry belongs to the class in question. The second number represents the odds that the sequence belongs to that class/category. The first number is used in this study to compare with the random dataset to get the information about differences in function.

Function	Total	Human	Nonhuman	Viruses	Bacteria	Parasites
Amino acid biosynthesis	-3.91	-3.51	-1.84	-2.37	0.08	-2.76
Biosynthesis of cofactors	-1.06	-0.36	-1.64	-1.28	-1.40	0.17
Cell envelope	9.70	10.23	0.91	0.09	-0.35	4.71
Cellular processes	-0.16	0.39	-1.31	-0.80	-0.56	-1.97
Central intermediary metabolism	0.71	1.01	-0.50	-2.52	0.74	1.75
Energy metabolism	-5.62	-5.09	-2.59	-5.14	1.41	-1.75
Fatty acid metabolism	-4.62	-2.40	-5.65	-5.08	-3.25	-0.34
Purines and pyrimidines	7.18	8.22	-0.54	-1.74	0.77	1.04
Regulatory functions	-7.78	-10.12	2.40	3.25	0.28	-1.70
Replication and transcription	-12.56	-14.29	0.83	1.78	0.34	-3.27
Translation	-8.11	-10.84	0.49	0.73	1.24	-3.45
Transport and binding	13.26	14.70	0.90	0.67	-0.97	5.45
Enzyme	-3.21	-2.23	-3.10	-2.70	-1.42	-0.92
Nonenzyme	3.21	2.23	3.10	2.70	1.42	0.92
Oxidoreductase (EC1.-.-.-)	-3.23	-2.09	-3.28	-5.36	0.44	-1.24
Transferase (EC2.-.-.-)	-4.41	-4.16	-1.58	-2.20	0.59	-2.16
Hydrolase (EC3.-.-.-)	3.68	3.93	-0.03	-2.35	1.15	2.95
Lyase (EC4.-.-.-)	-1.00	-0.65	-0.99	-1.61	0.78	-1.62
Isomerase (EC5.-.-.-)	-4.89	-3.83	-3.74	-1.46	-3.45	-1.76
Ligase (EC6.-.-.-)	-1.05	-0.52	-1.45	-2.09	-0.53	1.47
Signal transducer	6.82	7.35	-0.93	-1.80	-1.14	4.69
Receptor	10.09	10.98	-3.11	-2.77	-1.99	0.43
Hormone	0.34	0.40	-0.76	-0.14	-1.24	1.39
Structural protein	-0.39	0.07	-0.91	-0.83	-1.32	0.62
Transporter	-1.82	-2.70	1.59	2.13	-1.03	-1.33
Ion channel	0.99	0.70	1.03	1.22	-2.51	-2.06
Voltage-gated ion channel	-4.47	-4.98	0.15	0.59	-1.32	-1.75
Cation channel	1.92	1.49	1.43	1.47	-2.98	0.00
Transcription	-7.33	-8.28	1.19	1.60	-0.43	-1.95
Transcription regulation	-7.43	-8.83	2.95	3.10	2.43	-3.35
Stress response	9.03	9.54	-0.08	0.05	-1.58	4.37
Immune response	11.33	12.22	-1.54	-2.00	-0.50	2.25
Growth factor	1.04	0.91	0.50	0.79	0.49	-1.73
Metal ion transport	-0.18	-2.27	3.01	3.13	-0.42	1.64

Figure 1. The z-scores of comparing MHC class II binding proteins with the random proteins. The first column lists the 34 functional classes predicted in ProtFun server. The columns 2 to 7 demonstrated the E-scores for each group of proteins from Set 1. Each square represents the test result for each function in related group. The positive number goes red; the lower negative number goes blue.

For any function assignment method the ability to correctly predict the relationship depends strongly on the function classification scheme used. The approach to function prediction used by ProtFun is based on the fact that protein is not alone when performing its biological task. It will have to operate using the same cellular machinery for modification and sorting as all the other proteins do (post-translational modifications-PTMs). PTMs are essential for the prediction of several functional classes. In addition to attributes related to sub-cellular location the most important features for predicting if a protein is, for example, regulatory or not, are PTMs. Similarly PTMs are very important for correct assignment of proteins related to the cell envelope, replication and translation.

Statistics and analysis

Z-test and *t*-test: We compared the result from the ProtFun prediction for each of the 34 functional classes of MHC class II binding proteins with proteins in the random dataset. The viruses, bacteria and parasites group are compared using a

Function	Total	Human	Nonhuman	Viruses	Bacteria	Parasites
Amino acid biosynthesis	-3.91	0.00	0.00	0.00	0.00	0.00
Biosynthesis of cofactors	0.00	0.00	0.00	0.00	0.00	0.00
Cell envelope	9.70	10.23	0.00	0.00	0.00	4.71
Cellular processes	0.00	0.00	0.00	0.00	0.00	0.00
Central intermediary metabolism	0.00	0.00	0.00	0.00	0.00	0.00
Energy metabolism	-5.62	-5.09	0.00	-5.14	0.00	0.00
Fatty acid metabolism	-4.62	0.00	-5.65	-5.08	0.00	0.00
Purines and pyrimidines	7.18	8.22	0.00	0.00	0.00	0.00
Regulatory functions	-7.78	-10.12	0.00	0.00	0.00	0.00
Replication and transcription	-12.56	-14.29	0.00	0.00	0.00	0.00
Translation	-8.11	-10.84	0.00	0.00	0.00	0.00
Transport and binding	13.26	14.70	0.00	0.00	0.00	5.45
Enzyme	0.00	0.00	0.00	0.00	0.00	0.00
Nonenzyme	0.00	0.00	0.00	0.00	0.00	0.00
Oxidoreductase (EC1.-.-.-)	0.00	0.00	0.00	-5.36	0.00	0.00
Transferase (EC2.-.-.-)	-4.41	-4.16	0.00	0.00	0.00	0.00
Hydrolase (EC3.-.-.-)	0.00	3.93	0.00	0.00	0.00	0.00
Lyase (EC4.-.-.-)	0.00	0.00	0.00	0.00	0.00	0.00
Isomerase (EC5.-.-.-)	-4.89	-3.83	0.00	0.00	0.00	0.00
Ligase (EC6.-.-.-)	0.00	0.00	0.00	0.00	0.00	0.00
Signal transducer	6.82	7.35	0.00	0.00	0.00	4.69
Receptor	10.09	10.98	0.00	0.00	0.00	0.00
Hormone	0.00	0.00	0.00	0.00	0.00	0.00
Structural protein	0.00	0.00	0.00	0.00	0.00	0.00
Transporter	0.00	0.00	0.00	0.00	0.00	0.00
Ion channel	0.00	0.00	0.00	0.00	0.00	0.00
Voltage-gated ion channel	-4.47	-4.98	0.00	0.00	0.00	0.00
Cation channel	0.00	0.00	0.00	0.00	0.00	0.00
Transcription	-7.33	-8.28	0.00	0.00	0.00	0.00
Transcription regulation	-7.43	-8.83	0.00	0.00	0.00	0.00
Stress response	9.03	9.54	0.00	0.00	0.00	4.37
Immune response	11.33	12.22	0.00	0.00	0.00	0.00
Growth factor	0.00	0.00	0.00	0.00	0.00	0.00
Metal ion transport	0.00	0.00	0.00	0.00	0.00	0.00

Figure 2. The Bonferroni corrected z-test results of comparing MHC class II binding proteins with the random proteins. The first column lists the 34 functional classes processed in ProtFun serve. The columns 2-7 demonstrate the E-scores for each group of proteins of set 1. The significant results are assigned a color either in red or in blue, and the non significant one are in green.

t-test because the sample sizes do not exceed 50. Instead, the human, nonhuman and total groups applied with z-test.

Bonferroni correction: to avoid the false positive cases, we corrected the *p* values using a Bonferroni correction. Each *p* value is multiplied with the number of test in the function-comparing. If the corrected *p*-value is still below 0.05, we will accept the result as being significant.

Assigning the color scale according to the statistic result

For the purpose of comparing the MHC class II dataset with random dataset, the statistical outcome was visualized in a coloured table. The positive numbers appear in red; the negative numbers appear in blue.

Results

Human group

After Bonferroni correction, the human proteins, in the functional classes' cell envelope, purines and pyrimidines, transport and binding, signal transducer, receptor, stress

Function	Total	Human	Nonhuman	Viruses	Bacteria	Parasites
Amino acid biosynthesis	-0.86	-0.73	-0.50	-0.33	-0.08	-1.96
Biosynthesis of cofactors	-0.70	-0.43	-0.64	-0.27	-1.17	0.46
Cell envelope	4.70	4.93	0.55	0.30	-0.96	3.07
Cellular processes	0.25	1.10	-1.55	-0.93	-1.07	-1.00
Central intermediary metabolism	-0.32	-0.31	-0.12	-1.52	0.39	1.51
Energy metabolism	-2.65	-2.28	-1.47	-2.22	-0.03	-0.75
Fatty acid metabolism	-1.76	-0.13	-3.18	-2.62	-2.38	0.67
Purines and pyrimidines	4.66	5.04	0.60	-0.14	0.65	0.83
Regulatory functions	-2.94	-4.74	1.47	2.18	0.68	-1.80
Replication and transcription	-4.55	-5.49	0.24	1.13	0.44	-3.15
Translation	-4.10	-4.57	-0.85	-1.14	1.09	-2.90
Transport and binding	5.44	5.79	1.10	0.47	0.07	3.37
Enzyme	-2.25	-1.73	-1.60	-1.18	-1.23	0.01
Nonenzyme	2.25	1.73	1.60	1.18	1.23	-0.01
Oxidoreductase (EC1.-.-.-)	-1.36	-0.76	-1.45	-1.87	-0.33	-0.38
Transferase (EC2.-.-.-)	1.64	-1.55	-0.62	-0.65	0.02	-0.85
Hydrolase (EC3.-.-.-)	2.79	2.83	0.50	-1.66	1.33	2.57
Lyase (EC4.-.-.-)	-0.85	-0.96	-0.08	-0.15	0.47	-1.14
Isomerase (EC5.-.-.-)	5.00	-1.81	-0.83	-2.29	0.19	-2.80
Ligase (EC6.-.-.-)	-0.74	-0.70	-0.29	-1.31	0.47	1.71
Signal transducer	2.79	3.02	-0.07	-0.96	-0.33	2.77
Receptor	4.82	5.21	-0.73	-0.83	-0.90	2.17
Hormone	0.10	0.22	-0.99	-0.45	-1.09	0.73
Structural protein	-0.43	0.35	-1.28	-0.86	-1.56	0.18
Transporter	-1.38	-1.64	0.46	0.54	-1.33	-1.44
Ion channel	0.62	0.80	-0.42	-0.26	-2.36	-1.22
Voltage-gated ion channel	-0.79	-0.91	0.06	-0.14	1.19	-2.58
Cation channel	1.14	1.25	-0.08	-0.08	-1.49	0.00
Transcription	-2.32	-3.29	2.00	2.50	0.30	-1.72
Transcription regulation	-2.74	-3.74	1.94	2.00	2.23	-2.87
Stress response	5.03	5.29	0.38	0.47	-0.93	2.71
Immune response	5.73	6.07	0.14	-0.46	0.19	1.60
Growth factor	1.10	0.79	0.80	1.10	0.76	-1.68
Metal ion transport	-0.86	-1.78	1.72	1.41	1.29	1.51

Figure 3. The z-scores of comparing MHC class II binding proteins with the random proteins. The first column lists the 34 functional classes processed in ProtFun server. The columns 2-7 demonstrated the E-scores for each group of proteins of Set 2.

response and immune response, were significantly over-represented for proteins with MHC class II binding peptides ($p < 0.01$ after correction for multiple testing). Presented proteins were also less likely to play a role in energy metabolism, regulatory function, replication and transcription, translation or voltage-gated ion channel (Figures 1, 2 and 5).

Nonhuman group

In the nonhuman group the proteins involved in fatty acid metabolism were significantly under-represented ($p < 0.05$ after correction for multiple testing). This was especially found in virus proteins, which also had less relationship with energy metabolism. In parasites, several characters such as cell envelope, transport, binding, signal transducer and stress response were highly over-represented comparing with random dataset (Figures 2, 3 and 5).

Comparing human and nonhuman group

In general, there is no strong correlation between the proteins that are presented by MHC class II molecules in the human and the non-human group. The results from the parasite

Function	Total	Human	Nonhuman	Viruses	Bacteria	Parasites
Amino acid biosynthesis	0.00	0.00	0.00	0.00	0.00	0.00
Biosynthesis of cofactors	0.00	0.00	0.00	0.00	0.00	0.00
Cell envelope	4.70	4.93	0.00	0.00	0.00	0.00
Cellular processes	0.00	0.00	0.00	0.00	0.00	0.00
Central intermediary metabolism	0.00	0.00	0.00	0.00	0.00	0.00
Energy metabolism	0.00	0.00	0.00	0.00	0.00	0.00
Fatty acid metabolism	0.00	0.00	0.00	0.00	0.00	0.00
Purines and pyrimidines	4.66	5.04	0.00	0.00	0.00	0.00
Regulatory functions	0.00	-4.74	0.00	0.00	0.00	0.00
Replication and transcription	-4.55	-5.49	0.00	0.00	0.00	0.00
Translation	-4.10	-4.57	0.00	0.00	0.00	0.00
Transport and binding	5.44	5.79	0.00	0.00	0.00	0.00
Enzyme	0.00	0.00	0.00	0.00	0.00	0.00
Nonenzyme	0.00	0.00	0.00	0.00	0.00	0.00
Oxidoreductase (EC1.-.-.-)	0.00	0.00	0.00	0.00	0.00	0.00
Transferase (EC2.-.-.-)	0.00	0.00	0.00	0.00	0.00	0.00
Hydrolase (EC3.-.-.-)	0.00	0.00	0.00	0.00	0.00	0.00
Lyase (EC4.-.-.-)	0.00	0.00	0.00	0.00	0.00	0.00
Isomerase (EC5.-.-.-)	5.00	0.00	0.00	0.00	0.00	0.00
Ligase (EC6.-.-.-)	0.00	0.00	0.00	0.00	0.00	0.00
Signal transducer	0.00	0.00	0.00	0.00	0.00	0.00
Receptor	4.82	5.21	0.00	0.00	0.00	0.00
Hormone	0.00	0.00	0.00	0.00	0.00	0.00
Structural protein	0.00	0.00	0.00	0.00	0.00	0.00
Transporter	0.00	0.00	0.00	0.00	0.00	0.00
Ion channel	0.00	0.00	0.00	0.00	0.00	0.00
Voltage-gated ion channel	0.00	0.00	0.00	0.00	0.00	0.00
Cation channel	0.00	0.00	0.00	0.00	0.00	0.00
Transcription	0.00	0.00	0.00	0.00	0.00	0.00
Transcription regulation	0.00	0.00	0.00	0.00	0.00	0.00
Stress response	5.03	5.29	0.00	0.00	0.00	0.00
Immune response	5.73	6.07	0.00	0.00	0.00	0.00
Growth factor	0.00	0.00	0.00	0.00	0.00	0.00
Metal ion transport	0.00	0.00	0.00	0.00	0.00	0.00

Figure 4. The bonferroni corrected z-test results of comparing MHC class II binding proteins with the random proteins. The first column lists the 34 functional classes processed in ProtFun server. The columns 2-7 demonstrated the E-scores for each group of proteins of Set 2.

group were similar to those from human group. The negative association with proteins involved in energy metabolism was shared in both virus and human group (Figures 2, 4 and 5).

Comparing overlapping and nonoverlapping group

Due to the reduced size of the dataset, the only statistically significant result were over-representation of proteins associated with cell envelope, purines and pyrimidines, transport and binding, receptor, stress response and immune response, and under-representation of proteins associated with regulatory function, replication, transcription and translation (Figures 4 and 5).

Discussion

We found that human proteins that are presented by MHC II molecules are more likely to be related to the functions: cell envelope, purines and pyrimidines, transport and binding, signal transducer, receptor, stress response and immune response. The proteins are less likely involved in amino acid

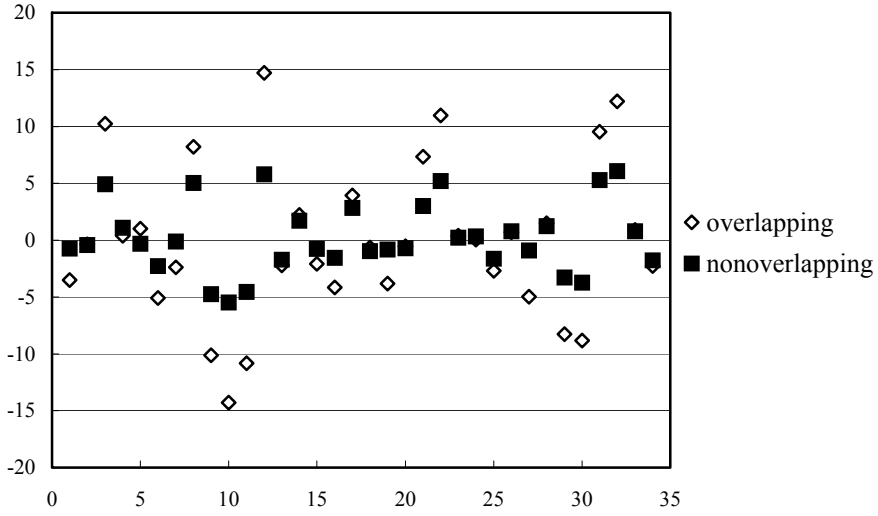

Figure 5. The comparison of the results of different MHC class II binding protein datasets. The x-axis represents the 34 functional classes on ProtFun server. The y-axis represents the results of z-test statistics. Each spot represents the E value for each function. The unfilled spots correspond to comparing the overlapping dataset (453 amino acid sequences) with random dataset. The fills spots correspond to comparing the nonoverlapping (159 amino acid sequences) dataset with random dataset. The variation direction of these spots tends to be identical and the different values between unfilled and filled spots due to the data size.

biosynthesis, energy metabolism, fatty acid metabolism, regulatory function, replication and transcription, translation, voltage-gated ion channel, transcription and transcription regulation

Therefore it seems that the MHC class II binding proteins apparently are those that appear on the surface of cells. The receptor or envelope functions may be over represented because they are the functions that the membrane proteins usually have. It seems likely that during the antigen processing involved with MHC class II antigen presentation, the antigen molecules are from the surface of self-cells and processed by antigen presenting cells. Still purines and pyrimidines remained over-represented in human group, which is an exception for the idea of membrane antigens because functions related to purines and pyrimidines are not typically found for membrane proteins.

Generally in humans the surface proteins or membrane proteins are found to be more likely to be presented by MHC class II molecules. These proteins may also be over-represented in thymus during the maturation of T lymphocytes. The thymocytes bearing high-affinity receptors for self-MHC molecules are deleted which results in self-tolerance. The thymocytes with that high affinity receptor are less likely to survive but if they do which can lead to strong autoimmune response.

Among the pathogens, the presented proteins from parasites tend to have functions very similar to those found in the human group, but for the viruses and bacteria the presented proteins were markedly different. The most over-presented functions were transcription regulation and the most under-presented one was fatty acid metabolism.

Acknowledgements

The authors would like to thank all of members in immunology group at CBS-BioCertrum in the Technical University of Denmark, especially Mette Børgesen, Morten Nielsen, Claus Lundegaard, and Peter Wad Sackett who helped start the project, and contribyted helpful scientific as well as statistic discussions.

References

1. Kuby J. Immunology, 3rd ed. New York: WH Freeman; 1997.
2. Cresswell P. Assembly, transport, and function of MHC class II molecules. Annu Rev Immunol. 1994;12:259-293.
3. Sette A, Chesnut R, Fikes J. HLA expression in cancer: implications for T cell-based immunotherapy. Immunogenetics. 2001;53: 255-263.
4. Yewdell J, Bennink J. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. Annu Rev Immunol. 1999;17:51-81.
5. Jensen LJ, Gupat R, Blom N, et al. Prediciton of Human Protein Function from Post-translatioal Modifications and Location Features. J Mol Biol. 2002;319:1257-1265.
6. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics. 1999;50:213-219.
7. Engelhard VH. Structure of peptides associated with class I and class II MHC molecules. Annu Rev Immunol. 1994;12:181-207.
8. Jensen LJ, Stoerfeldt HH, Brunak S. Prediction of human protein function according to Gene Ontology categories. Bioinformatics. 2003;19:635-642.