Identification of Viral and Bacterial Triggers for Human Autoimmune Diseases

by

Peter J. Bluvas Jr.

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology

May 24, 2002

Copyright 2002 Peter J. Bluvas Jr. All rights reserved.

The author hereby grants to M.I.T. permission to reproduce and distribute publicly paper and electronic copies of this thesis and to grant others the right to do so.

Peter J. Bluvas Jr.
Department of Electrical Engineering and Computer Science

Isaac Kohane MD, PhD
Thesis Supervisor

Arthur C. Smith
Chairman, Department Committee on Graduate Theses
Identification of Viral and Bacterial Triggers for Human Autoimmune Diseases

by

Peter J. Bluvas Jr.

Submitted to the
Department of Electrical Engineering and Computer Science

May 24, 2002

In Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Associations have been noted between infection with bacteria or viruses and initiation of human autoimmune diseases. One proposed explanation for this observation is that microbial antigens are similar to human (self) antigens, and that a molecular mimicry event is responsible for the initiation of an autoimmune response. T cells, part of the human immune system, recognize short linear stretches of proteins. Therefore identifying molecular mimics can be accomplished by finding local alignments between microbial and human proteins. Candidate human proteins specific to diabetes, autoimmune thyroid disease, arthritis, and myocarditis were compared to the proteomes of microbes with a known association with some human autoimmune disease or that are common in North America. This search identified a large number of potential molecular mimicry target sites. The alignments found were distributed rather uniformly across all human and microbial proteins included in the study.

Thesis Advisor: Isaac Kohane MD, PhD
Title: Research Affiliate, MIT Laboratory for Computer Science
Contents

Abstract ............................................................................................................................... 2
Introduction ...................................................................................................................... 4
Biology of the Immune System ...................................................................................... 5
Formulation as a Computer Science Problem ............................................................. 9
Construction of Human Query Lists ........................................................................... 10
Implementation ............................................................................................................. 19
Results ............................................................................................................................ 22
Discussion ...................................................................................................................... 27
References .................................................................................................................... 31
Introduction

The immune system has evolved to provide a diverse T cell repertoire that can effectively fight off infection from foreign invaders while maintaining regulatory mechanisms to prevent cross-reaction with self-peptides. A loss of self-tolerance, and the consequent immune destruction of host tissues, results in a state of autoimmunity. Examples of autoimmune diseases include rheumatoid arthritis, multiple sclerosis, insulin-dependent diabetes mellitus, and autoimmune thyroid disease. Some estimates indicate that more than 20% of the population suffers from an autoimmune disease.\(^1\)

Human autoimmune diseases are thought to develop through a complex combination of genetic and environmental factors. Genomic linkage studies have shown that there are specific genetic loci that contribute to corresponding autoimmune disorders.\(^2,3\) These results suggest that while these diseases produce very different effects in the body, they may be initiated by a similar mechanism. Therefore, techniques applied to the study of one particular disease can often be applied to other diseases as well.

Frequently, associations have been noted between infection with viruses or bacteria and onset of autoimmune disease.\(^4\) Although the exact mechanism for the development of a specific autoimmune response is not known, one proposed explanation is molecular mimicry. In this model, there is structural similarity between proteins of an infectious agent and those of the self, so that infection by a microbe leads to the activation of autoreactive T cells.\(^5\) This initial activation could then induce the expansion of autoreactive T cells or further expansion of a memory T cell population, thus leading to a state of autoimmunity.
The T-cell receptor (TCR) recognizes short stretches of peptides (8 to 11 amino acids)\(^5\) bound to major-histocompatibility (MHC) molecules on antigen-presenting cells (APCs). Therefore the mimic in this model is short, nearly identical linear peptide sequences shared between the host and the microbe. This type of similarity can be detected using current sequence alignment algorithms and publicly available databases of human and microbial proteins. A search for molecular mimicry candidates was performed, not based on a candidate gene or protein basis, but by comparing human proteins uniquely expressed in various tissues with entire microbial proteomes.

**Biology of the Immune System**

The human body maintains a set of immune cells called lymphocytes. Each lymphocyte bears an antigen receptor with a single specificity. The lymphocytes in the body, of which there are millions, collectively carry the antigen receptor specificities that make up the lymphocyte repertoire of the individual. This repertoire is available to the individual after birth, but only those lymphocytes that encounter their specific antigen will become activated.\(^6\)

The two major types of lymphocytes are called B and T cells. When activated, B cells differentiate into plasma cells, which secrete antibodies to fight infection. T cells are further divided into two main classes. On activation, one class differentiates into specialized cells for eliminating infected host cells. The second class differentiates into cells that coordinate the activities of other immune cells.\(^6\)
T cells do not recognize standalone antigens; rather, they recognize antigens that are displayed on the surfaces of specialized immune cells known as antigen-presenting cells (APCs). The function of APCs in the body is to take up suspicious cells and extracellular material, degrade them into small fragments, and display them on their surface for presentation to T cells. The T-cell antigen receptor (TCR) is specialized for the recognition of antigen as a small peptide fragment bound to a major histocompatibility complex (MHC) molecule on the surface of an APC.\[^{6,7}\]

---

**Figure 1:** APC digesting a bacterium and displaying it for inspection by a T cell. This demonstrates the processing that is done to an antigen before it is recognized by a T cell. The details of this mechanism will differ slightly for specific microbes and APC's.

---

When a naïve T cell (one that has never been activated) encounters the specific MHC:peptide complex to which its receptor binds, that lymphocyte becomes activated and initiates the immune system’s primary response. This activated T cell will then start to proliferate. Most of these activated T cell progeny differentiate to become short-lived effector cells capable of combating the infection. Other progeny become long-lived.
memory cells, specific for the same MHC:peptide complex of their single ancestor T cell.\[^{[6]}\]

These memory T cells are more sensitive to stimulation by antigen than are naïve cells, and they more quickly and vigorously respond to such stimulation. In addition, a large number of these memory cells remain in the body even after the primary response has eliminated the infection that triggered it. After the primary response, the number of T cells reactive to an antigen persists at a level that is 100 to 1000 times above the initial concentration.\[^{[6]}\]

Molecular mimicry is one mechanism that has been proposed for the onset of autoimmune disease. Consider the following scenario. An antigenic determinant on a foreign bacterial or viral protein is structurally similar to a determinant on one of the proteins made by the host, although different enough to be recognized as foreign by the host’s immune system. Upon recognition of this foreign microbial antigen the host’s immune system would generate a primary response. Since the memory cells created by the primary response persist in greater concentrations in the body and will now bind that antigen with a higher affinity than naïve cells, they may become activated by the host protein, thus creating an autoimmune response.\[^{[5]}\]
An autoreactive T cell exists

The autoreactive T cell is activated to fight off a microbial infection.

The activated T cell or its progeny cross-react with self antigens

Figure 2: Illustration of the molecular mimicry concept. Initially there exists some population of autoreactive T cells in the host, but they remain ignorant to host proteins because these proteins are expressed at very low levels. Infection by a microbe leads to the activation of these T cells in a normal immune response. Since the microbe shares an antigen with the host, these activated T cells then begin to attack host cells.
Formulation as a Computer Science Problem

The necessary property for this disease model is structural similarity between foreign and host proteins. Although public databases of human and microbial proteomes are readily available, the task of identifying structural similarity between proteins is very difficult. In this disease mechanism, T cells recognize linear peptide stretches of 8 to 11 amino acids. This unique biological property reduces the problem to identifying short, nearly identical portions of human and microbial protein. Computational algorithms and strategies can now be used to solve this simplified problem. Through the use of these techniques it is hoped that results can be obtained which will provide insight into biological truths that would have otherwise been too difficult to study in vivo. At the very least, results should be consistent with current biological knowledge.

Figure 3: A protein can be viewed as a linear sequence of amino acids chained together into one very large molecule. Candidates for molecular mimicry are stretches of 8 to 11 amino acids which are very similar in the human and microbial proteins. Such regions can be identified using computer science string alignment algorithms.
Construction of Human Query Lists

For a given autoimmune disease, there is a specific tissue in the body that is attacked by the immune system. Therefore the epitopes involved in a molecular mimicry event would derive from proteins expressed in these tissues. Each cell in the human body contains the entire genomic code. However, only a fraction of those genes are expressed in any tissue at any given time. Cells from very different tissue types utilize the same subset of the genome to carry out many basic biological processes such as metabolism, waste removal, and respiration. However, there is some degree of differential expression which is responsible for the numerous types of tissues found in the body. Since a given autoimmune response attacks only a single type of tissue, the immune system must be recognizing a protein that is expressed solely in that tissue type. Therefore to construct a candidate list of potential molecular mimics for a particular autoimmune disease, one would look for proteins expressed uniquely in the human tissue associated with that disease.

The National Institute of Health (NIH), through its National Center for Biotechnology Information (NCBI) maintains a database of gene sequences known as GenBank. GenBank contains the results of thousands of independent gene sequencing experiments. Entries in GenBank may be redundant, may contain only partial sequences of genes, or may be incorrect. NCBI also maintains an experimental system called Unigene which organizes these GenBank sequences into non-redundant gene-based clusters. In this manner all representations of the transcription product of a single human gene are collected into a unique Unigene cluster.
Figure 4: High-level organization of gene and protein databases available from NCBI. GenBank is a repository for raw data from sequencing experiments of genes and proteins. Unigene is a continually updated, experimental system that tries to organize GenBank submissions into non-redundant clusters which are believed to represent actual genes.

One way to construct a list of potential molecular mimicry targets is to search this Unigene database for keywords related to the disease of interest. For example to find proteins involved in diabetes one could search for gene names containing ‘pancreas’, ‘islet’, or ‘diabetes’. This method will identify a large number of candidate genes, but it has two major drawbacks: it misses genes that do not contain the keywords in their description and it does not find uniquely expressed genes. For example, actin is a protein found primarily in muscle cells, but it would not be found by this search because it contains no relevant keywords in its description. Likewise a keyword such as insulin would return many proteins expressed in the liver, and not just those from the islet cells in the pancreas.

An alternative search strategy involves making use of the mRNA expression profiles of cells from different tissues. In this method the mRNA from a homogeneous population of cells is isolated and then sequenced. Next these sequences are compared to the list of known genes in the genome. This provides a snapshot of what genes are being
actively transcribed in the cells at a given time. Unigene contains tissue-specific RNA expression data. This data is organized into libraries, where each library represents the results of a single experiment as described above. RNA from a single type of tissue is collected and sequenced. Each mRNA in the library is then compared to all the known Unigene clusters (representing known genes) to determine which cluster it aligns best with. Ideally each mRNA found in a given tissue library has been assigned to a known Unigene cluster, but these sequences can not always be aligned with a known cluster. This unique pattern of gene expression provides a fingerprint by which the library can be compared to other libraries.

Digital differential display is a computational method for comparing the patterns of gene expression between collections of these libraries. In previous studies this tool has been used to identify differential gene expression between cancer cells and their normal, healthy counterparts \cite{9}. In this project, DDD was used to identify genes uniquely expressed in a given tissue by performing an analysis of expression patterns among libraries from various tissue types.

DDD makes comparisons of the number of times sequences from different libraries are assigned to a particular Unigene cluster. Since there will be a large number of differences, a statistical test known as the Fisher Exact Test is used to determine which differences are significant. This test is employed to maximize the probability that the observed differences reflect true biological variation. For example, this test takes into account what fraction of sequences from each of the libraries was successfully assigned to some Unigene cluster.
In this study, lists of potential human molecular mimicry targets were constructed for insulin-dependent diabetes mellitus, autoimmune thyroid disease, arthritis, and myocarditis. The diabetes target set is composed of proteins from the “pancreas” library. For the autoimmune thyroid disease set the “thyroid” and “thyroid gland” libraries were used, for arthritis the “bone,” “cartilage,” and “connective tissue” libraries were used, and for myocarditis, the “heart” library was used. In addition, two control lists were created: the first consists of proteins from the “skin” library and the second from the “lung” library. Each of these six potential mimicry target lists was compared to an aggregate list consisting of libraries from a number of other human tissues available at the NCBI site. A detailed listing of the specific Unigene libraries used in these comparisons can be found in Table 1.

<table>
<thead>
<tr>
<th>Human Target Set</th>
<th>Unigene Libraries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>6344 229 2316 45 6949</td>
</tr>
<tr>
<td>Autoimmune Thyroid Disease</td>
<td>416 4885</td>
</tr>
<tr>
<td>Arthritis</td>
<td>5996 607 6356 1004</td>
</tr>
<tr>
<td>Myocarditis</td>
<td>241</td>
</tr>
<tr>
<td>Skin</td>
<td>3625 3583 936 6347 6346 4691 6330 6959</td>
</tr>
<tr>
<td>Lung</td>
<td>152 4774 1365 4913 3622 403 404 4875 3623</td>
</tr>
</tbody>
</table>

Table 1: There were six human target lists constructed in this study. The first four were composed of tissues that correspond to autoimmune diseases. The other two were of tissues that have associated autoimmune disorders, but these were used as control tissues in this study. The tissue libraries from the Unigene database used to construct each of the human target lists is given here.

Using the online DDD tool, these comparisons yielded a list of Unigene clusters with significant differential expression for each tissue type. Only those proteins from each of the six target lists that exhibited an expression level with a >10-fold difference as compared to the aggregate list, were considered in this study. The consensus protein
sequence identification number was then obtained for each of these clusters using the LocusLink database, also available for download from the NCBI site. This list of protein reference numbers was then entered into BatchEntrez, an NCBI data retrieval system which returned the amino acid sequences for each protein in FASTA format. This list of FASTA formatted human target lists is available online\(^1\).

\(^1\) http://web.mit.edu/pjbluvas/Public/thesis/results/intro_page.htm
Construction of Microbial Protein Databases

The next step was finding microbial proteins that could be potential molecular mimics of human proteins. In eliminating viruses and bacteria from the body, the immune system breaks down the microbes into small peptide fragments. It is these fragments, or determinants, which are then recognized by T cells during an immune response. This implies that any protein expressed by the microbe could contain the cross-reactive antigen which is recognized by the immune system. Therefore, the entire proteome, or all the known expressed proteins, of an organism need to be considered as candidates in a molecular mimicry search.

Information about entire genomes of organisms is collected at another NCBI database known as Entrez. This service contains all the known expressed genes for a number of organisms, as well as their corresponding protein transcription products. Viral and bacterial proteomes were downloaded from this service at the corresponding websites\(^2\)\(^3\).

Microbes that are common in North America or that have a noted association with a particular autoimmune disease were chosen because these are the most likely candidates for this model of disease onset. The following microbes were used: *borrelia burgdorferi*, *campylobacter jejuni*, *escherichia coli*, *haemophilus influenzae*, *helicobacter pylori*, *mycoplasma genitalium*, *coxsackievirus*, *foot-and-mouth disease virus*, *influenza*, *human herpesvirus*, *human papillomavirus*, *rubella virus*, *variola virus*, and *yersinia*.

**BLAST search algorithm**

All proteins are composed of chains of amino acids which may fold to produce very complicated three-dimensional structures. However, all proteins can be thought of as a linear sequence of amino acids, and therefore sequence alignment algorithms can be applied to a protein when viewed in this manner. There are many ways to align two sequences. Global alignment involves finding an optimal way to align two entire sequences. Local alignment (without gaps) simply refers to finding two same length subsequences from each of the two sequences being compared. This implies that there can be multiple alignments for a single comparison of two proteins. The Basic Local Alignment Search Tool (BLAST)\[^{10}\] is a standard local alignment algorithm used by computational biologists. This program was used to identify alignments between the human and microbial target lists.

![Global Alignment](image1)

![Local Alignment](image2)

**Figure 5: Generic depiction of global and local alignments between two sequences.**

The local alignment seeks only to align subsequences of any size from the two sequences.

Due to the biochemistry of amino acid side chain residues, certain amino acids can more easily be substituted without a major effect on biological function. BLAST makes use of an amino acid substitution matrix to capture these biochemical properties. This matrix assigns a score to each possible pairing of amino acids that could occur. Identities and common substitutions receive a positive score and unlikely replacements
are given a negative score. Using this matrix, the algorithm can calculate a raw score $S$ for a given alignment. The raw similarity score $S$ for an alignment of two same-length protein segments is the sum of the matrix similarity values for each pair of aligned amino acids. A local alignment is considered optimal if its score cannot be improved either by extending or trimming. The subsequence in such an alignment is known as a high-scoring segment pair (HSP).

It is possible for similar sequences of amino acids to arise simply by chance. Therefore a statistical measure is needed to evaluate the significance of alignments. BLAST uses a random sequence model based on the following assumptions: First, the amino acids are chosen independently according to their average background frequencies. Also the expected score for aligning any two random residues is expected to be negative. Given these assumptions, the expected number of HSP’s with a score of at least $S$ is:

$$E = Kmn \ e^{-\lambda S}$$

The parameters $K$ and $\lambda$ are specific to the database of sequences that is being searched. The p-value for a score $S$ is the probability of finding at least one such HSP in the sequences with a score of at least $S$. This p-value is related to the e-value by the following formula:

$$P = 1 - e^{-E}$$

Therefore the p-value and the e-value are interchangeable measures of the significance of an alignment. Most implementations of BLAST use the e-value since it more clearly delineates differences in significance over the practical range of values. Many of the results in this study are grouped according to the e-value that a given alignment received.
Likewise the raw score is normalized using these same parameters to produce a bit score:

\[ S' = \frac{\lambda S - \ln K}{\ln 2} \]

This normalized bit score allows results of separate searches to be directly compared since it takes into account factors such as the relative lengths of the sequences being aligned.

One additional complication that must be considered is the presence of gaps in the local alignments. When assigning a raw score, the presence of a gap is penalized in the same way as a mismatch between amino acids. Typically there is a Gap Penalty assigned for introducing a gap and a separate Gap Extension Penalty that is assigned for extending a gap.

BLAST does not directly calculate the parameters \( K \) and \( \lambda \) for each set of sequences that it is aligning. Instead the algorithm uses estimates of these parameters for a given combination of a substitution matrix and gap costs. BLAST searches for all HSP’s between two sequences with some maximum e-value.

Alignments between selected human proteins and the bacterial and viral proteomes were identified by using the BLASTp alignment program. This is an implementation of the BLAST algorithm available from NCBI at ftp://ftp.ncbi.nlm.nih.gov/ that compares protein sequences to protein sequences. A PAM30 scoring matrix, with gap penalty of seven and gap extension penalty of two was used for all searches. These parameters were used to bias the BLAST algorithm to identify highly conserved, relatively short peptide strings.
To determine if this algorithm was able to correctly detect alignments, it was tested on randomly generated databases with intentionally inserted matching subsequences. Five query peptide strings that were 50, 100, 200, 500, and 1000 amino acids in length were constructed. Each query peptide was comprised of random amino acids from a subset $S_1$ of all twenty amino acids. Each one of these query strings also contained a unique eight amino acid target string composed of random amino acids from subset $S_2$ of all twenty amino acids. The database strings of 50, 100, 200, 500 and 1000 amino acids were created in a similar manner. They too were comprised of random amino acids from a unique subset $S_3$ of all amino acids. Each one of these database strings contained all five target strings from the query strings. The sets $S_1$, $S_2$, $S_3$, were all disjoint. The query strings were aligned with the database strings using the BLASTp algorithm employing the same parameters as in the actual experiment (PAM30 Substitution Matrix, Gap Penalty of 7, Gap Extension Penalty of 2, and a minimum e-value of 10). This method was repeated using target strings containing six amino acids instead of eight. The BLASTp algorithm correctly identified all of the target strings in the artificial query and database strings that were constructed. It was successful in both the six and eight amino acids trials. Additionally, no false positive alignments were found.

**Implementation**

The BLAST algorithm was downloaded as a Linux executable with a command-line interface. The program requires inputs including the query and database files to be searched, the cutoff e-value of results to include, the substitution matrices to be used, and the gap penalties to be assigned. A single execution of the algorithm produces results for
a search between a single human target protein list and a single microbial proteome. The output of this execution is a text file containing information about all of the alignments that were detected in the search. A Perl program was written to automatically compare all combinations of human target lists and microbial proteomes.

Figure 6: BLAST searches were performed for every possible pairing of a human target protein list with a microbial proteome. Every pairing represents a single BLAST search whose output is a single text file. E-values and other search parameters are based on the statistics of this search design.

In this format the results are rather difficult to interpret since the raw output of BLAST searches are lengthy text files containing much redundant and extraneous information. Several other Perl programs were written to extract interesting features from these results. These programs all made use of a modified version of a BLAST results parsing program written by Nathaniel Strauss of the Whitehead Institute.

The first program converted the BLAST result files into HTML text files, allowing the alignments to be displayed on a web page. This helped in the initial exploration of the distribution of results across disease and microbes. It also helped to determine the types of alignments that were being found by the algorithm.

Next a program was written to count the number of alignments found for each microbe when compared to a single human target list. Since BLAST finds local alignments, there can be multiple alignments found when two protein sequences are
compared. A similar program returned the number of alignments found for each protein in each of the human target lists, when compared to all of the microbial proteomes. The outputs of both of these programs were tab-delimited text files. In calculating the number of alignments, both programs also grouped the results according to a series of threshold e-values (100, 10, 1, 0.1, 0.01, 0.001, 0.0001, 0.00001).

These tab-delimited files were opened using Microsoft Excel to create a spreadsheet of results. Excel was used to generate the graphs included in the Results section. Additionally, the human genes were sorted according to their biologic function using a gene classification database known as Gene Ontology which is available from Unigene. This database was entered in a Microsoft Access database, and queries were written to group the results according to their gene ontology.
Results

A complete listing of the results of these BLAST searches as well the human sequences that were used to perform them can be found online. An exploration of these results in a standard Web browser led to several interesting observations. First, there are a large number of alignments that were detected. Had there been only a few alignments found, then alternative methods would have been used to further study them. For example, given a large quantity of alignments it became more practical to tally the number of results returned instead of making detailed analyses of individual proteins and alignments.

A majority of the alignments fit the pattern that this study was designed to find; that is, they aligned short linear stretches of proteins. However, there were also some alignments that did not fit this blueprint. There were several human target proteins that were found to be homologous with microbial proteins. These proteins were very similar to long stretches (100 or more amino acids) of microbial proteins, and therefore returned a very large number of alignments. Since these alignments greatly affected the final values for totaling of results, they were omitted from those calculations. The actual alignments are still given, however, at the website listed above. These proteins are synaptotagmin V (Hs.23179) and ATP-binding cassette (Hs.54470) from the diabetes list and collagen type XI (Hs.82772) and collagen type V (Hs.82985) from the arthritis list.

The first view of the data looks at the total number of alignments found for each of the human target lists. The plots shown here give these totals for maximum e-values of 10 and 1. Alignments with lower e-values tended to be longer, and those with higher e-values tended to be very short. Again a striking feature of the data is the number of

4 http://web.mit.edu/pjbluvas/Public/thesis/results/intro_page.htm
potential molecular mimics that were identified. Looking at the number of alignments found for each of the six candidate human lists (Fig 7a) shows that even the controls had a number of successful matchings. Since the human target lists all contain a different number of proteins, this plot is somewhat misleading. Normalizing by number of proteins in each human candidate list provides another view of the data (Fig 7b). This plot can be interpreted as showing the potential for molecular mimicry that the associated tissue for a particular disease has. The density of alignments found across all human targets lists is relatively constant. It was expected that tissues associated with the most common autoimmune diseases would have the highest values on this plot.

Figure 7: (a) The number of alignments with e-values less than 10 and 1 are given for each of the six candidate human lists: diabetes, thyroid, arthritis, and myocarditis along with the two controls: skin and lung. (b) Plots the number of alignments normalized by the number of proteins in each of the candidate human lists.

An alternative view of the data shows the number of alignments found for each organism (Fig 8a) when compared against the human diabetes, arthritis, thyroid, and myocarditis target sets. Alignments were found for all the organisms included in this study. Since there were nearly one hundred microbe strains looked at in the study, it would be confusing to plot all of them on a single graph. Therefore, only the results for a
single representative of each microbial species are shown in this plot. A normalized view of the data (Fig 8b) is provided to account for the differing number of proteins in the various proteomes. The viruses, which have proteomes that are two orders of magnitude smaller than those of bacteria, tended to have a higher density of alignments.

Figure 8: (a) Plot of the total number of alignments found with e-value <= 1 in the diabetes, thyroid, arthritis, and myocarditis target sets. (b) Normalized plot showing the number of alignments found with e-value <= 1 divided by the number of proteins in that microbe’s proteome. For both plots, the order of organisms along the x-axis is (left to right): *borrelia burgdorferi*, *campylobacter jejuni*, *escherichia coli*, *haemophilus influenzae*, *helicobacter pylori*, *mycoplasma genitalium*, *coxsackievirus B5*, *coxsackievirus A*, *Ebola*, *foot-and-mouth disease virus*, *influenza A*, *influenza B*, *human herpesvirus*, *human papillomavirus*, *rubella virus*, *variola virus*, and *yersinia*. 
The third view of the data, is a breakdown of Figure 7. This shows the distribution of alignments found for each individual human protein in each of the four human disease target sets. It was expected that there would be a few human proteins which received many matches, and others that received very few. However, every human protein received at least one matching, and the hits were distributed relatively uniformly within and across the target sets.

![Figure 9: Plot of total number of alignments (e-value <= 10) found for each human protein in each of the four candidate sets. A hit is defined as an alignment with a microbial sequence that has an e-value less than 10. Each target list contains an evenly distributed number of hits across its constituent proteins with all of the human proteins in the study received at least one alignment with a microbial protein.](image)

These proteins were then grouped according to Gene Ontology as indicated in the Unigene database. The majority of proteins fell into five distinct ontology categories for cellular component. A plot of the total alignments for all human proteins which fell into
one of these categories is given in Figure 10, with the proteins sorted according to number of alignments within each category.

Figure 10: Human proteins were classified according to the cellular component they derived from, as indicated in the Unigene database. Proteins that belonged to any of the five most common categories are plotted here, sorted within each category by number of alignment (e-value ≤ 10).
Discussion

These results indicate that there are many potential molecular mimics between human and microbial T cell epitopes. Given the large number of candidates identified for this model, especially in the control tissues, one is surprised that autoimmunity is not a more common event. Any alignment of even a few amino acid residues between two proteins could allow for the occurrence of a molecular mimicry event. Therefore, any of the alignments detected in this study could potentially be biologically significant in that they are responsible for the triggering of a human autoimmune disease.

Even given the quantity of results found in this study, it is likely that the computational methods that were used overlooked many other potential molecular mimicry sites. Experiments using synthetic combinatorial peptide libraries have identified sets of peptides able to efficiently stimulate single T cell clones. These experiments found that sometimes only one or two conserved amino acid residues were necessary for a cross-reaction to occur. Cross-reactivity has been shown to occur without a high degree of sequence homology. Identifying motifs from such binding experiments could allow for a more refined search of protein databases that would better predict epitope binding affinity. The matrix substitution scores and statistical e-value calculations used by the BLAST algorithm may reject alignments that would cross-react in vivo. Using more refined alignment algorithms, it is expected that an even greater number of possible molecular mimicry sites will be identified.

It is important to realize that the pairings identified in these searches identify possible candidates for molecular mimicry. These results do not prove this model as a mechanism for disease onset. The regions identified in these alignments are not all
guaranteed to be T cell epitopes, or to be biologically active. Furthermore, linear sequence homology does not always guarantee cross-reactivity.\textsuperscript{[13]} Although linear amino acid sequence is involved in TCR recognition, this process is likely to involve more complicated secondary structure interactions as well.

A search algorithm could be designed which would identify all of these possible alignments. However, this would magnify the problem of having large number of results. Even given the rather limited search algorithm employed in this study, there were so many mimicry candidates identified that it was infeasible to perform detailed analyses of individual protein alignments. Instead computational methods were used in attempt to find a subset of these alignments that was most likely to be biologically significant. All of the alignments found in the study were counted, grouped, and viewed according the human target list they derived from (Figure 7), the microbe they derived from (Figure 8), or the individual human proteins they derived from (Figures 9 and 10). In each of these cases, the plots showed a relatively even distribution of the number of alignments found across whatever variable was located along the x-axis. The results are consistent with a scenario where there is some degree of noise in the number of alignments found. For example, in the normalized plot of alignments for each microbe (Figure 8b) the viral proteomes all report higher values than the bacterial proteomes. Viral proteomes are approximately two orders of magnitude smaller than bacterial proteomes. If there were a certain baseline level of extra “noisy” alignments found, these normalized viral plots would be the most sensitive, and it is expected that they would show higher values than their bacterial counterparts. Therefore, viewing the data in each of these ways was not able to identify any specific alignments which warrant further, detailed study.
The results of this study are greatly influenced by the set of proteins that are included in the human target lists, especially given the fact that all human proteins in the study had at least one alignment with a microbial protein. Better biological data would aid the computational process. Tissue library data of gene expression was used along with Digital Differential Display in attempt to find proteins uniquely expressed in a given tissue. This method did identify several unique proteins for each tissue, but these lists are not guaranteed to be complete or exact.

The results of more refined studies using modified human target lists or alignment algorithms would be the same: there will be a large number of plausible molecular mimicry candidates identified. In this study, these results were totaled and grouped according to the human and microbial proteins that they derived from. However, none of these analyses identified any alignments that were exceptionally likely to be biologically significant. Furthermore trends seen in such totaled results must be interpreted carefully. For example when looking at the total number of alignments found for a human protein, it must be remembered that for some microbes many different versions of that species were included in the search design. If a human protein matched with one papillomavirus, it most likely matched with the other 50 versions of this organism as well. Normalizing to remove these effects from the graph is also a very difficult task which only makes the results harder to interpret.

Further studies into this problem should focus on developing computational methods that can filter out the potential molecular mimics that have been identified to find those which are most likely to be biologically significant. If this list can be trimmed down to a reasonable number, then biological studies can be designed to verify that in
fact a molecular mimicry event is occurring and triggering an autoimmune disease. Such experiments would be difficult to design because autoreactive human T cells cannot be used to transfer disease to experimental animals since the TCR is MHC-restricted and humans and animals have different MHC genotypes. Therefore, to test this in animal models, an analogous computational search would have to be performed for the model animal to identify alignments between microbial and animal (self) proteins. Once identified, these microbial peptides could be synthesized and injected in the animal to see if an autoimmune disease is initiated. The eventual goal would be to gain enough insight into the mechanisms of human autoimmune responses to be able to manipulate the human immune system and prevent occurrence of human autoimmune disease.
References


