Summer 7-15-2014

THE PREDICTION OF B-CELL EPITOPE VIA BIOSTATISTICAL AND BIOINFORMATIC METHODOLOGY AND APPLICATIONS

Bo Yao
University of Nebraska-Lincoln, bo.yao1981@gmail.com

THE PREDICTION OF B-CELL EPITOPE VIA BIOSTATISTICAL AND BIOINFORMATIC METHODOLOGY AND APPLICATIONS

by

Bo Yao

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

For the Degree of Doctor of Philosophy

Major: Biological Sciences

(Bioinformatics)

Under the Supervision of Professor Chi Zhang

Lincoln, Nebraska

July, 2014

THE PREDICTION OF B-CELL EPITOPE VIA BIOSTATISTICAL AND BIOINFORMATIC METHODOLOGY AND APPLICATIONS

Bo Yao, Ph.D.

University of Nebraska, 2014

Adviser: Chi Zhang

By creating antibodies against antigens, B-cells, also named B-lymphocytes, play an important role in the immune system to fight against foreign invasion to the host body. Within the antigen specific to a certain B-cell antibody, the sections recognized and bound by antibody are called B-cell epitopes. As antigenic determinants, B-cell epitope identification is of vital importance in many immunological processes, such as vaccine design, immunodiagnostic tests, and antibody production. Towards this goal, biologists and immunologists have applied a variety of methods to identify B-cell epitopes through both experiments and bioinformatic predictions. Since the experiments for searching B-cell epitopes are time-consuming and expensive, bioinformatic methodologies have become important for the high-throughput study of B-cell epitopes.

There are two kinds of B-cell epitopes: linear (continuous) epitopes and conformational (discontinuous) epitopes. The methodologies and difficulties of bioinformatic predictions for the two categories are quite different. Due to more challenges of conformational B-cell epitope prediction, currently most of prediction tools aim to linear B-cell epitope.

The importance of B-cell epitopes has driven the development of faster and more precise tools in the past thirty years. Unfortunately, the limited success of these existing methods cannot match expectation because the achieved specificity and sensitivity leave room to be desired. In this dissertation, we developed new linear B-cell epitope tool SVMTriP with a sensitivity of 80.1% and a precision of 55.2%, which is higher than other tools such as BCPred and AAP (Chapter Two). We also developed new conformational B-cell prediction tool EPSVR and a meta server EPmeta based on Support Vector Regression (Chapter Three). Comparing to other conformational B-cell prediction tools such as DiscoTope, EPSVR shows a better prediction with AUC (Area Under receiver operating characteristic Curve) of 0.597. In addition, we are working on the tool SVMKER to predict epitopic residues on antigen (Chapter Four). To our knowledge, SVMKER is the first epitopic residue prediction tool just using protein sequence as input. These online tools will provide more choices for the identification of protein epitope by bioinformatic methodology.

ACKNOWLEDGMENTS

I would like to thank the following people who supported me during my graduate study with my sincere gratitude.

To my supervisor, Dr. Chi Zhang, for his careful instructions, detailed directions, and continuous encouragement and help during my whole study, he is forever my teacher and friend.

To my dissertation committee members, Dr. Etsuko Moriyama, Dr. Jeffrey P. Mower, Dr. Hideaki Moriyama, and Dr. Bin Yu, for their mentoring, professional guidance, and massive help on my proposal and writing.

To Dr. Chi Zhang's group, Dr. Tao Lu, Dr. Yongchao Dou, and Ms. Lin Zhang, for their advice and help on my projects. It is my pleasure to have worked with them.

To my dear wife, Junjie Xu, for her continuous support and care, and a most lovely baby, Eric. I love them forever.

To the collaborators on my various projects, Dr. Shide Liang and Dr. Xiaomei Guo, for their wonderful ideas, inspiring discussions, and great work to make the projects successful. It is my pleasure to have collaborated with them.

To my friend, Dr. Dandan Zheng, for her caring discussions and suggestions on my career and dissertation writing.

To my friends in Lincoln, Wenhu Guo, Meng Xie, Dante Placido, Agnes Yunyi Wu, Guodong Ren, and Kanika Jain, for their help during my study. To Noriko Inoguchi and Beth Whitaker for their support for my teaching in University of Nebraska. To Dr.

Cheng Cheng, Dr. Robert F. Diffendal, and Anne Diffendal, for their selfless help when I started with my settlement in Lincoln.

To my parents, for their continuous encouragement which keeps me always moving forward.

TABLE OF CONTENTS

CHAPTER ONE: THE IMMUNE SYSTEM AND B-CELL EPITOPES

1. Introduction to the Immune System

1.1 Immune system

Our immune system protects us throughout our lives against surrounding pathogenic factors, such as viruses, bacteria, pathogenic fungi and eukaryotic parasites. This powerful and profound system can usually be divided into two major categories: the innate immune system and the adaptive immune system (*1*). The innate immune system, also called the non-specific immune system, is the first defense line to fight against the invaders. The basic components of the innate immune system include barrier epithelial cells, *in vivo* sentinel cells for recognition and following removal, and Natural Killer cells for killing invader. The innate system demonstrates a generic immune response to varieties of pathogens. Sentinel cells, for example, present so-called pattern recognition receptors (PRRs) on their surfaces. PRRs recognize pathogen-associated molecular patterns (PAMPs) presented exclusively on microbe pathogens. Since PAMPs are shared by different pathogenic sources, the inflammatory responses followed by the recognition of PAMPs by PRRs are still non-specific (*1*). Furthermore, the innate immune system cannot confer memorable and long-lasting immunity to the host.

In contrast to the innate immune system, the adaptive system responds to pathogens in a completely different manner. During adaptive mechanisms, highly specialized white blood cells carry on adaptive immune responses to known or unknown pathogens on the second and third defense lines of the whole immune system. The adaptive immune system uses a small number of genes to express a large amount of

different antigen receptors. These receptors are uniquely expressed in different individual lymphocytes, and the offspring cells of lymphocytes inherit the same receptor specificity. As a consequence, the immunity acquired by the adaptive immune system can be learned, memorized, and then kept for long-lasting protection. Therefore, the adaptive immune system represents a more flexible evolutionary protection strategy against fugitive and unstable environment than the innate one (*2*). Among the entire defensive lines of the immune systems, T- and B-cells play the most important roles to recognize, lock, and eliminate the potentially harmful invaders. They will are discussed in the next section.

1.2 T-Cell and B-cell

Among immune components, T- and B-cells have been given the most attention by immunologists. Both T- and B-cells belong to one kind of white blood cells called lymphocytes, but the immune responses mediated by the two cell types are entirely different. T-cell immunity is based on cell-mediated attacks against foreign invaders into the host, while B-cell specifically plays an important role on so-called humoral immunity, which generates antibodies to search and recognize harmful antigens. Both T- and B-cells are initiated in bone marrow. However, T-cells are transferred to the thymus before their maturation, whereas B-cells stay in bone marrow until maturation. After mature T- or B-cells form, they will both immigrate to the bloodstream and be transferred throughout the host body.

According to their distinct roles, T-cells may be classified into multiple subgroups, including 1) Helper T-cells, 2) Cytotoxic T-cells, 3) Memory T-cells, and 4) Regulatory T-cells. Helper T-cells are also called CD4+ T-cells due to the expression of the protein

CD4 on the surface of mature helper T-cells. By secreting cytokines, helper T-cells function on the maturation of B-cells, activation of cytotoxic T-cells, and enhancement of the immune response of macrophages. Cytotoxic T-cells (also called CD8+ T-cells) can kill infected target T-cells directly, so sometimes they are also called T-killer cells. Cytotoxic T-cells may also kill tumor cells and even normal cells of a transplanted organ after surgery. Memory T-cells once experience the antigen-mediated infection, show a faster and stronger immune response when encountering the cognate antigen at the second time. The immunological memory comes from the proliferation and differentiation of naive T-cells undergoing the activation by antigen. 4) Regulatory T-cells (also known as suppressor T-cells) are a subset of specific T-cells involved in immunological tolerance. Regulatory T-cells suppress immune responses of other T-cells that can cause tissue damage. They induce tolerance to self-antigens and avoid autoimmune diseases (*3*). Despite the different roles of these T-cells, their immune responses are mediated by the cells themselves. It means that these cells are directly involved in recognition, binding, and destroying of foreign invaders.

Unlike the direct action of T-Cells, B-cells produce a variety of antibodies to obtain specific recognition to antigens. These antibodies, also called immunoglobulins, are a class of secreted proteins with a special Y shape. The Y shape of an antibody is composed of three components, one containing two C termini of heavy chains and the other two containing N termini of heavy chains and the whole light chains. Among a heavy chain, its C terminus is connect to N terminus by a flexible connecting chain (Figure 1.1). Therefore, the existence of these flexible connection chains can be adaptive

to the movement of N-terminus of Y shape, and enhance the binding of the light chains with antigens.

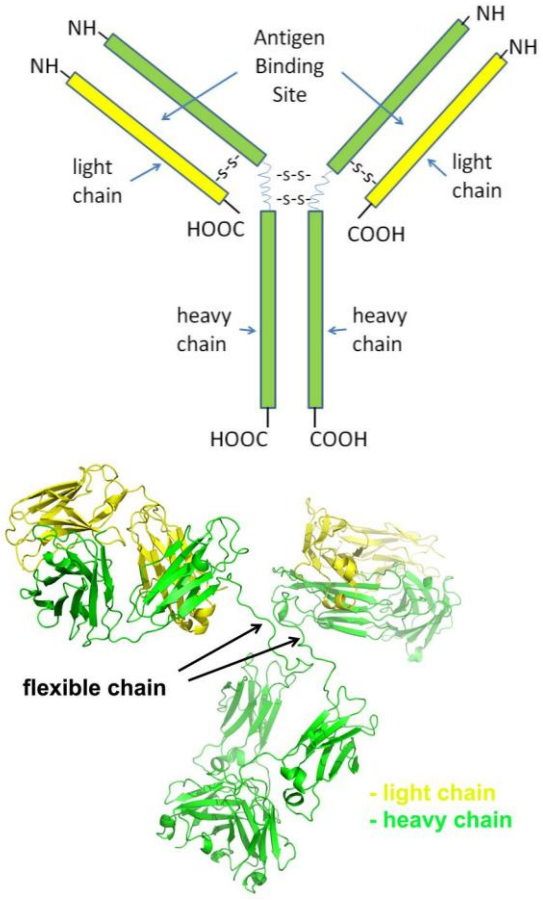

Figure 1.1 The Y shape of an antibody (top) and the 3D structure of the antibody (down, PDB ID: 1IGT) (*4*). Commonly, an antibody is composed of two light and two heavy chains. The flexible chains in the middle of the heavy chains determine the conformation of the antigen binding site of the antibody.

Although all the antibodies share the same Y shape, they can be categorized into five classes (IgA, IgD, IgE, IgG, and IgM) due to different C regions that are C terminus of heavy chain. Each class of antibody has its own *in vivo* location, acceptor, and specific biological function (*1, 2*). For example, IgA is the dominant antibody class in the

mucosal immune system such as the gut, respiratory tract and urogenital tract. IgA may prevent pathogen adherence. IgD may function to bind an antigen receptor when a B-cell meets with an unknown antigen. IgE is an especially interesting subclass of antibodies because it is mainly involved in allergic response. The interaction of allergen and IgE triggers the activation of mast T-cells, leading to a series of allergy-related characteristics. IgE is also found in autoimmune diseases. IgG is the major form in antibody-mediated immunity. It can search and recognize different antigens by taking different forms. In addition, IgG is the only antibody class that is transferred from mother to fetus through the placenta in passive immunity. IgM has a strong binding to pathogens. If there is no sufficient IgG, IgM may act instead to eliminate pathogens. Together, these five classes of antibodies can accomplish humoral immunity mediated by B-cells.

Studies on antibodies advance vaccine invention. A given kind of antibody may uniquely bind to its corresponding antigen. At the same time, the immuno-property of the antibody can usually be inherited and memorized by the host body as discussed above. The injection of vaccine (which actually is a ‘dead’ or inactive antigen) to the host body aims at the production of corresponding antibody by the *in vivo* immune system. The antibody may be memorized and kept effective against future invasions by the same active antigen for a period of time. Hence, understanding the interactions between antigens and antibodies becomes a key step to design novel vaccines against different kinds of diseases which patients are currently suffering with.

2. Antibody-antigen complex 3D structure

The most accurate way to study the antibody-antigen interaction is with a 3D complex structure. Unfortunately, there is a very limited number of antibody-antigen complex 3D structures in Protein Data Bank (PDB, http://www.rcsb.org/pdb/home/home.do) (*5*). For example, for *Homo sapiens*, there are only approximately 200 antibody-antigen complex 3D structures available in the PDB as of May 20th, 2013. To put this in perspective, it has been estimated that humans generate 10 billion different antibodies, therefore forming 10 billion unique interactions with their corresponding antigens (*6*). Among the 3D structures, most of them show a roughly similar appearance due to the special Y shape of the antibody. Using protein structure 4JAN as example (Figure 1.2), we may clearly find that the antigen-binding site of the antibody is located at the component which contains the entire light chain and a part of heavy chain. The same binding pattern is also found in other antibodies.

Although it is a kind of protein-protein interaction, antibody-antigen binding seems quite distinctive from other protein-protein interactions due to the specific roles of antibodies. First of all, the antibody-antigen complex is categorized as non-obligate which means the individual antibody and antigen also exist *in vivo* accompanying the antibody-antigen complex. Furthermore, this kind of non-obligate antibody-antigen complex is proven to be transient, which means the affinity between the antibody and the antigen is also not strong (*6*). Therefore, considering their specific binding patterns, antibody-antigen interactions cannot be studied by simply applying regular protein-protein interaction strategies, especially when trying to develop prediction tools for antibody-antigen interactions.

To investigate antibody-antigen complexes, much attention has been given to the antibody-binding site of antigens since this small portion of the antigen structure is in direct charge of binding with the antibody (*13*). For one thing, the identification of antibody-binding sites is a relatively easier task in structural biology due to their decreased sizes compared with the complete antigen proteins. Moreover, epitope identification is also the most vitally important part in antigen studies for medical applications, such as immuno-diagnosis and the development of novel vaccines. B-cell epitopes will be discussed in detail in the next section.

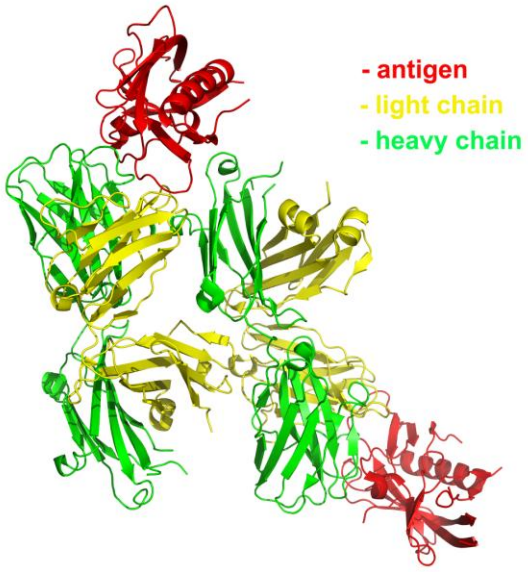

Figure 1.2 Crystal structure of broadly neutralizing antibody CH103 in complex with HIV-1 gp120 (PDB ID: 4JAN) (*8*)**.** An antibody binds to the antigen through its light-heavy overlap chain.

3. B-cell Epitope

3.1 Introduction to B-cell Epitopes

Epitopes are the antigenic determinants involved in antibody-antigen interactions. B-cells are commonly found in most mammals. By creating antibodies against antigens, B-cells play an important role in the immune system to fight against pathogenic invasions. Among the antigens related to B-cells, the sections recognized and bound by antibodies are called B-cell epitopes (Figure 1.3). Since they are antigenic determinants, the identification of B-cell epitopes helps biologists and immunologists investigate the pathways in the body's self-protection systems (*9*).

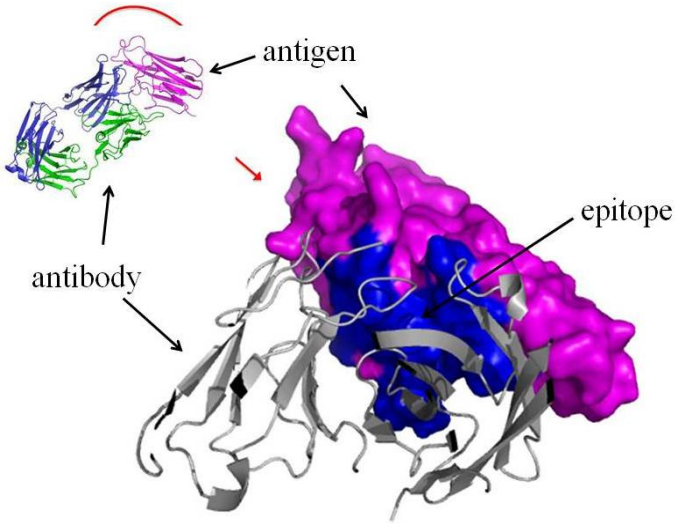

Figure 1.3 Epitope is the antibody-binding site on an antigen. The blue region of the antigen which binds to the antibody is called an epitope.

There are two types of B-cell epitopes. Linear (continuous) epitopes are short peptides containing a contiguous amino acid sequence, whereas conformational (discontinuous) epitopes are composed of amino acids that are not contiguous in sequence but close in the 3D structures (*10*).Some researchers have the view that the

boundary between linear and conformational epitopes is vague, because one conformational epitope may be seen as the combination of more than one shorter linear epitopes (*13*). The searching of conformational B-cell epitopes is much more difficult because the scarcely available 3D structure of an antigen is usually needed. Considering the complexity of conformational B-cell epitopes, most current research focuses instead on linear epitopes. The challenges of conformational B-cell epitope predictions will be discussed in detail in Chapter 3.

3.2 Applications of B-cell Epitopes

The identification of B-cell epitopes can be the foundation for many immuno-applications. For example, with the knowledge of an epitope sequence, we can synthesize peptides mimicking the epitope for diagnosing human diseases, such as tuberculosis. Another more exciting application lies in vaccine design. A vaccine is the assumed 'dead' or inactive form of a disease-caused pathogen, or any item can trigger immune response, hich is without the original activity or harmful impact on the host body. Vaccines can trigger the immune system and induce the production of correlated antibodies. For the host, the appearance of antibodies may be memorized and inherited so that the host body can be protected against the same 'true' pathogen for a period, or even the whole lifetime. Vaccines are widely used for human beings from infancy to adulthood against many different pathogens and infections. For example, some epidemics which were once fatal to the world, such as smallpox, have been controlled efficiently due to the invention of disease-specific vaccines (*11, 12*). Hence, vaccine design is a medical application of vital importance, and many more novel vaccines against fatal diseases for public health are constantly under investigation and testing.

One key step for vaccine design is the determination of the epitope on an antigen. With this knowledge, by protein recombinant technologies, we may produce epitope-based vaccines which reduces or eliminates the pathogenic sites of the real antigens. Hence, the epitope-based vaccines will keep the antigenicity and immunogenicity of the original antigens due to the presence of the epitopes, but lack the pathogenicity due to the absence of the active sites. When these epitope-based vaccines are injected into the host body to induce antibody genesis, they pose the least risk of adverse effects due to the minimal original antigen structures beyond the epitopes. At the same time, the immunity is triggered by epitope-based vaccines which then protects the host body against any future invasion of the same kind of antigen (*13*). Because of this unique advantage, the identification of B-cell epitopes is one of the major focuses of research for immunologists.

4. Immunology and immunoinformatics

4.1 Immunology

Immunology focuses on the components, responses, and mechanisms of the immune systems of all organisms, especially humans and other vertebrates. Its origin may date back to the 18th century or even earlier. In 1796, Edward Jenner firstly discovered the induced protection against human smallpox disease by cowpox (*14*). The induction procedure was then called vaccination, which continues to be used today. After the first successful demonstration of vaccination, more and more practical vaccines, such as rabies vaccine (1885) (*15*) and influenza vaccine (1945) (*16, 17*), were found and approved for public use. In the serum of a vaccinated individual, Emil von Behring and Shibasaburo Kitasato (1898) for the first time found antibodies, the important substances specifically binding to corresponding pathogens, which initiated the era with the humoral

theory of immunity (*18*). Thereafter, the causes of infectious diseases were gradually discovered, and the roles of pathogenic sources, such as viruses, bacteria, fungi, and parasites, were determined as infectious agents. It was accepted that the immune system, as the *in vivo* self-protection mechanism, can automatically reduce or even eliminate the effect of invading pathogens. Due to the important medical relevance, the cells and body tissues involved in the immune system have been systematically studied since then. By understanding how the immune system works, immunologists may design clinical strategies to control the immune responses to both disease-causing pathogens and non-harmful antigens. For example, antihistamine was designed to block histamine binding to histamine receptors and then inhibit the subsequent inflammatory responses. Hence, antihistamines may be used for allergy relief and to help patients reduce the pain from histamine-induced swelling and itching (*3*). Compared with most other biological branches, immunology is a rather new field with many immunological pathways and mechanisms still to be discovered.

The study of immunology is not only challenging but also highly valuable. The level of complexity in the immune system comes from life itself. The evolution of life always finds a way so that the immune system keeps developing against the evolution of the pathogens and their ever-changing invasion strategies. As a consequence, there are always new topics to investigate when people attempt to uncover the details of the immune system. The immune system is so complex that immunology is divided into many individual sub-areas, such as classical immunology, clinical immunology, developmental immunology, diagnostic immunology, and evolutionary immunology. Varieties of immune organs or tissues, such as T-cell, B-cell, spleen, thymus, bone

marrow, and lymph, may be the unique research focuses for individual immunological research groups. The achievements from these immunological studies can provide valuable insights and strategies for medical applications. As an example, the influenza vaccine, a most common immunological application, has saved millions of lives since it was introduced to the public. In the history of influenza pandemic, influenza viruses, such as H1N1, H2N2, and H3N2, have taken away numerous human lives. For example, Some report estimated that spanish flu caused by H1N1 killed at least 50 million people in 1918-1920 (*19*). Another H1N1 outbreak occurred in 2009 and ~185, 000 people died worldwide from April 2009 to August 2010. The lower mortality rate despite the more densely-populated world today was largely due to the invention and the widespread adoption of influenza vaccines (*20*). The induced antibody by the influenza vaccine may reduce the probability of virus infection and better protect us. As a consequence, due to their direct importance in medical applications, antibodies and their producer B cells are one major research focus for immunologists.

4.2 Immunoinformatics

The burgeoning immunological data drives a new field in immunology called immunoinformatics. Immunoinformatics is a branch of bioinformatics that has been developing quickly in the past decade. Using the principles of bioinformatics and computational biology, immunoinformatics researchers have begun to manage the collection, summary, data mining, and convenient Internet sharing of immunological resources. In addition to creating immune-related databases, a series of novel tools, such as data analysis, sequence alignments, and biophysical predictions, have been developed

to solve the questions on immunology. Below I summarize the development of immunoinformatics databases and prediction tools for B-cell epitopes.

4.2.1 B-cell Epitope Databases

With the growing resources of B-cell epitopes, it is necessary to collect and organize known information on reported B-cell epitopes with well-designed structures into databases. At present, there are several databases available online and free for public access. Some of these databases focus on linear B-cell epitopes, such as Bcipep (http://www.imtech.res.in/raghava/bcipep/data. html) which contains 2479 known linear B-cell epitopes (*21*). Another popular B-cell epitope database is Immune Epitope Database (IEDB, http://www.iedb. org/) which collects not only linear B-cell epitopes but also conformational B-cell epitopes and 3D structures of antibody-antigen structures (*22*). Actually, IEDB collects much more entries than other B-cell epitope databases. For example, as of May 30, 2013, there are 63452 B-cell linear epitope entries, 2056 conformational B-cell epitopes entries, and 646 3D structures of antibody-antigen complexes related to B-cell available in the IEDB database.

The B-cell epitopes collected in the databases above still represent only a small portion of all antigenic determinants which the immunologists are interested in. There is a need to search and identify more. There are many ways to determine B-cell epitopes, including experimental searching strategies and bioinformatic prediction tools based on biostatistical and bioinformatic technologies. Experimental results are reliable but conducting such experiments is time- and resource- consuming. Sometimes such experimental searches could be blind or random, in the hope of identifying something useful. On the other hand, bioinformatic predictions have the alternative advantages, such

as fast speed and low cost, comparing to the experiments. Although such predictions suffer from intrinsic false-positive rates, the quick and high-throughput properties are highly desirable, allowing researchers to develop novel methods with improved prediction performance. Even though 100% accuracy cannot be achieved, the predicted epitopes from bioinformatic methods can greatly narrow down the searching space for subsequent experiments and validations, and provide reasonable candidates to decrease the randomness of experimental searches. Therefore, the development of bioinformatic prediction tools has been a vital part of the study on B-cell epitopes. Linear and conformational predictions will be reviewed in more detail in the following two sections.

4.2.2 Principles of Linear B-cell Epitope Prediction

The prediction of linear B-cell epitopes is easier than that of conformational epitopes. Usually, the linear B-cell epitope prediction tool is at the protein-sequence level. That is, many previous studies on linear B-cell epitope prediction are based on the physicochemical property (or propensity scale) of constituent amino acids. With the input of protein sequences, the prediction tools are based upon the amino acid properties including hydrophilicity (*23, 24*), solvent accessibility (*25*), secondary structure (*26*), flexibility , and antigenicity (*27*). These known B-cell linear epitope prediction tools include but are not limited to: PEOPLE (*28*), BEPITOPE (*29*), BepiPred (*30*), ABCPred (*31*), AAP (*32*), BCPred (*33*), BayesB (*34*), BEOracle/BROracle (*35*), and BEST (*36*). In these various prediction tools, different properties of amino acid sequences were applied. For example, PEOPLE utilized four kinds of physicochemical properties: secondary structure, hydrophilicity, surface accessibility, and flexibility (*28*). BEPITOPE declared more than 30 propensity scales such as hydrophobicity scales and flexibility (*29*).

BepiPred did a similar analysis using a number of propensity scales but using a different training dataset (*30*). ABCPred specifically used a neural network to decrease the false positive rate in predicted linear B-cell epitopes (*31*). In more recent methods, string kernals such as spectrum, mismatch, local alignment, and subsequence were applied in BCPred (*33*) while amino acid pair propensity was used in AAP (*32*). BEOracle/BROracle was based on features from evolutionary, structural and compositional information of antigen sequences (*35*). In BEST, optimal models were trained by combining information from the epitope sequence, sequence conservation, secondary structure, and relative solvent accessibility (*36*).

Although different feature combinations based on antigen sequences were attempted by individual linear B-cell epitope prediction models, it is not yet known which combination provides optimal prediction performance. Furthermore, it seems that the "optimal" feature set may be biased by the training dataset chosen by individual research groups. For example, although BCPred tool received a higher AUC (Area Under receiver operating characteristic Curve) value than AAP using the Bcipep database, the two tools demonstrated approximately equal prediction performance when the IEDB database was used (*44*).

Another differentiating factor about linear B-cell epitope prediction is the application of different machine learning platforms. The most common platforms are Support Vector Machine (SVM) (*32*, *33*), Hidden Markov Model (HMM) (*30*), and Artificial Neural Network (ANN) (*31*). Support Vector Machine is one of the most popular supervised learning models used for classification, pattern recognition, and regression. A typical SVM is the binary linear classifier. With the input of two classes of

classified data, SVM defines the hyper plane as the boundary of two classes by calculating the maximum margin. The determination of the hyper plane is the main purpose of SVM algorithms. Derived changes, such as kernel trick in non-linear classification, may have effect on the hyper plane and be used to modify binary linear SVM. SVM has been widely applied, especially in the past decade, in the development of linear B-cell epitope prediction tools such as AAP (*32*), BCPred (*33*), BayesB (*34*), BEOracle/BROracle (*35*), and BEST (*36*). Other tools, such as BepiPred (*30*), applied HMM as their platform. HMM algorithms come from the field of statistics. Generally, HMM is deemed as a Markov model inside which there exists a Markov process containing unobserved middle states. The optimal model determination is similar to maximum likelihood estimation of the parameters involved in HMM given the training sequences. ANN is an information-processing paradigm involved in multiple lines of interconnected nodes (neurons) computing values from inputs, which has been applied in the reported linear B-cell epitope tool ABCPred (*31*). In contrast with SVM and HMM, ANN algorithms are much more complicated and require more calculation machine time.

4.2.3 Linear B-cell Epitope Prediction Tools

Below we will introduce and compare some popular linear B-cell epitope prediction tools. PEOPLE (Predictive Estimation Of Protein Linear Epitopes) was released in 1999 (*28*). It is a rather sample linear B-cell prediction tool. In PEOPLE, four kinds of basic properties of epitopes are used: secondary structure (mainly β-turns), surface accessibility, hydrophilicity, and flexibility. The four kinds of profiles are combined to calculate an antigenic index (AG). The AG is the final indicator to determine whether unknown small protein segments are epitopes. Although it is a rather

simple method, the four properties of epitopes used in PEOPLE are still popular in newly developed methods for the prediction of linear B-cell epitopes.

BEPITOPE is an updated version of the PREDITOP program developed by Michael Odorico and Jean-Luc Pellequer in 2003 (*29*). It is not an online tool and needs to be installed on local machines. In the BEPITOPE program the authors evaluated more than 30 propensity scales, such as hydrophobicity (where the negative values indicate hydrophilic regions, and positive scales stand for flexible regions), to search for potential epitopes. These propensity scales were calculated by the comparison of real linear B-cell epitopes and non-epitopes. Hence, for the BEPITOPE program, the collection of epitope set and non-epitope set and the following calculation of propensity scale is of vital importance for prediction performance. Although the limited database of linear B-cell epitopes may limit the success of BEPITOPE, the propensity scales used in BEPITOPE have proven useful in more recently developed tools.

BepiPred, released in 2006, applied a new strategy to predict linear B-cell epitopes (*30*). In BepiPred, the method relied on HMM. Like BEPTIOPE, propensity scales are the major indicator of epitope properties. The involvement of HMM improved prediction performance compared with naïve methods like PEOPLE. On its website (http://www.cbs.dtu.dk/services/BepiPred), the datasets used in BepiPred are also available. The sharing of datasets can be a good reference for other research groups when developing and benchmarking their methods, and promotes communication among the different research groups.

Like BepiPred, ABCPred also applied an interesting machine learning platform (*31*). Recurrent Neural Network is a powerful tool in the field of machine learning and it was used for the first time in linear B-cell epitope prediction by ABCPred. The dataset used by ABCPred is from Bcipep (*21*) after careful filtering of similar epitope sequences. The process of dataset construction in ABCPred was adopted by many later tools such as BCPred and AAP. In ABCPred, flexible epitope length was carefully discussed, and the suggested epitope size was no more than 20 amino acids (AA). To our knowledge, however, the optimal length of a certain linear B-cell epitope for prediction keeps unknown.

AAP and BCPred are very popular linear B-cell epitope prediction tools (*32*, *33*). Both tools are based on the SVM platform and filtered the training dataset from Bcipep database using a similar filtering process as in ABCPred. BCPred considered string patterns of epitope sequence while AAP applied amino acid pair antigenicity scale in the model. In addition, they both investigated and compared different window sizes of epitopes, such as 10, 12… and 20. AAP was released in 2007 and BCPred in 2008, and both are available online.

BayesB added new ideas from statistics (*34*). Bayes feature extraction is a popular feature in statistical methods, and it was applied in the BayesB tool in 2010 to enhance the performance beyond SVM-based methods. The two datasets in BayesB were borrowed from that of AAP and BCPred, again demonstrating the advantage of online tools and dataset-sharing. The basic scales in BayesB are relevant position-specific amino acid propensities, similar to those used in BEPITOPE.

The BEOracle/BROracle (*35*) tool uses classic strategies for linear B-cell epitopes, such as composition of protein sequence properties (including secondary structure, solvent accessibility and evolutionary conservation), the popular platform SVM, and cross-validation process. Despite this classical approach, it succeeded in generating a large training dataset due to the availability of the IEDB database (*22*). The size of a dataset is of vital importance for the success of the prediction tool. Hence, the IEBD database gradually replaced Bcipep and became a basic resource for linear B-cell epitope prediction.

BEST tool was released in 2013 (*36*). It considers evolutionary profiles generated by PSI-BLAST. The amino acid pair propensity scale developed for AAP and other protein sequence properties such as secondary structure and solvent accessibility were put into consideration by BEST. The increasing number of entries in the IEDB database offered a bigger training dataset, which benefited prediction performance. The success of the BEST tool demonstrated the utility of combining different traditional tools with a new and bigger training dataset.

4.2.4 Principles of Conformational B-cell Epitope Prediction

Conformational B-cell epitope prediction is more challenging than the task of linear prediction. Similar to linear B-cell epitope prediction tools, physicochemical properties of amino acid sequence are the major features used in the modeling algorithms. However, the difference between linear and conformational B-cell epitope predictions is the collection of training datasets. During the construction of datasets for known conformational B-cell epitopes, 3D structural information is usually required, which

presents a huge barrier for conformational B-cell epitope prediction. With more 3D structures of antigens, the latest conformational B-cell epitope prediction tools generally outperform their predecessors.

At present, there are a limited number of available conformational B-cell epitope prediction tools, such as DiscoTope (*37*), BEpro (PEPITO) (*38*), ElliPro (*39*), SEPPA (*41*), EPITOPIA (*42, 43*), and Bpredictor (*45*). Different tools used different combinations of features, which may be biased by their own training datasets. For example, DiscoTope integrated a linear combination of two scores, the hydrophilicity scale and the epitope log-odds ratio, the latter of which is one kind of epitopic residue propensity scores (*37*). BEpro (PEPITO) also applied linear combination to two scores: the epitopic residue propensity and the half sphere exposure values at multiple distances (*38*). ElliPro used only one score, the residue protrusion index (PI) (*39*). SEPPA employed the epitopic residue propensity and the compactness of neighboring residues around one residue (e.g., contact number which is the number of Cα atoms in the antigen within a distance of 10Å of the Cα atom of target residue), again using linear combination (*41*). EPITOPIA applied a naive Bayesian classifier to forty-four physicochemical and structural–geometrical attributes, including secondary structure, epitopic residue propensity, evolutionary conservation score, solvent accessibility to the surface, and hydrophilicity etc (*42, 43*). Bpredictor employed the Random Forest (RF) classifier to adjacent residue distance score, accessible surface area, conservation, secondary structure, and propensity etc (*45*).

In general, the features used by these predictors include conservation score, structural features such as secondary composition, geometry characteristics such as

protrusion index and planarity score, and amino acid features such as hydrophilicity and propensity (odd-ratio). These attributes can be integrated by linear combination or machine-learning algorithms, such as naive Bayesian classifiers, Support Vector Regression (SVR), and RF classifiers. Different numbers of features were used in a given predictor, from two scores to forty-four attributes. For small numbers of attributes, a simple linear combination can usually work well, whereas large numbers of features often require sophisticated machine-learning algorithms to optimally integrate the scores. Notably, some of these features may be mutually exclusive or overlapping. For example, the antigenic epitope is frequently located at either a protruding region or a flat surface. In such cases, linearly combining the two incompatible terms contradicts the physical situation and will degrade the performance of a predictor. In the next section, we will introduce these feature applications in several conformational B-cell epitope prediction tools released in the past decade.

4.2.5 Conformational B-cell Epitope Prediction Tools

Below the popular conformational B-cell epitope prediction tools are introduced following the order of release dates,

DiscoTope is the first conformational B-cell epitope prediction tool based on protein 3D structural information (*37*). DiscoTope uses linear combination to integrate two scores, the hydrophilicity scale and the epitope log-odds ratio, the latter of which is one kind of epitopic residue propensity scores (*37*). Its dataset was a group of 76 X-ray structures of antibody/antigen protein complexes. Compared with previous sequence-level tools, the involvement of protein 3D structures offered a more reliable prediction to

guide experimental epitope mapping. From the training dataset, a log-odds-ratio was calculated and then applied in the construction of the DiscoTope model. On the other hand, the tool used simple linear combination rather than popular machine learning technologies, such as SVM or HMM. Nevertheless, since the release of DiscoTope in 2006, the prediction of conformational B-cell epitopes shifted the focus to utilizing protein 3D structural data.

PEPITO, released in 2008, tried to improve prediction performance by applying more features of the 3D structure of antibody/antigen protein complexes, such as multiple distance thresholds and sphere exposure (*38*). The dataset of DiscoTope was used by PEPITO due to the limited availability of antibody/antigen protein complexes in the PDB database. Amino acid propensity scales, which are usually used in sequence-based linear B-cell epitope prediction, were also incorporated by PEPITO to enhance predictive performance.

ElliPro can accept either antigen 3D sequence and structure as input (*39*). If only a protein sequence is provided as input, ElliPro will firstly search protein structures in PDB based on sequence similarity and use the best-matched 3D structure as the input. ElliPro implemented Thornton's method which considers the shape of a protein as an approximate ellipsoid (*40*). ElliPro calculates the residue protrusion index (PI) and then clusters neighboring residues according to their PI values.

SEPPA introduced a concept of 'unit patch of residue triangle' to describe the local spatial context on the surface of a protein in 2009 (*41*). The unit patches were involved in the calculation of propensity indices and the following antigenicity scores. In

addition, the method constructed its own training dataset containing 84 structures. The test set with 119 antigens was from the training set of DiscoTope, IEDB, and Epitome. Machine learning platforms were not used in SEPPA.

Different from the three tools above, EPITOPIA initiated the application of machine learning technologies for the prediction of conformational B-cell epitopes (*42*, *43*). It applied a Naive Bayes classifier to the construction of an online tool. A searching antigen input was firstly divided into overlapping surface patches. For each middle residue of a patch, its immunogenicity score was calculated by combining physicochemical and structural properties of the patch. For EPITOPIA, the 3D structure of the antigen is still a must for online prediction.

We released the two conformational B-cell epitopes EPSVR (http://sysbio.unl.edu/EPSVR) and a meta server EPmeta (http://sysbio.unl.edu/EPmeta) in 2010 (*44*). EPSVR applied the six physicochemical attributes to its surface patch. The six attributes include epitope propensity, conservation score, side chain energy score, contact number, surface planarity score, and secondary structure. More details will be introduced in chapter three.

Bpredictor is a newer conformational B-cell epitope prediction tool released in 2011 (*45*). For this tool, a new concept of 'thick surface patch' was introduced to describe the local spatial context on a protein surface instead of 'surface patch'. It also compared the influences of different machine learning platforms. For example, SVM and ANN were found to be slower and more sensitive on parameter settings than RF. As to

dataset, Bpredictor borrowed the dataset from EPITOBIA and EPSVR (*44*). The latter is the conformational B-cell epitope prediction tool we released in 2009.

5. Summary

The importance of B-cell epitope drives us to develop new computational methods to predict epitopes on protein candidates. Currently the prediction of B-cell epitopes, including linear and conformational, still has a place to improve. In this dissertation, from chapter two to chapter four, we will introduce our newly developed methods to predict linear B-cell epitope (chapter two), conformational B-cell epitope (chapter three), and epitopic residues on antigen (chapter four) respectively.

References

1. Sercarz, E. E., Williamson, A., and Fox, C. F. (1974) The immune system: genes, receptors, signals; [proceedings], Academic Press, New York,.
2. Schindler, L. W., National Cancer Institute (U.S.), and National Institute of Allergy and Infectious Diseases (U.S.) (1993) The immune system : how it works, Rev. Dec. 1993. ed., U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health, [Bethesda, Md.?].
3. Lefkovits, I., Jerne, N. K., Steinberg, C. M., and Di Lorenzo, C. (1981) The Immune system, Karger, Basel ; New York.
4. Harris, L. J., Larson, S. B., Hasel, K. W., and McPherson, A. (1997) Refined structure of an intact IgG2a monoclonal antibody, Biochemistry 36, 1581-1597.
5. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures, European journal of biochemistry / FEBS 80, 319-324.
6. Fanning, L.J., Connor, A.M., Wu, and G.E. (1996) Development of the immunoglobulin repertoire. Clin. Immunol. Immunopathol. 79 (1): 1–14.
7. Ponomarenko, J. V., and Bourne, P. E. (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation, BMC structural biology 7, 64.
8. Liao, H. X., Lynch, R., Zhou, T., Gao, F., Alam, S. M., Boyd, S. D., Fire, A. Z., Roskin, K. M., Schramm, C. A., Zhang, Z., Zhu, J., Shapiro, L., Mullikin, J. C., Gnanakaran, S., Hraber, P., Wiehe, K., Kelsoe, G., Yang, G., Xia, S. M., Montefiori, D. C., Parks, R., Lloyd, K. E., Scearce, R. M., Soderberg, K. A., Cohen, M., Kamanga, G., Louder, M. K., Tran, L. M., Chen, Y., Cai, F., Chen, S., Moquin, S., Du, X., Joyce, M. G., Srivatsan, S., Zhang, B., Zheng, A., Shaw, G. M., Hahn, B. H., Kepler, T. B., Korber, B. T., Kwong, P. D., Mascola, J. R., and Haynes, B. F. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus, Nature 496, 469-476.
9. Jerne, N. K. (1960) Immunological speculations, Annual review of microbiology 14, 341-358.
10. Huber, R. (1986) Structural basis for antigen-antibody recognition, Science (New York, N.Y 233, 702-703.
11. Fulginiti, V. A., Papier, A., Lane, J. M., Neff, J. M., and Henderson, D. A. (2003) Smallpox vaccination: a review, part I. Background, vaccination technique, normal vaccination and revaccination, and expected normal reactions, Clin Infect Dis 37, 241-250.
12. Lofquist, J. M., Weimert, N. A., and Hayney, M. S. (2003) Smallpox: a review of clinical disease and vaccination, Am J Health Syst Pharm 60, 749-756; quiz 757-748.
13. Reineke, U., and Schutkowski, M. (2009) Epitope mapping protocols, 2nd ed., pp 1 online resource (xiii, 456 p., [416] p. of plates), Humana Press, New York.
14. Kanof, M. E., and United States. General Accounting Office. (2003) Smallpox vaccination review of the implementation of the military program, U.S. General Accounting Office, Washington, DC.

15. Briggs, D. J. The role of vaccination in rabies prevention, Current opinion in virology 2, 309-314.
16. Francis, T., Salk, J. E., Pearson, H. E., and Brown, P. N. (1945) Protective Effect of Vaccination against Induced Influenza A, The Journal of clinical investigation 24, 536-546.
17. Salk, J. E., Pearson, H. E., Brown, P. N., and Francis, T. (1945) Protective Effect of Vaccination against Induced Influenza B, The Journal of clinical investigation 24, 547-553.
18. Haas, L. F. (2001) Emil Adolph von Behring (1854-1917) and Shibasaburo Kitasato (1852-1931), Journal of neurology, neurosurgery, and psychiatry 71, 62.
19. Johnson, N. P., and Mueller, J. (2002) Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic, Bulletin of the history of medicine 76, 105-115.
20. Dawood, F. S., Iuliano, A. D., Reed, C., Meltzer, M. I., Shay, D. K., Cheng, P. Y., Bandaranayake, D., Breiman, R. F., Brooks, W. A., Buchy, P., Feikin, D. R., Fowler, K. B., Gordon, A., Hien, N. T., Horby, P., Huang, Q. S., Katz, M. A., Krishnan, A., Lal, R., Montgomery, J. M., Molbak, K., Pebody, R., Presanis, A. M., Razuri, H., Steens, A., Tinoco, Y. O., Wallinga, J., Yu, H., Vong, S., Bresee, J., and Widdowson, M. A. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study, The Lancet infectious diseases 12, 687-695.
21. Saha, S., Bhasin, M., and Raghava, G. P. (2005) Bcipep: a database of B-cell epitopes, BMC genomics 6, 79.
22. Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A., and Peters, B. The immune epitope database 2.0, Nucleic acids research 38, D854-862.
23. Parker, J. M., Guo, D., and Hodges, R. S. (1986) New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites, Biochemistry 25, 5425-5432.
24. Hopp, T. P., and Woods, K. R. (1981) Prediction of protein antigenic determinants from amino acid sequences, Proceedings of the National Academy of Sciences of the United States of America 78, 3824-3828.
25. Emini, E. A., Hughes, J. V., Perlow, D. S., and Boger, J. (1985) Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide, Journal of virology 55, 836-839.
26. Pellequer, J. L., Westhof, E., and Van Regenmortel, M. H. (1993) Correlation between the location of antigenic sites and the prediction of turns in proteins, Immunology letters 36, 83-99.
27. Kolaskar, A. S., and Tongaonkar, P. C. (1990) A semi-empirical method for prediction of antigenic determinants on protein antigens, FEBS letters 276, 172-174.
28. Alix, A. J. (1999) Predictive estimation of protein linear epitopes by using the program PEOPLE, Vaccine 18, 311-314.
29. Odorico, M., and Pellequer, J. L. (2003) BEPITOPE: predicting the location of continuous epitopes and patterns in proteins, J Mol Recognit 16, 20-22.

30. Larsen, J. E., Lund, O., and Nielsen, M. (2006) Improved method for predicting linear B-cell epitopes, Immunome research 2, 2.
31. Saha, S., and Raghava, G. P. (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network, Proteins 65, 40-48.
32. Chen, J., Liu, H., Yang, J., and Chou, K. C. (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale, Amino acids 33, 423-428.
33. El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2008) Predicting linear B-cell epitopes using string kernels, J Mol Recognit 21, 243-255.
34. Wee, L. J., Simarmata, D., Kam, Y. W., Ng, L. F., and Tong, J. C. SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction, BMC genomics 11 Suppl 4, S21.
35. Wang, Y., Wu, W., Negre, N. N., White, K. P., Li, C., and Shah, P. K. Determinants of antigenicity and specificity in immune response for protein sequences, BMC bioinformatics 12, 251.
36. Gao, J., Faraggi, E., Zhou, Y., Ruan, J., and Kurgan, L. BEST: improved prediction of B-cell epitopes from antigen sequences, PloS one 7, e40104.
36.
37. Haste Andersen, P., Nielsen, M., and Lund, O. (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures, Protein Sci 15, 2558-2567.
38. Sweredoski, M. J., and Baldi, P. (2008) PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure, Bioinformatics (Oxford, England) 24, 1459-1460.
39. Ponomarenko, J., Bui, H. H., Li, W., Fusseder, N., Bourne, P. E., Sette, A., and Peters, B. (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes, BMC bioinformatics 9, 514.
40. Thornton, J.M., Edwards, M.S., Taylor, W.R., and Barlow, D.J. (1996) Location of 'continuous' antigenic determinants in the protruding regions of proteins, EMBO Journal 5, 409-413.
41. Sun, J., Wu, D., Xu, T., Wang, X., Xu, X., Tao, L., Li, Y. X., and Cao, Z. W. (2009) SEPPA: a computational server for spatial epitope prediction of protein antigens, Nucleic acids research 37, W612-616.
42. Rubinstein, N. D., Mayrose, I., Martz, E., and Pupko, T. (2009) Epitopia: a web-server for predicting B-cell epitopes, BMC bioinformatics 10, 287.
43. Rubinstein, N. D., Mayrose, I., and Pupko, T. (2009) A machine-learning approach for predicting B-cell epitopes, Molecular immunology 46, 840-847.
44. Yao, B., Zhang, L., Liang, S., and Zhang, C. SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity, PloS one 7, e45152.
45. Zhang, W., Xiong, Y., Zhao, M., Zou, H., Ye, X., and Liu, J. Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature, BMC bioinformatics 12, 341.

CHAPTER TWO: PREDICTION OF LINEAR B-CELL EPITOPES

1. Introduction

By secreting antibodies against antigens, B-cells play an important role in the immune system to fight invading pathogenic organisms or substances. An antibody can specifically recognize and bind to an antigen, analogous to a key into a lock. Antigenic epitopes are regions of the antigen surface that are preferentially recognized by B-cell antibodies (*1*). Prediction of antigenic epitopes is useful for the investigation on the mechanism of body's self-protection systems and can help the design of vaccine components and immuno-diagnostic reagents (*2*).

B-cell antigenic epitopes are classified as either continuous or discontinuous (*3*). A continuous (also called linear) epitope is a consecutive fragment from the protein sequence; a discontinuous epitope is composed of several fragments scattered along the protein sequence, but still form an antigen-binding interface in three dimensions. A distinction between continuous and discontinuous epitopes is vague; a continuous fragment in a discontinuous epitope can be considered as a linear epitope. The majority of currently available epitope prediction methods focus on continuous epitopes due to the relative simplicity of the problem, in which the amino acid sequence of a protein is taken as an input. These prediction methods are based upon the amino acid properties including hydrophilicity (*4*, *5*), solvent accessibility (*6*), secondary structure (*7*), flexibility (*8*), and antigenicity (*9*). In addition, based on the epitope databases such as IEDB (*10*), Bcipep (*11*), and FIMM (*12*), some methods use machine learning approaches, such as hidden Markov models (HMM) (*13*), artificial neural networks (ANN) (*14*), and support vector machines (SVM) (*15*, *16*), to locate linear epitopes. Such methods include: PREDITOP

(*9*), PEOPLE (*17*), BEPITOPE (*18*), BepiPred (*13*), ABCPred (*14*), AAP (*15*), BCPred (*16*), BayesB (*19*), BEOracle/ BROracle (*20*), and BEST (*21*).

Currently available linear B-cell prediction tools show only a limited success. For example, one of the best available methods by 2011, BCPred, was reported as the accuracy and specificity of ~72% and ~79% using the five-fold cross-validation based on a dataset of 872 B-cell epitopes and 872 non-B-cell epitopes (*16*).To pursue more reliable and stable linear B-cell epitope prediction, immunoinformaticists need to develop new statistical models. The new models shall have lower false positive rates so that the prediction results can be more reliably used for experimental design.

Since more information including experimentally determined linear B-cell epitopes and 3D structures of antigens has been released in the past decades, development of new linear epitope prediction methods became more feasible. For instance, the IEDB database collects much more known epitopes than before. With the advance in bioinformatics technology, new algorithms have been developed for the prediction of active sites of proteins. These innovations stimulate the development of the prediction tools of linear B-cell epitopes.

In this chapter, we developed a new linear B-cell epitope prediction tool, SVMTriP, which uses a machine learning technique, SVM, with the tri-peptide similarity and propensity scores. SVMTriP was tested for varied epitope sequence lengths. With the five-fold cross-validation, SVMTriP achieves a sensitivity of 80.1% and a precision of 55.2% for sequences with 20 amino acids (AA), which are higher than those of AAP (sensitivity: 59.8%, precision: 58.5%) and BCPred (sensitivity: 54.0%, precision: 60.5%).

2. Materials and Methods

2.1 Datasets

The dataset was constructed by extracting non-redundant linear B-cell epitopes from IEDB (*10*), which is frequently updated and has the most complete set of linear epitopes. Total of 65,456 redundant B-cell linear epitopes were obtained from IEDB (version June 11th, 2012). The identical epitopes and those possibly related to T-cell were removed. The full-length sequences of corresponding epitopes were also collected. Then, the various lengths of epitope sequences, including 10AA, 12AA, 14AA, 16AA, 18AA, and 20AA, were extracted by trimming the long experimental measured epitopes or attaching more amino acid residues to both ends of short epitopes according to the full-length sequences. For a given length, 65,456 epitope sequences are filtered by a threshold of less than30% similarity, measured by BLASTP (*22*), were clustered together and only one of them was kept as a representative epitope sequence in the dataset. Finally, the positive dataset for each length had a total of 4925 non-redundant epitope sequences. For the negative dataset, the same number of non-redundant sub-sequences with each equal length is extracted from the non-epitopic segments in the corresponding antigen sequences.

2.2 Attributes

2.2.1 Tri-peptide Scores Matrix

The idea of tri-peptide score matrix was borrowed from the prediction method used in protein subcellular localization by Lei and Dai (*23*). A matrix D^k of high scored *k*-peptide pairs is of dimension $20^k \times 20^k$ (we did not consider X residue in BLOSUM or PAM matrix). The tri-peptide score matrix is defined as:

$$T^{(i)} = \sum \Phi^{(i)} \otimes \Omega_j \,, \qquad (1)$$

where $\Phi^{(i)}$ denotes the tri-peptide that represents the *i*-th attribute, Ω_j denotes the *j*-th tri-peptide in the tri-peptide subsequence space for the input sequence. The symbol "⊗" denotes getting the similarity score of any two corresponding tri-peptide, *i.e.*, sum of three similarity scores for three amino acid pairs from a BLOSUM/PAM matrix. For example, assuming the length of a given epitope candidate is 20 AA, the tri-peptide subsequence similarity kernel for the *i*-th attribute is generated by summing over similarity scores of the 18 pairs of tri-peptide; each pair consists of one tri-peptide from the input sequence and the tri-peptide represents *i*-th attribute from the tri-peptide subsequence space. Using BLOSUM62 as example, the steps are shown below:

a) A sliding window of 3AA along the sequence is used.

b) The score $T^{(i)}$ is defined as the sum of the score for the respective individual residue pair from BLOSUM62 between two pairs of tri-peptide. The score $T^{(i)}$ is zero if the sum of BLOSUM62 scores of three individual residue pairs is negative.

c) Each value in 20^3 features is calculated from the average score $T^{(i)}$s between the i^{th} tri-peptide and all of tri-peptides from step a.

A graphic interpretation is shown in Figure 2.1

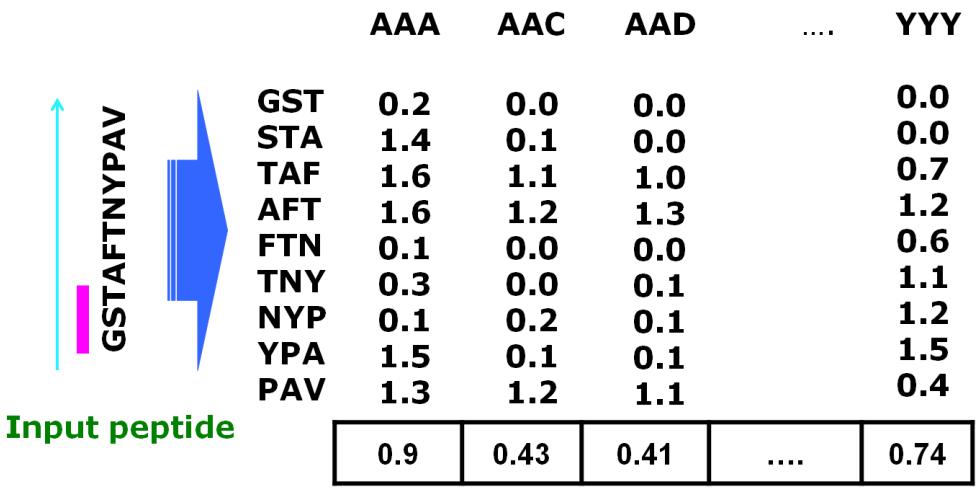

Figure 2.1 Illustration of tri-peptide score matrix construction. Each feature score is the average score of between the i^{th} tri-peptide and all the peptides from slide window of protein sequence.

To build the tri-peptide score matrix, BLOSUM and PAM matrices, the most popular score matrices used for protein sequence alignment, were tested. BLOSUM matrices are derived from residue-residue substitution probability (*24*) while PAM is based on observed mutations of closely related proteins (*25*). In our study, different BLOSUM matrices were tested, such as BLOSUM30, BLOSUM50, BLOSUM62, and BLOSUM75, where the number represents the percentage identity threshold that are used for determining closely related protein groups during the construction of BLOSUM matrices. Different PAM matrices were used as well, such as PAM120, PAM160, PAM200, and PAM250, where the number stands for the times of multiplication of the primary PAM matrix (PAM1) by itself when building a PAM matrix. The application of different BLOSUM or PAM matrices would influence the prediction result of final models.

2.2.2 Tri-peptide Subsequence Propensity

The propensity of tri-peptide subsequence representing the *i*-th attribute is calculated as in Equation (2):

$$P^{(i)} = \frac{f^{(i)}}{F^{(i)}}, \tag{2}$$

where $f^{(i)}$ is the frequency of *i*-th type of tri-peptide in the positive epitopes, and $F^{(i)}$ is the background frequency of *i*-th type of tri-peptide in 5×10^4 protein sequences randomly selected from the Refseq database (*26*).

2.2.3 Integrations of Tri-peptide Scores Matrix and Tri-peptide Subsequence Propensity

The tri-peptide subsequence space is used to encode the SVM attributes. This kernel has a space of 20^3 attributes for both tri-peptide substring and propensity. The score of *i*-th attribute, $K^{(i)}$, is defined as the tri-peptide subsequence similarity kernel modulated by its corresponding tri-peptide propensity. Please see Equation (3):

$$K^{(i)} = T^{(i)} \bullet P^{(i)}, \tag{3}$$

where $K^{(i)}$ denotes the score of the *i*-th attribute, $T^{(i)}$ denotes the *i*-th tri-peptide score matrix calculated by Equation (1), and $P^{(i)}$ denotes corresponding tri-peptide subsequence propensity of *i*-th tri-peptide subsequence calculated by Equation (2). Other features of physicochemical properties of amino acids, such as hydrophilicity and predicted secondary structure, had been also used to modify $K^{(i)}$. For hydrophilic and hydrophobic residue group, two fixed weights, *e.g.*, 0.5 and 2, will be used to change $K^{(i)}$.

2.2.3 Hydrophobicity

Hydrophobicity scale of residue usually is applied to linear B-cell epitope prediction (*17, 18*). The hydrophobicity profile implies a hydrophobic region that tends to be located away from the surface of antigen protein. A potential hydrophobic region represents a low probability to be an antibody-bind site. Hence, hydrophobicity scale was used here in order to evaluate the probability of a residue locating on protein surface.

Table 2.1 Hydrophobicity Scale Table

Ala	1.8	Glu	-3.5	Leu	3.8	Ser	-0.8
Arg	-4.5	Gln	-3.5	Lys	-3.9	Thr	-0.7
Asn	-3.5	Gly	-0.4	Met	1.9	Trp	-0.9
Asp	-3.5	His	-3.2	Phe	2.8	Tyr	-1.3
Cys	2.5	Ile	4.5	Pro	-1.6	Val	4.2

Amino acid hydrophobicity scales may be calculated by experimental biophysical methods. In our study, we used the hydrophobicity scale calculated by Jack Kyte shown in Table 2.1 (*27*).

2.2.4 Secondary structure

The structural information directly declares the spatial location where the residues stay. 3D structure is of the most interest but it is also challenging to predict. There are, however, many mature secondary structure prediction methods available and the results from those tools usually are more reliable. Hence, many linear B-cell epitope prediction tools incorporate secondary structural data to their models (*17, 20*). We also considered secondary structure information.

To obtain the secondary structural information of antigen, we applied one of the popular protein secondary structure prediction tools, PSIPRED, which is based on position-specific scoring matrices (PSSM) from PSI-BLAST. In order to run a PSIPRED tool on a local machine, the BLAST tool and the associated NR database must also be downloaded and installed correctly. PSIPRED tool package is available on http://bioinfadmin.cs.ucl.uk/downloads/psipred/ (*28*) and BLAST can be downloadable from National Center for Biotechnology Information (NCBI) website (ftp://ftp.ncbi.nlm.nih.gov/blast/ executables/blast+/LATEST/) (*22*).

2.3 Support Vector Machine Platform

Support Vector Machine (SVM) was selected as the machine learning technique for our new models. SVM has been proven powerful in various biological and immunological applications, such as epitope prediction (*15, 16*), disease diagnostic (*29, 30*), clinical outcome (*31*), and hub protein determination (*32*). SVM usually is associated

with learning algorithms for the application of classification and regression analysis. For example, a typical application is to generate binary linear classifier. SVM takes a set of two classes of objects as an input. By fitting the maximum-margin hyperplane, SVM determines the boundary between the two classes in a hyper-dimensional space. SVM then classifies an unknown object by checking where the unknown object is located besides the hyperplane in the hyper-dimensional space. Since developed by Vladimir N. Vapnik in 1963 (*33*), the theory of SVM has been differentiated into multiple branches, including linear model and nonlinear model, binary class SVM and multiclass SVM, and classic supervised learning SVM and semi-supervised transductive SVM. Many different SVM tools have been developed as listed in Table 2.2. Here we used SVM^{light} as a platform to train the optimal model. SVM^{light} is an easy standalone tool with customizable parameter options. It has been successfully applied in many SVM-based biological predictions, such as protein fold recognition.

Table 2.2 Some Available Support Vector Machine Tools

SVM Tool	Downloadable Access
SVM^{light} (*34*)	http://svmlight.joachims.org/
SVM^{struct} (*35*)	http://svmlight.joachims.org/svm_struct.html
mySVM (*36*)	http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html
TinySVM	http://chasen.org/~taku/software/TinySVM/
LIBSVM (*37*)	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
SVMTorch (*38*)	http://bengio.abracadoudou.com/SVMTorch.html
LS-SVMlab (*39*)	http://www.esat.kuleuven.be/sista/lssvmlab/

2.4 Model Training and Evaluation

2.4.1 Training Procedure

The SVM^{light} platform applies a series of parameters to obtain the optimal training models. Some important parameters include kernel type, c (trade-off between training error and margin), g (parameter γ of the radial basis function kernel), and p (fraction of unlabeled examples to be classified into the positive class). In our training procedure, we applied the popular kernel of the radial basis function. All SVM parameters were optimized by a grid search ($c=2^{-10\sim-1}$, $g=2^{-12\sim-3}$, and $p=2^{-5\sim-2}$). For each grid point of the triplets (c, g, and p), a five-fold cross-validation procedure was employed to evaluate the performance of the trained SVM model. To carry out the five-fold validation procedure, we ran pair-wise similarity comparison to the training dataset. The total of 4925 positive epitopes were split into five groups, and any two-epitope sequences from two different groups did not have sequence similarity more than 20%. At each triplet point, the F-measure was calculated as shown in equation (4). F-measure quantifies a tradeoff of sensitivity and precision in prediction performance. The optimal parameter set has the largest value in all points with the maximum F-measures. During the procedure of five-fold cross-validation, five test results were used to calculate the mean values and 95% confidence intervals of sensitivity, precision, and the maximal F-measure.

To optimize the parameter set during the process of SVM training, the three performance statistics, sensitivity, precision, and F-measure as defined in Equation (4) were used as the major criteria. Sensitivity, also called recall, is used to check the proportion of true B-cell linear epitopes identified as positives from all actual positives. Precision represents the proportion of true positives from the predicted linear B-cell epitopes. These two statistics are of the most interest of immunologists when predicting linear B-cell epitopes. However, they showed different trends with the change of SVM

parameter sets. One example is shown in Figure 2.2. During our training, when decreasing the value of g (parameter γ of the radial basis function kernel) precision would go down while sensitivity would go up. Therefore, to determine the optimal g value, we can select g value from the cross point in the graph. the optimal parameter set is decided by the maximization of F-measure.

We also attempted different types of kernel functions in SVM^{light}, such as linear, polynomial, and radial basis. The final kernel in our models focused on the radial basis function after comparing their performance.

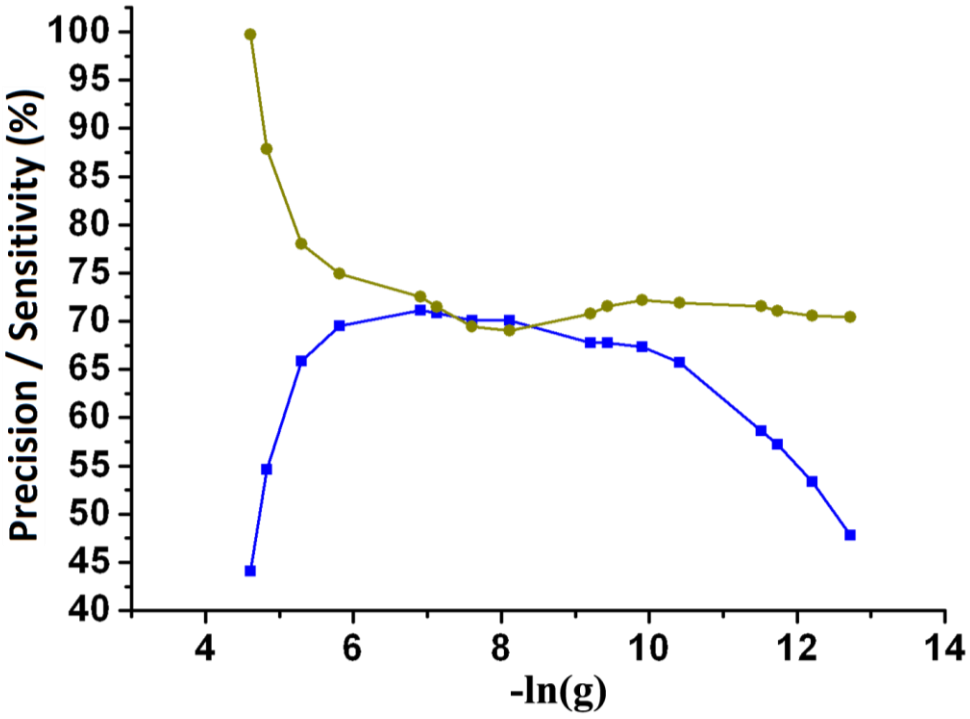

Figure 2.2 Curve of precision and sensitivity with different g values in SVM classifier training. g is parameter γ of the radial basis function kernel. The blue curve

stands for precision and the gray curve for sensitivity. Usually, the maximum F-measure can be found at the cross point of two curves where a optimal g parameter is determined.

2.4.2 Statistical Evaluation

To evaluate the prediction performance of linear B-cell prediction tool, performance statistics including sensitivity (Sen), specificity (Spe), precision (Pre), accuracy (Acc), Mattews correlation coefficient (MCC), and F-measure (F) are calculated by Equation (4) below,

$$\begin{aligned}
\text{Sen} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\% \\
\text{Spe} &= \frac{\text{TN}}{\text{TN}+\text{FP}} \times 100\% \\
\text{Pre} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100\% \\
\text{Acc} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \\
\text{MCC} &= \frac{(\text{TP})(\text{TN})-(\text{FP})(\text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \\
\text{F} &= \frac{2\times\text{Pre}\times\text{Sen}}{\text{Pre}+\text{Sen}},
\end{aligned} \tag{4}$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively. All of calculations above are based on five-fold cross-validation procedure.

Another statistical measure, AUC, is also calculated. AUC is the "area under the curve" where the curve is a receiver operating characteristic (ROC) curve. In the ROC curve, sensitivity, sometimes called TPR (True Positive Rate), is plotted against FPR (False Positive Rate), *i.e.* FPR = FP / (FP + TN). A higher AUC score represents higher

prediction performance. A java program available at http://pages.cs.wisc.edu/~richm/programs/AUC/ was used to calculate the AUC.

Although higher AUC values mean better prediction performance, the significance of two AUC score difference must be considered, too. Not significantly different AUCs mean that the two classifiers show equal performance when classifying unknowns. The online tool StAR was used to test whether the difference between ROC curves resulting from two models was statistically significant (*40, 41*).

2.5. Online Prediction Tool

We have also released an online tool for public use. The online tool SVMTriP is available on online (http://sysbio.unl.edu/SVMTriP). The major architect of this tool is shown in Figure 2.3. In the SVMTriP website, the online prediction tool contains three parts, 1) the identified optimal models based on SVM training with carefully-selected parameter sets; 2) a database used to store the requests from customers and the final prediction results after SVM classification by optimal models; and 3) a background server that implements the key process of SVM classification. The technologies involved in the SVMTriP online tool include Perl scripts, PHP, and My-SQL database. Blast Converting is applied to initial search of protein candidate in known epitope training dataset.

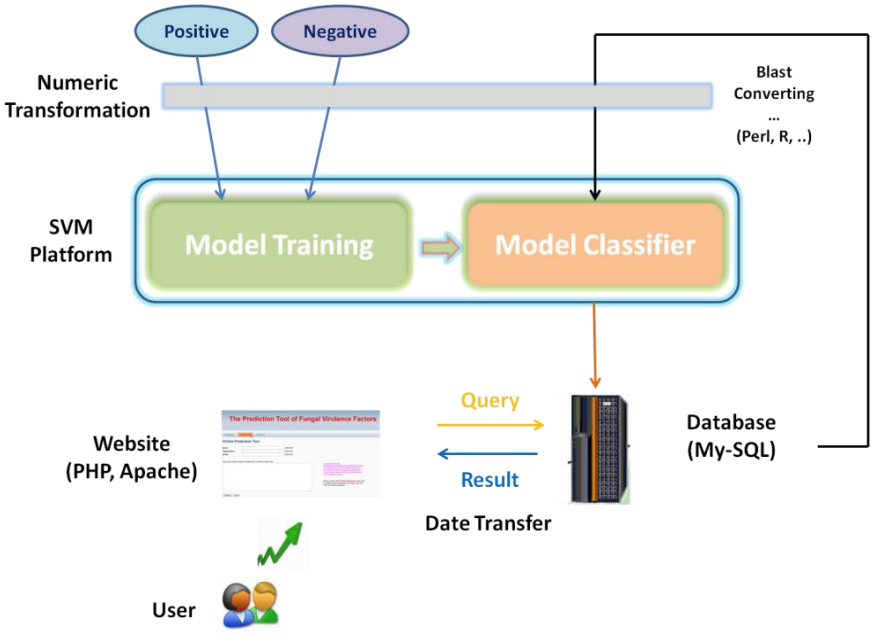

Figure 2.3: The illustration of the SVMTriP online tool

For the application on the online server, the prediction model is obtained by training SVMs using the datasets as described in the previous sections. To predict epitomes from a given full-length protein sequence, the sliding window method is employed to obtain subsequences with variable window sizes including 10AA, 12AA, 14AA, 16AA, 18AA, and 20AA with a step size of 2AA. For each subsequence, SVMTriP calculates its score, and a positive score indicates that the subsequence is a putative antigenic epitope. Based on the testing results, 20AA was set as the default epitope length for SVMTriP to search for putative epitopes on the web server.

3. Results

3.1 Prediction performance

SVMTriP is trained and tested with different epitope lengths, and for each length, the SVM parameters have their independent optimal values. For example, for 20AA-

length cases, SVMTriP reaches its optimal performance at c=32, g=0.05, and p=0.5 for the SVM model with Sn=80.1%±2.1% and P=55.2%±1.0% at the point with the maximal F-measure, 0.693. All results are shown in Table 2.3. Though, for different lengths of epitope sequences, SVMTriP has various points with the maximal F-measure, the precision values for different lengths are similar. The sensitivity increases significantly as the length of the epitope sequences increases. The range of the values of areas under the receiver operating characteristic curves (AUC) is from 0.674 to 0.702. Based on results of the performance assessment, SVMTriP for 18AA- and 20AA-length cases have the best performance. However, one may note a fact that most of experimentally determined epitopes from IEDB have less than 20 AA residues. A possible reason why SVMTriP favors long length of sequences is that a long sequence may have more tri-peptides to show a detectable frequency tendency. Another possibility is that the epitopic amino acid residues in experimentally determined epitopes are subsets of all real epitopic residues. Based on the testing results, 20AA is set as the default epitope length for SVMTriP to search for putative epitopes on the web server.

3.2 Comparison with AAP and BCPred

For comparison, AAP and BCPred are implemented locally based on their method descriptions, trained/tested with the same dataset and the five-fold cross-validation procedure for 20AA case. The results are listed in Table 2.4. Compared with BCPred and AAP, SVMTriP has a similar precision value, but significantly improved sensitivity at the point with the maximal F-measure. Figure 2.4 shows the ROC curves for the three methods. One may notice that SVMTriP has significantly larger sensitivity than BCPred and AAP in the region of low false positive rate. The AUC values are 0.667, 0.667, and

0.702 for AAP, BCPred, and SVMTriP, respectively. The AUC value of SVMTriP is significantly higher than those from the other two methods; the p-values of comparison against AAP and BCPred are 2.17×10^{-5} and 1.58×10^{-5}, respectively.

Table 2.3 Performance of SVMTriP models with different epitope lengths

Length (AA)	Sn (%)	P (%)	F-measure	AUC
10	68.5 ±2.5	55.5 ±1.5	0.615 ±0.020	0.674
12	67.5 ±3.5	57.0 ±2.0	0.620 ±0.030	0.681
14	64.8 ±4.9	56.5 ±2.5	0.605 ±0.030	0.689
16	63.5 ±5.5	57.1 ±3.0	0.601 ±0.045	0.685
18	79.0 ±1.9	54.1 ±1.1	0.641 ±0.015	0.666
20	80.1 ±2.1	55.2 ±1.0	0.693 ±0.060	0.702

Table 2.4 Performance of different linear B-cell epitope prediction methods

Methods	Sn (%)	P (%)	F-measure	AUC
AAP*	59.8 ±0.9	58.5 ±6.5	0.590 ±0.040	0.667
BCPred*	54.0 ±7.1	60.5 ±2.5	0.572 ±0.055	0.667
SVMTriP	80.1 ±2.1	55.2 ±1.0	0.693 ±0.060	0.702

*The results for AAP and BCPred, are obtained by the software implemented locally.

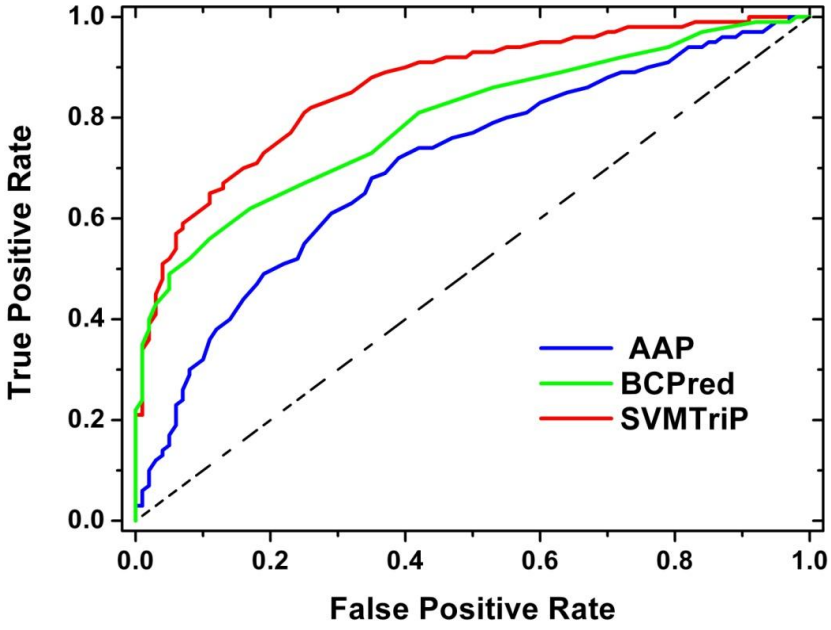

Figure 2.4 ROC curves for AAP, BCPred, and SVMTriP

4. Discussion

4.1 Determination of Different Models

The determination of linear B-cell epitope prediction models relies on different factors. The first consideration is the length of epitopes. The range of epitopes determined experimentally usually varies from 5 to 30AA. The optimal length of B-cell epitopes for computational prediction is unknown. Results obtained by ABCpred suggested the epitope length in statistical models should be no more than 20AA (*14*). In this study, we set up individual models with different epitope lengths. Another factor is the combination of features used in models. The optimal combination of features is also not clear. Based on different training datasets, the optimal feature set in different linear B-cell epitope prediction models may be quite different. Hence, in this study, we also determined the optimal combination of specific features with statistical evaluation.

The optimal window size for linear B-cell epitope is yet another unknown. We noticed that among the reported linear B-cell epitopes, over 60% were with the length of 12-18AA. During the process of learning, there can be a minor deviation between the optimal window size for prediction and the average length of real linear B-cell epitopes. For example, the function of B-cell linear epitopes is potentially influenced by the neighboring residues. Hence, a longer sequence segment embedding linear epitopes may contain more useful information. Therefore, in our learning procedures, we attempted various lengths of slide window, including 10AA, 12AA, 14AA, 16AA, 18AA, and 20AA.

Our training showed the optimal model came from the combination of tri-peptide score matrix and tri-peptide subsequence propensity. The hydrophobicity and predicted secondary structure scales did not contribute to final model SVMTriP. The optimal SVMTriP model was based on tri-peptide score matrix and tri-peptide subsequence propensity.

4.2 The Influence of Different Kernels on SVMTriP Models

4.2.1 Prediction with tri-peptide propensity alone

The propensity of tri-peptide alone is tested and the result is shown in Table 2.5. The prediction sensitivity is 56.5%, which is little smaller than 59.8% of AAP, a method based on bi-peptide propensity. On the other hand, the precision of tri-peptide propensity is 61.0%, which is similar with AAP's precision of 58.5%. This result indicates that combining similarity scores is essential for the tri-peptide model to achieve a better performance.

4.2.2 Prediction with tri-peptide similarity alone

The tri-peptide similarity scores can be calculated with either Blosum62 or PAM160 matrixes (the results of Blosum and PAM metrics are not shown here). The performance of two different matrices for the tri-peptide model is evaluated with the same procedure of the five-fold cross-validation for 20AA-length epitopes. The results are shown in Table 2.5. Without the propensity score, using Blosum62 matrix shows similar performance as using the PAM160. However, when combined with the propensity score, Blosum62 matrix leads to a higher prediction performance.

4.2.3 Discrete tri-peptide subsequence models

We also implement a method that uses the space of tetra-peptide subsequence with one mismatch, *i.e.,* discrete tri-peptide subsequences. For this case, the subsequences are considered in patterns either A_AA or AA_A, where 'A' represents the amino acid residue to be considered and '_' represents the residue position that will be ignored in the comparison. The number of SVM attributes is still 20^3, which is identical to that of the tri-peptide model. Interestingly, as shown in Table 2.5, without considering propensity scores, the subsequence models of A_AA and AA_A patterns have similar sensitivity and precision with the tri-peptide model. However, the combination of similarity and propensity of the tri-peptide model significantly enhances the performance, while addition of the propensity does not increase sensitivity and precision for A_AA and AA_A patterns. This finding indicates that the propensity is more important for the tri-peptide model than the discrete tri-peptide subsequence model.

Table 2.5 Comparison among the tri-peptide subsequence models with or without propensity

Kernels			Sn (%)	P (%)	F-measure
Tri-peptide	Propensity only	N.A.	56.5 ±12.5	61.0 ±6.3	0.584 ±0.085
Tri-peptide	w./o. Propensity	Blosum62	54.5 ±6.5	60.5 ±1.5	0.573 ±0.035
		PAM160	55.0 ±7.2	61.1 ±1.8	0.578 ±0.040
	w./ Propensity	Blosum62*	80.1 ±2.1	55.2 ±1.0	0.693 ±0.060
		PAM160	69.3 ±10.0	58.5 ±3.5	0.633 ±0.050
AA_A pattern	w./o. Propensity	Blosum62	54.8 ±6.8	60.5 ±1.5	0.579 ±0.040
		PAM160	55.2 ±7.1	61.3 ±2.0	0.577 ±0.045
	w./ Propensity	Blosum62	60.5 ±5.5	57.5 ±2.5	0.589 ±0.040
		PAM160	59.5 ±5.5	57.5 ±1.5	0.585 ±0.035
A_AA pattern	w./o. Propensity	Blosum62	55.5 ±8.5	60.6 ±2.2	0.581 ±0.050
		PAM160	55.2 ±8.1	60.5 ±1.5	0.577 ±0.055
	w./ Propensity	Blosum62	60.5 ±6.5	57.5 ±1.5	0.590 ±0.040
		PAM160	59.5 ±5.5	57.5 ±1.5	0.585 ±0.025

* The parameter set with Blosum62, Tri-peptide, and propensity were chosen to determine the optimal model of SVMTriP

4.2.4 Top weighted tri-peptide

The prediction model relies on the occurring-frequency distribution of tri-peptides in the tri-peptide space, *i.e.*, all combinations of any three amino acids. In Table 2.6, tri-peptide with top 20 weights in the optimal SVM model of 20AA-length epitopes are listed. All of the top ranked tri-peptides contain Glutamine or Proline, whereas the occurring frequencies of Glutamine and Proline in known linear epitopes (20AA) are only 8.1% and 6.84%, respectively. In the background of overall proteins, the occurring frequencies of Glutamine and Proline are 3.84% and 3.44%, which is not significantly different to the values in linear epitopes. The tri-peptide containing Glutamine or Proline may play an important role in epitope recognition by B-cell antibodies. The algorithm of SVMTriP successfully utilized this difference to distinguish linear epitopes from other parts of protein peptides.

Table 2.6 Weights of tri-peptides in the optimal SVM model

Tri-peptide	Rank	Weight Score*		Tri-peptide	Rank	Weight Score*
QQP	1	503251.79		GQQ	11	121677.62
PQQ	2	488627.71		QPY	12	116598.60
QPQ	3	367386.40		YPQ	13	113237.37
QPF	4	246462.39		QQF	14	81709.59
FPQ	5	234868.65		PYP	15	79191.37
PQP	6	231353.73		FQQ	16	77357.97
QGQ	7	153161.76		PPP	17	76320.05
PFP	8	151840.02		QPP	18	64756.05
QQQ	9	128930.20		QFP	19	63814.16
QQG	10	122291.90		PPQ	20	63173.33

*Weight scores are calculated by the formula $w = \sum \alpha_i x_i$. Here α_i is dual representation of the decision boundary; and x_i (i=0, 1, 2…n) is vector described in SVM model. Both α_i and x_i are available in the model file.

4.3 Independent Test to Compare SVMTriP and Other Linear B-cell Epitope Prediction Tools

Another independent test was developed as a tendency test between virus and human proteins by BCPred, AAP, and SVMTriP. Independent testing of different epitope prediction methods is challenging because of the limited number of known epitopes. In this study, we devise an alternative independent test method. In the training set, most epitopes are from virus or bacteria, and their corresponding antibodies are mainly human antibodies. A basic property of the human immune system is the capability to distinguish any pathogenic agents, viral or bacterial, from the innate structures of the human being. All known B-cell epitopes in the training set came from the response of whole immune system, including the response of CD4 T helper cells. Trying to simulate the human immune system, a successfully trained epitope prediction method should act the same, *i.e.*, be able to distinguish pathogenic proteins from human proteins. In other words, the virus proteins should be preferentially more highly scored than human proteins by a successful

prediction algorithm. To implement this test, 105 20AA-length peptides are collected from virus and human proteins: 5×10^4 peptides are randomly selected from 391,466 virus proteins and others from 81,967 human proteins in the NCBI Refseq protein database. AAP, BCPred, and SVMTriP are applied to these virus and human peptides, and top-ranked peptides are returned. The fractions of virus peptides in different numbers of top-ranked peptides are shown in Figure 2.4. All three methods returned more virus peptides than human peptides within the top-ranked peptides. SVMTriP, however, selected higher percentage of virus peptides than both AAP and BCPred. For example, in total 400 top-ranked peptides returned by SVMTriP, 90.5% of them, *i.e.* 362, are virus peptides. There are 47.8% (191) and 56.5% (226) virus peptides returned by AAP and BCPred, respectively. This indicates the exceptional ability of SVMTriP to distinguish epitopic and non-epitopic peptides.

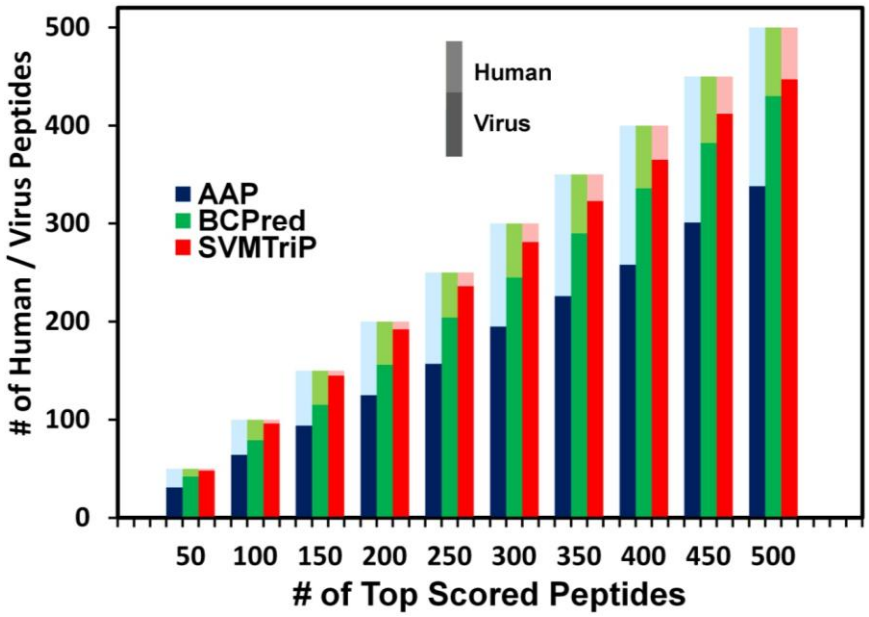

Figure 2.4 Tendency test for BCPred, AAP, and SVMTriP. Three bars at the same point on the *x*-axis are the results for APP (blue), BCPred (green), and SVMTriP (red),

respectively. In the same bar, the light part is for the number of returned human peptides, and the dark part is for virus. For example, at the point of 400 returned peptides, the dark part in the red bar is 362, which means that 362 viral peptides are return in all 400 peptides by SVMTriP, and the light red part represents 38 human peptides.

4.4 The Challenge of Linear B-cell Epitope Prediction

Although the prediction of linear B-cell epitope has achieved some success, we are still not fully satisfied with currently available tools. One of major issues is still rather high false positive rates. The main reason is that antibody-binding sites on antigens are much less conserved than other types of binding sites. The predator-prey game between antibody and antigen has changed the triumphal side many times during the evolutionary history. Antigen tried to avoid the recognition and clearance by antibody by changing its active sites, while antibody does all the best to figure it out and bind to the antigen. As a consequence, the binding sites on antigen, *i.e.*, epitopes, show less conservation and lower sequence similarity. At present, almost all of linear B-cell epitope prediction tools, including SVMTriP, are based on the assumption of conservation and sequence similarity among known linear B-cell epitopes. This is a dilemma between evolutionary immune-pressure and computational algorithms. One possibility is that B-cell antibodies show certain favorite binding pattern on antigen considering their special Y-shapes.

Another uncertainty in these available prediction tools is from the training datasets used. Obviously the datasets are of vital importance for prediction performance. The known linear B-cell epitopes mostly were determined by well-designed experiments. However, such experimental determination is slow and fund-consuming. Moreover, these

experiments usually focus on specific disease-related antigens which people are more interested in. The bias in datasets used for training linear B-cell epitope prediction methods influences the identification accuracy. With a limited number of available linear B-cell epitopes, Blythe and Flower (*42*) showed that propensity based methods cannot be used reliably for predicting B-cell epitopes, which could only yield success rate marginally better than random prediction. A bigger non-redundant dataset should be one of key factors in future development of improved linear B-cell epitope prediction methods.

We also compare the specificity of linear B-cell epitope prediction compared to other protein-protein interaction prediction. For example, given the interaction between antibody and antigen is usually transient non-obligate, the binding tends to be one kind of rather weak interaction (*43*). It means that residues of antibody and antigen proteins are involved in temporary and unstable adjacency. Special properties of these residue-residue bonds potentially affect the performance of algorithms which based on conserved and stable residue-residue interaction, such as some protein-protein binding site prediction tool (*44*). Hence, it is difficult to simply use prediction methods for regular protein-protein interaction in epitope prediction.

We developed a new method, SVMTriP, to predict linear antigenic epitopes. Applied to non-redundant B-cell linear epitope data extracted from IEDB, SVMTriP achieves a sensitivity of 80.1%, a precision of 55.2%, and AUC of 0.702 with five-fold cross-validation. The combination of similarity and propensity of tri-peptide subsequences can improve the prediction performance for linear B-cell epitopes.

Moreover, SVMTriP is capable of recognizing viral peptides from human protein sequences effectively.

The SVMTriP website collects the queries from users and stores them as the job queue in the database. The local service checks these queries and completes the searching along the input antigen sequence one by one. The predicted linear B-cell epitopes will be stored in the database and sent to the Result webpage. Generally, under a normal job load, a complete search on 200AA-length antigen needs about 20 minutes.

On the SVMTriP website, we also released the training dataset extracted from IEDB database as of June 11th, 2012 (http://sysbio.unl.edu/SVMTriP/download.php). This dataset is a non-redundant linear B-cell epitope set containing 4925 entries. For each entry, we obtained the real epitope segments and the full-sequence of the corresponding antigen. We extended or subtracted real epitope segments to construct 10AA, 12AA, 14AA, 16AA, 18AA, and 20AA subset. These datasets may be used to similar model training for new B-cell linear epitope prediction tools using other training features or algorithms.

References

1. Getzoff, E. D., Tainer, J. A., Lerner, R. A., and Geysen, H. M. (1988) The chemistry and mechanism of antibody binding to protein antigens, *Advances in immunology 43*, 1-98.
2. Milich, D. R. (1989) Synthetic T and B cell recognition sites: implications for vaccine development, *Advances in immunology 45*, 195-282.
3. Reineke, U., and Schutkowski, M. (2009) Epitope mapping protocols, 2nd ed., pp 1 online resource (xiii, 456 p., [416] p. of plates), Humana Press, New York.
4. Hopp, T. P., and Woods, K. R. (1981) Prediction of protein antigenic determinants from amino acid sequences, *Proceedings of the National Academy of Sciences of the United States of America 78*, 3824-3828.
5. Parker, J. M., Guo, D., and Hodges, R. S. (1986) New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites, *Biochemistry 25*, 5425-5432.
6. Emini, E. A., Hughes, J. V., Perlow, D. S., and Boger, J. (1985) Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide, *Journal of virology 55*, 836-839.
7. Pellequer, J. L., Westhof, E., and Van Regenmortel, M. H. (1993) Correlation between the location of antigenic sites and the prediction of turns in proteins, *Immunology letters 36*, 83-99.
8. Karplus, P. A., and Schulz, G. E. (1985) Prediction of Chain Flexibility in Proteins - a Tool for the Selection of Peptide Antigens, *Naturwissenschaften 72*, 212-213.
9. Kolaskar, A. S., and Tongaonkar, P. C. (1990) A semi-empirical method for prediction of antigenic determinants on protein antigens, *FEBS letters 276*, 172-174.
10. Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A., and Peters, B. The immune epitope database 2.0, *Nucleic acids research 38*, D854-862.
11. Saha, S., Bhasin, M., and Raghava, G. P. (2005) Bcipep: a database of B-cell epitopes, *BMC genomics 6*, 79.
12. Schonbach, C., Koh, J. L., Sheng, X., Wong, L., and Brusic, V. (2000) FIMM, a database of functional molecular immunology, *Nucleic acids research 28*, 222-224.
13. Larsen, J. E., Lund, O., and Nielsen, M. (2006) Improved method for predicting linear B-cell epitopes, *Immunome research 2*, 2.
14. Saha, S., and Raghava, G. P. (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network, *Proteins 65*, 40-48.
15. Chen, J., Liu, H., Yang, J., and Chou, K. C. (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale, *Amino acids 33*, 423-428.
16. El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2008) Predicting linear B-cell epitopes using string kernels, *J Mol Recognit 21*, 243-255.
17. Alix, A. J. (1999) Predictive estimation of protein linear epitopes by using the program PEOPLE, Vaccine 18, 311-314.

18. Odorico, M., and Pellequer, J. L. (2003) BEPITOPE: predicting the location of continuous epitopes and patterns in proteins, J Mol Recognit 16, 20-22.
19. Wee, L. J., Simarmata, D., Kam, Y. W., Ng, L. F., and Tong, J. C. SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction, BMC genomics 11 Suppl 4, S21.
20. Wang, Y., Wu, W., Negre, N. N., White, K. P., Li, C., and Shah, P. K. Determinants of antigenicity and specificity in immune response for protein sequences, BMC bioinformatics 12, 251.
21. Gao, J., Faraggi, E., Zhou, Y., Ruan, J., and Kurgan, L. BEST: improved prediction of B-cell epitopes from antigen sequences, PloS one 7, e40104.
22. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic acids research 25, 3389-3402.
23. Lei, Z., and Dai, Y. (2005) An SVM-based system for predicting protein subnuclear localizations, BMC bioinformatics 6, 291.
24. Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks, PNAS 89, 10915-10919.
25. Dayhoff, M.O., Schwartz, R. and Orcutt, B.C. (1978) A model of Evolutionary Change in Proteins. Atlas of protein sequence and structure (volume 5, supplement 3 ed.). Nat. Biomed. Res. Found. pp. 345–358. ISBN 0-912466-07-3.
26. Pruitt, K. D., Tatusova, T., Klimke, W., and Maglott, D. R. (2009) NCBI Reference Sequences: current status, policy and new initiatives, Nucleic acids research 37, D32-36.
27. Kyte, J., and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein, Journal of molecular biology 157, 105-132.
28. Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, Journal of molecular biology 292, 195-202.
29. Wang, H., and Huang, G. Application of support vector machine in cancer diagnosis, Medical oncology (Northwood, London, England) 28 Suppl 1, S613-618.
30. Zhang, M. M., Yang, H., Jin, Z. D., Yu, J. G., Cai, Z. Y., and Li, Z. S. Differential diagnosis of pancreatic cancer from normal tissue with digital imaging processing and pattern recognition based on a support vector machine of EUS images, Gastrointestinal endoscopy 72, 978-985.
31. Schramm, A., Schulte, J. H., Klein-Hitpass, L., Havers, W., Sieverts, H., Berwanger, B., Christiansen, H., Warnat, P., Brors, B., Eils, J., Eils, R., and Eggert, A. (2005) Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling, Oncogene 24, 7902-7912.
32. Andorf, C. M., Honavar, V., and Sen, T. Z. Predicting the binding patterns of hub proteins: a study using yeast protein interaction networks, PloS one 8, e56833.
33. Vapnik, V., and Lerner, A. J. (1963) Pattern recognition using generalized partrait method, Automation and Remote Control 24, 774-780.
34. Joachims, T. (1999) Making Large-Scale SVM Learning Practical, Advances in Kernel Methods - Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed.), MIT-Press.

35. Schölkopf, B., Burges, C. J. C., and Smola, A. J. (1999) Advances in kernel methods : support vector learning, MIT Press, Cambridge, Mass.
36. Ruping, S. (2000) mySVM-Mannual, University of Dortmund, Lehrstuhi Informatik 8, http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM.
37. Chang, C. C., and Lin, C. J. (2011) LIBSVM: a library for support vector machine, ACM Transaction on Intelligent Systems and Technology 2:27.
38. Collobert, R., and Bengio, S. (2001) SVMTorch: Support Vector Machines for Large-Scale Regression Problems, Journal of Machine Learning Research 1, 143-160.
39. Pelchmans, K., Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., Lukas, L., Hamers, B., Moor, B. D., and Vandewalle, J. (2002) LS-SVMlab: a Matlab/C toolbox for Least Squares Support Vector Machines, Internal Report 02-44, ESAT-SISTA, K.U. Leuven (Leuven Belgium) Lirias number: 21472.
40. DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44, 837-845.
41. Vergara, I. A., Norambuena, T., Ferrada, E., Slater, A. W., and Melo, F. (2008) StAR: a simple tool for the statistical comparison of ROC curves, BMC bioinformatics 9, 265.
42. Blythe, M. J., and Flower, D. R. (2005) Benchmarking B cell epitope prediction: underperformance of existing methods, Protein Sci 14, 246-248.
43. Ponomarenko, J. V., and Bourne, P. E. (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation, BMC structural biology 7, 64.
44. Neuvirth, H., Raz, R., and Schreiber, G. (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites, Journal of molecular biology 338, 181-199.
45. Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids, Nucleic acids research 38, W529-533.

CHAPTER THREE: PREDICTION OF CONFORMATIONAL B-CELL EPITOPES

1. Introduction

Conformational B-cell epitopes are discontinuous segments along antigen sequence and they are responsible to the interaction between antigen and antibody. A conformational epitope is usually composed of several shorter antibody-binding regions physically separated on antigen sequence. Therefore, a conformational B-cell epitope can be considered as the union of multiple distinct shorter linear B-cell epitopes. These short linear B-cell epitopes are located independently on the surface of antigen. Therefore, it is also possible each of these linear epitopes belongs to different chains of the antigen if a quaternary structure of multiple polypeptide chains exists in the antigen protein (*1*, *2*). The distribution of subunits of conformational B-cell epitopes gives a hint for the 3D structural shape of the entire antigen protein. Determining conformational B-cell epitopes approximately equals to identifying the binding surface structure of antigen proteins. Usually in the process of identification conformational B-cell epitopes experimentally, resolving 3D structures of antigen proteins is the primary strategy. Protein 3D structures can be determined by the nuclear magnetic resonance (NMR) or the X-ray crystallography, both highly time-consuming. In the Protein Bank Database, therefore, there is only a very limited number of structures of antigen-antibody complexes (*3*).

Although accurate prediction of antigenic epitopes is needed for immunological research and medical applications, it is still a challenging task. Prediction of conformational B-cell epitopes seems more difficult so currently there are only a few

such methods available. The lack of 3D structures of antigens is the main barrier. Without 3D structural information of antigen binding to antibody, it is difficult to identify discontinuous subunits of conformational B-cell epitopes and to construct a positive training dataset for prediction methods. As we discussed, the reliable and unbiased positive training dataset is the key for the success of conformational B-cell epitope prediction. All discontinuous epitope prediction methods require the three-dimensional structure of the antigenic protein. The small number of available antigen-antibody complex structures limits the development of reliable discontinuous epitope prediction methods and an unbiased benchmark set is very much in demand (*4*, *5*).

Although discontinuous epitopes dominate most antigenic epitope families (*6*), due to their computational complexity, only a very limited number of prediction methods exist for discontinuous epitope prediction". It also reads much better. Currently only several conformational B-cell epitope prediction tools are available: CEP (*7*), DiscoTope (*8*), BEpro (PEPITO) (*9*), ElliPro (*10*), SEPPA (*11*), EPITOPIA (*12*, *13*) and EPCES (*4*), and Bpredictor (*14*). Due to unsatisfactory performance of currently available methods, we developed new tools aiming to improve prediction reliability of conformational B-cell epitopes. In this section, we introduce an antigenic Epitope Prediction method by using Support Vector Regression (EPSVR) with six attributes: residue epitope propensity, conservation score, side chain energy score, contact number, surface planarity score, and secondary structure composition. With an independent test dataset, we compare EPSVR against other conformational B-cell prediction tools. EPSVR and its related resources are available on http://sysbio.unl.edu/EPSVR.

The idea of consensus results to improve confidence of prediction is popular in bioinformatic research. For example, ensemble method for gene prediction may integrate multiple single predictions. Among known conformational B-cell epitope prediction tools released, most of them have their own biased results due to its specific training datasets. To decrease the biased influences, a meta server can be used to obtain a consensus output from multiple prediction tools.

The success of the meta server also depends on the prediction performance of each member. We first consider the prediction performance of each method reported in literatures. The attributes used in prediction must be taken into consideration. Prediction would be biased if some attributes are used repeatedly in multiple methods. We also have to pay attention to the training sets used in those tools. Use of similar training datasets in multiple tools may cause bias in prediction. A good meta server should include methods based on a wide range of training sets. Finally, the availability of tools is also a requirement if we want to install it locally or obtain the results online for further consensus analysis. For example, because CEP tool is no longer available online, we removed it from our candidates. Similarly, we did not include ElliPro because it was not available for download.

Our EPmeta server incorporates EPSVR, EPCES, EPITOPIA, SEPPA, PEPITO, and Discotope1.2. Prediction is done by each tool individually first. The outputs from the six tools are combined together to generate a single consensus output. With an independent test dataset, EPmeta showed a more confident prediction than its each member. The EPmeta server is available at http://sysbio.unl.edu/EPmeta.

2. The Development of Novel Conformational B-cell Epitope Tool, EPSVR

2.1 Dataset collection

a) Training Dataset

The training set was gathered and screened from three protein datasets: 1) 22 antigen-antibody complexes and their unbound structures from protein docking Benchmark 2.0 (*15*); 2) 59 representative antigen-antibody complexes compiled by Ponomarenko and Bourne (*5*); 3) 17 antigen-antibody complex structures released between February 2006 and October 2008 with available unbound antigen structures, which was the test set use in EPCES server (*4*). Any antigen-antibody complex was discarded if its antigen had no available unbound structure because the unbound structures were required for prediction. A complex structure was not used if its antigenic epitope consisted of amino acid residues located on multiple chains. A complex was included if the sequence identity between its antigen and all other antigens from the other complex structures was less than 35% following local sequence alignment. For an antigen with a sequence identity in the range of 35~50%, we accepted the antigen-antibody complex if the binding topology was not the same as its homologous complex. For an antigen with more than one antigenic epitope, only one was used in order to avoid confusion in subsequent application of support vector regression methods. As a result, a total of 48 complexes and their unbound structures meeting the above criteria were used as a training set.

b) Testing Dataset

The test set was curated from 293 entries of the Conformational Epitope Database (CED, Release 0.03) with the following criteria (*16*). We only considered entries that had unbound antigen structures, but no complex structures. Multiple entries with the same antigen structure were combined and considered as one target, and antigenic residues from multiple entries were mapped onto one protein structure. The sequence identity between any two selected proteins was also required to be less than 35%. All selected antigens were also screened against the rest of CED database and our training set; the sequence identity between a selected antigen and other antigens with complex structures in the CED or in the training set was less than 35%. A total of 22 antigenic proteins in the CED met all the above criteria; these were: 1www, 1hgu, 1eku, 1mbn, 1av1, 1pv6, 1al2, 2gmf, 1a7c, 1y8o, 1og5, 1jeq, 1dab, 1w7b, 1ly2, 1rec, 1nu6, 2b5i, 2gib, 1p4t, 1xwv, and 1qgt. Three antigenic proteins, 1www, 1hgu, and 1xwv, were excluded since they had multiple antibody-binding sites and the mapped antigenic residues were evenly distributed on the protein surfaces. Therefore, the final test set contained 19 antigen structures.

2.2 Attributes

Six attributes were used to antibody binding site prediction.

a) Epitope propensity

Epitope propensity at the amino acid position *i*, $E_{\text{propensity}}(i)$, is defined as

$$E_{\text{propensity}}(i) = \left(\ln \frac{P_r^{\text{interface}}}{P_r^{\text{surface}}}\right) \times \frac{S_r}{S_r^{\text{ave}}} \quad (1)$$

where $P_r^{interface}$ and $P_r^{surface}$ represent the probabilities of residue type *r* located on antibody-binding interface or just on the surface of antigen protein, respectively. They were calculated based on the residue types from the antigen proteins in the training set. All of epitopic residues were considered to be at the antibody-binding interface. Parameters S_r and S_r^{ave} are the relative accessible surface area of residue *r* at the sequence position *i* and the average relative accessible surface area of surface residues of type *r*, respectively. The C_α atom of Gly is considered as a side chain atom for convenience (*17*). Since antigen-antibody interfaces have different residue composition compared with other protein-protein interfaces, we used the training dataset to derive residue-specific antibody binding site propensities in epitopes and background proteins.

b) Conservation score

A residue conservation score relies on position-specific substitution matrix (PSSM), which is obtained by three rounds of searches using PSI-BLAST (*18*) starting with the BLOSUM62 substitution matrix. The conservation score at the position i is defined as

$$E_{conserv}(i)=\begin{cases}|M_{ir} - B_{rr}|, \text{ if } M_{ir} - B_{rr} < 0, \\ 0, \text{ if } M_{ir} - B_{rr} > 0\end{cases} \quad (2)$$

where M_{ir} is the position-specific score in PSSM for the residue type *r* at sequence position *i*, and B_{rr} is the diagonal element of BLOSUM62 for residue type *r*. Conservation score is set to 0 if the position-specific score after three rounds of PSI-BLAST search is larger than the original position-specific score in BLOSUM62 (*17*). As we discussed in chapter two, epitopic residues show lower conservation than other

functional residues of most of non-antigenic proteins. In contrast to regular protein-protein interaction where conserved surface residues in the unbound structure are considered as interface residues, the poorly conserved residues of antigen are considered as the putative antibody-binding site residues due to adaptive evolutionary pressures for antigen proteins.

c) Side-chain energy score

Side-chain energy can influence protein structural conformation and further function on the spatial context of protein surfaces. Side-chain energy score is calculated from the side-chain energies of all possible rotamers for a given residue type at a sequence position whereas other sequence positions have native residue types and observed atomic coordinates. The weights of the energy function are optimized so that the native residue was predicted energetically favorable at each position of the training proteins. The assumption is that the residues at the antibody binding site have a higher energy score than other surface residues so that the free energy of the antigen-antibody system could go down significantly upon association.

The definition of side-chain energy score was given in Liang and Grishin (*19*). The energy unit is kcal·mol^{-1}. The side-chain energy score of amino acid *i* is defined as,

$$E_{\text{side chain}}(i) = -f \ln\{\sum_{R} \exp[-E_{\text{side chain}}(R_i)]\} \quad (3)$$

where the summation is over all the rotamers available for a given residue type and the constant prefactor f = 1/2.41, which is from the slope of the regression line between the calculated and experimentally measured unfolding ΔΔG of a set of point mutation data. R_i is a given rotamer of residue *i* and $E_{\text{side chain}}(R_i)$ is defined as,

$$E_{\text{side chain}}(R_i) = -0.143\,S_{\text{contact}} + 0.724\,V_{\text{overlap}} + 1.72\,E_{\text{hbond}} + 28.6\,E_{\text{elec}} - 0.0467S_{\text{pho}} + 0.0042S_{\text{phi}} + 1.14\,(F_{\text{phi}})^{30} + 7.95\,V_{\text{exclusion}} - 0.919\ln(f_1 f_2) - 4.3\,N_{\text{ssbond}} - G_{\text{ref}} \quad (4)$$

where S_{contact}, V_{overlap}, E_{hbond}, E_{elec}, ΔS_{pho}, and ΔS_{phi} represent atom-contact surface area, overlap, H-bond energy, electrostatic interaction energy, buried hydrophobic solvent accessible surface, and buried hydrophilic solvent accessible surface between the rotamer of residue *i* and the rest of protein, respectively. F_{phi}, $V_{\text{exclusion}}$, N_{ssbond}, and ΔG_{ref} are defined as the fraction of the buried surface of non-hydrogen-bonded hydrophilic atoms, the normalized solvent exclusion volume around charged atoms, the flag of disulfide bridge(1 or 0), and the difference between the free energy of the rotamer in solvent and denatured protein, respectively. f_1 is the observed frequency of the rotamer and f_2 is the observed frequency of the amino acid residues in a given backbone conformation.

d) Contact number

The residue contact number is the number of Cα atoms in the antigen within a distance of 10 Å of the Cα atom of residue *i* (*8*). A residue with a small contact number was considered as an antibody binding site residue.

e) Surface planarity score

The planarity of each surface patch was calculated by evaluating the root mean squared deviation (rmsd) of all the Cα atoms in the surface patch from the least squares plane through the atoms. The rms deviations were inverted such that a high planarity score for a patch was interpreted as a planar patch and antibody binding site (*20*).

f) Secondary structure composition

This score was defined as the fraction of patch residues forming turns or loops in all 20 patch residues. Based on Chou and Fasman's method (*21*), the α-helix and β-sheet were defined as four or more consecutive residues having φ and ψ angles within 40° of (-60 °, -50°) and three or more residues having φ and ψ angles within 40 °of (-120 °, 110 °) or (-140 °, 135 °), respectively. The remaining regions were considered turns and loops.

2.3 Training Procedure for EPSVR

For each surface patch, the number of epitopic residues could be any integer value between 0 and the patch size (20 for this study), and each surface patch had six Support Vector Regression (SVR) attributes as described above. The residue epitope propensity, conservation score, and side-chain energy score were calculated at the residue level and averaged over all residues in the patch. The six scores and the number of observed epitopic residues in the patch were scaled to 0~1.

All SVR parameters were optimized by a grid search ($c = 2^{-10\sim-1}$, $g = 2^{-12\sim-3}$, and $p = 2^{-5\sim-2}$) where c is trade-off between training error and margin, g is parameter gamma for radial basis function kernel, and p is the fraction of unlabeled examples to be classified into the positive class (*22*); and for each grid point of triplets, a leave-one-out procedure was applied to evaluate the trained SVR model. Specifically, the patch score of each surface patch for a target in the training set was predicted by the SVR model trained with the other 47 antigen-antibody complexes, from which the residue epitope propensity score was also derived. After this procedure was repeated 48 times, the mean AUC value of 48 predictions represents the performance of the current grid point for SVR parameters. The triplet of parameters that reached the highest value of mean AUC was chosen and

used for the test set, and the final support vector machine model was trained with all 48 targets.

2.4 Prediction Procedure for EPSVR

A surface patch is defined as a central surface residue and its 19 nearest surface neighbors in space, where a surface residue is defined if the relative accessibility of its side chain is greater than 6% with probe radius = 1.2Å. First, we searched for all surface residues and enumerated all surface patches of a given antigen structure, and calculated their six SVR attributes. For each surface patch, we predicted the number of putative epitopic residues by the trained SVR model. Here, a patch score was defined as the fraction of the number of putative epitopic residues to the total number of amino acid residues in the patch, *i.e.*, 20. One surface residue was assigned a residue score by averaging patch scores of all patches in which this amino acid residue is included. Finally, we sorted surface residues according to their residue scores and the top-ranked ones were considered as epitopic residues. The assumption here is that a residue frequently appearing in top- scoring patches is likely an epitopic residue.

Patch analysis was used in all existing B-cell discontinuous epitope studies. In the examples of EPCES and EPITOPIA, a patch score was derived by averaging the scores of all residues in the patch, and the central residues of top scored patches were predicted as epitopic residues. However, the value of the patch score was actually correlated with the number of epitopic residues in the patch rather than the central residue. Here, we used SVR to predict the number of epitopic residues in a surface patch and residues frequently located in the top scored patches were predicted as epitopic residues. For this case, the

SVR model is more suitable than a support vector classifier. In this study, we used an SVR package, called LIBSVM, obtained from http://www.csie.ntu.edu.tw/~cjlin/ libsvm.

2.5 Results

2.5.1 Prediction for the Training Set for EPSVR

When $c=2^{-6}$, $g=2^{-5}$, $p=2^{-3}$, the mean value of AUC for the 48 targets in the training set reached its maximum, 0.670, in the leave-one-out test. As a comparison, the mean AUC value is 0.644 predicted by EPCES, whose residue interface propensity was derived from the other 47 targets using the same leave-one-out procedure as described. The improvement of EPSVR could be attributed to the machine learning method because EPSVR and EPCES used the same six scoring terms. In another study, Rubinstein *et al.* applied support vector classifier (EPITOPIA) to predict B-cell epitopes and obtained a mean AUC value of 0.65 for a similar non-redundant set of 47 antigen-antibody complex structures in cross validation (*13*). Our algorithm showed slightly better performance for a somewhat different training set.

2.5.2 Prediction for the Test Set for EPSVR

We applied our algorithm, with the optimally trained parameters, to the independent test set, and achieved a mean AUC value of 0.597, which was lower than that of the training set. Nevertheless, 6 out of 19 targets were predicted with an AUC value greater than 0.7. Note that the interface residues of antigens in the test set were identified by point mutations, overlapping peptides, and ELISA, which are not as accurate as the crystal complex structure method.

Six antigens in test proteins (1eku, 1av1, 1al2, 1jeq, 2gib, and 1qgt) contained multiple chains, but we only used a single chain where the experimental antigenic epitope was located for prediction. If the whole protein was used for prediction, the mean AUC value of the six proteins decreased from 0.672 to 0.623. Unlike antigenic epitopes, the interfaces of protein-protein complexes, especially non-transient complexes, are usually more hydrophobic and conserved than protein surfaces; it makes the exposed protein-protein interfaces relatively-easily distinguishable from both the antigenic epitopes and other protein surfaces. In other words, a single chain protein that has both protein-protein binding interfaces and epitope made the epitope prediction task easier.

3. Development of Conformational B-cell Epitope Meta Tools EPmeta

3.1 Selection of Conformational B-cell Epitope Prediction Tools

Selection of conformational B-cell epitope prediction tools is a key step in constructing the meta server. Before EPSVR, there were only a limited number of methods available, *i.e.*, CEP (*7*), DiscoTope (*8*), PEPITO (*9*), ElliPro (*10*), SEPPA (*11*), EPITOPIA (*12, 13*), and EPCES (*4*). These tools applied different physicochemical properties of epitopes in their corresponding model-training, such as the hydrophilicity scale and the epitope log-odds ratios in DiscoTope (*8*), the epitopic residue propensity and the half sphere exposure values at multiple distances in PEPITO (*9*), residue protrusion index (PI) in ElliPro (*10*), the epitopic residue propensity and the compactness of the neighboring residues around one residue in SEPPA (*11*), forty-four physicochemical and structural–geometrical attributes in EPITOPIA (*12, 13*), and six physicochemical properties used in EPCES (*4*) and EPSVR (*23*). These attributes have

different effects on determining conformational B-cell epitopes. Hence, in our EPmeta server, different servers should have different weights in constructing the consensus result.

3.2 The Architecture of EPmeta

An open question for any meta server is how to quickly obtain prediction results from multiple tools. One choice is install these tools locally. Local installation requires standalone packages. Unfortunately, DiscoTope (*8*), PEPITO (*9*), ElliPro (*10*), SEPPA (*11*), or EPITOPIA (*12*, *13*) cannot provide a standalone package for local installation.

With the idea of remote searching technology, a machine-simulation will obtain the searching results on web browser through accessing remotely available servers. In contrast to using locally installed programs, a meta server can directly access to remotely available services, run a query, and download the results from the remote servers. This is the strategy used for EPmeta.

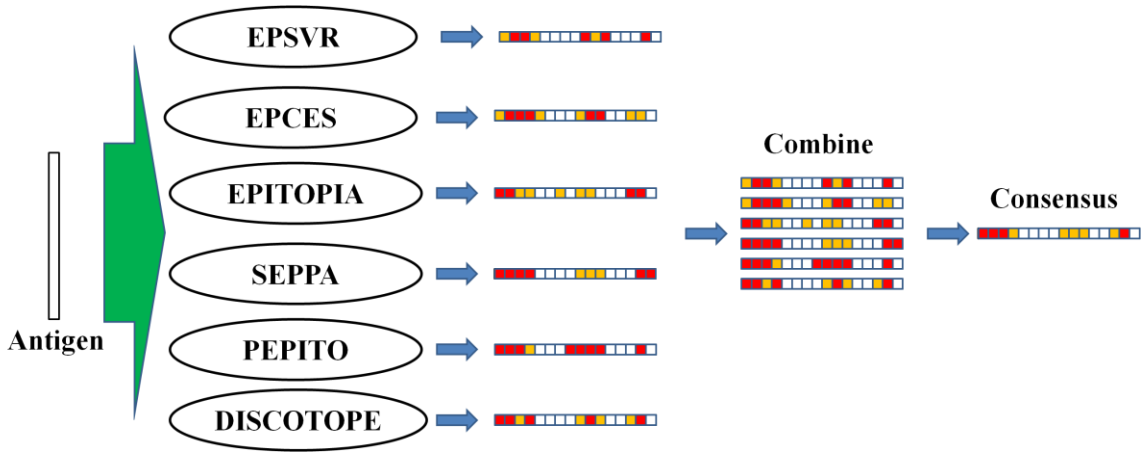

Figure 3.1 Architecture of the EPmeta server.

3.3 The programming technologies to complete the EPmeta server

The main technological blockage in using remote servers is how to communicate between the meta server and other online tools. Unfortunately, we solved this issue by carrying out a type of machine-simulation of automated web-browser navigation. Scripts can automate completing a series of customized operations on web browsers, such as to open a web browser, input a Uniform Resource Locator (URL) address, fulfill the form on web page, upload a PDB file (3D protein structures), and download a web source page. All the operations including human-like activities with mouse and keyboard in front of a computer are simulated. Hence, the meta server does not need an API or a web server to communicate with online tools. It just relies on the successful access to the remote servers by a web browser, such as Internet Explorer or Firefox.

There are many ways to carry out machine-simulated automated navigation of a web browser. Similar technologies are applied on automated testing during the development of web applications. To decrease human labor on repeated software testing, software development engineers in testing (SDET) developed the platform of automated testing. For automated testing of web applications, machine-simulated navigation of a web browser is one of the fundamental functions of the platforms. Although many such automated testing platforms are commercial products, there are a few open-source free platforms on Microsoft Windows with Internet Explorer and fewer on Linux with Firefox. In this study, we chose a platform based on a programming language Ruby and the Watir platform to our EPmeta. Watir is an open-source free platform providing the Ruby library for automating web browsers (http://watir.com/). Using Ruby and Watir, we completed

the automated access of online tools by our meta sever and achieved the query on these tools.

Figure 3.1 illustrates the architecture of the EPmeta server. We choose the six tools because their reports showed the best predictions when we prepared EPmeta server in 2009. After completing six queries to EPSVR, EPCES, EPITOPIA, SEPPA, PEPITO, and Discotope1.2, we generate a consensus result.

For the Meta server, the basic idea was that a surface residue is predicted as an epitopic residue if two or more single servers voted for it. In this naive sense, the mean AUC values of the 19 testing proteins was calculated to be 0.562, 0.618, 0.627, 0.621, and 0.612 predicted by the top 2, 3, 4, 5, and 6 servers, respectively (Table 3.1). To adopt a more sophisticated strategy, the top 25% of surface residues were returned as predicted epitopic residues by EPSVR, EPCES, and EPITOPIA. When the number of the predicted residues was increased from 25% to 50%, from 50%to 75%, and from 75% to 100%, SEPPA, PEPITO, and DiscoTope1.2 were, respectively, included in the voting. For example, the new antigenic residues predicted by EPSVR, EPCES, EPITOPIA, and SEPPA were added to the top 25% residues predicted by EPSVR, EPCES, and EPITOPIA. The prediction started with 1% of the surface residues for each of the four servers and increased insteps of 1% until 50% of surface residues were predicted as antigenic residues. Then we added PEPITO and used five servers to predict the top 50%~75% surface residues and so on. With this method, we achieved a mean AUC value of 0.638, which is higher than all single servers, especially, Discotope1.2 and PEPITO (p-value < 0.05). The reason that we used this strategy to integrate the various predictions results from our finding that a single server had better prediction accuracy when only a

small fraction of the surface residues were predicted as epitopic residues. If 50% of surface residues, for example, were predicted as epitopic residues by the Meta server, the prediction accuracy was 14.4% for the Meta server with a voting set including EPSVR, EPCES, and EPITO-PIA, where each server output the top 51% surface residues as candidates of antigenic residues. As a comparison, the prediction accuracy was slightly higher (15.3%), if the Meta server also returned 50% of the surface residues as epitopic residues, but got votes for those returned residues from all of the six servers, where each server output their own top 32% surface residues as candidates of epitopic residues.

The following is the final algorithm:

```
[
BEGIN
N = the total number of surface residues;
E = the number of predicted epitopic residues;
if E ≤25% * N then,
        return Predictor (0, E, EPSVR, EPCES, EPITOPIA);
else if E > 25%*N AND E ≤50%*N then,
        return Predictor (R25, E, EPSVR, EPCES, EPITOPIA, SEPPA);          // Rp = p% of surface
residues already predicted as epitopic residues;
else if E > 50% AND E ≤75%*N then,
        return Predictor (R50, E, EPSVR, EPCES, EPITOPIA, SEPPA, PEPITO);
else if E > 75% AND E ≤100%*N then,
        return Predictor(R75, E, EPSVR, EPCES, EPITOPIA, SEPPA, PEPITO, Discotope1.2)
endif.

Function Predictor(Rp,E,SERVER1,SERVER2,SERVER3,...)
```

```
Begin
        set the prediction of each single server to 0;
do {
        Increase the prediction of each single server at the step of 1%;
        Collect residues predicted by at least two of the servers;
} While(Rp + collected epitopic residues other than Rp < E);
Return total epitopic residues;
END
]
```

3.4 Results

Although EPSVR and EPCES used the same six scoring terms, we found that it was necessary to include both of them in the Meta server. When we used a voting server set including EPCES, EPITOPIA, and SEPPA, *i.e.* excluding EPSVR, the average AUC value decreased to 0.587 for the test set. The average AUC value predicted by EPSVR, EPITOPIA, and SEPPA (0.611) was also lower than that predicted by EPSVR, EPCES, and EPITOPIA in the standard procedure (0.618). We also tried to increase the threshold of votes from two to three for a voting server set, but the results did not improve.

4. Discussions

We introduced a SVR method to integrate six attributes for discontinuous epitope prediction and a server, EPSVR, which can be accessed online. The AUC of EPSVR is 0.597, which is higher than that of any other existing single server. Although they used the same scoring functions, EPSVR exhibited improved performance over EPCES. This was attributed to the fact that EPSVR searched the six-dimensional parameter space of all

scores more broadly than the voting method we previously used. Furthermore, a Meta server, EPmeta, combining EPSVR and the other existing single servers together, had an AUC value of 0.638, which is higher than any single server, especially, DiscoTope and PEPITO. We also found that the use of both EPSVR and EPCES, which use the same 6 scoring terms, resulted in a higher performance for EPmeta than if only one was used. The AUC results for different methods are shown in Table 3.1.

Table 3.1 List of the Conformational B-cell Epitope Prediction Methods and Their Obtained AUC Results

Method	URL of web server	AUC	Accuracy[b] (%)
DiscoTope (*8*)	http://www.cbs.dtu.dk/services/DiscoTope/	0.567	15.5
BEpro(PEPITO) (*9*)	http://pepito.proteomics.ics.uci.edu/	0.570	17.0
ElliPro (*10*)	http://tools.immuneepitope.org/tools/ElliPro/iedb_input	0.585	14.3
SEPPA (*11*)	http://lifecenter.sgst.cn/seppa/index.php	0.576	17.2
EPITOPIA (*12*, *13*)	http://epitopia.tau.ac.il/index.html	0.579	17.8
EPCES (*4*)	http://sysbio.unl.edu/EPCES/	0.586	18.8
EPSVR (*23*)	http://sysbio.unl.edu/EPSVR/	0.597	24.7
Bpredictor (*14*)	http://code.google.com/p/my-project-bpredictor/downloads/list	0.598[a]	24.0[c]
EPmeta (*23*)	http://sysbio.unl.edu/EPmeta/	0.638	25.6

[a)] The AUC value is obtained from the Reference (*14*). [b)] 10% of surface residues are returned as predicted epitopic residues. [c)] Estimated based on the Figure 4 in the Reference.

To assess and compare prediction performance of these predictors, we carried out an independent test by the testing set containing 19 protein monomer structures with epitope information derived from experimental methods other than crystal structures. AUC score is the major criterion for each method. A receiver operating characteristic (ROC) curve represents a dependency of true positive rates (sensitivity) and false positive

rates (1-specificity), plotted at various thresholds. To change the thresholds, the number of predicted residues is increased in steps of 1% of total surface residues. The mean AUC values are calculated using a java program available at http://pages.cs.wisc.edu/~richm/programs/AUC/ (*24*), except for Bpredictor. For Bpredictor, the AUC value is directly obtained from the manuscript (*14*) where the same benchmark by Liang *et al.* (*4*) was applied as in the current work. Among single servers, EPSVR and Bpredictor have the best performance according to the AUC values. Although EPSVR has the highest mean AUC value, the differences between EPSVR and other servers are not statistically significant (p-value >0.05), according to the pairwise t-student tests. The Meta server, EPmeta, achieves a mean AUC value of 0.638, which is significantly higher than all single servers.

We also calculated the accuracy by the same independent test because the accuracy, *i.e.* positive prediction rate, is useful for experimental testing. When each server returns 10% of surface residues as predicted epitopic residues, the accuracy was 14.3%, 15.5%, 17.0%, 17.2%, 17.8%, 18.8%, 24.7%, and 25.6% for Ellipro, DiscoTope1.2, BEpro (PEPITO), SEPPA, EPCES, EPITOPIA, EPSVR, and EPmeta, respectively, as shown in Figure 3.2. The accuracy is around 24% for Bpredictor based on Figure 4 in the Reference (*14*). The rationale of selecting 10% surface residues to be predicted as positive is because the average length of antigen proteins is around 200 amino acids and the average size of epitopic patch is about 20 amino acid residues. The current level of accuracy of all predictors is not yet satisfactory. Even the highest accuracy, 25.6% achieved by EPmeta, leaves room for further improvement. If 3% of

surface residues are returned as predicted epitopic residues, the accuracy of EPmeta is 31.6%, which is the overall highest value by all conditions and methods.

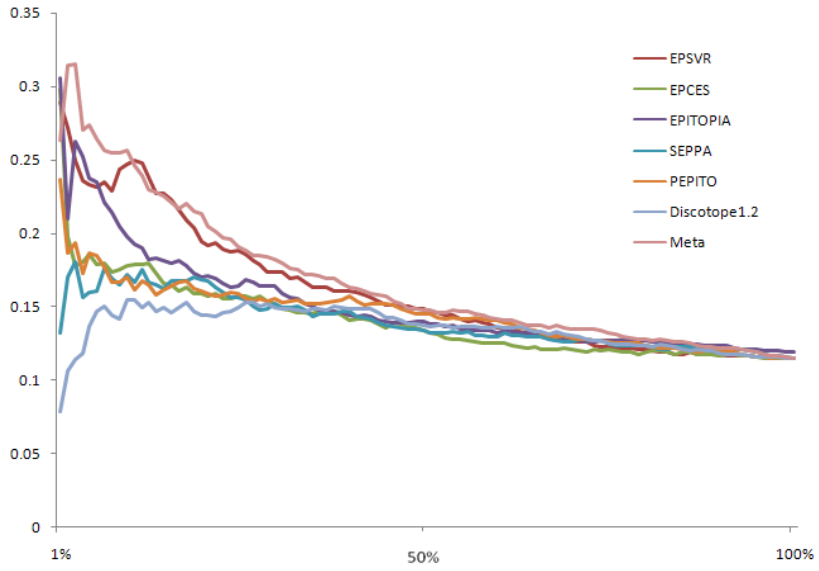

Figure 3.2 Prediction accuracy of six antigenic epitope prediction servers and Meta server on 19 independent testing proteins. Y-axis is AUC score and x-axis is the threshold of prediction score. The prediction accuracy was averaged for 19 independent testing proteins, except for EPITOPIA because it failed to assign scores for the antigenic residues of 1jeq and the prediction accuracy was averaged over the other 18 proteins.

For the EPmeta server, its architecture of using remote web servers lowers the pressure of the local meta server. The computational time is mainly based on the speed of online tools and the transferring speed throughout the internet. While with steady internet connection, communication between the meta server and online tools is no longer an issue, computational time with each online tool is the major time consuming part. Most of conformational B-cell epitope prediction tools require more than ten minutes, even half an hour, to complete one query for 200-aa sequence. For example, the average computing

time for EPITOPIA is over thirty minutes based on our tests accessing from the University of Nebraska – Lincoln. The distributed architecture of the meta server allows the parallel operations involved in multiple queries to different online tools. Therefore, unlike local implementation of different tools, we can run these queries online simultaneously and we do not have to care about the machine pressures since these online servers are independent. Distributed system has been widely applied for large-scale or intensive computation, usually shared among multiple physical machines.

This distributed architecture also lowers the risk of the EPmeta server for installation and maintenance of multiple programs. Since we did not intend to install the tools locally, proper installation was not an issue. It is also easier to track the newest version of tools by just accessing them directly online. As a result, the distributed architecture of the meta server guarantees always the good query of results from different tools.

5. Challenge of Conformational B-cell Epitope Prediction

In recent years, a number of new conformational B-cell epitope prediction algorithms have been developed. While the prediction performance has been improved, it is still far from satisfactory. Compared with other bioinformatic problems, antigenic epitope prediction is especially difficult due to the lack of properties that are universally but uniquely observed for the antigenic epitopes but not for other protein surfaces. Additionally, regular binding-site prediction methods are not suitable for antigenic epitope prediction because they focus on the conservation of surface residues.

5.1 Single Chain or Multiple Chains

The recognition of antibody to antigenic epitopes has high specificity; the epitopic surface is not as conserved as other functional protein binding sites, which comes from the conserved functions of protein-protein interactions during evolution. The interfaces of regular protein-protein binding are usually more conserved and have more hydrophobic amino acid residues than non-binding protein surfaces. This makes the exposed protein-protein interfaces relatively easy to distinguish from both the antigenic epitopes and non-binding protein surfaces. In other words, the prediction task for a single chain protein that has both protein-protein binding interfaces and an antigenic epitope is easier than that of a protein complex.

In the benchmark dataset, six of the proteins (PDB IDs: 1eku, 1av1, 1al2, 1jeq, 2gib, and 1qgt) possess multiple chains. Therefore, in our evaluation all methods were tested with two different scenarios for these six proteins: prediction on a single chain where the experimental antigenic epitope is located and prediction on the whole protein including all chains. When multiple chains were examined, all chains were considered, and the total number of surface residues was counted for the intact complex structure. As a result, some methods, such as EPSVR, showed decreased performance with lower mean AUC values for the 6 proteins when the whole protein was used for prediction compared with those based on the single chain containing the antigenic epitope. Therefore, in the future, if sufficient data exist, a range of test datasets shall be compiled for different cases, *i.e.*, single chain antigens, single chains from antigen complexes, and antigen complexes. A good antigenic epitope predictor shall have satisfying performance on all types of benchmarks.

5.2 Protein Binding Site Prediction Methods

Due to the lack of many epitope prediction methods for analysis and comparison, potein-binding site prediction methods are frequently used for conformational epitope prediction (*5, 25*) since epitopic patches can be considered as a type of protein binding sites. The methodologies used by protein binding site prediction and epitope prediction are similar; both integrate some amino acid scoring functions with a machine learning algorithm or other platform to train a prediction model on known data. The major difference is their distinct training sets; while protein binding site prediction uses all known protein-protein binding complexes, an epitope prediction method is trained with antibody-antigen complexes only. Therefore, we also applied the benchmark epitope dataset to test some binding site prediction methods. For this we selected binding-site prediction methods that have both demonstrated good performance and convenient web servers for public use. The AUCs achieved by these methods for the epitope benchmark are shown in Table 3.2. One can see that the performances of the binding-site prediction methods to predict B-cell epitopes are significantly lower than all conformational epitope prediction methods (shown in Table 3.1). This is not surprising because all binding-site prediction methods are designed based on the conservation and hydrophobicity of binding patches. B-cell epitopic patches are neither conserved nor more hydrophobic compared with other protein-protein binding surfaces. Instead, the residues on the antigenic epitopes are more diverse than regular surface residues due to the evolution pressure from the host immune system. Therefore, we conclude that the general binding-site prediction methods are not suitable for antigenic epitope prediction. Any epitope prediction methods developed in the future is not recommended to claim performance improvement by simply compared with binding-site prediction methods.

Table 3.2 List of the Protein Binding Site Prediction Methods and Their Obtained AUC Results

Method	URL of web server	AUC
ProMate (*26*)	http://bioinfo.weizmann.ac.il/promate/	0.530
ConSurf (*27*)	http://consurf.tau.ac.il/index_proteins.php	0.460[a]
PINUP (*17*)	http://sysbio.unl.edu/services/PINUP	0.562
PIER (*28*)	http://abagyan.ucsd.edu/PIER/pier.cgi?act=dataset	0.537

[a)] Conserved residues are selected as for common binding site prediction.

5.3 Future Directions

Currently, various sets of attributes and classifiers have been applied by different existing epitope prediction algorithms. It naturally leads to one question: which combination of attributes is optimal for the prediction? To answer this question, one may systematically evaluate different machine-learning algorithms on all non-redundant attributes and allocate the optimal set among them. Also of great importance to the epitope prediction research is the growth of the training data, especially the antigens that have both bounded and unbounded structures. It is also important to collect high quality independent testing data, such as the ones compiled by Liang *et al.* (*23*), that contain experimentally measured epitopic residues but no complex structures. We also recommend that all future researchers implement their developed algorithms as free accessible web servers or downloadable software packages, because B-cell epitope prediction algorithms will likely become more and more complicated and meta-methods usually have better prediction accuracy than any of the single algorithms.

REFERENCES

1. Lefkovits, I., Jerne, N. K., Steinberg, C. M., and Di Lorenzo, C. (1981) *The Immune system*, Karger, Basel ; New York.
2. Reineke, U., and Schutkowski, M. (2009) Epitope mapping protocols, 2nd ed., pp 1 online resource (xiii, 456 p., [416] p. of plates), Humana Press, New York.
3. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures, *European journal of biochemistry / FEBS 80*, 319-324.
4. Liang, S., Zheng, D., Zhang, C., and Zacharias, M. (2009) Prediction of antigenic epitopes on protein surfaces by consensus scoring, BMC bioinformatics 10, 302.
5. Ponomarenko, J. V., and Bourne, P. E. (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation, BMC structural biology 7, 64.
6. Walter, G. Production and use of antibodies against synthetic peptides, Journal of Immunol. Methods. 1986;88:149–61. doi: 10.1016/0022-1759(86)90001-32.
7. Kulkarni-Kale, U., Bhosle, S., and Kolaskar, A. S. (2005) CEP: a conformational epitope prediction server, Nucleic acids research 33, W168-171.
8. Haste Andersen, P., Nielsen, M., and Lund, O. (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures, Protein Sci 15, 2558-2567.
9. Sweredoski, M. J., and Baldi, P. (2008) PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure, Bioinformatics (Oxford, England) 24, 1459-1460.
10. Ponomarenko, J., Bui, H. H., Li, W., Fusseder, N., Bourne, P. E., Sette, A., and Peters, B. (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes, BMC bioinformatics 9, 514.
11. Sun, J., Wu, D., Xu, T., Wang, X., Xu, X., Tao, L., Li, Y. X., and Cao, Z. W. (2009) SEPPA: a computational server for spatial epitope prediction of protein antigens, Nucleic acids research 37, W612-616.
12. Rubinstein, N. D., Mayrose, I., Martz, E., and Pupko, T. (2009) Epitopia: a web-server for predicting B-cell epitopes, BMC bioinformatics 10, 287.
13. Rubinstein, N. D., Mayrose, I., and Pupko, T. (2009) A machine-learning approach for predicting B-cell epitopes, Molecular immunology 46, 840-847.
14. Zhang, W., Xiong, Y., Zhao, M., Zou, H., Ye, X., and Liu, J. Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature, BMC bioinformatics 12, 341.
15. Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J., and Weng, Z. (2005) Protein-Protein Docking Benchmark 2.0: an update, Proteins 60, 214-216.
16. Huang, J., and Honda, W. (2006) CED: a conformational epitope database, BMC immunology 7, 7.
17. Liang, S., Zhang, C., Liu, S., and Zhou, Y. (2006) Protein binding site prediction using an empirical scoring function, Nucleic acids research 34, 3698-3707.

18. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic acids research 25, 3389-3402.
19. Liang, S., and Grishin, N. V. (2004) Effective scoring function for protein sequence design, Proteins 54, 271-281.
20. Jones, S., and Thornton, J. M. (1997) Analysis of protein-protein interaction sites using surface patches, Journal of molecular biology 272, 121-132.
21. Chou, P. Y., and Fasman, G. D. (1978) Empirical predictions of protein conformation, Annual review of biochemistry 47, 251-276.
22. Joachims, T. (1999) Making Large-Scale SVM Learning Practical, Advances in Kernel Methods - Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed.), MIT-Press.
23. Liang, S., Zheng, D., Standley, D. M., Yao, B., Zacharias, M., and Zhang, C. EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results, BMC bioinformatics 11, 381.
24. Davis, J., and Goadrich, M. (2006) The Relationship Between Precision-Recall and ROC Curves, 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, USA, 26th - 28th June.
25. El-Manzalawy, Y., and Honavar, V. Recent advances in B-cell epitope prediction methods, Immunome research 6 Suppl 2, S2.
26. Neuvirth, H., Raz, R., and Schreiber, G. (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites, Journal of molecular biology 338, 181-199.
27. Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids, Nucleic acids research 38, W529-533.
28. Kufareva, I., Budagyan, L., Raush, E., Totrov, M., and Abagyan, R. (2007) PIER: protein interface recognition for structural proteomics, Proteins 67, 400-417.

CHAPTER FOUR: PREDICTION OF EPITOPIC RESIDUES WITH PROTEIN SEQUENCES

1. Introduction

In the previous chapters, we discussed the two kinds of B-cell epitopes: linear and conformational epitopes. It was reported that about 90% B-cell epitopes are conformational (*1*). Therefore, we developed prediction tools for conformational epitopes based on known 3D structures of given antigens. However, the small number of solved structures of antigens limits the application of our epitope prediction. In this chapter, we will tackle this difficulty by developing a new method to analyze antigen protein sequences to predict epitopic residues. Currently the study on protein-sequence-level prediction of epitopic residues of B-cell epitope is still under development. One possible reason is a limited resource of known epitopic residues of B-cell epitope. Another reason lies in the complexity of epitope binding patterns. At present, to our knowledge, there are a very few epitope-antibody binding patterns reported. So it is still difficult to extract the principles of key residues from known epitope-antibody binding patterns. Due to these facts, we must find a new way to predict the key residues of B-cell epitopes.

In chapters two and three, machine learning methods have been proven powerful in terms of classification. By the analysis of known B-cell epitopes and non-epitopes, a machine learning method draws a border to separate epitopes and non-epitopes. In this chapter, we will apply the Support Vector Machine (SVM) as a tool of machine learning to predict epitopic residues based on input of protein sequences.

To set up a training process, the first step is to collect known epitopic residues of B-cell epitopes. The immune epitope database (IEDB) contains the updated information

of B-cell epitopes most of which are still hypothetical (*2*). For our study, we employ a more reliable way to collect key epitopic residues of antigens. In our study only 3D structures of antigen-antibody complexes precisely showing the residues of epitope directly bound to antibody are used as the data resource. To construct the positive dataset, we have investigated 561 antigen-antibody complexes whose antigens come from 11 species. For these antigen sequences, we filtered them by a threshold of 30% similarity to generate an unbiased dataset. After the similarity filtering, we extracted 2682 key residues which come from 134 unique antigens. The details about building the positive dataset containing known key epitopic residues are described in section 2.1.

For each epitopic residue, the subsequence surrounding it with a certain length is extracted from the original antigen, and the sequence pattern and physicochemical characteristics of any given subsequence are quantified for epitopic residue prediction. The sequence pattern and physicochemical properties of these segments include Shannon Entropy (SE), Relative Entropy (RE), Position Specific Scoring Matrix (PSSM), Predicted Secondary Structure (SS), Protein Disorder (DIS), Solvent Accessible Area, Overlapping Properties (OP), Sequence Complexity (SC), and Averaged Cumulative Hydrophobicity (ACH). They are considered and optimized as the segment features for SVM training in our study. Similar strategy and features can also be applied to the point residue prediction, such as the prediction of phosphorylation sites (*3*, *4*).

In this study, we developed a novel tool, named SVMKER, to predict the epitopic residues with an input of a protein sequence only. With carefully designed feature characterization of the sequence segments, SVMKER has shown a precision of 59.5% and a sensitivity of 52.2% using five-fold cross-validation process. It is the first time that

a prediction tool has been developed for epitopic residue prediction at the protein sequence level. This tool can provide a preliminary search for epitopic residues on antigens before an experimental design.

2. Materials and Methods

2.1 Datasets

The original information about epitopic residues of B-cell epitopes was extracted from PDB (*5*). As of June 2012, we downloaded 561 3D structures of antigen-antibody complexes involved in 11 species. From these 3D structures, we extracted 134 low-similarity antigen sequences with a threshold of 30% similarity. An amino acid residue of the antigen is considered as one epitopic residue if it has at least one atom that has less than 6 Å distance to any atom of antibody (*6*). The criterion is illustrated with an example in Figure 4.1 below.

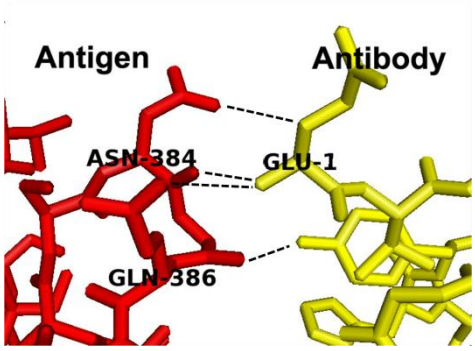

Figure 4.1 The distance of atom-atom distance between antigen and antibody. The epitopic residue on an antigen must contain at least one atom with less than 6 Å distance away from an antibody's atom. For example, in the 3D structure of human IgM rheumatoid factor Fab bound to its autoantigen IgG Fc (PDB ID: 1ADQ) (*7*), the amino acid residue ASN-384 is considered as an epitopic residue because it contains two atoms less than 6 Å from the antibody. The same happens to GLN-386.

After identifying all known epitopic residues from the 3D structures of the antigen-antibody complexes, we created the positive datasets containing 2682 protein sequence segments with epitopic residues in the middle positions. The segment length includes 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, and 23 amino acids (AA). Figure 4.2 shows protein sequence segments with different sizes. The negative dataset was constructed by including the non-redundant segments with the same length of positive dataset segments from the sections of antigen sequences where no epitopic residues exist. The total numbers of positive and negative segments are 2682 each; the positive and negative segment ratio is 1:1.

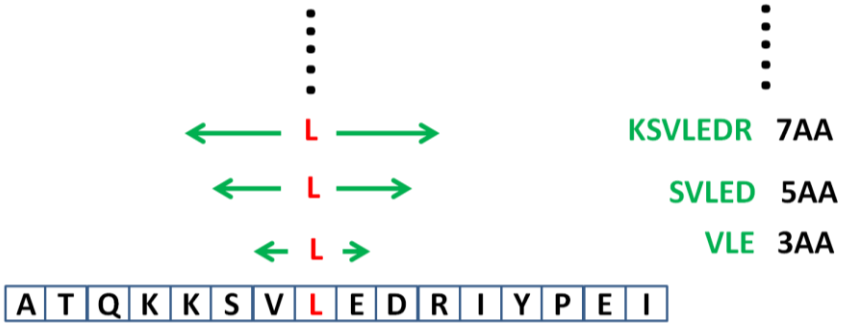

Figure 4.2 A segment in the positive dataset determined by extending the key epitopic residue in both sides along the antigen sequence.

2.2 Attributes

For a machine-learning technology, we need to quantify the sequence pattern and physico-chemical properties of these protein segments in the training set. We used multiple attributes including Shannon Entropy (SE), Relative Entropy (RE), Position Specific Scoring Matrix (PSSM), Predicted Secondary Structure (SS), Protein Disorder (DIS), Solvent Accessible Area, Overlapping Properties (OP), Sequence Complexity

(SC), and Averaged Cumulative Hydrophobicity (ACH). The details of calculation for these attributes are described in the following sections.

2.2.1 Shannon Entropy (SE)

SE score, a widely used sequence conservation measure, is calculated by weighted observed percentages (WOP) extracted from the results of PSI-BLAST (*8*) with Non-redundant protein sequence database. The WOP vector for a position in a given protein sequence shows the position-specific distribution of 20 amino acids. The SE score for the given position is defined as:

$$SE = -\sum_{i=1}^{20} p_i \log(p_i), \quad (1)$$

where $p_i = a_i / \Sigma a_j$, a_j is the j-th value in the WOP vector for this given position. If a position has complete conservation, the SE score has the smallest value, 0.

2.2.2 Relative Entropy (RE)

RE measures the amino acid background distribution, and also requires the WOP matrix. The RE score of one type of amino acids is calculated as:

$$RE = \sum_{i=1}^{20} p_i \log\left(\frac{p_i}{p_0}\right), \quad (2)$$

where $p_i = a_i / \Sigma a_j$, a_j is the j-th value in the WOP vector for this given position and p_0 is the protein BLOSUM62 background distribution.

2.2.3 Position Specific Scoring Matrix (PSSM)

PSSM is commonly used for the representation of motifs or patterns in biological sequences. To identify the pattern of neighbors besides key epitopic residues, we calculated the PSSM matrix using the BLASTP tool. PSSM can provide 20 bits for vector features.

2.2.4 Secondary Structure (SS)

The most accurate way to obtain the information of secondary structures would be from the 3D structures of proteins. However, for a given protein sequence, the secondary structures can only be predicted. In this chapter, the SS attribute of each residue has three bits to show the possibility scores of three types of secondary structures (H: helix, E: β-sheet, and C: coiled coil) which is predicted by PSIPRED (*9*).

2.2.5 Protein Disorder (PD)

PD is important for protein function. Previous works suggest that protein disorder information is helpful to improve the discrimination between active sites and non-active sites (*10*). In our study, protein disorder areas are predicted by DISOPRED (*11*). The prediction result provides a score for each residue between 0 and 1, corresponding to more structured to more disordered status.

2.2.6 Accessible Surface Area (ASA)

All epitopic sites are on the surface of an antigen, and hence, large solvent accessibility is also an important feature of the catalytic residues. To improve the prediction accuracy, the solvent ASA information of each residue is included into the algorithm as well. The ASA attribute needs to be predicted from protein sequences. In our study, RVP-net is used to predict the relative solvent ASA for each residue in a given protein sequence (*12*). Each residue has a real value in (0, 1) for the ASA attribute.

2.2.7 Averaged Cumulative Hydrophobicity (ACH)

ACH has been demonstrated to be an important attribute for protein functional residues. The attribute is quantified by computing the average of the cumulative hydrophobicity indices over the segment sizes of 3, 5, 7, ... , 21, and 23AA. There are

ACH scores for 10 bits in the feature vector. The hydrophobicity index proposed by Sweet and Eisenberg is used in this chapter (*13*).

2.2.8 K-nearest neighbor profiles (KNN)

KNN usually is applied in the prediction of active sites, such as phosphorylation sites (*4*). A KNN score for one given sequence is the proportion of positive key epitopic residue in its k nearest neighbors in the training set where the distance between the two sequences is proportional to their sequence similarity; a pair of similar sequences has a short distance. The parameter k of KNN is set as 0.25%, 0.5%, …, and 5.and the KNN profile attribute has 20 bits.

2.3 Training and Five-fold Cross Validation

The training process is based on the Support Vector Machine tool, SVM^{light} (*14*). The kernel function used is the radial basis function, $\exp(-\gamma\|a-b\|^2)$. To obtain the optimal training performance, we did a grid searching in the range of ($c=2^{-10\sim1}$, $g=10^{-8\sim0}$, and $p=2^{-5\sim0}$) where c is the trade-off between training error and margin, g is the parameter γ in the radial basis function kernel, and p is the fraction of unlabeled examples to be classified into the positive class. For each segment length of the training dataset, the grid searching was independently executed and completed for different optimal parameter sets.

Five-fold cross-validation is executed for SVM training in the absence of an independent testing set. We first split the training dataset into five subgroups with the same size. Each group contains the same number of positive segments (epitopic residue inside) and negative segments (non-epitopic residue) as any other fold. During the procedure of making five groups, we make sure that a segment has more similarity with sequences in the same group than one from other group. The segments with more

sequence similarity are grouped in the same fold, which can significantly reduce the potential bias in the validation. During the five-fold cross-validation process, in turn the four folds are used to calculate the model and the last one is to evaluate the accuracy of the model calculated.

The related statistical evaluation is listed below,

$$\text{Sen} = \frac{\text{TP}}{\text{TP+FN}} \times 100\%$$

$$\text{Pre} = \frac{\text{TP}}{\text{TP+FP}} \times 100\%$$

$$\text{F} = \frac{2 \times \text{Pre} \times \text{Sen}}{\text{Pre+Sen}},$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively. All of calculations above are based on five-fold cross-validation procedure. We also generated the receiver operating characteristic (ROC) curve for statistical evaluation of SVMKER. In the ROC curve, false positive rate, *i.e.* FPR = FP / (FP + TN), is x-axis while sensitivity (Sen or true positive rate, as shown above), is y-axis. Area under the curve (AUC) has been widely accepted as a performance index, with a higher AUC score representing a higher prediction performance. A java program available at http://pages.cs.wisc.edu/~richm/programs/AUC/ was used to calculate the AUC (*15*).

3. Results

The length of segments extending from epitopic residue will greatly impact the prediction. The shorter lengths, such as 3, 5, and 7AA, are difficult to generate a stable prediction performance. In current version of SVMKER, the eight attributes were applied, and they are SE (1 bit), RE (1 bit), PSSM (20 bits), SS (3 bits), ASA (1 bit), PD (2 bits),

ACH (10 bits), and KNN (20 bits). Using 3AA as example, the length of a feature vector for a given sequence segment is $7 \times (1+1+20+3+1+2+10+20) = 174$. The small numbers of features for training might be the major reason to lower the prediction performance for 3, 5, and 7AA. Under current conditions, 19AA is the most accommodative option, and hence, 19AA is used for the actual search of unknown proteins.

Each segment lengths, *i.e.*, 3AA, 5AA, 7AA, …, 19, 21, and 23 AA, showed the specific optimized parameter sets (Table 4.1). With the five-fold cross-validation, the statistical evaluation was obtained and it is also listed in Table 4.1. The predictions for 3, 5, or 7AA length show very low performance. The reason may be the fact that short segments enclosing epitopic residue hardly provide enough information for classification. The best prediction is based on 19AA segment training dataset, and 17AA yields close prediction performance based on the F-measure. Furthermore, with longer segments as 21AA and 23AA, the statistical evaluation showed worse prediction performance than 19AA and 17AA (shown in Table 4.1). The influence of segment lengths on the prediction of key epitopic residue is similar to what we observed in sliding-window lengths on linear B-cell epitopes. For our previous tool, SVMTriP, the prediction on 20AA is better than that on the shorter epitope lengths (*16*).

Table 4.1 Statistical Evaluation of SVMKER with Different Lengths

Segment	c	g	p	Sen	Pre	F
3AA	0.25	0.01	0.0625	0.240	0.275	0.256
5AA	0.25	0.001	0.125	0.232	0.270	0.250
7AA	0.25	0.001	0.125	0.267	0.315	0.289
9AA	0.125	0.0001	0.25	0.335	0.389	0.360
11AA	0.25	0.001	0.125	0.387	0.445	0.413
13AA	0.125	0.0001	0.125	0.468	0.434	0.450
15AA	0.125	0.0001	0.25	0.522	0.483	0.502

17AA	0.125	0.001	0.25	0.572	0.535	0.553
19AA	**0.125**	**0.001**	**0.25**	**0.595**	**0.522**	**0.556**
21AA	0.125	0.0001	0.125	0.469	0.481	0.475
23AA	0.125	0.0001	0.25	0.442	0.493	0.466

We also compared SVMKER with our two conformational B-cell epitope prediction tools, EPCES (*17*) and EPSVR (*18*). The ROC curves of the three tools are shown in Figure 4.3 using a length of 19AA. EPCES shows the best prediction performance, EPSVR the second, and SVMKER the third. The AUC scores for EPCES, EPSVR and SVMKER are 0.632, 0.582, and 0.549, respectively. The fact that EPCES received a higher AUC score may be from different test dataset usage. It is in accordance with the facts that both EPCES and EPSVR are 3D structure-level tools while SVMKER is based only on protein sequences. All structural level information of a given antigen is known for EPCES and EPSVR. For EPCES and EPSVR, only surface residues are considered for prediction. However, SVMKER needs to predict epitopic residues from all amino acids of a given antigen.

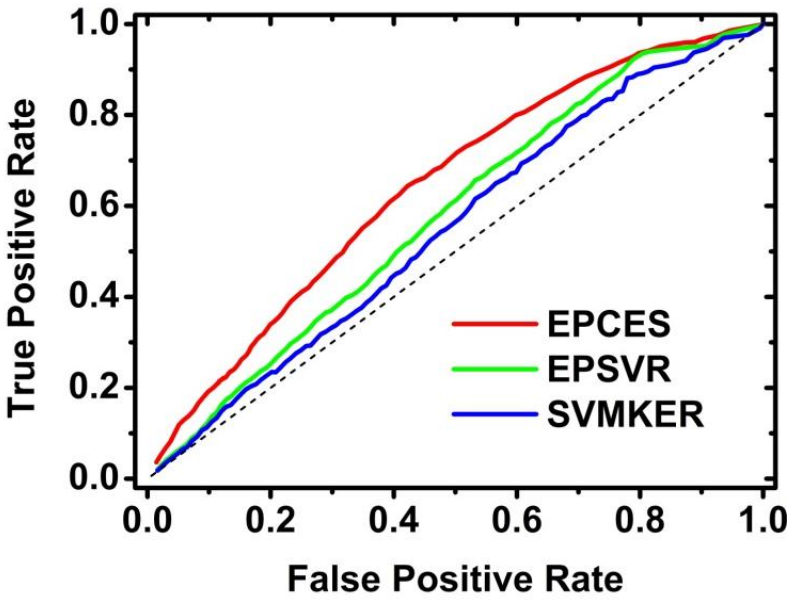

Figure 4.3 ROC curves for EPCES, EPSVR, and SVMKER throughout five-fold cross-validation. The area under each ROC curve (AUC) is 0.632, 0.582, and 0.549, respectively.

4. Discussions

Compared with conformational B-cell epitope tools, SVMKER may possess more potential for the study of epitopes. For example, although EPCES and EPSVR have better performance than SVMKER, both require 3D structures of antigens, which greatly limit their applicability. After all, only a very small proportion of antigens have solved 3D structures. The determination of protein 3D structures using X-ray diffraction or nuclear magnetic resonance spectroscopy is time- and fund-consuming. SVMKER requires only a simple input, *i.e.*, the amino acid sequence of protein candidate, and therefore, it is more practical for the immunologists. SVMKER has much wider applicability without the knowledge of 3D structure of protein candidates. Moreover, our comparison shows that the performance of SVMKER is close to that of EPSVR. Therefore, in despite of the room for accuracy improvement, as a sequence-level tool, SVMKER may be very useful in epitopic residue search as a preliminary filter for subsequent experimental design.

SVMKER may not have yet reached the most optimal model. One of the reasons is the optimization of attributes. The eight attributes mentioned above are usually applied to predicting binding patterns and motifs. Note that some attributes are dependent to others. For example, PSSM is often a pre-condition considered in the prediction tool of protein secondary structure (*9*). PSSM and SS thus require to be treated as associated indices if they are both considered in SVMKER. Another issue is the normalization of

multiple attributes. In the current version of SVMKER, all the attributes have been normalized into the range of [0, 1]. To improve SVMKER, we will add more attributes into further consideration. The dependence and normalization of these attributes must be kept in mind.

We calculated the weight scores of eight attributes in the 19AA optimized model. The weight scores are calculated by the formula $w = \sum \alpha_i x_i$. Here α is dual representation of the decision boundary; and x_i (i=0, 1, 2…n) is vector described in the SVM model. Both α_i and x_i are available in the model file. After normalizing weight scores of eight attributes, we illustrated their weight scores in the optimal model in Figure 4.4. Among these eight attributes, PSSM has the largest weight score (0.218) while PD is the least (0.049). That means that PD (the protein disorder feature) only weakly contributes to the determination of the SVMKER optimized model. The reason could be because of the low performance of the protein disorder prediction tool that we used. Considering the high weight scores of PSSM, SS, and KNN, the sequence similarity is still the major factor to determine the boundary between epitopic residue and non-epitopic residue.

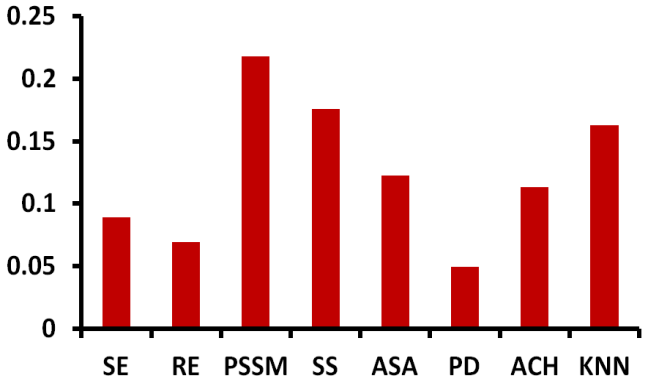

Figure 4.4 The Weight Scores of Eight attributes in 19AA optimal model.

Another way to improve SVMKER comes from the update of the training dataset. In the current version of SVMKER, we only applied the 3D structures of the antigen-antibody complex deposited in the PDB database by June 2012. The 3D protein structures of antigen-antibody complex may precisely provide epitopic residues but the limited number of available 3D structures constraints the size of the training dataset. More resource should come from the IEDB database. By the end of 2013, there are more than 50K B-cell epitopes reported in IEDB database. For most epitope entries, the corresponding binding sites or epitopic residues are recorded or suggested. Some of epitopic residues were determined by the determination of protein structure or point mutation experiments. However, most of epitopic residues in IEDB are still hypothetical from prediction tool and not suitable for training dataset. Obviously, our model cannot be simply based on these hypothetical epitopic residues since they are predicted and contains many false positives. It is necessary to carefully check and filter by excluding those hypothetical epitopic residues in IEDB before they are added to the training dataset. We believe that the SVMKER performance will be improved when more known epitopic residues are used to optimize the model.

5. Conclusions

In this chapter, we developed a new tool, SVMKER, to predict epitopic residues in antigens. The determination of epitopic residues greatly benefits the application of B-cell epitope, such as vaccine design. It can act as the pre-condition for point mutation experiments on the validation of antigen. The current version of SVMKER reaches a precision of 59.5% and a sensitivity of 52.2% using five-fold cross-validation. In further evaluation, we compared SVMKER and EPCES and EPSVR. The AUC score of the

three tools is 0.549, 0.632, and 0.582, respectively. SVMKER shows a close prediction performance with EPSVR. Considering its significant advantage as a sequence-level tool, SVMKER meets much broader needs than EPCES and EPSVR, which are structure-level tools. We will keep improving SVMKER through further optimization of the attributes and the continuous enrichment of the training dataset using the IEDB database. All of optimal models, datasets, and online tool will be released for public usage in the future.

References

1. Walter, G. Production and use of antibodies against synthetic peptides, Journal of Immunol. Methods. 1986;88:149–61. doi: 10.1016/0022-1759(86)90001-32.
2. Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A., and Peters, B. (2010) The immune epitope database 2.0, Nucleic acids research 38, D854-862.
3. Gao, J., Thelen, J. J., Dunker, A. K., and Xu, D. (2010) Musite, a tool for global prediction of general and kinase-specific phosphorylation sites, Mol Cell Proteomics 9, 2586-2600.
4. Dou, Y., Yao, B., and Zhang, C. (2014) PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine, Amino acids.
5. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures, European journal of biochemistry / FEBS 80, 319-324.
6. Nussinov, R., and Schreiber, G. (2009) Comformational protein-protein interactions, CRC Press, Boca Raton, FL.
7. Corper, A. L., Sohi, M. K., Bonagura, V. R., Steinitz, M., Jefferis, R., Feinstein, A., Beale, D., Taussig, M. J., and Sutton, B. J. (1997) Structure of human IgM rheumatoid factor Fab bound to its autoantigen IgG Fc reveals a novel topology of antibody-antigen interaction, Nature structural biology 4, 374-381.
8. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic acids research 25, 3389-3402.
9. McGuffin, L. J., Bryson, K., and Jones, D. T. (2000) The PSIPRED protein structure prediction server, Bioinformatics (Oxford, England) 16, 404-405.
10. Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., and Dunker, A. K. (2004) The importance of intrinsic disorder for protein phosphorylation, Nucleic acids research 32, 1037-1049.
11. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, Journal of molecular biology 337, 635-645.
12. Ahmad, S., Gromiha, M. M., and Sarai, A. (2003) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences, Bioinformatics (Oxford, England) 19, 1849-1851.
13. Sweet, R. M., and Eisenberg, D. (1983) Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure, Journal of molecular biology 171, 479-488.
14. Joachims, T. (1999) Making Large-Scale SVM Learning Practical, Advances in Kernel Methods - Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed.), MIT-Press.
15. DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44, 837-845.

16. Yao, B., Zhang, L., Liang, S., and Zhang, C. (2012) SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity, PloS one 7, e45152.
17. Liang, S., Zheng, D., Zhang, C., and Zacharias, M. (2009) Prediction of antigenic epitopes on protein surfaces by consensus scoring, BMC bioinformatics 10, 302.
18. Liang, S., Zheng, D., Standley, D. M., Yao, B., Zacharias, M., and Zhang, C. EPSVR and EPmeta: prediction of antigenic epitopes using support vector regression and multiple server results, BMC bioinformatics 11, 381.

CHAPTER FIVE: SUMMARY AND FUTURE WORK

1. Summary

In this dissertation, we investigated several computational methods to identify the antigenic epitopes of protein candidate. First, a new linear B-cell epitope prediction tool SVMTriP was developed based on the combination of tri-peptide similarity and their propensity (*1*). Being compared with other linear B-cell epitope prediction tools, such as BCPred (*2*) and AAP (*3*), SVMTriP showed a better prediction performance with AUC score as 0.702 (BCPred: 0.667; AAP:0.667). Then, we developed a conformational B-cell epitope prediction tool called EPSVR and a meta server, EPmeta (*4*). EPSVR was developed based on six attributes with Super Vector Regression. EPmeta integrated multiple single servers, such as DiscoTope (*5*), PEPITO (*6*), SEPPA (*7*), EPITOPIA (*8*), EPCES (*9*), and EPSVR (*4*), to find a consensus result. The statistical evaluation based on an independent test dataset showed that EPSVR has the AUC score as 0.597 and EPmeta as 0.638. The third tool that we developed is SVMKER that can predict epitopic residues of an antigen sequence. To our knowledge, SVMKER currently is the first prediction tool of epitopic residues using protein sequence as input. These tools we developed provide more choices for immunologists to identify the antigenicity of protein candidate in a quick and cheap way.

1.1 Linear Epitope Prediction

Currently, Users can access SVMTriP by http://sysbio.unl.edu/SVMTriP. Generally, it takes 20-30 minutes to complete the prediction process for a protein sequence with a length of 200 AA. If any linear B-cell epitopes are found, they will be

listed with their scores and locations on the original protein sequence. Users can select different epitope lengths, from 10AA to 20AA, for their specific cases. In addition, the training dataset collected from IEDB (*10*) in SVMTriP is also shared online (http://sysbio.unl.edu/SVMTriP/Download). These dataset may benefit other B-cell prediction groups as references to develop new tools.

Since May 12th 2012 when the online server was set up, SVMTriP has been visited more than 17,000 times, and more than 26,000 jobs were submitted for prediction. Moreover, we also helped other research groups to predict 12205 protein candidates by offline operation of SVMTriP.

The application of SVMTriP helps immunologists to quickly narrow the range of protein candidates. A real case was reported from Dr. Yurij Innov's group, Department of Cancer Genetics, Roswell Park Cancer Institute, NY (private communications). In their investigation, two non-redundant linear epitopes on the human PAP protein were discovered by their experiment. One of these two linear epitopes was successfully predicted by SVMTriP. Other applications of SVMTriP were reported as well, such as meta-analysis of IgE-binding allergen epitopes (*11*) and prediction of IL4 Inducing Peptides (*12*).

1.2 Conformational Epitope Prediction

We also constructed the online servers for EPSVR (http://sysbio.unl.edu/EPSVR) and EPmeta (http://sysbio.unl.edu/EPmeta) in 2010. EPSVR requires an input of 3D structure of protein candidate when submitting a query. The average running time for EPSVR is 10-20 minutes. Same as EPSVR, EPmeta requires 3D structure of protein

when opening a new query. Usually, EPmeta requires much longer time that EPSVR because it needs to get the prediction results from all single servers, and then integrate into a consensus result. The statistics showed the average running time for a job is 1.5-2 hours.

The development of EPSVR and EPmeta give rise of new members for the very limited set of available conformational B-cell epitope prediction tools. EPSVR has been applied to the epitope prediction by many other groups. For example, a read case came from the identification of epitopes on D8 antigen (*13*). The prediction of D8 antigen using EPSVR generated 13 potential epitopic residues. After experimental validation by site-directed mutations, 6 of 13 variants indeed showed a significant drop in antibody-antigen interaction (*13*).

More antigen-antibody complex structures may increase the prediction performance for EPSVR. When we released EPSVR in 2010, only 98 antibody-antigen complex 3D structures were involved into the training of model. In the past four years, more and more antibody-antigen complex 3D structures were released. According to IEDB records, till June 2014, there were 591 antibody-antigen complex 3D structures reported in PDB database (*14*). Hence, it is possible to update the training dataset for EPSVR.

1.3 Epitopic Residue Prediction

Currently, the SVMKER approaches our preliminary requirement but still needs a further improvement. SVMKER reached a precision of 59.5% and a sensitivity of 52.2%. The five-fold cross-validation process showed the AUC value of SVMKER was 0.549,

which was lower than EPCES (0.632) and EPSVR (0.582). Considering the input of SVMKER was protein residue sequence while that of EPCES and EPSVR was protein 3D structure, SVMKER is more valuable for real studies because most proteins do not have known 3D structures yet.

The improvement of SVMKER is under our consideration. A possible way is to find more known epitopic residues and then increase the size of training dataset. Another strategy is to optimize the vector features of SVMKER model. By determining optimal weights of vector features, SVMKER is expected to show a higher AUC score.

2. Future Work

2.1 Importance of Food Allergen Prediction

We will extend the epitope prediction tools that we developed to food allergen prediction. In chapter one, various types of B-cell antibodies, including IgE, which is mainly involved in allergy (*15*), were described. For atopic people, a specific allergen elicits T helper lymphocyte type 2 (Th2) responses (*16*), and leads to synthesis of allergen-specific IgE antibodies, which bind to mast cells and basophils. Further exposure to the antigen may stimulate the release of vasoactive mediators such as histamine and leukotrienes that cause the symptoms of allergy (*17, 18*). An IgE can recognize and bind to the allergen, which leads to a series of allergy-related symptoms. IgEs can bind to allergens and then stimulate the release of chemical materials, such as histamine and cytokinase, from the mast cells (*19*). Hence, allergen should contain at least one IgE binding site, *i.e.*, epitope.

Among the variety of allergies troubling human beings, food allergy is one of the most prevalent diseases affecting almost all races in all geographic locations. Food allergy is a non-protective immune response triggered by food(s) specific to the affected individual, such as eggs, cow's milk, peanuts, certain tree nuts, soybeans, wheat, and other complex foods. Allergy is estimated to occur in over 20% of the population in industrialized countries, primarily as airway allergy, but with between 2-4% of adults and 6-8% of children experiencing allergic symptoms following exposure to specific foods (*20*).

The risk of food allergy is increasing synchronously due to the quick development and application of transgenesis in agriculture. Transgenesis can introduce one or more exogenous genes, so called transgenes, into a living target organism, such as agricultural crop or other interesting species. The transgenesis process means new proteins are expressed in the target organism which potentially increases the risk of food allergen.

Our study on B-cell epitope prediction stimulates our interests on the prediction of food allergens. Considering the relationship between regular antigen dominant epitopes and allergens, we proposed to apply our developed tools, such as SVMTriP, EPSVR, and EPmeta to the prediction of allergens. In the following sections, we will explain our primary research to assess the probability of an unknown protein as an allergen.

2.2 Design of the Food Allergen Prediction Pipeline

The allergen prediction pipeline includes sequence similarity searching, protein structure modeling, sequence-level epitope predication, and 3D structure-level epitope prediction. Figure 5.1 shows the flowchart of the pipeline.

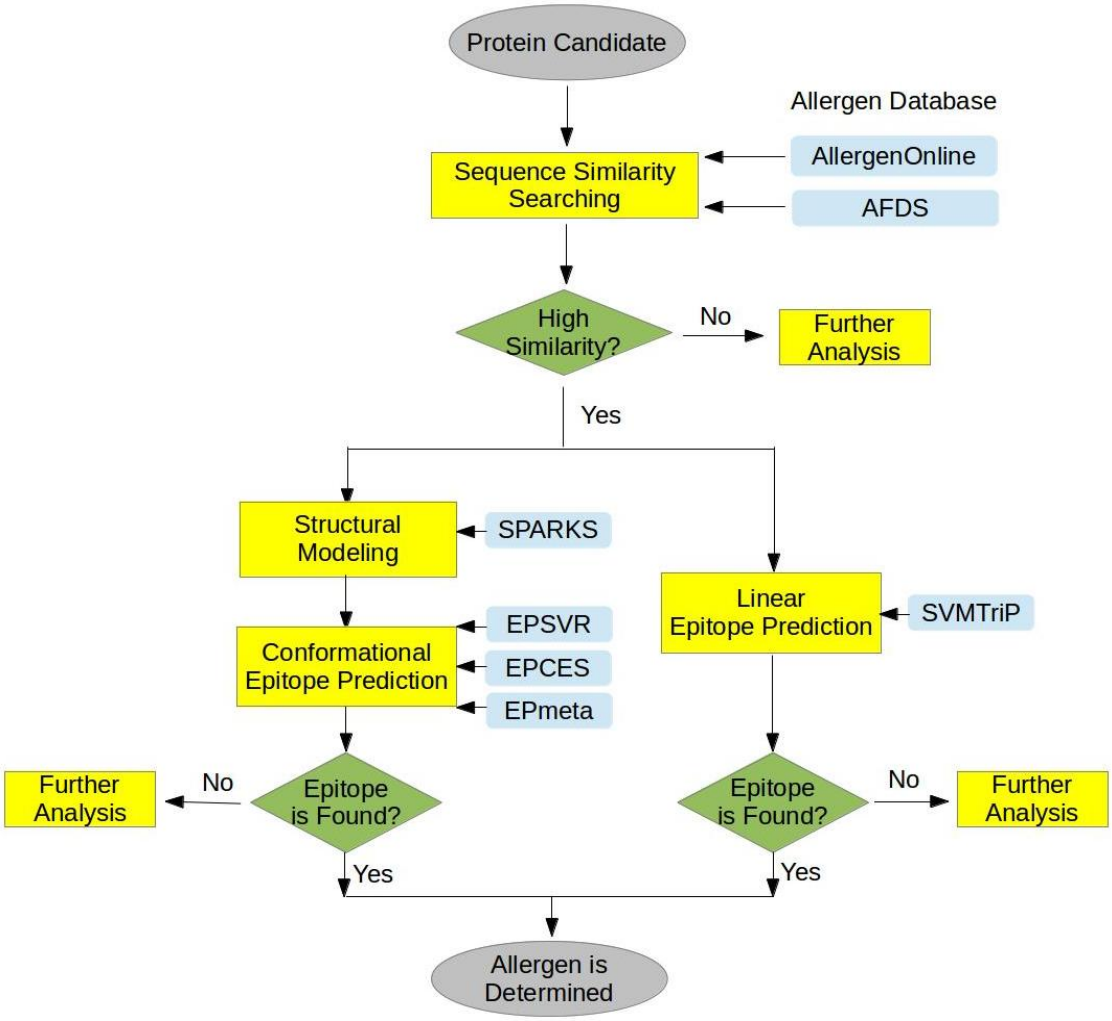

Figure 5.1: Allergen prediction pipeline using epitope tools

The initial sequence similarity searching is done to broadly search for the allergen protein candidates including those with even low probabilities to be allergens. The allergen databases, such as AllergenOnline.org and AFDS, will be used in our pipeline for this step. For the result of sequence similarity search, the criteria include: 1) moderately long stretches of amino acid (80 amino acids) with a minimum of 35% identity and 2) very short stretches (6-8 amino acids) of 100% identity. The search result matching either of the two criteria will be treated as a potential allergen and therefore

qualified for the next check. In our preliminarily study, these initial hits from sequence similarity searching possibly contained rather many false positives, which must be eliminated by further refinement as discussed in the next sections.

If protein candidate has no available 3D structure, we will apply linear B-cell epitope prediction tool such as SVMTriP and SVMKER. By SVMTriP, sliding-window search along the protein sequence is executed to determine the epitopic probability. With the pre-defined cutoff (the default cutoff is 0.2), once the epitopes are picked up, these potential antibody-binding sites, usually with 8-20 amino acids, will have another round of similarity search against allergen databases, such as AllergenOnline.org and AFDS. If the high similarity hits (>80%) find in allergen databases match the epitopes from SVMTriP, the protein candidate is thought of one allergen with a high confidence.

If the 3D structure of protein candidate is available, conformational epitope prediction tools such as EPSVR, EPCES, and EPmeta can be used. If the protein structure is not available, we will conduct the structural modeling to predict the protein structure with SPARKS-X, a tool to predict 3D protein structures (*28, 30*). After the 3D structure is predicted, conformational B-cell epitope tools will continue the rest of prediction processing.

2.3 Summary

Our developed B-cell epitope prediction tools are the good supplement for the prediction of allergens. Considering the very limited resource of known allergens and even more limited knowledge regarding IgE-binding sites on these allergens, it is quite difficult to apply any machine learning technologies to the classification and

identification of IgE binding site on proteins. Currently the most common strategy still relies heavily on the traditional sequence similarity search using FASTA and BLASTP tools. However, the results from FASTA and BLASTP contain an unsatisfactory level of false positive rates. In the pipeline we designed, we apply the similarity search method as well as epitope prediction methods, both linear and conformational, to predict the potential epitopic sites on a protein candidate. The discovery of antibody epitopes will increase the confidence level for the allergen prediction. The candidates with low scores that do not pass the checking point of epitopic sites are eliminated to decrease the false positive rate from the sequence similarity search.

References

1. Yao, B., Zhang, L., Liang, S., and Zhang, C. (2012) SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity, PloS one 7, e45152.
2. El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2008) Predicting linear B-cell epitopes using string kernels, J Mol Recognit 21, 243-255.
3. Chen, J., Liu, H., Yang, J., and Chou, K. C. (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale, Amino acids 33, 423-428.
4. Liang, S., Zheng, D., Standley, D. M., Yao, B., Zacharias, M., and Zhang, C. EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results, BMC bioinformatics 11, 381.
5. Haste Andersen, P., Nielsen, M., and Lund, O. (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures, Protein Sci 15, 2558-2567.
6. Sweredoski, M. J., and Baldi, P. (2008) PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure, Bioinformatics (Oxford, England) 24, 1459-1460.
7. Sun, J., Wu, D., Xu, T., Wang, X., Xu, X., Tao, L., Li, Y. X., and Cao, Z. W. (2009) SEPPA: a computational server for spatial epitope prediction of protein antigens, Nucleic acids research 37, W612-616.
8. Rubinstein, N. D., Mayrose, I., Martz, E., and Pupko, T. (2009) Epitopia: a web-server for predicting B-cell epitopes, BMC bioinformatics 10, 287.
9. Liang, S., Zheng, D., Zhang, C., and Zacharias, M. (2009) Prediction of antigenic epitopes on protein surfaces by consensus scoring, BMC bioinformatics 10, 302.
10. Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A., and Peters, B. The immune epitope database 2.0, Nucleic acids research 38, D854-862.
11. Lollier, V., Papini, S.D., Brossard, C., and Tessoer, D. (2014) Meta-analysis of IgE-binding allergen epitopes, Cinical Immunology 153, 31-39.
12. Dhanda, S.K., Gupta, S., Vir, P., and Raghava, G.P.S. (2013) Prediction of IL4 inducing peptides, Clinical and Developmental Immunology 2013.
13. Culang, I.S., Benhnia, M., Matho, M., Kaever, T., Maybeno, M., Schlossman, A., Nimrod, G., Li, S., Xiang, Y., Zajonic D., Crotty S., Ofran, Y., and Peters, B. (2014) Using a combined computational-experimental approach to predict antibody-specific B cell epitopes, Structure 22, 646-657
14. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures, European journal of biochemistry / FEBS 80, 319-324.
15. Schindler, L. W., National Cancer Institute (U.S.), and National Institute of Allergy and Infectious Diseases (U.S.) (1993) *The immune system : how it works*, Rev. Dec. 1993. ed., U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health, [Bethesda, Md.?].

16. Kapsenberg, M. L., Jansen, H. M., Bos, J. D., and Wierenga, E. A. (1992) Role of type 1 and type 2 T helper cells in allergic diseases, *Current opinion in immunology 4*, 788-793.
17. Metzger, H. (1991) The high affinity receptor for IgE on mast cells, *Clin Exp Allergy 21*, 269-279.
18. Nadler, M. J., Matthews, S. A., Turner, H., and Kinet, J. P. (2000) Signal transduction by the high-affinity immunoglobulin E receptor Fc epsilon RI: coupling form to function, *Advances in immunology 76*, 325-355.
19. Gould, H. J., Sutton, B. J., Beavil, A. J., Beavil, R. L., McCloskey, N., Coker, H. A., Fear, D., and Smurthwaite, L. (2003) The biology of IGE and the basis of allergic disease, *Annual review of immunology 21*, 579-628.
20. Sicherer, S. H., and Sampson, H. A. (2010) Food allergy, *The Journal of allergy and clinical immunology 125*, S116-125.
21. Goodman, R. E. (2008) Performing IgE serum testing due to bioinformatics matches in the allergenicity assessment of GM crops, *Food Chem Toxicol 46 Suppl 10*, S24-34.
22. Ladics, G. S., Knippels, L. M., Penninks, A. H., Bannon, G. A., Goodman, R. E., and Herouet-Guicheney, C. (2010) Review of animal models designed to predict the potential allergenicity of novel proteins in genetically modified crops, *Regul Toxicol Pharmacol 56*, 212-224.
23. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic acids research 25*, 3389-3402.
24. Pearson, W. R., and Lipman, D. J. (1988) Improved tools for biological sequence comparison, *Proceedings of the National Academy of Sciences of the United States of America 85*, 2444-2448.
25. Nakamura, R., Teshima, R., Takagi, K., and Sawada, J. (2005) [Development of Allergen Database for Food Safety (ADFS): an integrated database to search allergens and predict allergenicity], *Kokuritsu Iyakuhin Shokuhin Eisei Kenkyujo hokoku = Bulletin of National Institute of Health Sciences*, 32-36.
26. Goodman, R. E. (2006) Practical and predictive bioinformatics methods for the identification of potentially cross-reactive protein matches, *Molecular nutrition & food research 50*, 655-660.
27. Ladics, G. S., Bannon, G. A., Silvanovich, A., and Cressman, R. F. (2007) Comparison of conventional FASTA identity searches with the 80 amino acid sliding window FASTA search for the elucidation of potential identities to known allergens, *Molecular nutrition & food research 51*, 985-998.
28. Yang, Y., Faraggi, E., Zhao, H., and Zhou, Y. (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates, *Bioinformatics (Oxford, England) 27*, 2076-2082.
29. Lu, T., Yang, Y., Yao, B., Liu, S., Zhou, Y., and Zhang, C. (2012) Template-based structure prediction and classification of transcription factors in Arabidopsis thaliana, *Protein Sci 21*, 828-838.

30. Faraggi, E., Xue, B., and Zhou, Y. (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network, *Proteins 74*, 847-856.