DEVELOPMENT AND TESTING OF ALGORITHMIC SOLUTIONS FOR PROBLEMS IN COMPUTATIONAL GENOMICS AND PROTEOMICS

by

Thiruvarangan Ramaraj

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Computer Science

MONTANA STATE UNIVERSITY
Bozeman, Montana

May 2010

APPROVAL

of a dissertation submitted by

Thiruvarangan Ramaraj

This dissertation has been read by each member of the dissertation committee and has been found to be satisfactory regarding content, English usage, format, citation, bibliographic style, and consistency and is ready for submission to the Division of Graduate Education.

Dr.Brendan Mumey

Approved for the Department of Computer Science

Dr. John Paxton

Approved for the Division of Graduate Education

Dr. Carl A. Fox

STATEMENT OF PERMISSION TO USE

In presenting this dissertation in partial fulfillment of the requirements for a doctoral degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for extensive copying or reproduction of this dissertation should be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Thiruvarangan Ramaraj

May, 2010

DEDICATION

To My Parents for their Unconditional Love, Support, and Sacrifices

ACKNOWLEDGEMENTS

First and foremost, I offer my sincerest gratitude to my advisor, Dr. Brendan Mumey for his supervision, advice, excellent guidance, and support and being extremely patient with me this entire process. I have no words to express my deep gratitidue and I am greatly indebted to him more than he knows. I would like to thank Dr.Al Jesaitis for his valuable insights and comments with Immunology related aspects of my research. Many thanks go to Dr. Ed Dratz for his valuable advice and discussion in Biochemistry related questions. I thank Dr. Denbigh Starkey for his constructive comments on my research work and thesis. I like to thank Dr. Joann Mudge, NCGR, Santa Fe, NM for her guidance, support and time on my genomics related research and also gracefully agreeing to be on my committee. Also I would like thank Dr. Tom Angel for providing thoughtful discussion on my work.

I specially like to thank Ms. Jeannette Radcliffe, Ms. Kathy Hollenback and Mr. Scott Dowdle of Computer Science department for providing me with great support. I would like to thank Montana INBRE, Dept. of Computer Science, MSU-Bozeman, and NCGR, Santa Fe, NM for their kind financial support. Several students have helped me with my research work, I would like to give my special thanks to Robbie Lamb, Richard MacAllister, Illai Karen, Anoop Sendamarai, and Anburaj Muthumani.

Last but not least I thank Anitha Sundararajan for all her moral support. I owe a great debt of gratitude to everyone who helped me make this happen.

TABLE OF CONTENTS

TABLE OF CONTENTS - CONTINUED

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure

Page

LIST OF FIGURES CONTINUED

Figure Page

LIST OF FIGURES CONTINUED

Figure Page

LIST OF FIGURES CONTINUED

Figure Page

LIST OF FIGURES CONTINUED

Figure Page

ABSTRACT

This dissertation covers three subjects: (i) computational characterization of Antigen (Ag)-Antibody (Ab) interactions (ii) a novel and effective algorithm to predict the epitope of a protein based on an antibody imprinting technique (iii) a comparison of existing *de novo* genome assembler algorithms targeted specifically at the assembly of data generated by Illumina (Solexa) short-read sequencing technology, and suggestions for their improvement.

The first part focuses on identification, characterization and understanding the ways in which the antibodies and antigens interact. We analyze Epitope/Paratope region using a large dataset of Ag - Ab complex structural data taken from the PDB. Epitope/Paratope regions in our dataset have been characterized in terms of their size, average amino acid residue composition, residue-residue pairing preferences, and residue dispersion in the epitope and paratope regions. This analysis provides a more up-to-date picture of the Ag-Ab interface and provides new insights into the role of residue composition and distribution in Ag-Ab recognition. The above analysis helps in obtaining a refined substitution matrix optimized for antibody imprinting technique and used to improve the effectiveness of the epitope prediction algorithms that have also been developed and are the second focus of the thesis.

The third and the final part focus on the de novo genome assembly problems. The genome assembly programs takes the short reads generated by Whole genome shotgun sequencing technology and computationally reconstructs the genome. For the genome assembly problem the connections between read length, read type, repeat complexity, quality score and coverage and how these parameters help in improving or diminishing the capability of the assembly programs to assemble the sequence data were studied in depth. At the end of this experimental process it gives us a better understanding of the impact of the above mentioned parameters on the complexity of genome assembly and helps ascertain margins on these parameters of sequence data that enable efficient and accurate assembly by the programs.

CHAPTER 1

INTRODUCTION

Proteins are large organic compounds composed of linear polypeptide chains made of 20 different amino acids residues. To fully understand the biological role of a protein one requires knowledge of its structure and function. There are several different proteins in human cells and each protein has its own folded functional structure and whenever the three dimensional structure of a linear protein sequence could be determined, the information has provided important insights into mechanisms of action and may be extremely useful in drug design. With the increased number of proteins available, traditional methods of protein structure determination are often times is not feasible. So computational approaches to predicting the structure of proteins are becoming increasingly popular. One of the main aims in biology is to describe how cells work and define the rules by which they live. A main concept is “form defines function”, if this is quoted, where from which means that if we know the shape of the shape of a molecule then we can better understand the function of that molecule. Antibodies that bind to protein surfaces of interest can be used to report the three dimensional structure of the protein. The general structure of all antibodies is very similar, but a small region at the tip of two identical arms of the protein is extremely variable. This allows more than 10^8 – 10^9 antibodies with slightly different tip structures to exist. This region is known as the hypervariable region. Each of these variants can bind to a different target, known as an antigen. This huge diversity of antibodies allows the immune system to recognize virtually any molecular surface.The unique part of the antigen recognized by an antibody

is called an epitope. The alignments of the antibody epitopes to the discontinuous regions of the one dimensional amino acid sequence of a target protein indicates how segments of the protein sequence must be folded together and provide long range constraints for solving the 3-D protein structure.

Antibodies can recognize either continuous or discontinuous epitopes. Discontinuous epitopes provide the most useful structural information in antibody imprinting because they can reveal distant segments of primary sequence that are in close proximity on the native, folded protein. This notion that an antibody binds a protein antigen might be exploited to derive structural information about the protein of interest.

In chapter 2 the PDB (Protein Data Bank), [Berman et al, 2000], was mined for unique antigen-antibody complexes to learn as much as possible about the interface region amino acid composition and structure and the substitutability of antigen residues when bound to an antibody. The interaction region amino acid characteristics and insights is used to improve the epitope predictions in the next chapter.

Chapter 3 focuses on improving EPIMAP, a method for predicting the antibody binding site, or epitope of a protein using multiple sequence alignment approach and refine the alignment scoring and improve on epitope prediction considerably using the Ag-Ab interface analysis and new insights into the role of residue composition and distribution in Ag-Ab recognition.

Chapter 4 delves into the genome assembly problem. In Bioinformatics, genome assembly refers to the process of taking a large number of short DNA sequences which are generated by shotgun sequencing project and putting them back together to create a

representation of the original chromosomes from which the DNA originated. High quality de novo assembly using illumina (solexa) genome analayzer short reads is possible using many publicily available short read assemblers. Several challenges faced in terms of assembly process were discussed by summarizing several de novo bacterial genome assembly experiments.

CHAPTER 2

ANTIBODY/PROTEIN ANTIGENS INTERACTIONS: COMPUTATIONAL SUMMARY OF 62 PDB STRUCTURES

Introduction

The antibody – antigen interface determines the specifity and avidity of antibody immune function. We present a generalized picture of the interfaces captured from the PDB, database of published structures of proteins, interactions identified by our analysis may be significant for binding and were used for improving epitope alignments discussed in the following chapter.

Proteins are linear polypeptide chains with a wide variety of amino acid sequences, typically comprised of hundreds of the 20 different amino acid residues [Baker, Sali 2001]. Protein tertiary substructures or folds are determined implicitly by their amino acid sequences and the local amino acid composition is predictive of the secondary structural content and to some extent the complex fold adopted [Eisenhaber et al. 1996, Dubchak et al. 1993, Chou 1995]. Full understanding of biological role of proteins requires knowledge of function, structure, multi-protein complex formation, and mechanism of action. There are about 100,000 different protein amino acid sequences and perhaps 1,000,000 different modified protein forms in human cells and each protein form has a characteristic folded 3D structure that is necessary for proper function, localization, and association with interactive partners. With the increased number of proteins under investigation, it is clear that traditional methods like X-ray crystallography

or Nuclear Magnetic Resonance (NMR) are often not feasible for protein structure investigation and determination.

Prediction of protein structure given knowledge of amino acid sequence alone is not yet reliable, however with certain structural constraints, positional information about a limited number of amino acid residues in the three dimensional fold of a protein, computational predictions of structure is now a reality [Dandekar et al. 1997, Bystroff et al. 1998, Bystroff et al. 2002, Yuan et al. 2003]. Such information can come from the protein surface in terms of side chain surface accessibility [Bennett et al. 2008], nearest neighbor distance information from cross-linking [Jacobsen et al. 2006, Jin et al. 2008], and NMR [Burritt et al 1998], and identification of proximity of different regions of the protein sequences based on their participation in an antibody antigen interface [Jesaitis et al, 1996, Burritt, et al, 1998, Bailey et al 2000]. Most proteins do not act alone, but function as components of protein-protein complex [Dhungana et al. 2009]. Surface structure drives protein association and the intrinsic information present in structural form is used by proteins to establish contacts and functionally productive interactions. Thus, determining the structure of one protein surface at an interface can provide structural information about the other protein surface, which inturn can provide enough information from which the protein structures could be determined.

Composition of AA Residues Involved in the Antigen-Antibody Interface

Protein-antigen-antibody (Ag-Ab) complexes constitute a relatively large group of protein-protein interfaces that have been characterized structurally. The size of a typical protein Ag-Ab combined interface is approximately 1400-2300 $Å^2$ [Amit, Mariuzza et al. 1986, Conte et al. 1999] based on certain types of calculations of molecular surface or solvent accessible surface area. The antibody amino acid residues involved in contact with antigens are contained in 6 loops in the antibodies that are called the Complementarity Determining Regions or CDRs: 3 from the 25 kDa light chain CDR_{L1-3} and 3 from the 50 kDa heavy chain CDR_{H1-3} [Chothia et al. 1989]. The amino acid residues in the CDR loops form surfaces that make intimate contact with the antigen. Earlier extrapolation of a limited number of structures of Ag-Ab complexes indicated that a major fraction of the antibodies recognize discontinuous epitopes (i.e. widely spaced regions from the primary amino acid sequence of the antigen) on protein surfaces [Barlow et al. 1986]. When available, the structures of the antibody alone and the antigen alone most often indicate that these complexes form in a lock and key manner with little or no structural change induced upon complex formation, especially for the higher affinity antibodies [Van Regenmortel 1996]. Thus, the antibody carries a 3 dimensional "imprint" of the protein contact surface in the fold of its variable light and heavy chain domains and this surface represents the 3-dimensional complement equivalent to a 3 dimensional photographic negative of the antigen surface structure contacted by the antibody. The Ag-Ab interface structures also represent a relatively well defined model

subset of all protein-protein interfaces, where one protein of the complex has a very well studied secondary structure.

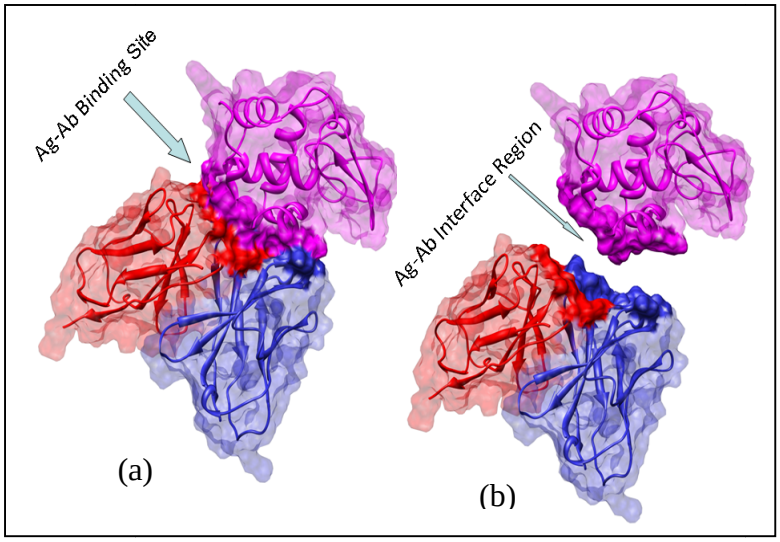


Figure 1**:** 3-D Structure of 1JHL (Ribbon Structure). The antigen chain A is shown in magenta, antibody heavy chain in blue, and antibody light chain in red. (a) shows the binding site, and (b) shows the interfaces separated so that the surface is better visualized.

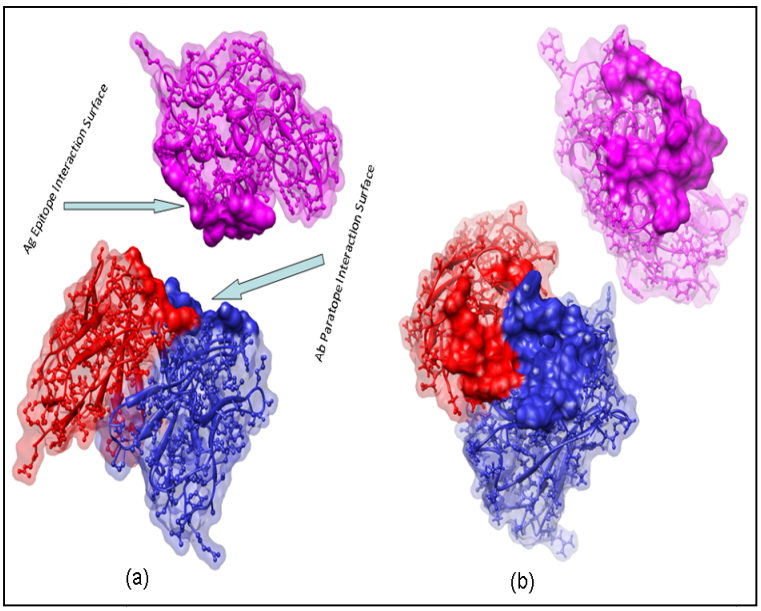


Figure 2**:** (a) The epitope surface of the antigen and the antibody in the interaction region is shown separated by an arbitrary translation imposed on the complex. (b) The epitope surface of the antigen and the antibody interface is shown but with the surfaces of both molecules facing upwards.

To better understand structural parameters involved in Ag-Ab interactions we carried out an examination of the amino acid residue composition and distribution in antigen and the antibody as well as the interactive pairing of residues between the antigens and the antibodies. To date the number of complexes so examined has been limited. A review by MacCallum et al. in 1996 considered 10 complexes, Davies and Cohen in 1996 reviewed three additional anti-idiotype complexes, and LoConte et al. 1999 studied 19 antibody – antigen complexes of which 7 were lysozymes and [Sundberg and Mariuzza 2003] listed the structures of 30 complexes but generalizations made from this entire group were not discussed. Although the former studies considered the size, shape, planarity, and CDR residue contacting propensities of the Ab residues in exceptional detail, generalizations about the properties of the antigens was more limited. Furthermore, the relatively small number of complexes examined limits gains in general understanding regarding such a diverse group of interactions. We have examined the contact regions of 62 unique Ag-Ab complexes currently available from the protein data bank (PDB). Although, there are approximately 101 Ag-Ab complexes in the PDB, of those 39 were redundant owing to studies involving site-directed mutagenesis of single amino acid residues which we felt would bias the studies giving higher weight for such protein antigens. We, therefore, sought to expand our view of the Ab-protein Ag interface, to facilitate extraction of general structural information about the antigen surface from the antibody contacts. For this study, we calculated the average values of the following Ag-Ab interface parameters: size, eccentricity, planarity, discontinuity, secondary structure, hydrogen bonding, amino acid composition, and the amino acid

interactions between the antibodies and the antigens. We then attempted to present a generalized picture of the interfaces and the interactions that may be significant for binding.

Definitions & Methods

Antigen Epitope and Non-Epitope Region

A protein antigen epitope is the part of the protein macromolecule that is recognized by the antibody. It is also called the antigenic determinant. Figure 2 (a) represents the 3-D ribbon structure of 1JHL antigen-antibody complex. The epiope surface region is highleted in magenta. Epitopes recognized by antibodies can be taught as 3-D surface features of an antigen molecule. These features fit precisely and thus bind to the antibodies.

Antibody Paratope and Non-Paratope Region

The paratope is the antigen binding part of the antibody, i.e the part that recognizes the antigen. Figure 2(a) shows the paratope surface of the 1JHL structure in blue (heavy chain) and red (light chain).

Surface Residues Delineation

The Ag-Ab data set was grouped based on the number of amino acid residues in the antigen for each complex. Group I of "*peptide*" antigens had fewer than 25 amino acid residues, Group II, of "*small size*" proteins, had more than 25 but less than 130 residues, while Group III, of "*large size*" proteins, had greater than 130 residues. This grouping helps examine how interactions differ with varying antigen size. The complete

list of the complexes analyzed is given in Table A1 of the Appendix. We defined surface residues in the epitope and non-epitope regions of the antigen as those residues with a solvent-accessible surface area (SAS) of > 50Å^2. Since the calculated surface area for the amino acid residue with the smallest side chain, glycine, is 75 Å^2 (http://www.fli-leibniz.de/IMAGE_AA.html) and for the largest (tryptophan) is 255 Å^2, our cutoff value represents 2/3 of the maximum amino acid residue surface that would be necessary for classification of a glycine to be included in the Ag-Ab contact surface. For all the analyses presented, we used > 50Å^2 as a cut-off value. This surface calculation was achieved using the UCSF Chimera molecular visualization program *(http://www.cgl.ucsf.edu/chimera/)*. The Chimera program calculates the molecular surfaces with embedded software from the MSMS *(http://www.scripps.edu/~sanner/html/msms_home.html/)* package [Sanner et al. 1996].

Epitope and Non-Epitope Region Classification

There are two main approaches to describe epitope residues in Ag-Ab complexes. The first approach uses the Solvent Accessible Surface Area (SASA) between two atoms of an interactive pair of molecules to calculate proximity [McConkey et al. 2002], while the second approach uses distance cut-off between antigen and antibody atoms in the complexes. For our work, we used the second approach and defined epitope and non-epitope regions by the contacting residues. The theoretical maximum separation distance between two contacting atoms is 6.6 Å, albeit in practice the majority of contact residues are < 5Å apart [McConkey et al. 2003]. A 5Å cutoff for interface definition has been employed recently by [Hafenstein et al. 2009] in defining the "footprint" of an antibody

on and antigen surface. Thus we define the antigen epitope and antibody paratope as the collection of amino acid residues of an antigen or antibody, in which any atom of the epitope residue is separated from any antibody atom by a distance ≤ 5Å.

Estimation of Surface Residues in
Epitope/Paratope and Non-Epitope/Non-Paratope Regions

To calculate the surface residues in the interface regions, the number of atoms in the interface region is counted explicitly. For example, two residues, one having five solvent-accessible atoms in the interface region and the other having two solvent accessible atoms in the interface region would both be considered as contributing to the interface We identify all the antigen and antibody residue solvent-accessible atoms that were separated by a distance of ≤ 5Å from each other. After this computation for all the complexes, we identified and defined an epitope region and paratope region for the antigen and antibody in each complex, respectively. Since some atoms of the antibody or antigen are less that 5 Å distant from the opposing surface but are not on the surface of their respective protein, we added another filter process, where we included only the residues that were also on the surface of the uncomplexed protein as defined above.

Amino Acid Composition of Epitope/Paratope and Extra-Interface Surface

We calculated the raw frequency of occurrence of each amino acid residue for the set of interface surfaces (epitope and paratope) and the entire protein antigen and antibody surfaces of all the Ag-Ab complexes in our data set.

Molar Fraction. For each epitope and paratope surface, we calculated the Molar Fraction of an amino acid residue in that surface by dividing the raw frequency of

occurrence of that amino acid in that surface by the total number of residues in that surface.

$$Molar\ Fraction\ (x, i) = \frac{Total\ number\ of\ x\ in\ surface\ i}{Total\ number\ of\ amino\ acid\ residues\ in\ surface\ i}$$

x represents a particular amino residue type and i is the i^{th} interface surface (epitope and paratope)

The molar fraction values for all the epitope paratope pairs provide a better way of comparing the occurrence of any residue in the epitope/paratope surface to its occurrence on the surface outside the epitope/paratope. For a relative measure of occurrence, we defined the Occurrence Propensity as the ratio of the average molar fractions of any amino acid over all epitopes or paratopes and its average molar fraction over the entire surface of their respective protein (antibody or antigen)

Average Molar Fraction. We calculate the average molar fraction for each amino acid residue type in the average epitope and paratope surfaces by summing all molar fractions for a particular residue over all epitopes (or paratopes) surfaces and dividing by the total number of surfaces.

$$Average\ Molar\ Fraction\ (x) = \frac{\sum_{i=1}^{n} Molar\ Fraction\ (x, i)}{n}$$

Occurrence Propensity. The average occurrence propensity for a particular residue type in an interface is calculated as the ratio of its average molar fraction in the interface surface and the average molar fraction of the residue over the entire surface of the protein bearing that interface

$$Average\ Occurrence\ Propensity(x) = \frac{Average\ Molar\ Fraction\ (x\ in\ Epitope\ Surface)}{Average\ Molar\ Fraction\ (\ x\ in\ Entire\ Surface)}$$

This average Occurrence Propensity speaks to the likelihood of finding a particular residue in the epitope surface versus the likelihood of finding it anywhere on the protein surface. A high Occurrence Propensity suggests a higher probability that a particular amino acid residue occurs in the epitope/paratope surface than on the surface outside the interface. Average Occurrence Propensities < 1 indicate that the particular amino acid residue is less likely to occur in the epitope/paratope surface than in the extra-interface surface.

Average Epitope Occurrence Probability. In this section we calculate the estimated probability that each residue belongs to the epitope given that it is in the antigen. For each complex in our data set we calculated the epitope occurrence fraction.

$$Epitope\ Occurrence\ Probability\ (x, i) = \frac{Number\ of\ x\ in\ epitope\ surface\ i}{Number\ of\ x\ in\ the\ entire\ surface\ Antigen}$$

The average epitope occurrence fraction then can be calculated as follows,

$$Average\ Epitope\ Occurrence\ Probability = \frac{\sum_{i=1}^{n} Epitope\ Occurrence\ Fraction(x, i)}{n}$$

This value will be useful for giving an a priori score to each protein target position as its likelihood of belonging to the epitope. The Average epitope occurrence probability is presented in table A6 (See Appendix). We consider group II and III combined and graphically represented in figure 8.

The average epitope occurrence probability indicates the probability of an amino acid residue occurs on the epitope surface.

Antigen-Antibody Interaction Surface

We characterized the Ag-Ab interfaces in terms of surface planarity, eccentricity, size, and epitope discontinuity. ProtorP, a protein-protein interaction analysis server [Reynolds et al. 2009] was used to calculate the surface planarity and eccentricity. The planarity of the surfaces between Ag-Ab complexes is calculated by computing the root mean square deviation of the all the interface atoms from the least-squares plane through the interface atoms. If all the atoms would exactly fit the same plane, the planarity index would be zero [Bahadur, and Zacharias, 2008, Jones, and Thornton, 1996]. As such, the planarity can be viewed as an indication of how deep and rough the surface of the interface is.

Another parameter that we examined was the eccentricity (also known as circularity) of the interface. The eccentricity is a measure of the shape of the interface [Reynolds et al. 2009]. The eccentricity is calculated as the ratio of the length of the principal axes of the least-squares plane through the atoms in the interface. A ratio of near 1.0 indicates that an interface is approximately circular.

We also calculated the maximum dimension of the epitope and paratope, the largest distance between any two residues in a particular surface. This was determined by doing a pair- wise Euclidean calculation of the distance between each pair of atoms in the epitope or paratope surfaces.

Lastly, to understand the secondary structure of the interfaces, we also examined the continuity of sequence in the protein antigen surface as well as the content of secondary structural elements. We calculated the epitope discontinuity, defined as the

number of segments of the Ag sequence within the epitope that were separated from their neighbor regions by minimum gaps of 3 and 4 amino acid residues. Also, the α-helical and β-sheet content information of the interface regions were extracted from the PDB file.

Epitope/Paratope Site Amino Acids Frequency of Interaction Matrix

Calculating Actual Frequency of Interaction Matrix. To obtain a measure of the importance of a particular residue type to the epitope and paratope, we also calculated the raw frequency of interactions between particular residues on the epitope surface to those on the paratope surface and vice versa. A pair of amino acid residues i and j was considered to be in contact if the distance between at least one of their atoms was at most 5Å (our defined cutoff distance). The number of pair wise interactions C_{ij} between amino acid residue type i in the epitope surface and j in the paratope surface is calculated. The computed C_{ij} values are represented in the 20 × 20 matrix (Table 3).

Calculating Actual to Scaled Expected Ratio as a Measure of Strength of Association. The best way to understand the involvement of the amino acids in the interaction region, protein antigen epitope and the antibody paratope is to study the ratio of actual to adjusted frequency of interaction for each complex in our data set and then find the average of the all ratios. So in this section the Actual Frequency of Interaction, adjusted expected frequency of interaction, and the ratio of actual to adjusted frequency of interaction was calculated.

Actual Frequency of Interaction Matrix: For Each complex the actual frequency of interaction was calculated. The actual pair wise interaction can be written as

$$C_{ij}^{k} = \sum_{i,j=1}^{20} A_{ij}^{k}$$

where C_{ij} is the number of interactions between residues of type i on the epitope and j on the paratope in the complex *k*. This is specified as a 20 × 20 matrix, which represents the actual frequency of interaction matrix for a particular complex.

Expected Frequency of Interaction Matrix: For each complex the expected frequency of a pair of amino acid interaction is proportional to the product of a constant value and the product of the raw frequency of occurrence of each amino acid in their respective interface regions, epitope and paratope.

$$E_{ij}^{k} \approx \ c \ \times \ f_{i}^{k} \times f_{j}^{k}$$

The expected frequencies are the frequencies that we would predict (expect) in each cell of the matrix.

$$c = {\sum C_{ij}^{k}}\Big/{\sum E_{ij}^{k}}$$

where E_{ij}^{k}, is the expected frequency of interaction of amino acid *i* in the epitope and amino acid residue *j* in the paratope of complex *k*, and c, is a constant value, and f_{i}^{k} is the frequency of amino acid i in the epitope surface and f_{j}^{k}s the frequency of amino acid j in the paratope surface of complex *k* and C_{ij}^{k} is the total sum of all the actual pair wise interactions, and E_{ij}^{k} is the total sum of all the expected pair-wise interactions

This is also specified as a 20 × 20 matrix, which represents the expected frequency of interaction matrix for a particular complex.

Ratio of Actual to Scaled Frequency of Interaction: For each amino acid pair wise interaction, the ratio of the actual to scaled frequency of interaction is calculated only if the expected frequency of interaction, $E_{ij} > 0$ as follows

$$R_{ij}^{k} = \frac{C_{ij}^{k}}{E_{ij}^{k}}$$

R_{ij}^{k} ratio of actual to scaled frequency of interaction,C_{ij}^{k} is the actual pair wise frequency of interactions, and E_{ij}^{k} is the scaled expected pair wise frequency of interaction. For each complex a 20 × 20 matrix is computed which represents the ratio of actual to scaled expected frequency of interaction for each amino acid pair wise interaction for a particular complex. Finally, the average of all ratios (entire data set) is calculated and represented as a 20 × 20 matrix in table 4.

Programming & Statistics

Perl scripting language was used for all our data generation and processing. R (http://www.r-project.org/index.html) and Excel were used for statistical analysis.

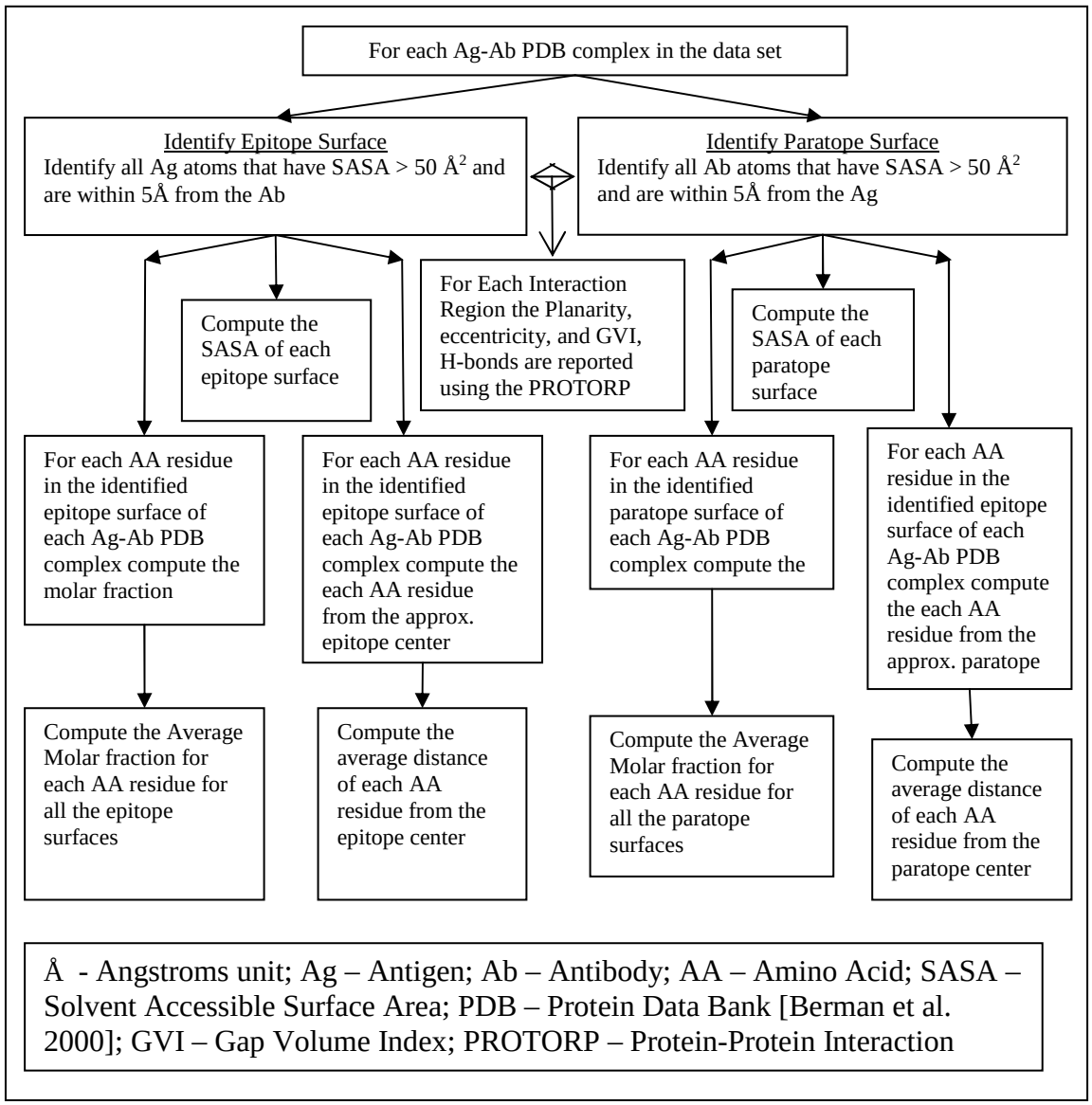


Figure 3: Antigen-Antibody Interaction Region Characterization Work Flow

Results & Discussion

General Epitope features

We used 62 non-redundant published structures of distinct protein or peptide Ag-Ab complexes to gain a more generalized understanding of the Ag-Ab interface region than currently exists. The identification, with PDB codes, of the antibody paratopes and antigen epitopes analyzed for all the Ag-Ab complexes investigated are listed in Appendix Table A1.

The total solvent accessible surface area of a molecular surface is computed by summing all the solvent accessible surface area of all the atoms in that surface. We calculated the epitope and paratope solvent accessible surface area (SASA) as well as their sum, i.e. the combined interface region surface area. The average area of the solvent-accessible molecular epitope surfaces (Table 2), is 1135 ± 350Å^2 for the 15 Group I antigens, 1075 ± 179Å^2 for the 26 Group II antigens, and 1125 ± 233Å^2 for 21 Group III antigens (Table 2). These surfaces have maximum dimensions of 21.4 ± 5.9 Å, 29.3 ± 9.3 Å, 29.9 ± 5.6 Å, respectively (Table 2). For all the protein antigens of greater that 25 amino acids (i.e. Group II and Group III combined) these values are 1097 ± 204 Å^2 and 29.6 ± 7.8 Å, respectively. Correspondingly, for the paratope, the average surface area values are 749 ± 263 Å , 1015 ± 202 Å, 1063 ± 226 Å, respectively and suggest that the areas of the epitope and paratope are very close to one another except for the group I peptide antigens.

The Group I (peptides) epitope and paratope solvent accessible surface area values are interesting. The average surface area ratio (epitope vs paratope) is ~1.5. This differential indicates that the epitope surface is 50% bigger than the paratope surface and might suggest that a paratope "ridge", as was suggested by MacCallum et al. (1996) for small antigens, which might wedge between two epitope peptide stretches much like the interaction between three pipes of equal diameter, i.e. the buried area of one pipe being less than that of the other two combined. Considering all groups combined, the values of the epitope plus paratope surface areas also confirm Sundberg and Mariuzza's (2003) estimate of ~1,400-2,300 $Å^2$ as the range of the combined Ag-Ab surface buried in an interface based on a more limited set of structures (see above). Averaged over all 62 structures presented here, our value for the combined Ag-Ab surface area is 2073 ±459 $Å^2$.

When viewed from an axis perpendicular to its least squares calculated plane, the antigen antibody interface is not circular but has an eccentricity value of between 0.6 to 0.8, where the most a circular value belongs to the more diverse Group II antigens (Table 2). The Ag-Ab interface is also irregular in the vertical plane as evidenced by the planarity index which is the root mean square deviation of interface atoms from the average plane. The planarity index ranges from 2.0 to 2.2 Å from Group I to Group III and its overall average is 2.2 ± 0.2 Å. These values suggest that the side chains, of either paratope or epitope, which can be as long as 7 Å in extended conformation lie relatively flat on this surface and that the surfaces probably don't inter-digitate more that 2-3 Å. Also, the number of H – Bonds ranges from 18.14 $\pm$ 10.10 for Group I, 23.88 $\pm$ 19.36

for Group II, 19.71 ± 18.93 for Group III, 21.98 ± 18.86 for all protein groups (Group I and II) combined and 21.80 ± 17.22 for all groups (Group I, II, and III) combined. The gap volume index, another measure of the closeness of the interaction, is obtained by calculating the quotient of the gap volume and the interface surface area and is given in Table 2 for the different groups. Its values for the three Groups range from 1.3, to 2.2, and 3.6 for Groups I through III, in that order. These values suggest that there is relatively little space between antibody and antigen structures, but that the fit is tighter for the smaller antigens and supportive of the presence of small voids which could contain water molecules [Sundberg and Mariuzza, 2003] between the larger antigens and their respective antibody interactive surfaces.

Table 1: Characteristics of the antigen groups

	$\mu^{\dagger}$	σ^{*}	Total # of Residues	Total # of Surface Residues	% of Residues on the Molecule Surface
Antigen Data					
Group I (15)	9.67	4.482	145	145	100%
Group II (26)	112.92	63.96	2936	1454	49.5%
Group III (21)	349.86	145.16	7347	3221	43.8%
All Protein Groups Combined(47)	218.78	193.11	10283	4675	45.5%
All Groups Combined (62)	168.19	195.37	10428	4820	46.2%
Antibody Data (62)	433.44	269.61	26873	10804	40.2%

† - Mean, * - Standard Deviation

Table 2: Properties of the antigen epitope groups

	Group I		Group II		Group III		All Proteins Combined (Grp. II & III)		All Peptides and Proteins Combined (Grp. I & II & III)	
	μ†	**σ***	**μ**†	**σ***	**μ**†	**σ***	**μ**†	**σ***	**μ**†	**σ***
AA	6.90	4.28	17.60	13.40	14.10	8.18	31.70	19.72	38.60	22.00
Epitope Maximum Dimension(Å)	21.36	5.93	29.31	9.33	29.92	5.59	29.58	7.81	27.69	8.16
Hydrogen Bonds	18.14	10.10	23.88	19.36	19.71	18.93	21.98	18.86	21.08	17.22
Epitope Surface Area ($Å^2$)	1134.7	349.6	1074.5	178.5	1124.7	233.4	1096.9	203.9	1106.0	244.3
Gap Volume Index (Å)	1.30	0.74	2.36	1.02	3.55	2.99	2.90	2.21	2.53	2.08
Planarity (Å)	1.97	0.42	2.20	0.52	2.23	0.64	2.21	0.57	2.16	0.54
Eccentricity	0.79	0.17	0.66	0.11	0.74	0.13	0.70	0.13	0.72	0.15

† - Mean, * - Standard Deviation

AA – # of amino acid residues in the epitope surface

Gap Volume Index Definition: The gap volume is used to give a measure of the complementarity and closeness of packing of the interface between the two subunits. This is accomplished by measuring the volume of empty space between the atoms. The gap volume index is measured in angstroms, and is computed by dividing gap volume in $Å^3$ by the Interface Area (ASA) in $Å^2$ [Reynolds et al 2009]

Amino Acid Composition

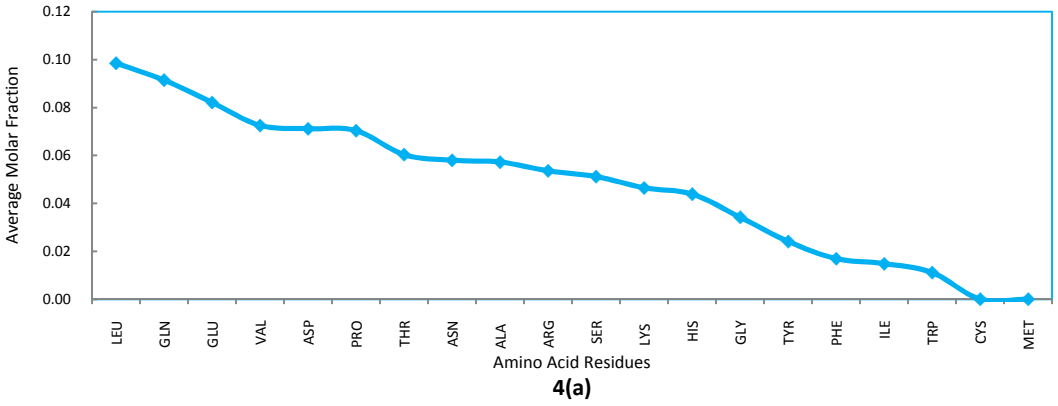
To determine the biochemical properties of the protein interfaces, we examined the amino acid compositions of the epitopes and paratopes of all 62 complexes and compared them with the compositions of the protein surfaces outside the epitope/paratope interface regions. Based on the total number of residues exposed to the surface in each Group, the percentage of the protein antigen residues on the surface in Groups I-III, were 100%, 49.5%, 43.8% individually, 45.5% for the small and large proteins combined (Groups II & III), and 46.2% for all peptides and proteins combined (Groups I & II & III)

(Table 1). The antigen epitopes contain 8.9 ± 5.5, 13.4 ± 12.0, 13.5 ± 7.7, amino acid residues for the three Groups respectively (Table 2). Adding all the residues of each group as the total, the molar fraction of each type of the 20 amino acids was calculated. These results are presented in Table A2 of the Appendix for all the groups and their combinations in alphabetical order of residue name. And the same results are presented in descending order by molar contribution for each amino acid residue in Figures 4(a, b, c) 5(a, b, c), 6 (a, b, c) and 7 (a, b, c) for the epitopes and paratopes, respectively.

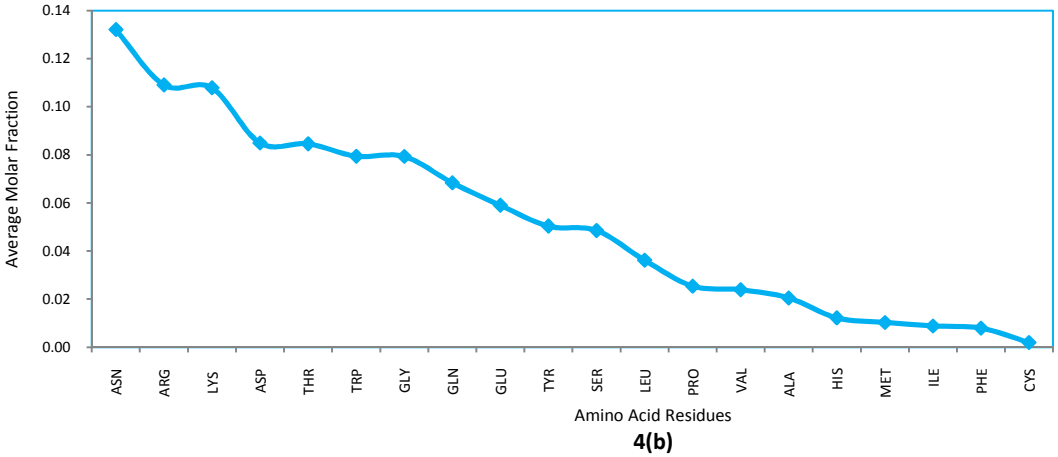
Inspection of the average molar fraction of the 20 amino acid residues in the epitope surface of each class is revealing and is shown in Figures 4 (a, b, c). There are no occurrences of MET and CYS in the Group I (Figure 4a) epitopes and a less than 2.5 mole percent occurrence of mostly aromatic TRP< ILE< PHE<TYR. Most abundant (> 7 mole %) in this group are ASP< VAL<GLU< GLN < LEU, a mixture of negatively charged polar, and hydrophobic amino acids consistent with peptide solubility. In Group II (Figure 4b) the low abundance order of less than 2% occurrence is CYS< PHE< ILE< MET< HIS, essentially hydrophobic and aromatics and the two sulfur containing groups. The most abundant residues in Group II (Figure 4b) with greater than 8.5 mole percent occurrence are THR< ASP< LYS< ARG<ASN. Lastly, in Group III (Figure 4c) the low abundance residues are (< 3 mole %) are the sulfurous and aromatic as well as the smallest, least rotationally constrained residue, CYS<PHE<GLY<MET<TRP. The most abundant (> 7.5 mole %) are the four charged residues ARG<ASP<GLU<LYS.

Amino acid residues are differentially expressed on protein surfaces depending on their intrinsic properties. These properties, have been almost universally applied in what

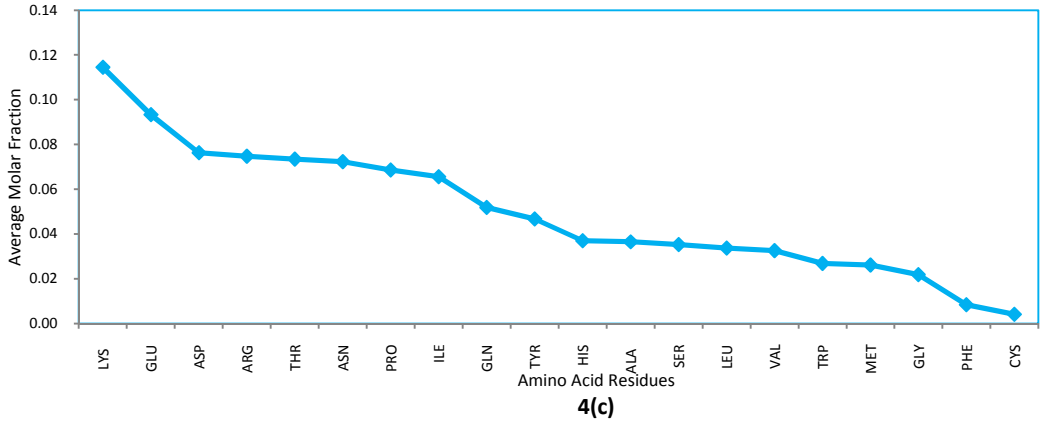
recently has been suggested as dubious attempts [Blythe M. J et al 2005] at predicting antigenicity of sequences of proteins. However, they are very useful in identifying the significance of the above amino acid occurrences, if we consider a parameter that describes the amino acid epitope/paratope expression relative to its overall expression on the protein surface. We calculated the Occurrence Propensity (ratio of frequency in the interface to frequency overall) for each group to give a measure of the significance of finding a particular amino acid in the epitope vs the overall surface of the protein. Figure 6(a, b, c) and Appendix table A2 clearly shows that, for the protein antigens, TRP, TYR, MET, ILE, GLN (which except for GLN are low abundance residues) occur in the epitope at a much higher than expected frequency (>1.5) suggesting that they play a special role on the recognition process. Indeed Nussinov and colleagues identified surface TRP, PHE, and MET as residues that identify binding interfaces (Ma et al. 2003). Furthermore, Bogan and Thorn [Bogan et al. 1998] identified TRP, TYR, ARG as enriched in distributed hotspots of binding energy surrounded by solvent occluding residues that figure importantly in dimer interfaces of proteins. These differentials in average occurrence propensities may suggest that a set of amino acid residues, with higher average occurrence propensities may be more important for an Ab-Ag interaction while those with less average occurrence propensity may not contribute much to the interactions. Although highly informative, one also needs to consider how "well" the various interface residues interact with amino acid residues on the opposing interface surface.



4(a)



4(b)



4(c)

Figure 4: Average Molar Fraction of Epitope surface (a) Group I, (b) Group II, and (c) Group III

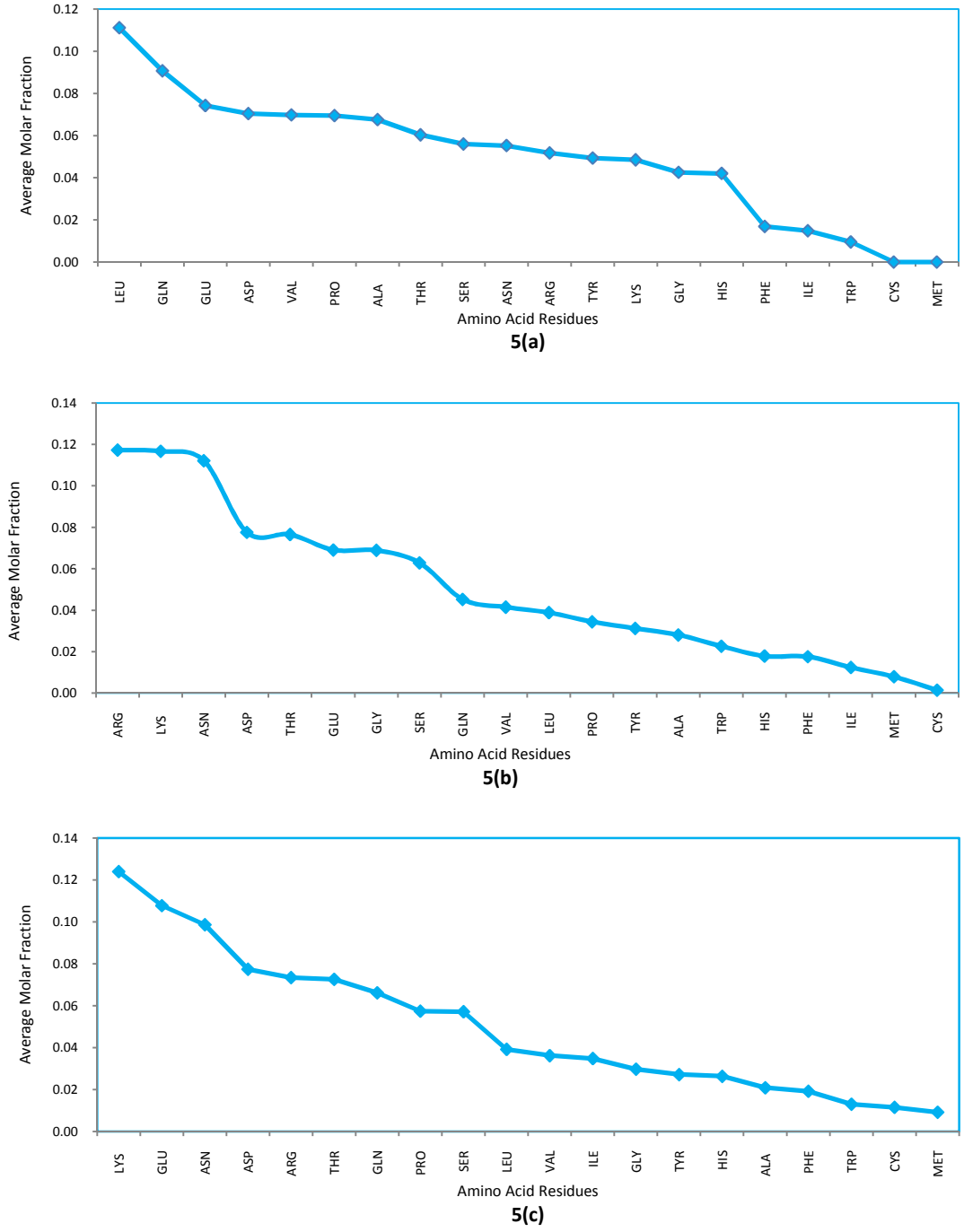
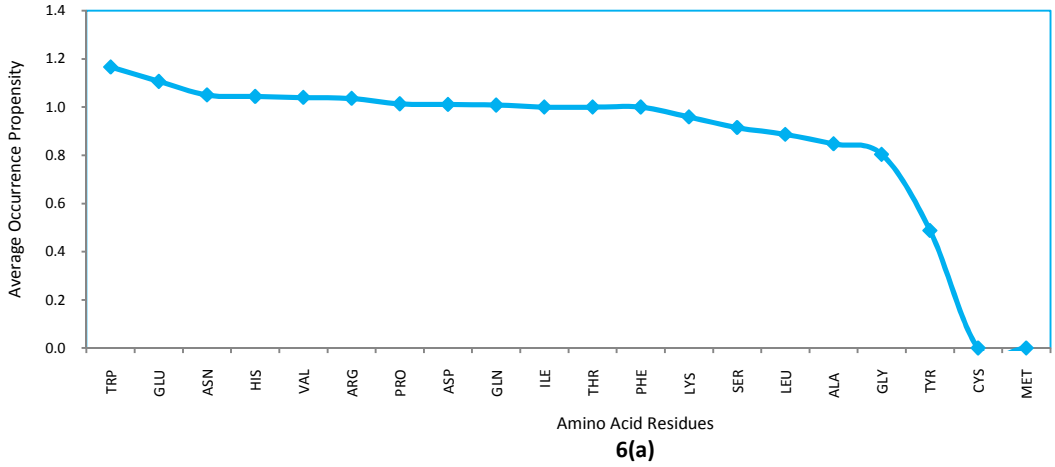
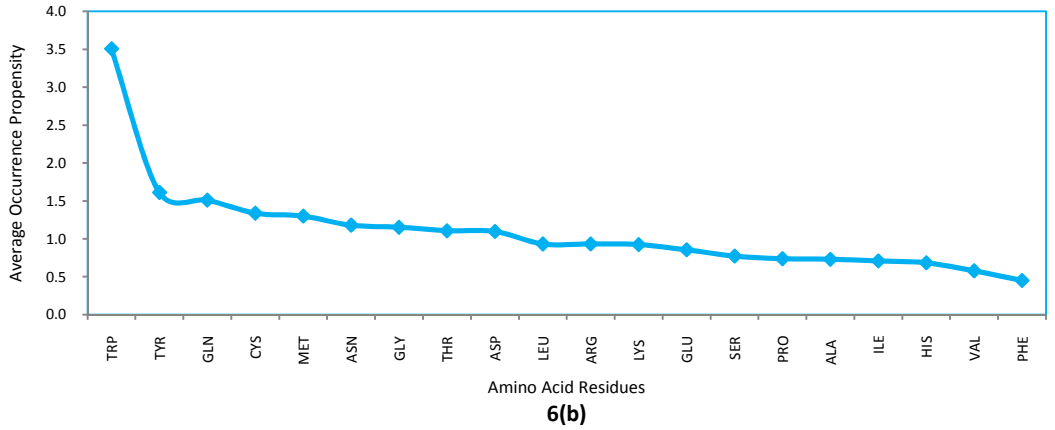


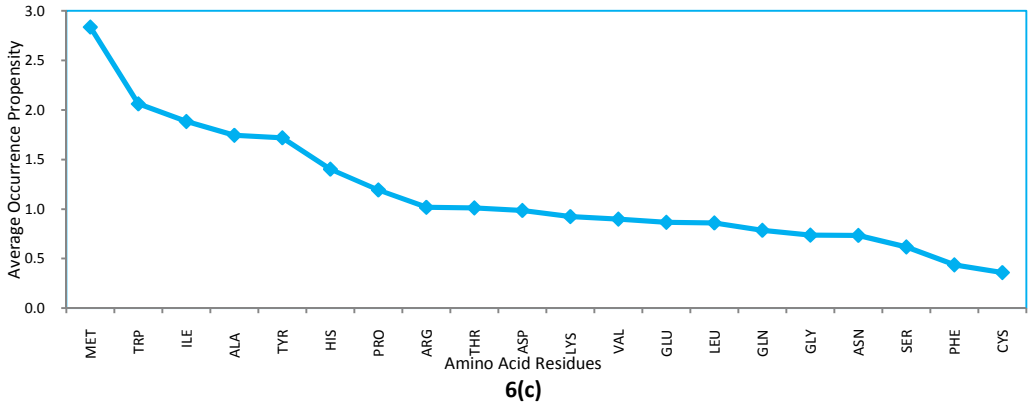
Figure 5 : Average Molar Fraction of Entire surface (a) Group I, (b) Group II, and (c) Group III



6(a)

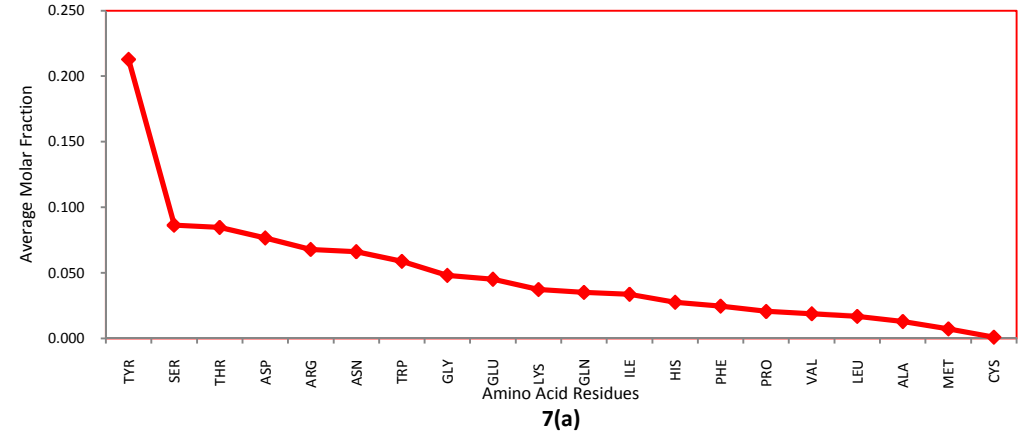


6(b)

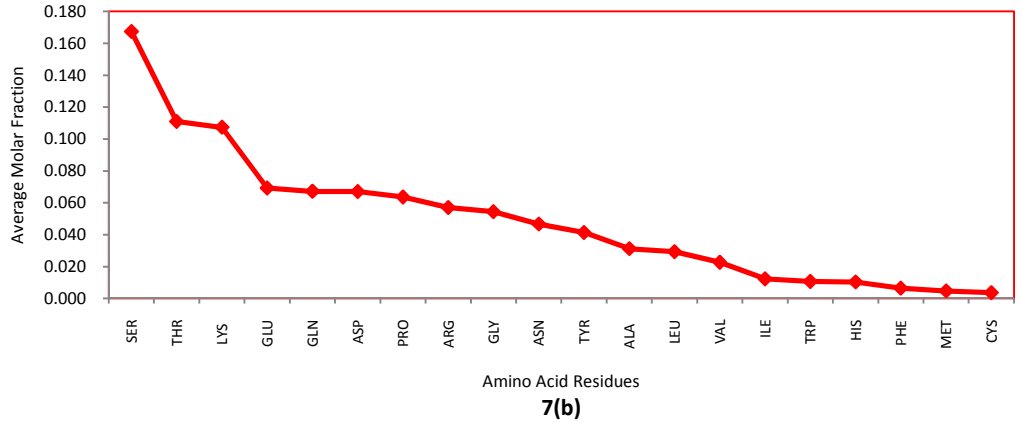


6(c)

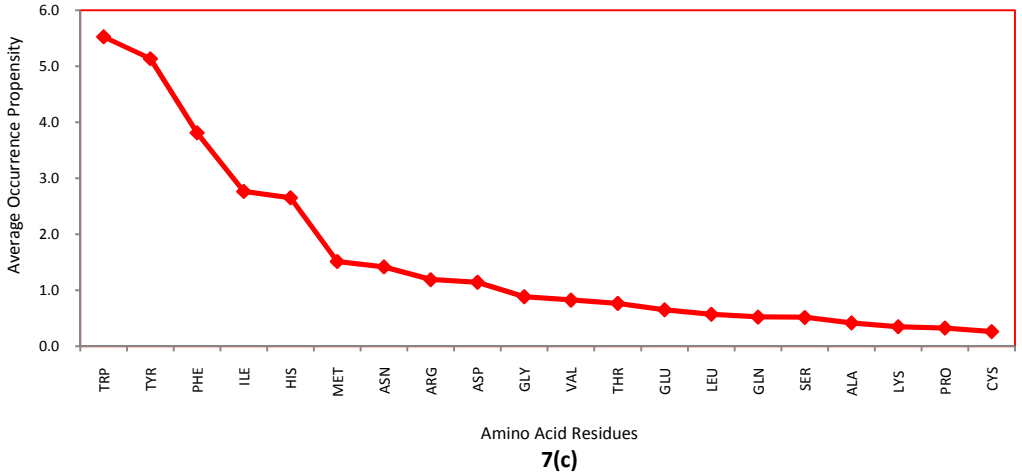
Figure 6: Occurrence Propensity of each amino acid residue type in the epitope to the whole (epitope plus the nonepitope) surfaces. (a) Group I, (b) Group II, (c) Group III.



7(a)



7(b)



7(c)

Figure 7: (a) Average Molar Fraction of Each AA in the Paratope Surface and (b) Average Molar Fraction of Each AA in the Entire Antibody Surface (c) Occurrence Propensity of Each AA in the Paratope to the whole antibody (Paratope plus the Non - Paratope) Surface.

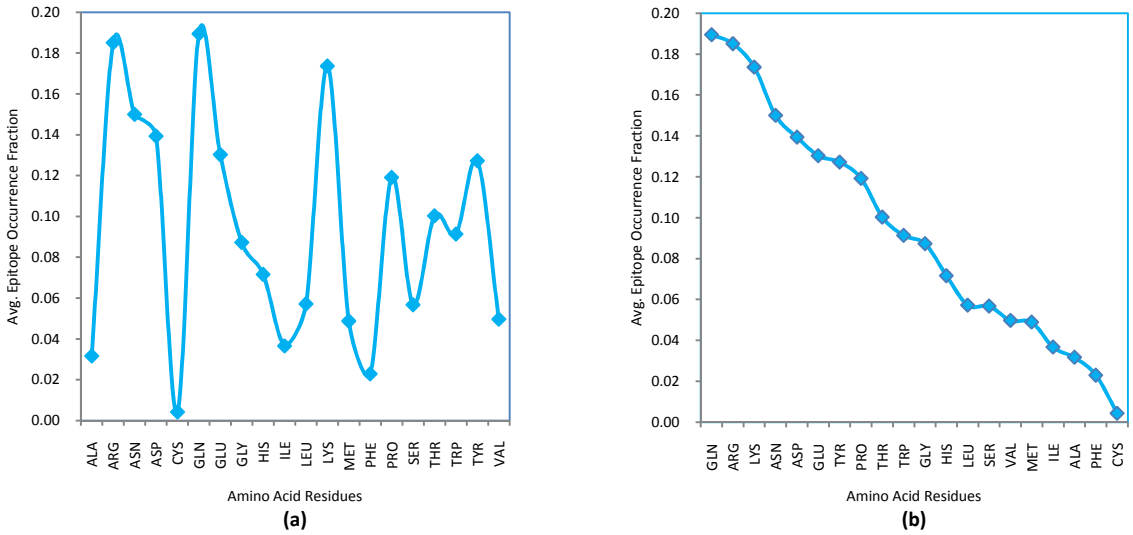


Figure 8: Epitope Average Occurrence Probability (a) Group II & III proteins combined and (b) Presented in descending order, values of each AA Residues.

Additional insight emerges from considering the antibody paratope surface. The amino acid average Occurrence Propensity values for the paratope regions of the antibodies are shown in Figure 7c and numerically in the Appendix Table A3. The Occurrence Propensities are quite high for some types of residues on the antibody paratope (Figure 7c), whereas for small and large protein antigens the occurrence propensities of different amino acids tend to be less distinctive (Figure 6b, 6c). The highest ratios were TRP>TYR>PHE>ILE>HIS>MET, ranging from 5.5 to 1.5 suggesting these residues to be very high value for antibody antigen interaction and especially the strong dominance of TRP and TYR in this interface. Interestingly, the highest average molar fractions amino acid residue occurrences in the paratopes (≥ 0.059, appendix table A3), are TYR(0.21)>SER>THR>ASP>ARG>ASN>TRP (0.059) shown in Figure 7a. It is worthy to note that TYR is 3X times more abundant than the other high abundance residues and nearly 5X more abundant in the paratope surface than on the entire surface

of the antibody. These differences suggest special functional roles for these residues in Ag-Ab interfaces.

Interactions of Antibody/Antigen Amino Acid Residues

Clearly, some amino acids are more represented than others in the epitopes and paratopes. This probably means that they are of correspondingly higher importance to the Ab-Ag interaction, yet it argues against their role in specificity, i.e. less abundant residues could imply a higher degree of specificity. However, if they have fewer interactions with the paratope residues their contribution might be more important for positional spacing in structure rather that amino acid side chain recognition [Pinilla C et al 1993]. To get another measure of the significance of particular residue types for Ag-Ab binding, we sought to identify the residues that are the most frequently involved in the interactions of the antigen and antibody pairs. We thus calculated the number of contacts that each residue on the epitope makes with specific residues on the antibody and vice versa. The interacting residues were scored if the distance between at least one of the atoms of the residue to the atoms of the complementary member was below the 5Å cutoff, consistent with our epitope/paratope site definition. We also made the corresponding calculation for the antibody paratope residues. This parameter therefore, is a combination that includes a component that depends on the number of times a particular residue occurs in the epitope and paratope as well as component that depents on side chain properties (i.e. size, hydrophobicity, etc.). This calculation is tabulated in a 20 X 20 matrix showing the raw interaction number for residues in either the paratope or epitope with residues in the opposing surface as is shown in Table 3. In Table 4, the

average of all ratios (entire data set) is calculated and represented as a 20 × 20 matrix. This ratio explains the strength of association between amino acid pairs in the interaction region. The higher the ratios the higher is the strength of association between the AA pairs.

Table 3: Frequency of interaction matrix

		Antibody Paratope Surface (Ab)																				
Antigen Epitope Surface(Ag)		**A**	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**	
	A	8	14	11	1	0	4	15	0	11	8	2	6	1	0	6	8	7	16	32	0	**A**
	R	4	41	52	34	0	3	13	11	14	2	6	0	3	3	0	50	29	97	123	4	**R**
	N	3	29	54	36	2	11	11	12	18	18	8	9	0	17	5	23	23	41	142	5	**N**
	D	11	47	38	1	0	7	6	11	12	10	3	14	0	8	0	19	31	19	174	0	**D**
	C	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	4	0	8	0	**C**
	Q	0	23	16	23	0	10	9	26	12	6	3	6	0	3	8	26	30	40	144	4	**Q**
	E	5	40	8	14	0	7	4	18	14	7	5	9	1	14	11	41	22	16	129	9	**E**
	G	0	12	9	30	1	2	15	6	14	0	4	0	0	1	6	7	21	7	75	0	**G**
	H	0	4	10	12	0	0	0	3	11	2	0	0	0	2	0	8	11	14	46	5	**H**
	I	2	4	17	8	0	0	4	7	6	6	4	9	1	10	0	6	3	17	55	3	**I**
	L	6	17	13	11	0	13	9	2	6	9	10	9	1	24	2	21	16	21	62	14	**L**
	K	3	10	28	46	1	12	22	13	5	15	0	4	0	6	10	19	12	63	134	12	**K**
	M	0	0	11	0	1	0	6	3	5	3	1	0	8	16	3	0	0	7	21	3	**M**
	F	3	5	6	0	0	1	0	0	6	0	2	0	0	8	0	2	7	0	14	0	**F**
	P	3	6	9	16	0	4	6	0	9	5	0	0	1	21	5	0	10	8	72	5	**P**
	S	0	4	10	15	3	1	8	5	7	4	0	5	0	0	10	7	19	9	48	0	**S**
	T	0	30	40	18	0	11	14	1	9	16	4	5	3	4	3	36	14	38	65	10	**T**
	W	0	12	21	7	0	1	1	3	0	16	0	2	0	1	0	3	4	5	19	0	**W**
	Y	1	38	31	0	0	1	2	3	5	7	9	2	1	7	12	20	9	22	26	2	**Y**
	V	0	8	6	2	0	1	8	2	10	4	1	6	1	0	3	7	2	8	70	2	**V**
		A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**	

Ab – Amino Acid Residues on the Antibody Paratope

Ag – Amino Acid Residues on the Antigen Epitope

Table 4: Average of All Ratios (Actual to Expected Frequency) Interaction Matrix

Antigen Epitope Surface(Ag)	Antibody Paratope Surface (Ab)																				
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
A	0.76	1.46	0.65	0.55	0.00	0.29	0.20	0.00	1.35	1.12	2.40	0.73	3.22	0.00	0.79	1.26	0.56	2.21	0.94	0.00	A
R	1.07	1.23	1.18	1.01	0.00	0.33	0.63	1.35	0.71	0.16	1.89	0.00	1.51	0.79	0.00	0.92	1.24	2.02	1.03	0.70	R
N	0.40	1.10	1.13	0.92	6.04	0.38	0.55	0.78	1.61	1.11	1.27	0.68	0.00	1.26	0.41	0.75	1.44	1.10	1.35	0.72	N
D	0.98	1.10	1.56	0.12	0.00	0.23	0.28	1.27	0.99	0.85	0.45	0.93	0.00	0.71	0.00	0.48	0.51	0.91	1.51	0.00	D
C	0.00	0.00	0.00	0.00	6.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.02	0.00	10.06	0.00	C
Q	0.00	1.01	0.57	1.04	0.00	1.07	0.75	1.35	0.51	0.90	2.12	1.58	0.00	0.25	0.55	1.50	0.67	1.03	1.15	0.55	Q
E	0.23	1.29	0.46	1.00	0.00	1.21	0.58	0.70	1.18	0.70	0.93	0.48	0.48	0.92	1.46	1.52	1.12	0.70	1.35	1.48	E
G	0.00	0.67	1.01	0.86	6.04	1.44	1.00	1.33	1.79	0.00	2.15	0.00	0.00	1.50	1.16	0.22	1.26	0.23	1.21	0.00	G
H	0.00	0.43	0.94	2.19	0.00	0.00	0.00	0.81	1.96	1.51	0.00	0.00	0.00	1.43	0.00	1.23	0.77	1.61	1.44	1.03	H
I	0.80	0.49	0.95	0.31	0.00	0.00	1.65	1.92	1.14	1.51	2.64	3.19	1.28	1.04	0.00	0.37	0.28	1.82	1.79	1.56	I
L	0.69	0.44	0.80	0.97	0.00	0.48	0.51	0.36	0.69	1.44	3.49	0.80	2.41	1.97	0.48	0.89	2.44	1.26	1.32	3.06	L
K	0.68	0.24	1.04	1.53	1.21	0.43	1.46	0.80	0.32	0.36	0.00	0.43	0.00	0.60	0.12	0.62	0.58	1.78	1.66	0.17	K
M	0.00	0.00	2.79	0.00	3.02	0.00	1.13	0.58	1.44	0.64	4.07	0.00	5.09	5.27	0.20	0.00	0.00	2.84	1.70	1.24	M
F	7.24	0.79	2.05	0.00	0.00	3.17	0.00	0.00	2.48	0.00	1.44	0.00	0.00	2.48	0.00	0.93	1.27	0.00	1.66	0.00	F
P	0.54	0.25	0.77	1.25	0.00	3.24	1.02	0.00	1.83	0.53	0.00	0.00	1.07	1.68	2.68	0.00	0.51	0.40	1.61	0.93	P
S	0.00	0.07	1.17	1.78	3.02	0.00	0.34	0.73	0.65	0.18	0.00	0.92	0.00	0.00	1.16	1.05	1.13	0.19	1.02	0.00	S
T	0.00	1.35	1.32	0.55	0.00	1.25	2.22	0.17	1.01	1.66	0.70	0.63	0.80	0.93	1.02	0.76	0.52	1.18	0.99	2.29	T
W	0.00	0.45	2.30	0.90	0.00	0.28	0.63	1.37	0.00	3.10	0.00	0.40	0.00	0.34	0.00	0.38	0.26	0.50	0.87	0.00	W
Y	0.92	2.55	1.25	0.00	0.00	0.47	0.64	0.73	0.71	1.01	2.34	0.38	1.47	1.77	2.27	0.73	0.77	1.90	0.72	0.49	Y
V	0.00	1.08	0.68	0.47	0.00	0.83	0.17	0.26	2.35	1.72	1.03	1.61	3.61	0.00	0.73	1.46	0.27	0.45	1.46	1.02	V
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	

>= 0 < 1 | > 1 < 2 | > 2 < 3 | > 3 < 4 | > 5

Ab – Amino Acid Residues on the Antibody Paratope
Ag – Amino Acid Residues on the Antigen Epitope

Of all the residues on either surface, the paratope tyrosine contacts *all* epitope residues with high frequency, whereas only GLU and LEU on the epitope have such a broad range of interactions but with much lower frequency. This observation suggests that for the paratope at least TYR has a very special role

To display a measure of the overall interaction frequencies of each residue of one surface with with all the residues of complementary surface, we summed rows or colums from Table 3 and plotted the values for each amino acid type. Figure 9a plots the absolute frequency of interactions of each amino acid residue type in the epitope with any residue in the paratope and Figure 9b plots the absolute frequency of interactions of each amino acid residue type in the paratope with any residue in the epitope. These results show differences and similarities for these parameters on the two surfaces. For the epitope residues the number of interactions can be grouped into three categories < 200 (low: CYS < PHE < MET < TRP < HIS < VAL < ALA < SER < ILE < PRO < TYR), > 200-400 (intermediate: GLY < LEU < THR < GLU < GLN < ASP < LYS) and > 400 (high: ARG > ASN). The groups on the frequencies of the paratope residues are substantially different with < 200 (CYS < MET < ALA < LEU < VAL < PRO < LYS < GLN < GLY < ILE < PHE < GLU < HIS) , > 200-400 (THR < ASP < SER < ARG < ASN), and > 400 (TRP < TYR). The most striking result from this analysis is that the antibody TYRs made ~1500 contacts with antigen residues. The corresponding value for the epitope TYRs, was about 200, suggesting the antibody TYRs make 7.5 times the number of contacts than the antigen TYRs.

The remainder of the the paratope residues have approximately the same range of contact numbers as the epitope residues, suggesting that these interactions may be more similar and perhaps characterisitic of interactive protein surfaces in general.

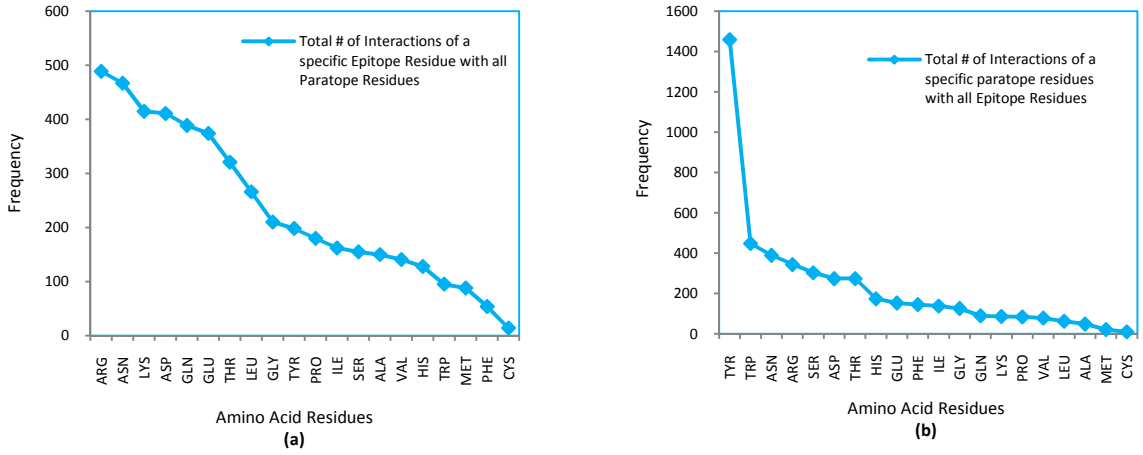


Figure 9: Interaction frequencies (a) AAs in the epitope pairing with AAs in the paratope, (b) AAs in the paratope pairing with AAs in the epitope.

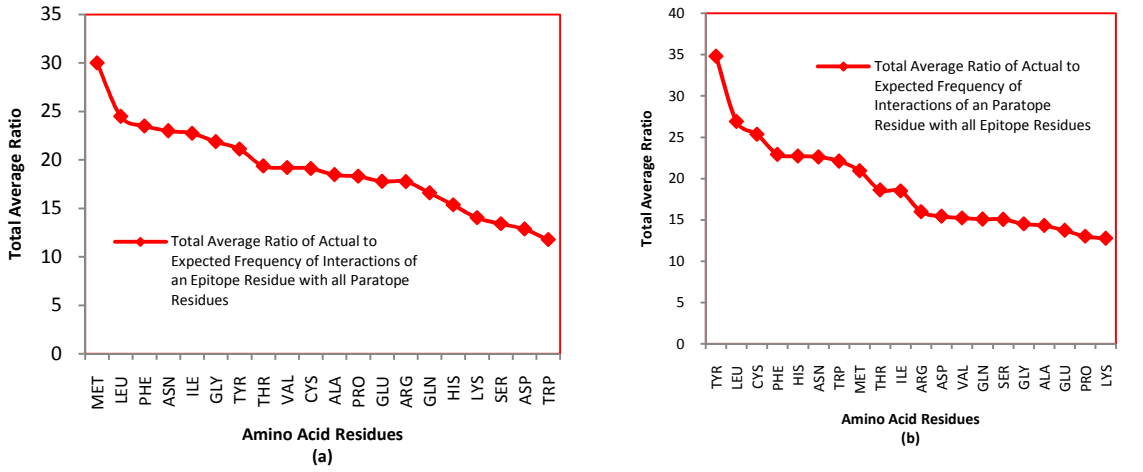


Figure 10: Total Average Ratio of Actual to Expected AAs frequency of interactions (a) Epitope pairing with AAs in the paratope and (b) Paratope pairing with AAs residues in the epitope.

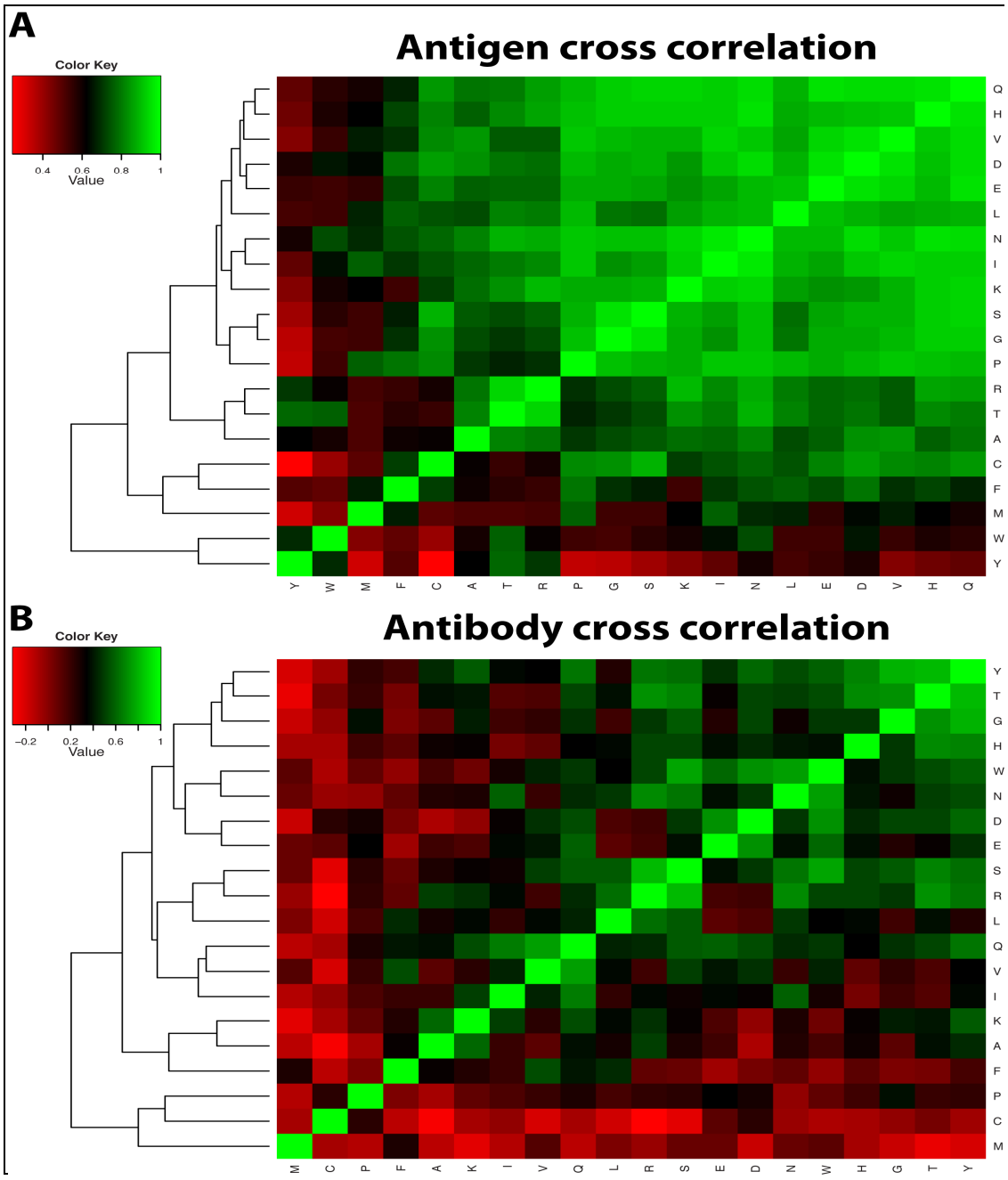


Figure 11: A hierarchical cluster analysis of the Pearson product-moment correlation coefficient of the eptitope (A) and paratope (B) amino acid interaction frequencies.

Because the occurences of the different residues in each type of surface varies considerably, the frequency of interaction of each residue type will reflect the number of times it occurs in the epitope and paratope as well as any special property it might have that increases its interaction with another residue (i.e. size, charges, hydrophobicity etc.). Figure 10a and 10b shows the average ratio of the actual to expected frequencies. This average ratio is calculated for each residue type in the epitope and paratope surface by taking the actual number of interactions of a particular residue type with another and dividing it by the expected number of interactions of each in its respective surfaces. Figure 10a and 10b show these values total for all interactions of one particular residue type with all the others on the opposing surface. Thus in this distribution the relative "importance" of each residue for making contacts on the corresponding surface is displayed. This distribution evens out the disparities, but shows that the paratope residues 34.84 = TYR > LEU > CYS > PHE > HIS > ASN > TRP > 20.94 have the most interactions on a per residue basis with the epitope while epitope residues span the range from 30.01 = MET > LEU > PHE > ASN > ILE > GLY > TYR > 19.20 for their interactions. Only LEU, ASN are shared by both lists suggesting that each type of surface mediates special complementary interactions with the exception of those relying on LEU and ASN, the two amino acids with large side chains.. Again TYR acts in a special role for the paratope "intensively sampling" many points on the epitope surface.

Additionally, one could interpret this analysis to suggest that the higher the ratios the higher is the strength of association between the amino acid pairs. Table 5 represents the top 10 ratios.

Table 5: Top 10 Actual to Scaled Expected Frequency of Interaction Ratio

	Ag-Ab	Average Ratio
1	CYS-TYR	10.06
2	PHE-ALA	7.24
3	ASN-CYS CYS-CYS GLY-CYS	6.04
4	MET-PHE	5.27
5	MET-MET	5.09
6	MET-LEU	4.07
7	VAL-MET	3.61
8	LEU-LEU	3.49
9	PRO-GLN	3.24
10	ALA-MET	3.22

To examine the issue of complementarity further and to obtain another measure of residue contribution to antibody recognition of antigen, a Pearson product-moment correlation coefficient analysis was performed on this data in Table 3. Such an analysis provides a measure of which residues on each surface act like one another relative to their interactions with the apposing surface. The result of this analysis is shown graphically in Figure 11 as heat maps for antibody and antigen cross correlations.

The numerical values for these matrices are given in Table A5 (a,b) of the Appendix. Red thru black indicates that amino acids are more anti-correlated (i.e. the residues do not share that same contact specificity) while black through green suggests they are uncorrelated or positively correlated, respectively.

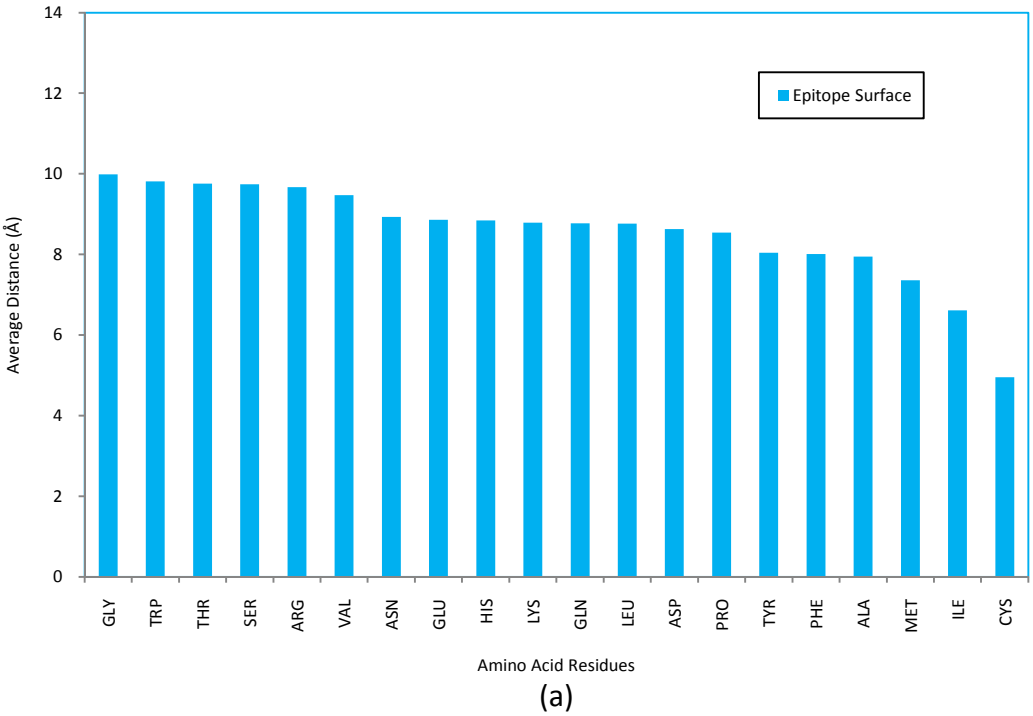
For the antigen map, the residues TYR, TRP, MET, and CYS show these least cross correlation. This means that they do not interact with the same residues on the paratope as most of the other epitope residues. The large green area in the upper right suggests that these residues are more promiscuous in their interactions with the paratope residues. Examining the antibody map suggests the the least correlated residues are Met, CYS, PRO, PHE but that there is a more steady gradation of interaction propensities amoung the antibody residues but with TYR showing the most promiscuity. These results confirm the special role of TYR and further suggests that it interacts with a broad range of epitope residues based on its intrinsic properties. They also suggest that the least correlated residues engage strongly in the specifity of the antibody antigen interaction. Additionally, the large number of highly positively correlated residues in the epitope suggest that a higher level of complexity may be required to achieve full spcificity, which may utilize units of two or three residues with specific sequence [Chen, J et al 2007, Wang, L et al 2009]. This latter notion is supported by the fact that in phage display epitope mapping visual recognition is usually achieved by spotting doublet or triplets of sequence that match the antigen being analyzed.

Spatial Distribution of Amino Acids in the Interfaces

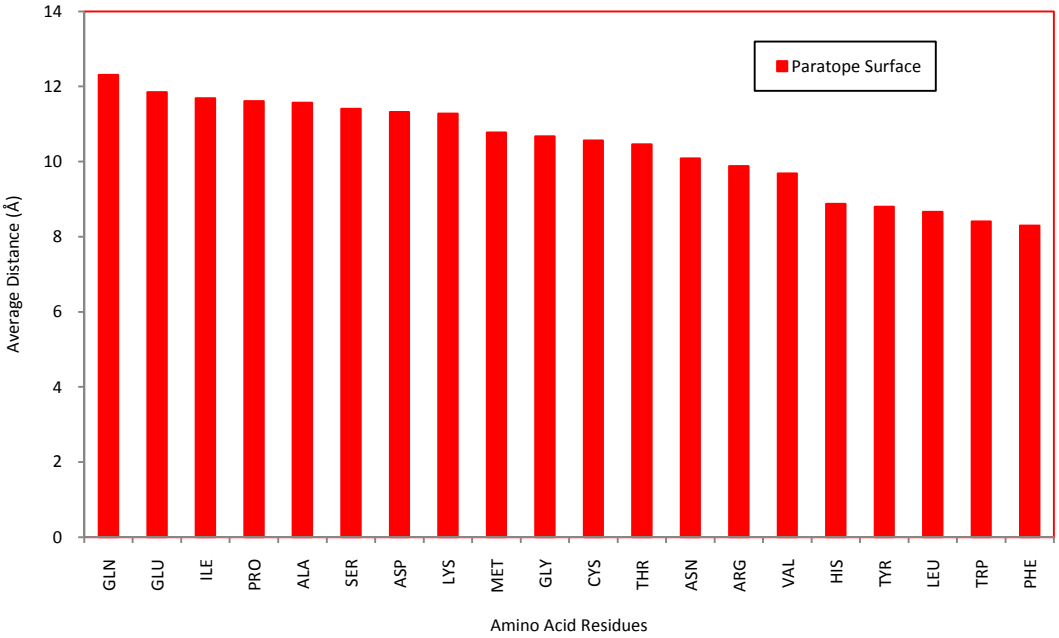
Another contrasting compositional feature of epitopes and paratopes are the differences in the distribution of residue types in the plane of the Ag-Ab interface. The current notion of antibody epitope structure is that the crucial contacts are made near the center of the epitope but that for some antigens especially larger ones this role becomes more distributed (MacCallum et al, 1996, Sundberg, E. J. and Mariuzza, R. A., 2002, Janin J. H. et al. 2003). Table A4 (see Appendix) gives the distances in angstrom units from the "geographic" center of the epitope and the paratope region derived from our calculation for each residue type. This value and its standard deviation provide a measure of how broadly and and uniformly the residues are distributed in the interfaces. Figure 12 shows the distribution of residue types from the average center of the epitope (Figure 12a) and paratope (Figure 12b) surfaces. It is of interest that the residues that are distributed most broadly or closest to the center of either the epitope or paratope span the full spectrum of specific frequency of interaction, suggesting that is no preferential distribution of frequency of interaction. Comparing the values in Table 3 and figure 12b, it is interesting to see that the farther the residues are from the center of the paratope the lower is their interaction frequency with antigen residues, and the closer the residues are to the paratope center they greater is the frequency supporting McCallum at al's hypothesis[MacCallum, et al 1996]. The antibody paratope residues ARG, ASN, HIS, TRP and TYR that are most commonly used by the antigen epitope residues to bind are distributed in closer proximity (9–10 Å, Table A4) to the center of the paratope.

The other paratope residues ALA, ASP, GLN, GLU, ILE, LYS, PRO, and SER do not interact as much with the antigen epitope amino acid residues and are located farther (> 11Å, Table A4) from the paratope center, suggesting that there is more binding towards the center of the paratope and less towards the exterior.

This overall difference suggests that the outer antibody residues play a unique role around the periphery of the antigen. The most distant residues (~9.5-10Å table A4,) on the epitope are GLY > TRP > THR > SER > ARG > VAL while those of the paratope (~11-12Å, table A4,) are GLN > GLU > ILE > PRO > ALA > SER > ASP > LYS. Such differences might suggest that part of the binding mechanism, involving charge pairing and a mix of polar and aliphatic interactions, occur at the periphery of the interface.



(a)



(b)

Figure 12: (a) Average distance in Å of AAs from the center of the average center of the epitope surface and (b) Average distance in Å of different AAs from the average center of the paratope surface.

Secondary Structure of the Interface

We also examined the epitope for generalized secondary structural features. The Ab interface clearly "reads" the Ag surface with a biased set of amino acid residues contained in the 6 CDRs [MacCallum et al. 1996]. This arrangement suggests that the antibody probes the Ag surface with at least six discontinuous segments of the Ab light and heavy chains. We examined the degree of discontinuity of the antigen sequence in the average epitope.

Figure 13 shows the bimodal distribution of complexes having minimum gaps of 3 or 4 in their sequences in the epitope regions and shows that they generally include complexes with 1 or 2 discontinuous segments (20-40%) but with approximately half having 3-5 discontinuous segments. This observation suggests that conformational constraints on the Ag structure may be coded within the conformation of these discontinuous segments. Additionally, it suggests that antigen epitope mapping strategies, such as phage display, need to take into account the contribution of discontinuous regions of the antigen in formulating identification of correspondence in phage display peptide sequences with antigen sequences.

These discontinuous epitope regions are contained in random coil configuration as shown in Figure 14. Wilson and Stanfield suggested that most peptide antibody interactions involved beta turns and beta strands. The histogram clearly indicates that Group I peptide antigens are mostly composed of random coils (87%) with less than 10% contribution from α–helix and β–sheet. In Groups II and III epitopes have higher α–helical and β – sheet content representing about 50% of the interaction. . The relative contribution of these two structures was 17% vs. 26 % for group II antigens 23%/27 % for group III.. In cases where these segments can be mimicked by peptides captured from random phage display peptide library sequences by antibodies, the retrieved sequences may provide the structural constraints need to model the epitope as well as the Ag surfaces [Jesaitis et al. 1999, Mumey et al. 2003].

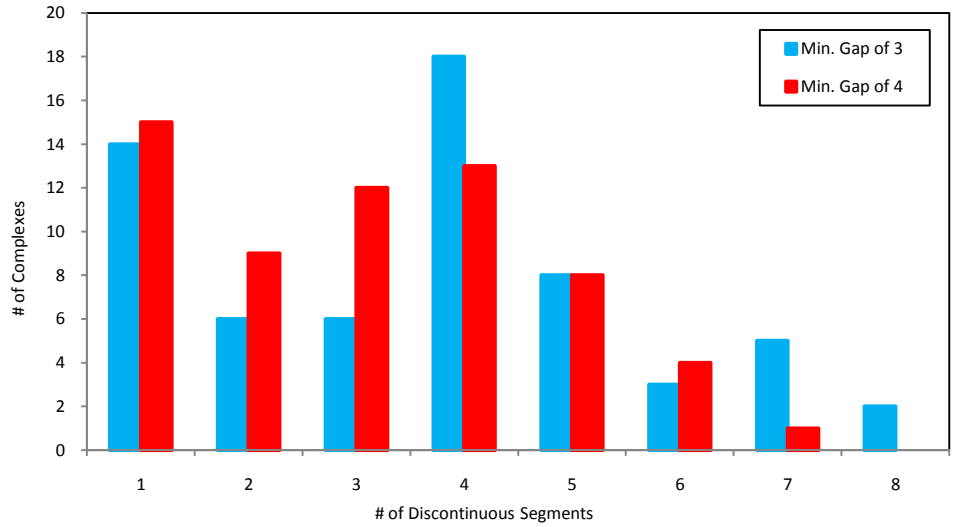


Figure 13: Epitope Discontinuity with a minimum gap distance of 3 (Blue) AAs and a minimum gap distance of 4 (Red) AAs.

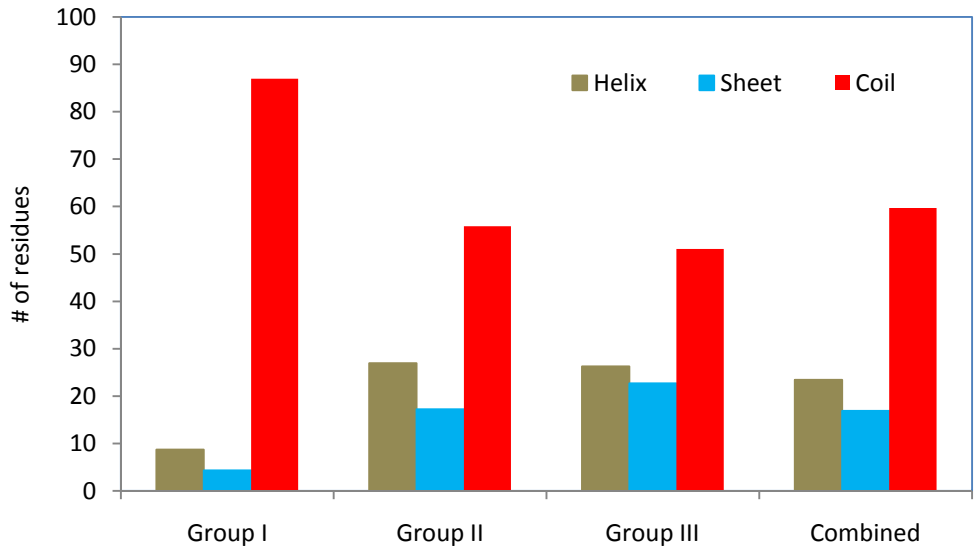


Figure 14: Percentage composition of α-helices(Tan) and β-sheets(Blue), and random coil (Red) on the epitope region of peptides, small proteins and large proteins and all combined.

Conclusions

Protein antigen-antibody (Ag-Ab) interfaces are a rich source of information about protein-protein interactions and provide important structural information about protein antigens and how they are recognized and bound by antibodies. A data set comprised of the 3D structures of 62 non-redundant Ag-Ab complexes, from the Protein Data Bank (PDB), was assembled and used to determine the general physical as well as biochemical features of the Ag-Ab interfaces. For the entire group of structures, we found that the average Ag-Ab interface had: 1) a combined solvent accessible surface area of 2073± 459 $Å^2$ with approximately equal contributions from both the antigen epitope and antibody paratope surfaces; 2) a maximum dimension of 28 ± 8Å, gap volume index 2.5 ± 2.1 Å , planarity of 2.2 ± 0.5 Å, eccentricity of 0.7 ± 0.2, 12.3 ± 19.1 amino acid residues, and 21 ± 17 hydrogen bonds; and 3) epitopes consisting of primarily discontinuous regions with at least 3-4 residue gap lengths with 10-20% beta sheet or alpha helical secondary structural elements. We also found that the least abundant (<3.5 mole %) epitope residues were CYS, PHE, MET, ILE and HIS were all distributed closest (5-9 Å) to the center of the epitope. The most abundant epitope residues (>7.5 %), on average, were ASP, GLU, ARG, ASN, and LYS and were found 9-10 Å away from the epitope center. The five greatest normalized interaction frequencies of epitope residues with paratope residues were found with ARG>ASN>LEU> GLN>GLU. Paratope residues CYS, MET, ALA, LEU, VAL had the lowest molar abundance while TYR>TRP>ASN>ARG>PHE>HIS>SER had the greatest specific frequency of interaction with epitope residues. The amino acid residues PHE, TRP, LEU, TYR AND

HIS were distributed closest (8-9 Å) to the center of the paratope. The most abundant paratope residues (>7.5 %), on average, were ASP, SER, THR, and TYR and were found from 8.5 to 11.5 Å away from the epitope center. Because of the high molar abundance, high specific frequency of interaction, low substitutability, and relatively tight and centric distribution of Tyr we conclude that it plays a central role in Ag-Ab binding. We also conclude that interfaces of epitopes and paratopes use a different set of amino acid residues to establish their highest frequency contacts. Both have more hydrophobic residues nearer the interface centers. The highest contact epitope residues include both hydrophobic and positively charged species while corresponding paratope residues were primarily aromatic. Since most of the low abundance high frequency of interaction residues have been previously identified as components of protein-protein interaction hotspots, we surmise that antigen epitopes rely on one or at most two hotspots for their interaction in the context of an interacting distributed network of complementary but non-substitutable amino acids from the epitope and paratope with TYR playing an especially important role in the paratope surface.

We have generalized the physical and biochemical characteristics of antibody antigen interfaces from the structures of 62 non-redundant complexes. Our analysis suggests that the interfaces have maximum dimensions of 29Å by 19Å with sufficient convolutions to double the surface area suggested by the product of the dimensions without producing significant dead volume between surfaces. Over this fairly large epitope surface there are significant discontinuities in sequence segments with minor to moderate secondary substructure. There is also little interdigitation of side chains as the

planarity of the surfaces is relatively high. For the peptide antigens, the antibody probably presents itself as a ridge that can contact multiple peptide structures such as alpha helicies while for larger antigens the antibody appears to form a cup-like or concave surface that utilizes its edges to help hold on to the antigen with multiple interactions suggesting that the affinity of antibody for epitope results from a distributed structure including its edges. The composition of the surfaces is such that at most one or two residues especially MET, CYS, PHE, THR, and TYR provide hotspots of high interaction frequency with high specificity that may contribute importantly to antigen recognition. TYR plays an especially important role in the paratope surface as it has the highest relative representation and interaction frequency but lowest Pearson cross correlation of interaction and may function as peg or probe that anchors the epitope to the paratope and multiple contact points. We also surmise that higher order structures, involving groups of amino acid residues probably play a very important role in recognition. Based on the interaction frequencies of different residues it may be possible to predict antigen epitope structure from limited knowledge of proximities of component amino acid residues.

CHAPTER 3

EPIMAP APPROACH: NEW ALIGNMENT SCORING MECHANISMS AND MODIFIED DYNAMIC MULTIPLE SEQUENCE ALIGNMENT

Introduction

A large fraction of protein structures of interest cannot be solved by traditional structural biology techniques such as X-ray crystallography and NMR (Nuclear Magnetic Resonance). Antibodies can either recognize continuous or discontinuous epitopes, but virtually all epitopes that have been analyzed in detail are discontinuous. Discontinuous epitopes can potentially provide extremely useful structural information because with suitable analysis they could reveal distant segments of primary sequence that are in close proximity on the native, folded protein and could reveal changes in protein structure in different functional states when appropriate antibodies are available [Padlan, E 1996]. In this approach peptide probes selected experimentally from a random peptide library to have a high affinity to antibodies of interest. The computational problem addressed in our previous work on this issue was to align each probe individually to the target protein. These alignments were performed with a program called EPIMAP [Mumey et al., 2003].

EPIMAP Approach - Background

The core idea of the antibody imprint method is that “probe” peptides that bind to the active region of a specific antibody are expected to be highly similar to the binding site of a protein that also binds to the same antibody. The computational problem seeks to align

the probe amino acid sequence, ***s***, to one or more regions of the "target" protein amino acid sequence, ***t***. Typically, ***s*** is in the range of 7-20 amino acids long and the target protein sequence, ***t*** is several hundred amino acids. There is a probability that interacting amino acids in the contact region might be substituted with chemically similar amino acids. In addition, unlike traditional string alignment problems, we allow for localized sequence rearrangements. Possible rearrangements capture cases where loops of the linear protein sequence may be pinched together with sequence inversions to form the antibody epitope binding site. Additionally, it is possible for local rearrangements of amino acids to occur, reflecting the fact that the binding site of an antibody may be a complex surface, not just a linear sequence. As such, the problem is outside the scope of classical string alignment algorithms such as Smith-Waterman.

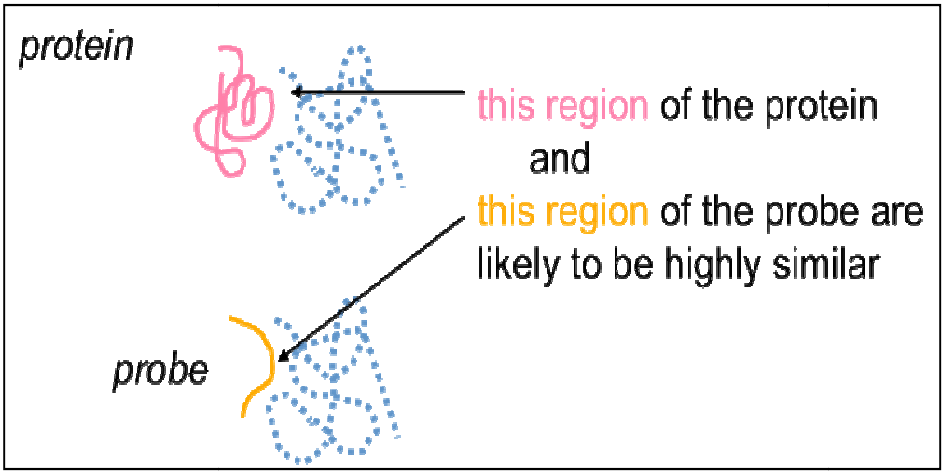


Figure 15: Strongly binding peptide probes are sequenced from selected phage DNA clones. These probes serve as "witnesses" to the structure of the target protein.

We have chosen an initial approach based on a general combinatorial alignment problem. In general, we will allow any permutation of the probe sequence to align to the underlying protein sequence. Furthermore, gaps will be permitted in both probe and target sequences. Large gaps can occur when aligning the probe to the target sequence when as frequently found experimentally, the epitope is discontinuous. We also allow unaligned probe residues, reflecting the possibility of a non-specific residue insertions in the probe (such as might occur if there were an extended sheet conformation and some amino acid side chains would point away from the binding site on the protein surface. To be a valid alignment, each probe position and target position can be used at most once per mapping. Formally, an alignment A consists of a sorted set $P_A = \{i_1 < i_2 < \ldots < i_k\}$ and another set $T_A = \{j_1, j_2, \ldots, j_k\}$ with the interpretation that the i_p-th probe residue $s(i_p)$, is aligned to the j_p-th target residue, $t(j_p)$, for $1 <= p <= k$. We adopt a two-part scoring system to evaluate the quality of alignments. The scoring system is composed of a substitution score and an epitope gap cost and break point cost,

$$score(A) = S(A) - G(A) - B(A).$$

The $S(A)$ component is calculated with a substitution matrix M, similar in principle to a Dayhoff matrix, used in other protein alignment contexts. The substitution matrix is also used to score unaligned probe residues; if the probe residue in position i is not aligned to any target position it is charged a penalty according to the character c occurring in position i of the probe sequence.

This cost coded in the substitution matrix, in the entry *M(c, -)* and must be parameterized. We have:

$$S(A) = \sum_{p=1}^{k} M\,(s(i_p), t(j_p)) + \sum_{probe\ positions\ i \notin PA} M(s(i), -)$$

$$G(A) = \sum_{j=1}^{k-1} d(|j_{p+1} - j_p|)$$

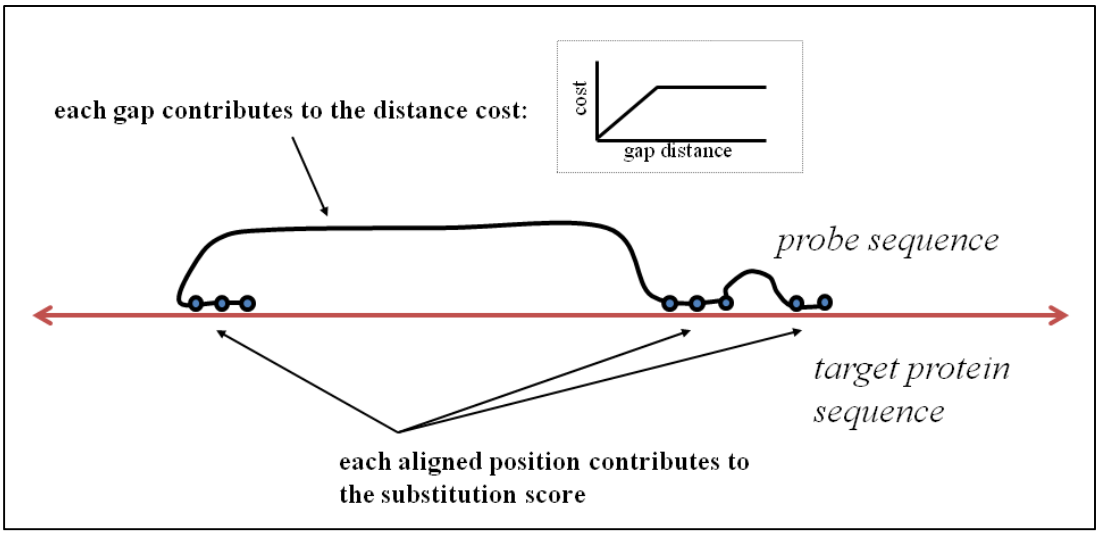


Figure 16: EPIMAP Approach Scoring Mechanism.

The epitope gap cost *G(A)* is calculated by examining the number of amino acid residues skipped along the target protein sequence between successive aligned probe positions: where *d(x)* is the cost of skipping x amino acids along the target between successive mapped probe positions.

We introduce a break point score *B(A)* fit each probes aligned to the target. We add break points on to a probe based on the alignment. Circular probes with constrained conformations (disulfide bonded loops) are sometimes used in random sequence libraries and so we also include the term $d(|j_k - j_1|)$ in the above sum in the circular case. In some cases, e.g. membrane spanning proteins, it may be known or surmised that certain regions of the target protein are inaccessible to antibodies and thus might be excluded from consideration as potential alignment positions. The computational problem is thus to find finding an alignment *A* that maximizes *score(A)*.

As a point of departure to systematically seek the best alignments of each peptide probe sequence to the target protein sequence, we developed an algorithm and pilot software in initial studies (Mumey et al., 2003), called EPIMAP. The gap penalty function and the substitution matrix using in EPIMAP were initially assessed by analyzing Antibody Imprinting data previously derived from actin [Jesaitis et al., 1999] and analyzed at that time by visual inspection. We used a substitution matrix developed by Bordo and Argos for surface-exposed residues [Bordo and Argos, 1991] and have experimented with other common matrices used for sequence comparison (Dayhoff, PAM, etc).

Investigation of the Specificity and Substitutability of Antigenic Epitope Residues

One of the key components was to the mine the PDB (Protein Data Bank, Berman et al, 2000) for unique antigen-antibody complexes to learn as much as possible about the substitutability of antigen residues when bound to an antibody (Chapter 1). This should allow us to refine the alignment scoring and improve on epitope prediction considerably. So we analyzed 62 unique co-crystallized antibody-antigen complexes from the PDB (Table A1, See Appendix for our Data Set). In chapter 1 we computed the pairwise interactions in the interface area for the same data set. Specifically we can count the number of times C(X, Y) residue pairing occurs where X is antigen residue and Y is an antibody residue. This will let us build a better substitution-scoring matrix as follows:

Let P(i, j) be the probability that an epitope residue i pairs with antibody residue j,

$$P(i,j) = {C(i,j)}\Big/{\sum_{k=1}^{20} C(i,k)}$$

Then we can define the overall likelihood that a given probe i residue maps to the same unknown residue k on the antibody that a target residue j does as

Table 6: Derived Substitution Matrix

Antigen Epitope Surface(Ag)	Antibody Paratope Surface (Ab)																			
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
A	0.005	0.024	0.020	0.104	0.131	0.060	0.040	0.039	0.067	0.038	-0.039	0.044	-0.063	-0.035	0.058	0.009	-0.016	-0.038	-0.064	0.149
R		0.160	0.115	0.171	0.210	0.160	0.118	0.117	0.181	0.131	0.041	0.156	-0.001	0.010	0.132	0.085	0.083	0.035	0.022	0.205
N			0.140	0.214	0.280	0.178	0.148	0.162	0.198	0.163	0.058	0.160	0.051	0.072	0.199	0.120	0.062	0.072	-0.022	0.255
D				0.334	0.421	0.273	0.259	0.248	0.276	0.242	0.135	0.221	0.109	0.177	0.291	0.202	0.115	0.124	0.042	0.372
C					0.632	0.373	0.341	0.368	0.379	0.308	0.201	0.299	0.155	0.268	0.400	0.332	0.122	0.101	-0.051	0.465
Q						0.251	0.220	0.224	0.254	0.206	0.102	0.214	0.062	0.097	0.250	0.181	0.085	0.051	-0.023	0.325
E							0.220	0.194	0.219	0.173	0.095	0.166	0.037	0.106	0.228	0.145	0.065	0.028	-0.010	0.305
G								0.248	0.242	0.176	0.073	0.193	0.036	0.100	0.253	0.187	0.051	0.031	-0.083	0.306
H									0.286	0.223	0.114	0.229	0.092	0.131	0.271	0.196	0.107	0.078	-0.015	0.331
I										0.207	0.085	0.189	0.108	0.089	0.234	0.137	0.067	0.069	-0.037	0.297
L											0.038	0.079	0.003	0.032	0.133	0.033	0.000	-0.041	-0.075	0.172
K												0.217	0.064	0.027	0.226	0.151	0.082	0.053	-0.051	0.284
M													0.154	0.054	0.157	0.002	-0.045	-0.036	-0.115	0.166
F														0.163	0.170	0.051	-0.028	-0.016	-0.046	0.184
P															0.319	0.196	0.070	0.065	-0.064	0.350
S																0.154	0.023	0.012	-0.094	0.253
T																	0.049	0.036	-0.001	0.142
W																		0.164	0.018	0.116
Y																			0.061	0.013
V																				0.443

$$S(i,j) = \sum_{k=1}^{20} P(i,k)P(j,k) = \sum_{k=1}^{20} P(i,k)P^T(k,j)$$

Written as a matrix equation, $S = P \cdot P^T$. The matrix S (Appendix Table A8) should be a good candidate for a substitution matrix since it expresses the *a priori* likelihood that an aligned probe and target residue would bind to same (unknown) antibody residue. We present this matrix S in logarithm form and add a constant λ for calculations of sequence similarity.

$$M(i,j) = log_{10}(S(i,j) + \lambda), we\ chose\ \lambda\ to\ be\ equal\ to\ 1$$

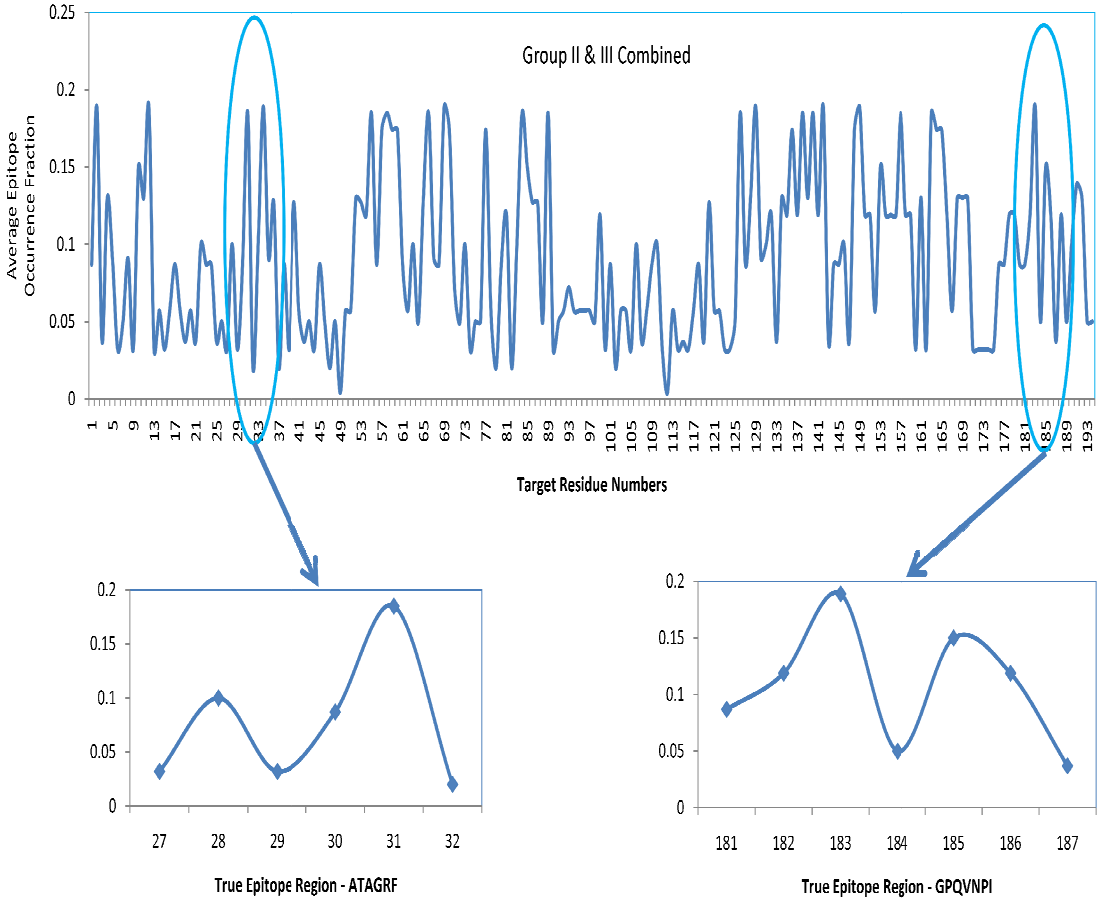
M (Table 6) is a symmetrical matrix and used in our alignment mechanisms.

Investigation of The Average Epitope
Amino Acid Residue Occurrence Probability

In this section we use the Average Epitope Occurrence Probability computed in Chapter 2 presented in table A6 (See Appendix). We consider the group II and III combined average occurrence probability and is graphically represented in figure 6 (a,b) . This occurrence fraction is used against target protein amino acids. Each target position gets a score, which is its average occurrence fraction in the epitope surface. Specific positions along the target that has high scores will be picked up suggesting that these regions might be part of the true discontinuous epitope.

We tested this on p22 (phox) protein to see if the published epitope, 28-ATAGRF and 182-GPPQVNPI was identified and also on the primary amino acid sequence of recombinant human IL-10, interleukin protein which is 160 amino acids long. The IL-10 epitope is discontinuous and the two main epitope segments are considered to be from 71-83 and 125-137. The results are shown in figures 14, and 15.

(a)



(b)

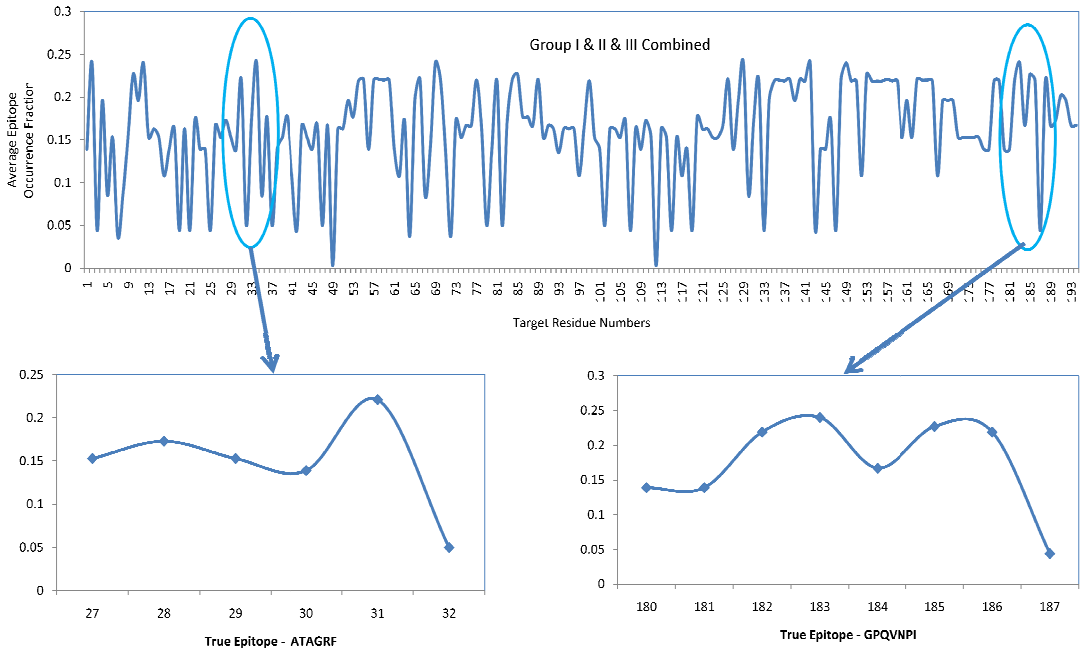
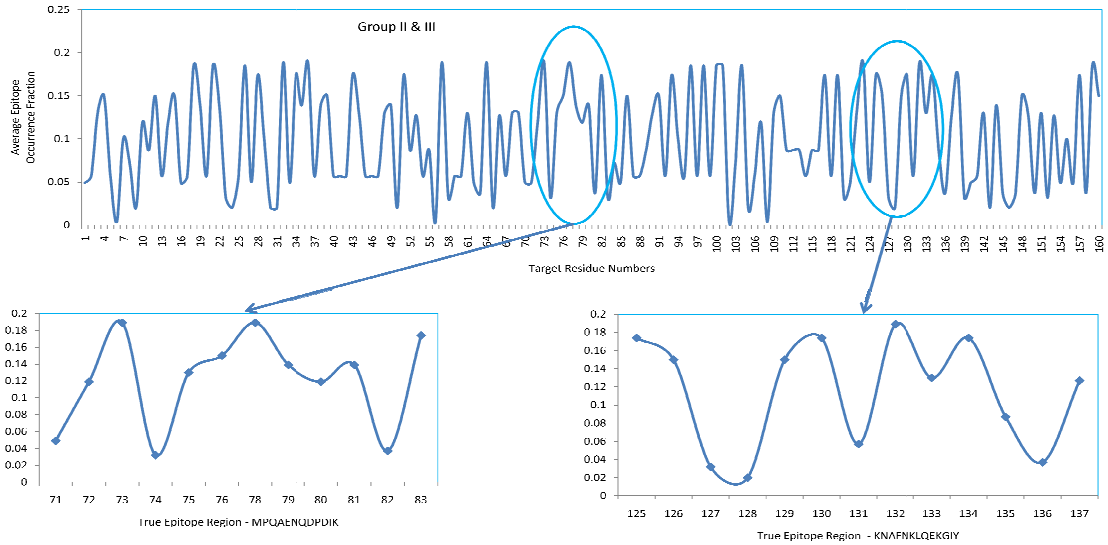


FIGURE 17 : p22 (phox) protein target sequence where each AA position in the target sequence is plotted with its average epitope occurrence Probability values. (a) Using values from Group II & III Combined, (b) Using Values from groups I, II, & III Combined

(a)



(b)

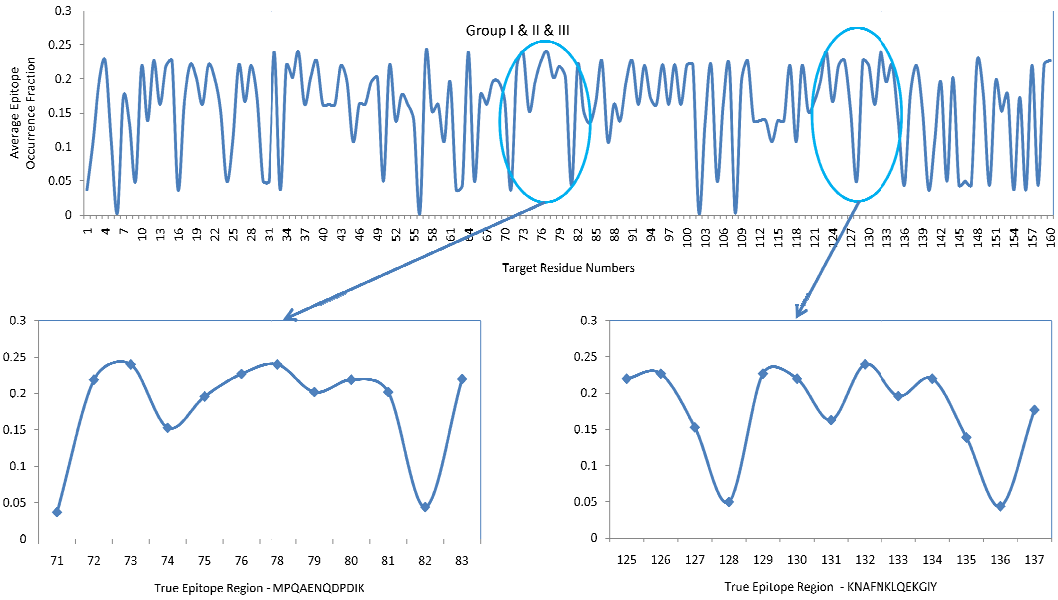


FIGURE 18: IL-10 protein target sequence where each AA position in the target sequence is plotted with its average epitope occurrence fraction value. (a) Using values from Group II & III Combined, (b) Using Values from groups I, II, & III Combined

Different approaches in Improving Epitope Alignment and Mapping Algorithms

Simple Scoring Mechanism

Here, a simple antigen epitope prediction method using the above derived substitution matrix from the frequency of interaction matrix (Table 4, Chapter 1) is described. Each probe sequence is divided into -tuple (={1, 2, 3, 4, 5, 6, 7, 8, 9}) compositions (Figure 16). Each -tuple and its reversal of the probe sequence are aligned to the target sequence. The scoring of the alignments come from the substitution matrix (Table 3) For =1, the residue in the probe that has highest substitution score is aligned

to the target residue and the target residue gets the substitution score. For k = {2, 3, 4, 5, 6, 7, 8, 9} tuples, each k tuple and its reversal is aligned to the target. Each k tuple score is the sum of each amino acid residue substitution score, and the highest score is assigned to the k tuple position of the target sequence. For a target of sequence of length l, when k =1 each target position will get a score, for k > 1, each overlapping k tuple along the target sequence will get a maximum score, in this case the score length will become $(l - (k - 1))$. Also for k >1 each position on the target sequence the score is replaced with the average of the overlapping k tuple alignment, this will bring back the score length to the original target length which is l. An example of the scoring for k =1 and k =2 is shown in figure 17.

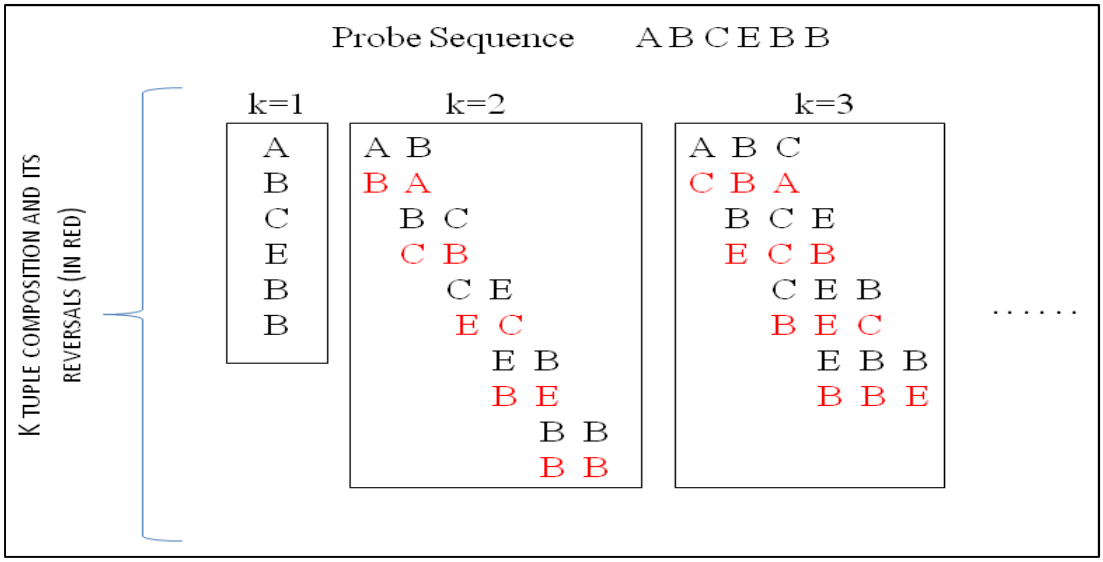


Figure 19: Example showing a kmer spectrum of a probe sequence for k=1, 2, 3

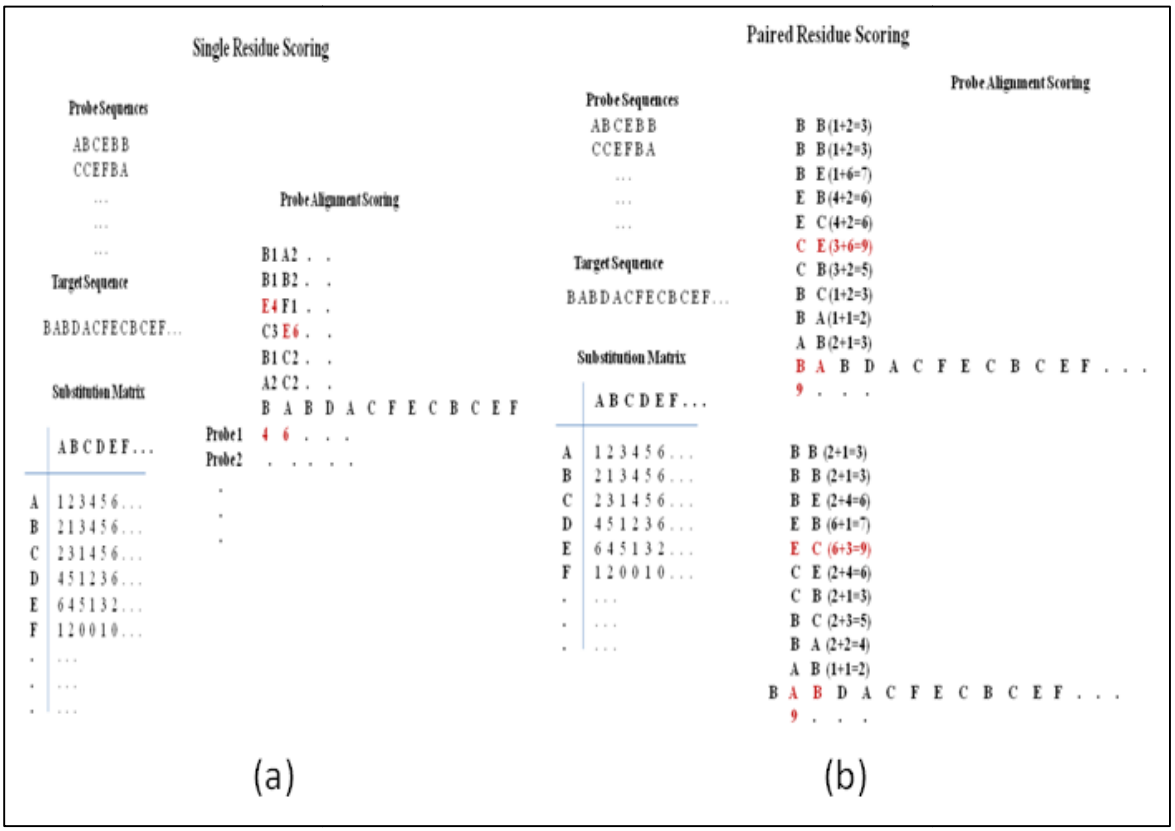


Figure 20: (a) Single residue scoring mechanism and (b) Paired residue scoring mechanism.

This simple scoring mechanism and the substitution matrix was used on p22 (phox) protein and antibody 44.1 to see if the published epitope for this antibody, 28-ATAGRF and 182-GPPQVNPI was identified. The program clearly identified the region 182- GPPQVNPI region. The results are presented in the following figure 21 and 22. Also this simple scoring mechanism and the substitution matrix was used on IL-10 protein and antibody 9D7 to see if the epitope for this antibody, 33, 55, 59, 60, 74 - 75, 78 - 79, 82 - 83, 117, 119, 125 - 126, 129 - 135, 136 - 137 was identified. The results are presented in the following figure 23 and 24.

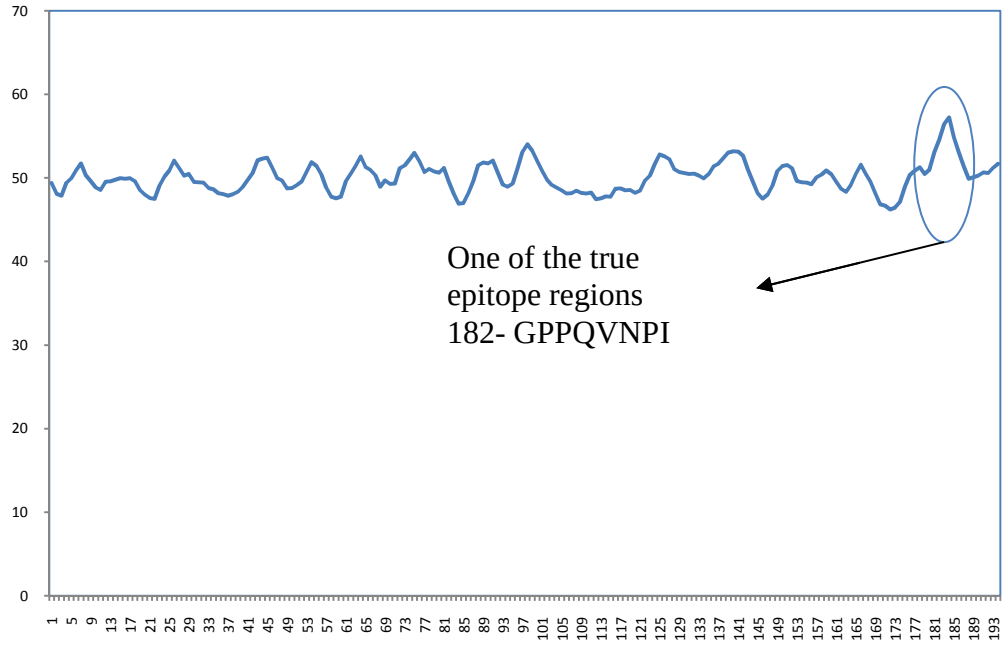


FIGURE 21: 44.1 antibody probes aligned to p22 (phox) target protein using the scoring mechanism with k tuple size of 4 and finding the average of the overlapping k tuples. The graph clearly indicates a spike in the epitope region 182 - 190

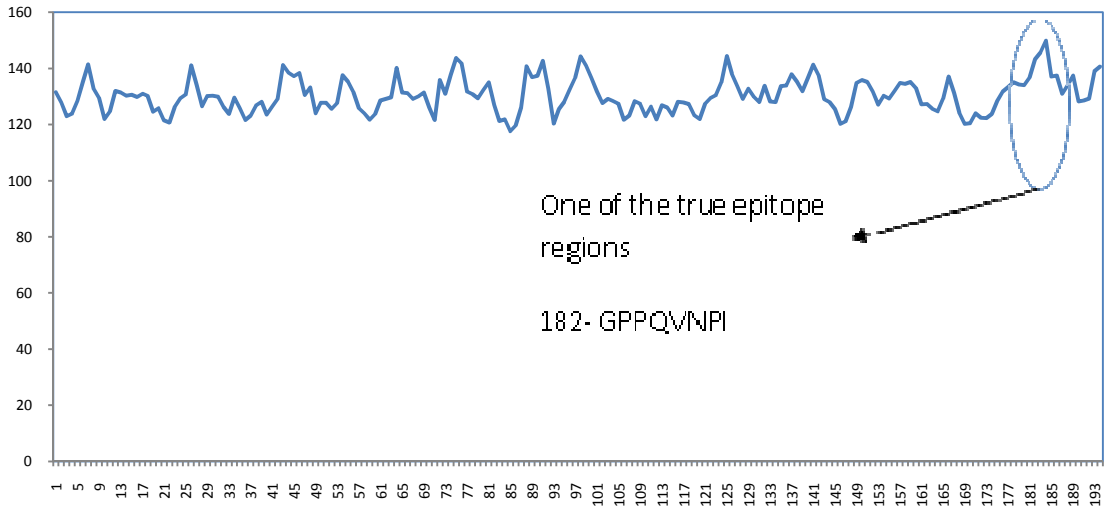


FIGURE 22: 44.1 antibody probes aligned to p22 (phox) target protein using k values 1, 2, 3, and 4 and then summing all the values at each position in the target. This approach did not produce any better result than using a k tuple of 4, but still showed a spike in the true epitope region (182 – 190).

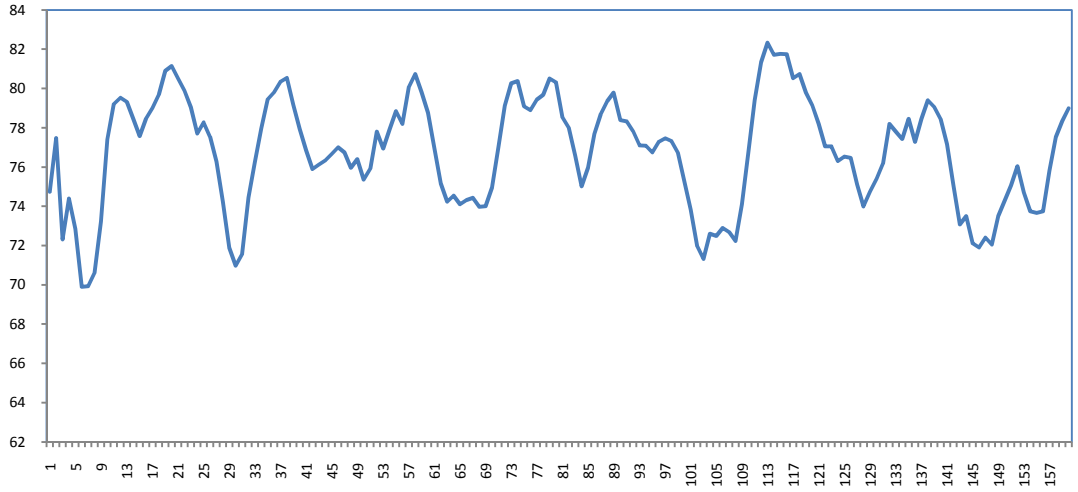


Figure 23 : 9D7 antibody probes aligned to IL-10 protein target using the scoring mechanism described above with k tuple size of 4 and finding the average of the overlapping k tuples.

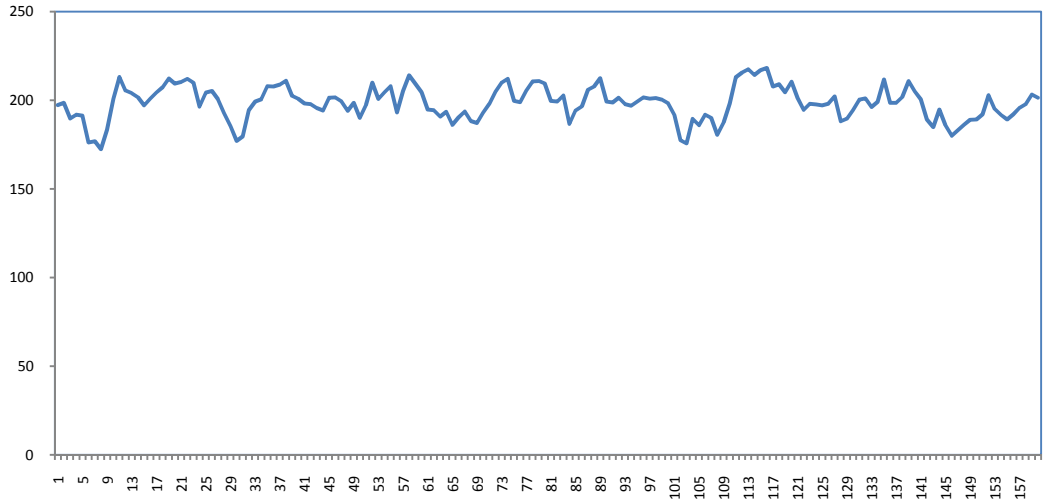


Figure 24 :9D7 antibody probes aligned to IL-10 protein target using k values 1, 2, 3, and 4 and then summing all the values at each position in the target.

Even though Epitope mapping using simple scoring ideas mentioned in the above sections, 1) using the average epitope amino acid residue occurrence fraction, and 2) the overlapping k tuple alignment method slightly picks up certain regions of the epitope does not cleary predict the epitope regions. Hence in the following section we provide an improved EPIMAP algorithm apparoach MSA – EPIMAP that uses multiple sequence alignment.

Modified Dynamic Multiple Sequence Alignment Approach

As mentioned previously the principal goal of protein sequence alignment is to discover biological similarities among proteins. Multiple sequence alignment can be a useful technique for studying and analyzing sequence-structure relationships. So this approach remains an important area of research as biological inferences can be made from the conservation or variation within the aligned positions, especially with reference to the structure of at least one of the aligned sequences. In protein sequence alignment, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region is among lineages. The absence of substitutions, or the presence of only very conservative substitutions in a particular region of the sequence, suggests that this region has structural or functional importance. In the previous EPIMAP approach only two sequences are aligned at a time, one probe sequence at a time against the target sequence. Multiple sequence alignment incorporates more than two sequences at a time. In our approach the goal would be to align all the probes sequences against the target sequence at once. This method should allow improved identification of the conserved sequence across the target.

This conserved sequence can be used in conjunction with the structural information of the target protein.

Multiple sequence alignment problems are computationally difficult to produce and most formulations of the problem are NP complete combinatorial problems. We modify the EPIMAP approach to solve this probe-target alignment to find optimal and suboptimal alignments.

Methodology

In order to improve the efficiency of the alignment, in this new approach a probe is randomly selected from the probe list and is aligned to the target sequence. Once a random probe is aligned it is removed from the probe sequence list and the target gets appended with the aligned probe. In the next step another random probe is selected from the probe list and is aligned to the new appended target. This process is repeated until all the probes in the list are aligned to the target. For each random probe alignment to the target, the same scoring mechanism from EPIMAP is used (described above in the EPIMAP approach section) except at each target position we take the sum of pairs score of the probe to the target plus a constant α (refered to as the *sop Factor*) times the probe to the other probes aligned to the target at that position.

$$\{target\ sop\ score + \alpha\ (probes\ sop\ score)\} - gap\ penality - bp\ cost$$

$\{target\ sop\ score + \alpha\ (probes\ sop\ score)\}$, for each target position is stored and printed out at the end for each random alignment. The main idea is to build up the alignment in a scaffolding manner based on the probes that are already aligned to the target sequence. At the end this alignment gets a score, which will be

$$\sum sop - \sum gp - \sum bp$$

Where sop score is the sum of pairs at each position in the alignment, and gp is the total gap length cost, and bp is the total break point cost. This random ordering of alignments is repeated over several (usually 80 – 100) times and the alignment with the highest score is selected. This method of construction of alignments of probe sequences to the target protein sequence would lead to better epitope predictions.

Searching Best Parameters

The program takes in several parameters, maximum gap cost, gap extension cost, cost for deletion in a probe, sum of pairs factor, break point cost. We run our program for different combination of these parameters to get the best alignment.

APX – HARDNESS of (MSA) EPIMAP Problem

A reduction from MAX - 3SAT is used to show the APX hardness of the original EPIMAP alignment of a probe to a target problem [Garey, and Johnson, 1979]. We show that MAX – 3SAT $\leq_p$ EPI – ALIGN $\leq_p$ MSA – EPI - ALIGN

To begin with, we formally define the original EPI - ALIGN problem:

EPI- ALIGN Problem:

Input: A probe string s, a target string t (over a common alphabet), a substitution M, a distance penality function d, an objective score Q.

Output: A decision on whether there exists an alignment with score at least Q.

Theorem: MAX – 3SAT $\leq_P$ EPI – ALIGN $\leq_P$ MSA – EPI - ALIGN

Proof: We show that EPI - ALIGN is APX – HARD via a polynomial time reduction from MAX - 3SAT. Consider an instance of MAX – 3SAT, $I_{MAX\text{-}SAT}$ consisting of a collection of clauses $\mathrm{C} = \{\mathrm{c}_1, \mathrm{c}_2, \mathrm{c}_3, \ldots, \mathrm{c}_\mathrm{m}\}$ on a finite set of variables $U = \{x_1, x_2, x_3, \ldots, x_k\}$. We will construct an instance I_A (u) of EPI - ALIGN such that *(m – u)* clauses in $I_{MAX\text{-}SAT}$ are satisfiable then there exists an alignment for I_A (u). We construct I_A as follows: The string alphabet used is

$$A = U \cup \{\neg x_1, \ldots, \neg x_k\} \cup \{c_1, \ldots, c_m\} \cup \{y_1, \ldots, y_k\} \cup \{\#, *, @\}$$

All entries of M are set to - ∞ except the following: $M(\alpha, c_i) = 0$ is a literal in clause c_i, $M(x_i, y_i) = M(\neg x_i, y_i) = k$ for all $1 \leq i \leq k$, and $M(., \text{'*'}) = 0$ (here "." represents any symbol).

For each literal α, let $[\alpha]$ be the multiplicity of α among all clauses in C. The probe string used is

$$s = @\, B_1\, B_2 \ldots B_k\,, where\ B_i = \underbrace{x_i \ldots x_i}_{[x_i]+1}\ @\ \underbrace{\neg x_i \ldots \neg x_i}_{[\neg x_i]+1}\ @$$

The target string used is

$$t = \underbrace{**** \ldots *}_{|s|-(m+k)+u} \underbrace{\#\#\#\#\# \ldots \#}_{|s|-(m+k)+u} c_1\, c_2 \ldots c_m\, y_1\, y_3 \ldots y_k$$

The distance penalty function used is

$$d(l) = \begin{cases} 0, & \text{if } l \leq |s| - (m + k) + u \\ 1/2, & \text{otherwise} \end{cases}$$

Observe that $m + k < |s|-(m+k)+u$, so only jumps across the central gap of #'s, referred to as the *bridge*, will contribute to the gap cost. The leading @ of *s* forces any finite-score alignment to begin on the left side of the bridge. Note that every non-# letter in the target must be matched in order to completely align the probe (all probe positions must be matched as $M(\cdot, -) = -\infty$). In order to match all of the y_i 's, at least one literal from each **B***i* must be used. Thus, each $\boldsymbol{B}_i$ contributes at least one return jump across the bridge. If a literal is matched against a clause symbol c_i , then any truth assignment that makes this literal true will satisfy c_i . We choose $Q = k(k-1)$ to insist that each $\boldsymbol{B}_i$ binds its corresponding literal and contributes *two* jumps across the bridge. Because the positive and negative literals in each block $\boldsymbol{B}_i$ are separated by an @, only literals of a single polarity can be matched to symbols to the right of the bridge. This ensures a consistent truth assignment. Thus, any alignment with score exactly $k(k-1)$ will produce a satisfying assignment for $I_{MAX\text{-}SAT}$ and vice versa. ■

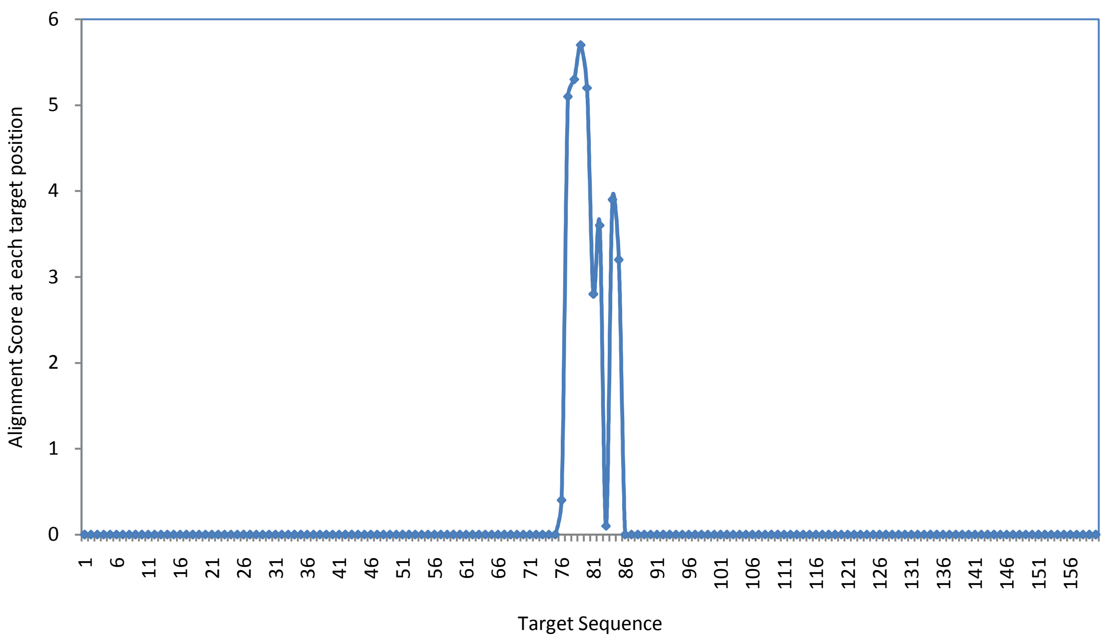
Experimental Results

To validate our new MSA – EPIMAP approach, we selected a model system where the structure of the antibody epitope was previously known. We chose the interleukin protein IL10 and antibody 9D7. The primary amino acid sequence of recombinant human IL-10 is 160 amino acids long. The structure of the antibody-antigen complex of 9D7-IL10 was determined by x-ray crystallography (PDB: 1lk3.pdb). From the x ray crystal structure of the antibody (9D7) bound to its antigen (IL-10), the molecular contacts were determined using CPP4 (Padlan 1996). The epitope for antibody 9D7 was then mapped employing the antibody imprinting method. Peptides that mimic the epitope on IL-10 that the antibody 9D7 binds to were selected from a random peptide phage display library. The 9D7 epitope on IL-10 is discontinuous; it is composed of two regions in the primary sequence that are close together in the folded protein but are not contiguous in the primary amino acid sequence of IL-10. Antibody 9D7 binds to two regions of IL10 composed of residues 71-83 and 125-137. The amino acid sequences of the selected peptides (probes) were then aligned onto the primary sequence of the target IL-10 using MSA - EPIMAP. We ran this data on MSA – EPIMAP for a combination of different parameter set. For Each parameter set on the search space grid we did a number of random ordering of probe alignments and picked up the alignment with the highest score (Figure 25).

```
Producing mapInfo
EPIMAP3 Alignment Results
Parameters:
costSlope: 0.5 costMax: 1.0 bpCost: 0.5 delCost: 1.0 sopFactor: 0.25
sub matrix: Latest Scoring Matrix Derived From The FIM Matrix
                                                                          WRIPDNGMG                                                                              probe GMGNDPIRW bp 0
                                                                           PKQQDSRRW                                                                             probe WRRSDQQKP bp 0
                                                                          WRQTDRSAT                                                                              probe TASRDTQRW bp 0
                                                                           EKEYWRIQW                                                                             probe EKEYWRIQW bp 0
                                                                           SIGWLRHLN                                                                             probe NLHRLWGIS bp 0
                                                                          WGTVS                                                                                  probe SVTGW bp 0
                                                                           VKGLPHLNY                                                                             probe YNLHPLGKV bp 0
                                                                           QKGQIQ                                                                                probe QIQGKQ bp 0
                                                                           SKSFAALFS                                                                             probe SFLAAFSKS bp 0
                                                                           SWFNSRGAS                                                                             probe SWFNSRGAS bp 0
                                                                           PLERGSAGA                                                                             probe AGASGRELP bp 0
                                                                           LQNTARARL                                                                             probe LQNTARARL bp 0
                                                                           SNVISIKNL                                                                             probe LNKISIVNS bp 0
                                                                           DQRMARLTL                                                                             probe DQRMARLTL bp 0
                                                                          WRIPDNGNG                                                                              probe GNGNDPIRW bp 0
                                                                          WRIPDNRIG                                                                              probe GIRNDPIRW bp 0
                                                                           QSFGLNTSI                                                                             probe ISTNLGFSQ bp 0
                                                                           KAWN                                                                                  probe KAWN bp 0
                                                                           KVIRNSHHE                                                                             probe KVIRNSHHE bp 0
                                                                           RLERGSAGA                                                                             probe AGASGRELR bp 0
                                                                          TLSLWWH                                                                                probe TLSLWWH bp 0
                                                                           DGSTLRLGF                                                                             probe FGLRLTSGD bp 0
                                                                           GTNEATEGR                                                                             probe RGETAENTG bp 0
                                                                           RQIEWMTAL                                                                             probe RQIEWMTAL bp 0
                                                                           GRVNRSTFR                                                                             probe GRVNRSTFR bp 0
MSENSCTHFPGNLPNMLRDLRDAFSRVKTFFQMKDQLDNLLLKESLLEDFKGYLGCQALSEMIQFYLEEVMPQAENQDPDIKAHVNSLGENLKTLRLRLRRCHRFLPCENGGGSGGKSKAVEQVKNAFNKLQEKGIYKAMSEFDIFINYIEAYMTMKIRN
                                      ^               ^   ^^             ^^  ^^  ^^                                 ^ ^     ^^  ^^^^^^ ^^                        <- known epitope pos:
1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
         1         2         3         4         5         6         7         8         9         0         1         2         3         4         5         6
                                                                                                   1

0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.4 5.1 5.3 5.7 5.2
2.8 3.6 0.1 3.9 3.2 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
Final Best SOP Score: 35.352999999999998
Final Best Gap Length Score: 0.0
Final Best Break Point Score: 0.0
Final Best Score: 35.352999999999998
```

Figure25: Actual Alignment of 9D7 Probes to IL10 Protein Target



(a)

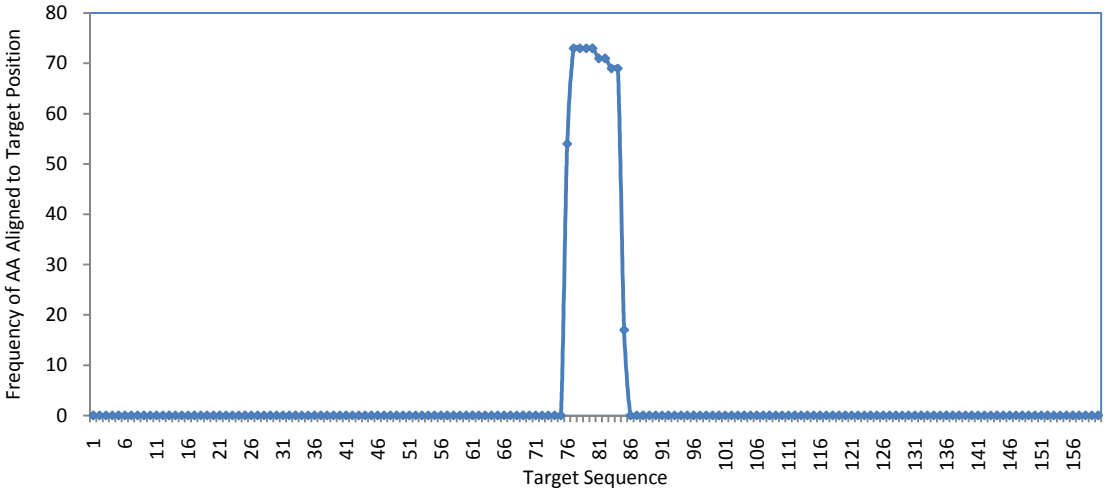
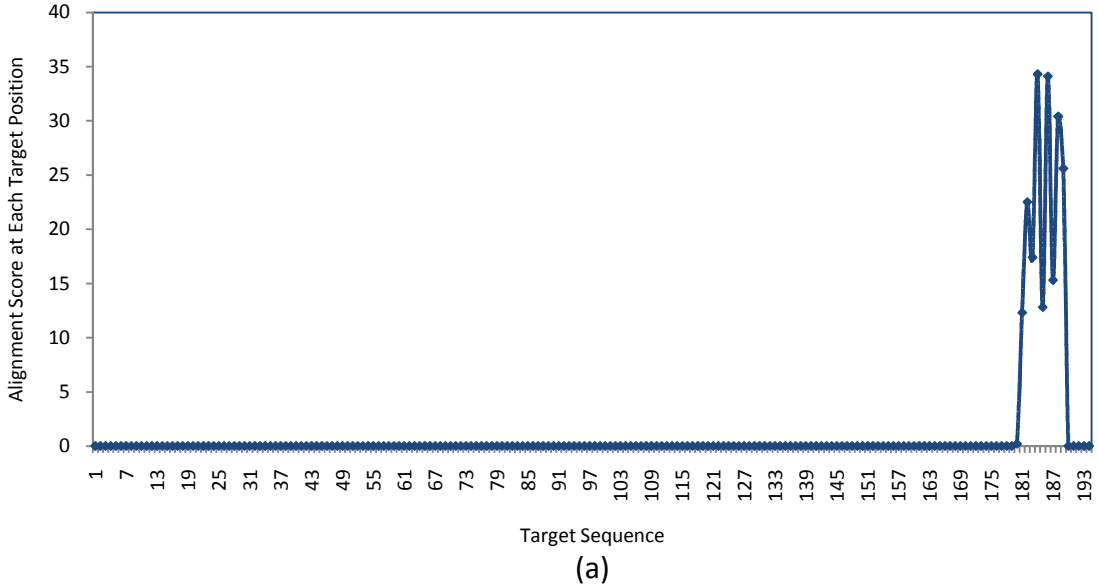


Figure 26: 9D7 Antibody Probes against IL10 Protein (a) Plot representing the scores at each target position. (b) Plot representing the frequency of amino acids aligned at each target position

We also aligned another data set, 44.1 antibody probes to p22 phox data. We know the trtue epitope region which is 28-ATAGRF and 182-GPPQVNPI. We present results in the following figure 27.



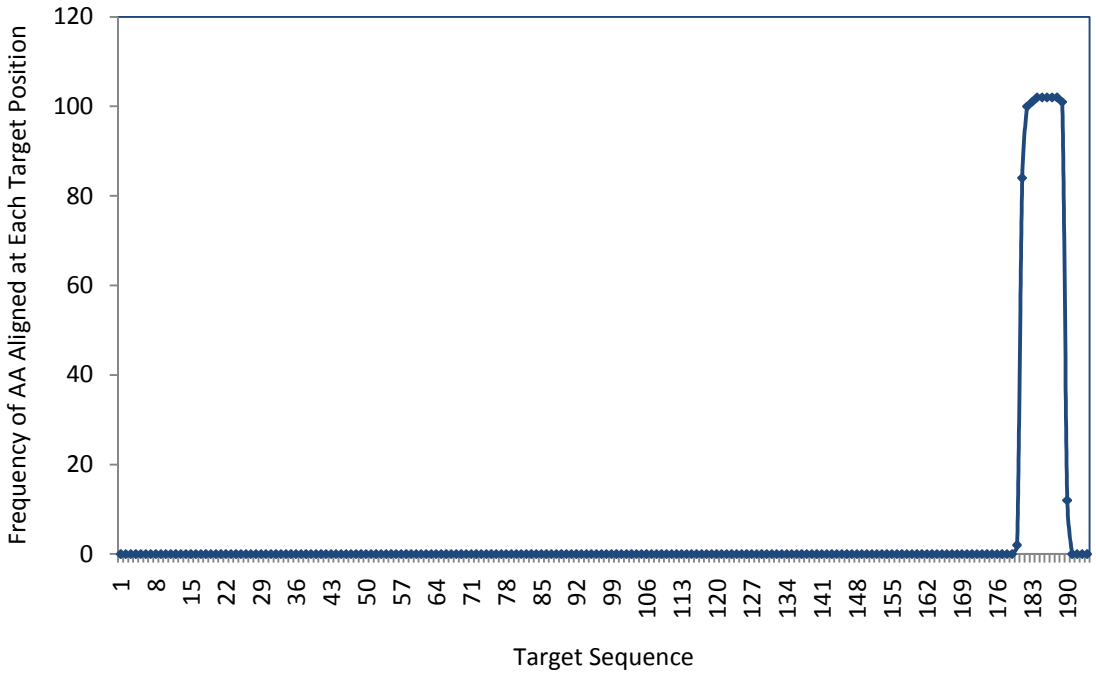


Figure 27: 44.1 Antibody Probes against p22 phox data (a) Plot representing the scores at each target position. (b) Plot representing the frequency of amino acids aligned at each target position

Alignment Comparison of MSA – EPIMAP with Original EPIMAP. Here we compare 9D7 Antibody Probes against IL10 Protein alignment using the original EPIMAP and our new improved MSA – EPIMAP. The result is shown in figure 28. Similar comparison is performed for 44.1 Antibody Probes against p22 phox data and the results are shown in figure 29. The results clearly indicate that MSA – EPIMAP approach improved the alignment and helps in better epitope identification.

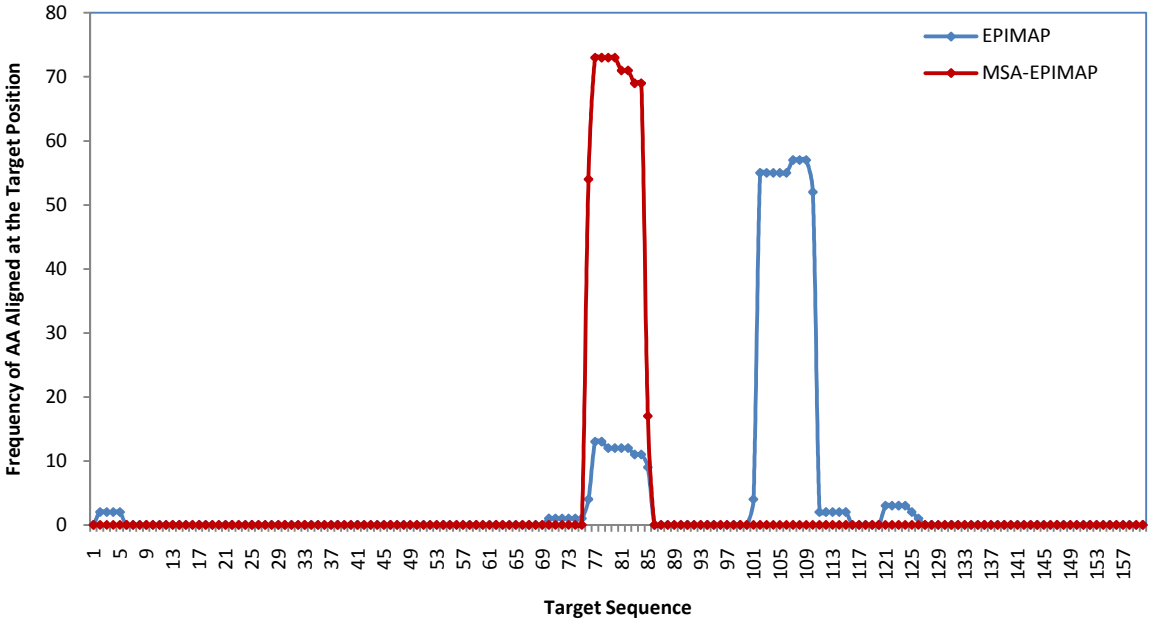


Figure 28: 9D7 Antibody Probes against IL10 ProteinTarget – Comparison between Original EPIMAP to MSA - EPIMAP

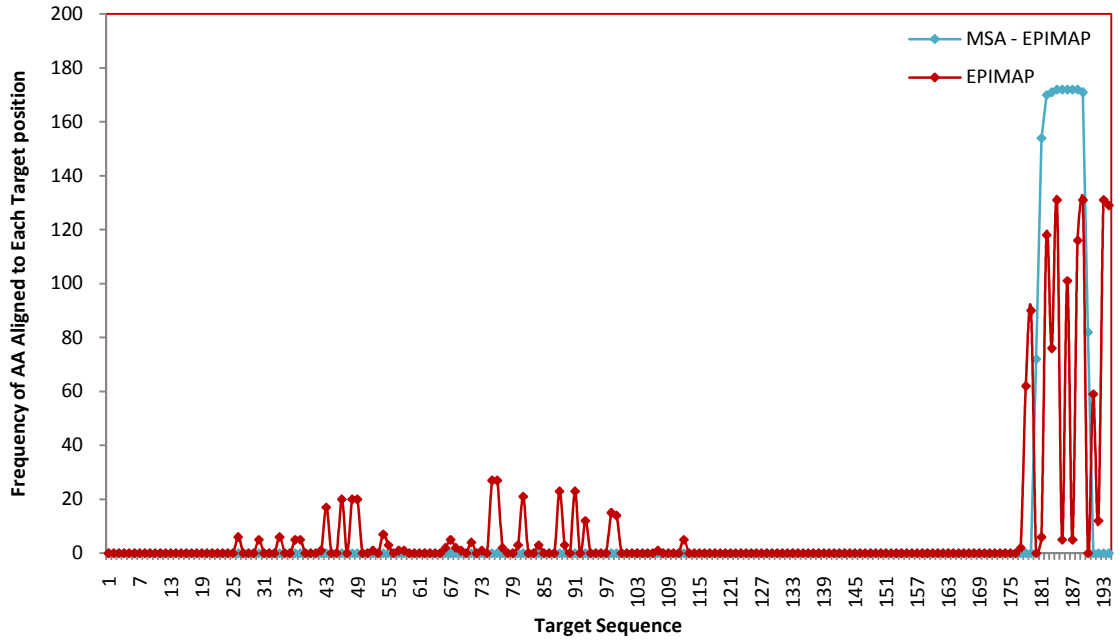


Figure 29: 44.1 Antibody Probes against p22 phox Protein Target – Comparison between Original EPIMAP to MSA - EPIMAP

Alignment Evaluation of MSA - EPIMAP. We took the best alignment from the MSA – EPIMAP approach and evaluated it. We identify the the highest scoring alignment and the parametrers for the corresponding alignment. For this alignment the false positives, and the false negatives are indentified. False positive: Predicting a region as a possible true epitope site when it is not, and False Negative: Not predicting a region as a possible epitope site when in fact it is a true epitope site. This can done only when the true epitope site is known.

$$E[i] = \begin{cases} 0, & if\ i\ \notin to\ the\ true\ epitope\ region \\ 1, & if\ i\ \in to\ the\ true\ epitope\ region \end{cases}$$

$$where, i\ goes\ from\ 1\ to\ n, and\ n\ is\ the\ length\ of\ the\ target\ sequence$$

The $\{target\ sop\ score + \alpha\ (probes\ sop\ score)\}$ score $S[i]$ for each taget position is computed, we define τ, where τ_{min} is zero, and τ_{max} is equal to the maximum score (ρ) plus a small constant ξ (1.0001).

$$\tau_{min} = 0;\ \tau_{max} = \rho + \xi$$

Interval[$\{\ \tau_{min}, \tau_{max}\}$] represents 100 equal range of values between τ_{min}, τ_{max}. For each interval point if make

$$P[i] = \begin{cases} 1, & if\ E[i] \geq \tau \\ 0, & otherwise \end{cases}$$

$$where, i\ goes\ from\ 1\ to\ n, and\ n\ is\ the\ length\ of\ the\ target\ sequence$$

Once we have $E[i]$, and $P[i]$, each position of i is compared,

$$\begin{cases} if\ E[i] = P[i] = 0\ or\ 1, correct\ prediction\ at\ position\ i \\ if\ E[i] = 0\ and\ P[i] = 1, \quad report\ as\ false\ positive \\ if\ E[i] = 1\ and\ P[i] = 0, \quad report\ as\ false\ negative \end{cases}$$

So for each interval point we be track of the number of false positives and false negatives. Finally, the false positives and false negatives are plotted as a scttor plot on the x axis and y axis and the area under the plot is calculated. This area is reported as the final alignment score and smaller the area better the alignment.

For the actual alignment shown in figure 25 for the 9D7 antibodies and IL10 protein target we evaluated the alignment and computed the area under the plot and is represented in figure 30.

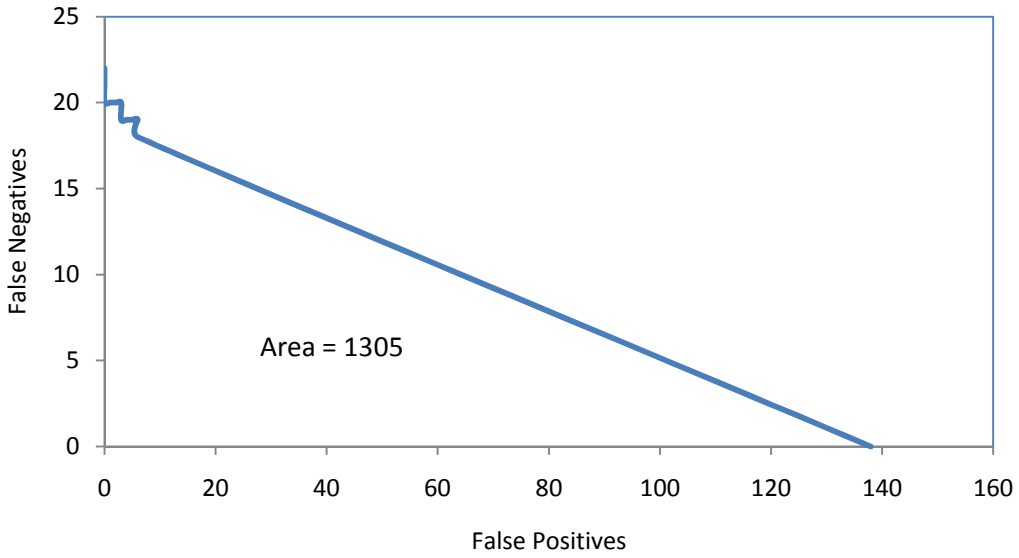


Figure 30: Plotting the false positives and false negatives as a scatter plot and the area under the plot is shown for 9D7 anibody probes against IL10 protein target.

Conclusions & Future Work

The MSA – EPIMAP problem improves upon our existing methods for epitope prediction based on antibody imprinting. By solving the alignment problem using the new derived substitution matrix (Table 6) for all probes simultaneously increased the size of the search space but lead to better epitope predictions. In our MSA approach we use a simple random localized serach approach. To futher improve the alignment we could still extensively search for the input parameters. Also trying to implement more sophisticated MSA approaches like Expectations – Maximization (EM) method, Genetic Algorithms (GAs), Simulated Annealing, Hidden Markov Models (HMM) and other available methods.

CHAPTER 4

DE NOVO GENOME ASSEMBLY

Introduction

Recent technological advances have dramatically improved next generation sequencing throughput and quality. Due to these advances it is now much cheaper to sequence entire genomes for different organisms. Genomes vary in size from millions of nucleotides in bacteria to billions of nucleotides in humans and most animals and plants. With the rapid advances in the next generation sequencing technology, many algorithmic advances have been made in de novo assembly using next generation sequence reads. There have been a lot of assemblers developed targeted solely in reproducing the best possible assembly from the data generated by the new short read sequencing technologies.

In Bioinformatics, genome assembly refers to the process of taking a large number of short DNA sequences which are generated by a shotgun sequencing project and putting them back together to create a representation of the original chromosomes from which the DNA originated. Some of the challenges faced in terms of assembly process include errors in the data due to limitations in sequencing technology, human mistakes during laboratory work and most notably the presence of repetitive sections called repeats that can be thousands of nucleotides long and occur in different locations especially in large genomes. The DNA reads belonging to the repeats are difficult to position and assemble correctly. Moreover some DNA fragments from a genome are

impossible to sequence resulting in gaps in coverage which further complicates the assembly process.

Whole genome shotgun sequencing, the basic strategy for all genome sequencing projects, randomly shears DNA fragments to produce short reads and allow for the generation of mate-pairs where reads come in pairs with a known approximate distance (insert size) between them. To guarantee that every domain of the genome is expressed in the generated reads and to compensate for sequencing error, the genome is oversampled several times, a number referred to as the coverage of the genome. The assembly programs use this information to computationally reconstruct the genome. For the genome assembly problem, connections between read length, read type, repeat complexity, quality score and coverage were studied in detail and assessment was made as to how these parameters helped in improving or diminishing the capability of the assembly programs while assembling the sequence data. Sequence data was presented from different organisms in detail in terms of number of sequence reads, read length, read type, mate-pair information and genome coverage and results were compared from the different assembly programs in the following sections. Here results of computational challenges inherent with genome assembly and algorithms are summarized and results of several de novo assembling experiments are discussed. At the end of this experimental process a better understanding of the impact of the above mentioned parameters on the complexity of genome assembly can be gained as these experiments help ascertain margins on the parameters of sequence data that enable efficient and accurate assembly by the programs.

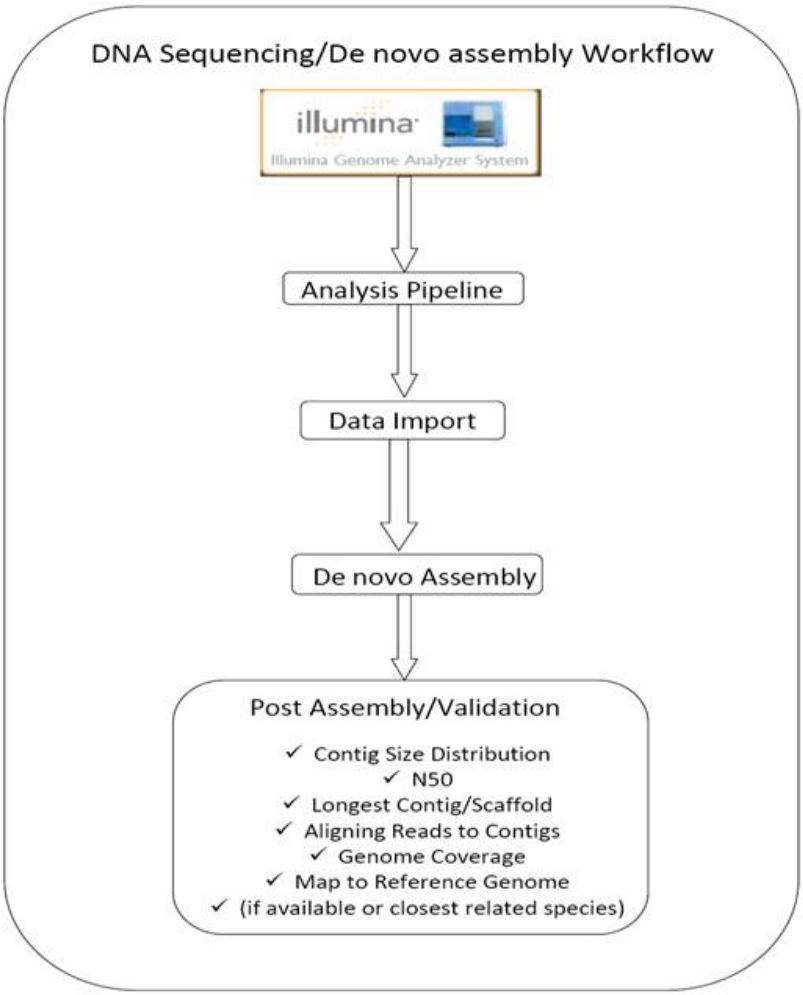


Figure 31: Sequencing and Genome Assembly Work Flow

DNA Sequencing Technology

Until recently, the main sequencing DNA methodology has been Sanger sequencing. This sequencing method has major limitations and remains prohibitively costly and time consuming for many genome projects. Recently there have been many sequencing technologies developed that have the potential to overcome these limitations, but at the same time the data produced by these technologies pose new challenges in

assembling the sequence reads. These sequencing technologies called next-generation sequencers have the ability to process millions of sequence reads in parallel. Some of the commercially available sequencers are 1. Roche (454) GS FLX sequencer, 2. Illumina genome analyzer, and 3. Applied Biosystems SOLiD sequencer. All these sequencers produce shorter read lengths 35 – 400 Base Pairs depending on the platform than capillary sequencers which produce 650 – 800 base pairs. The second generation reads can impact the utility of the data for various applications such as de novo assembly and genome resequencing. The sequencing technology used here at NCGR's Genome Sequencing Center is Illumina Genome Analyzer. Illumina Genome Analyzer System is based on the Solexa sequencing technology providing a high-speed, massively parallel genetic analysis system for genetic analysis and functional genomics. Some of the highlights this technology are that, it has scalable ultra-high throughput and it requires sample input as low as 100 ng - 1µg enabling a host of applications where sample is limited. Also it is simple, fast and automated.

Comparison: Sanger Reads Vs Solexa Short Reads

The main issue between the sanger and solexa reads is the read length. The Sanger sequencing technology generates reads that are routinely 800-1000 nucleotide base pairs long, referred to as the first generation reads. The next generation technologies (Illumina, 454, ABI SOLiD, etc) produce humungous quantities of sequence data in the form of smaller reads, ranging from 36 – 400 nucleotide base pairs long depending on the technology.

The following table 4 shows how much the number of sequence reads increases from Sanger technology to the Solexa short read technology for different organisms

Table 7: Read Difference Between Sanger and Next Generation Technologies

Organism	Genome size	Sanger Reads with 8X Coverage (1000 nt)	Solexa Short Reads with 100X Coverage (100nt)
Virus, Phage Φ-X174	5,400	43	5400
Bacterium, *Escherichia coli* (million reads)	4,000,000	32,000	4,000,000
Nematode, *Caenorhabditis elegans* (million reads)	98,000,000	784,000	98,000,000
Plant, *Arabidopsis thaliana* (million reads)	157,000,000	1,256,000	157,000,000
Mammal, *Homo sapiens* (billion reads)	3,200,000,000	25,600,000	3,200,000,000

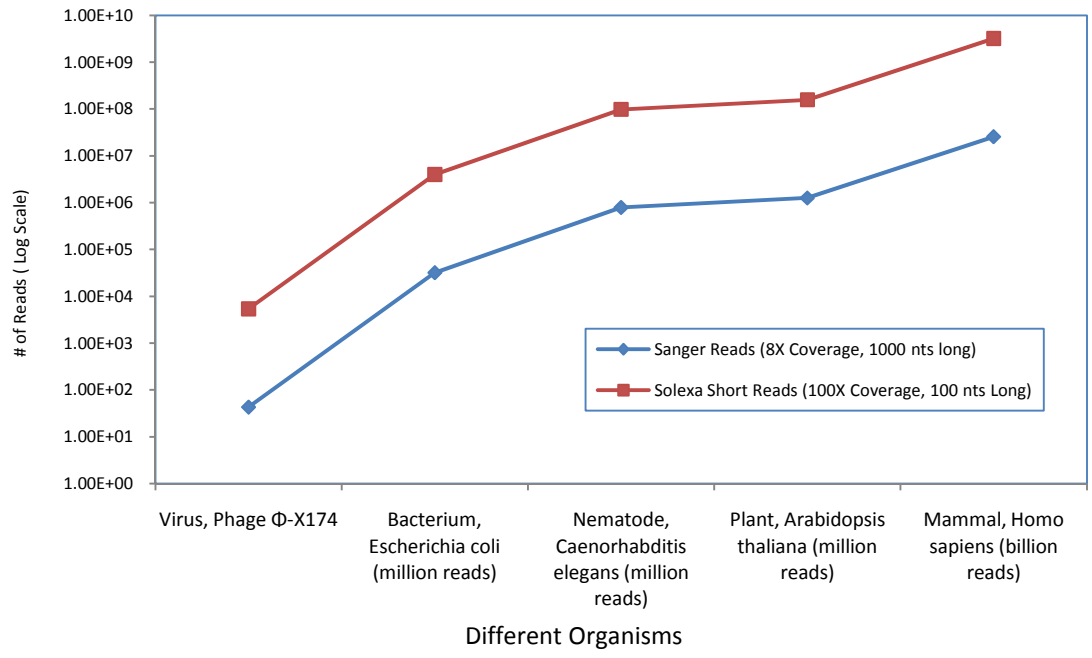


Figure 32: Read difference between Sanger and Solexa technology reads

De novo Sequence Assembly Process of Next Generation Data

An important and a critical step in the sequencing of different genomes is to assemble all the short reads or fragments that are extracted from the sample to form a set of contiguous sequences (contigs) that represents the DNA in the sample. Read length is very crucial when it comes to assembling accurate sequence, especially for genomes as complex and repetitive as the human and plants genome. Assembling a genome using reads generated by the next generation sequencing technologies requires a different approach than the methods that were developed for the long reads generated by the Sanger sequencing technology. Whole genome shotgun sequencing is the basic strategy for most genome sequencing projects today; it randomly shears DNA fragments to produce short reads and allows for the generation of mate-pairs where reads come in pairs with a known approximate distance (insert size) between them.

Assembly Algorithms

Greedy Approach

Greedy approach is the simplest solution to the assembly problem. In this approach, starting with the reads that overlap best, and ending when there are no more reads to be combined, the assembler greedily combines reads that are most similar to each other in an iterative fashion. Two reads are considered to overlap with each other if the prefix of one read shares sufficient similarity with the suffix of another sequence read. The definition of an overlap read is commonly the length of the overlap and the percentage of nucleotides that is shared between the reads. This method provides the

most intuitive solution but the disadvantage is that at each step of the assembly process only local information is considered and can be easily confused with complex repeats throughout the sequence data that can lead to misassemblies thus not leading to a globally optimal solution.

Several assemblers like PHARP, TIGR, CAP3 developed for the First generation (Sanger technology generated sequence reads) sequence data use greedy algorithmic approach.

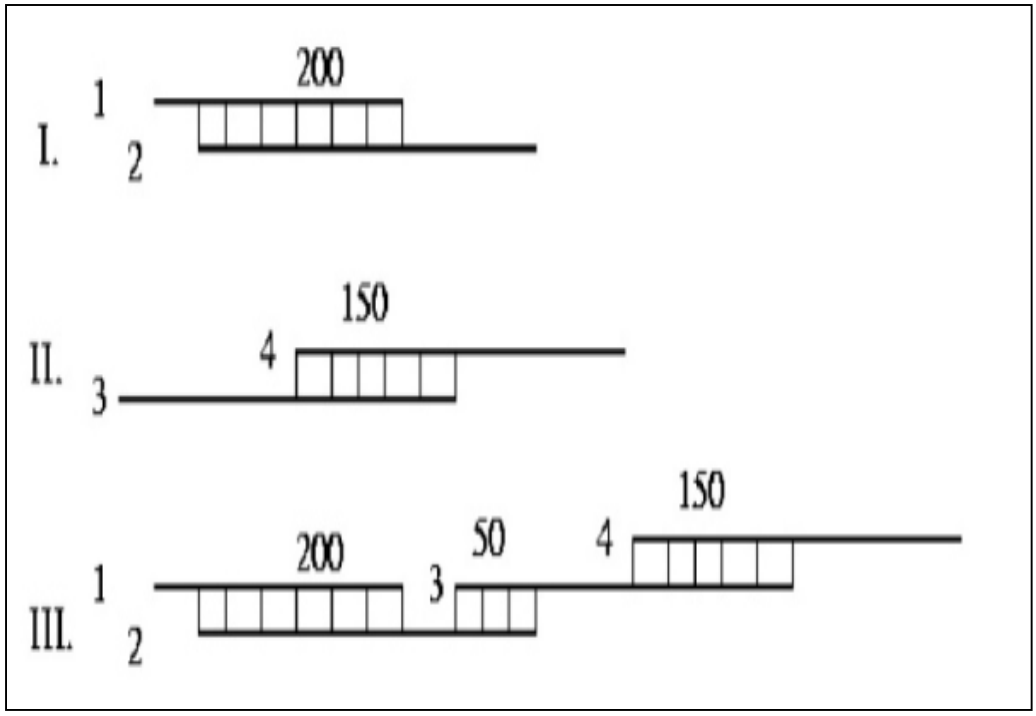


Figure 33: The assembler joins, in order, reads 1 and 2, then reads 3 and 4, then reads 2 and 3. [http://www.cbcb.umd.edu/research/assembly_primer.shtml]

Overlap-Layout-Consensus Graph Approach

This classical approach to sequence assembly is based on graph theory with the associations between the sequence reads to be assembled illustrated as a graph. The nodes represent each of the reads and an edge connecting the two nodes symbolizes that the corresponding reads overlap.

Contigs are generated by identifying a path through the graph that contains each node at most once; in graph theory this path is called the Hamiltonian path (also called traceable path, a path in an undirected graph which visits each node (or vertex) exactly once). The assemblers that follow this paradigm go through three phases, the Overlap phase, Layout phase and finally the Consensus phase.

The Overlap Phase: The assembler builds a graph structure by computing all pair wise alignments between the sequence reads.

The Layout Phase: The graph structure is cleaned and simplified by removing all its redundant edges thus resolving ambiguities. This refined graph will comprise of a set of nonintersecting simple paths. Each path corresponds to a assembled contig.

The Consensus Phase: In the third and final phase the assembler builds a multiple alignment of the reads consistent with the chosen path covering the whole genome inferring a consensus sequence.

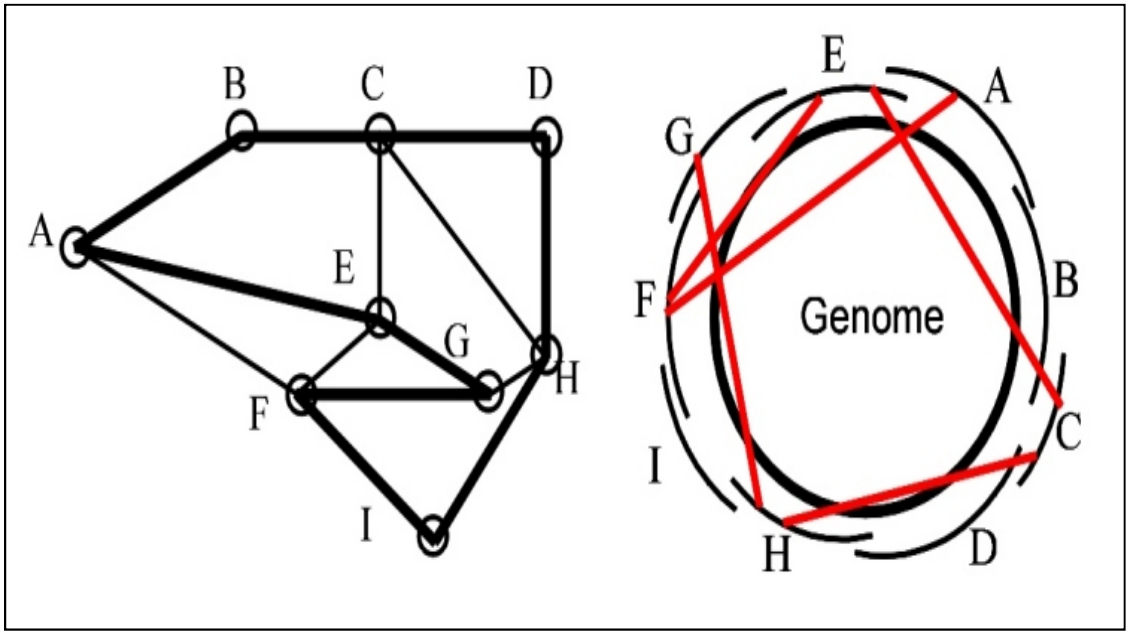


Figure 34: The thick edges in the picture on the left (a Hamiltonian cycle) correspond to the correct layout of the reads along the genome (figure on the right). The remaining edges represent false overlaps induced by repeats (exemplified by the red lines in the figure on the right) [http://www.cbcb.um.edu/research/assembly_primer.shtml]

This approach works best for limited number of reads with significant overlaps. The overlap graph could be extremely large making a de novo assembly very computational intensive. Few assemblers for next generation sequence data implement the Overlap-Layout-Consensus approach

Eulerian Path Graph Approach

Pevzner et al 1989 proposed an elegant formulation for contig reconstruction as an euler path problem on a deBruijn graph.

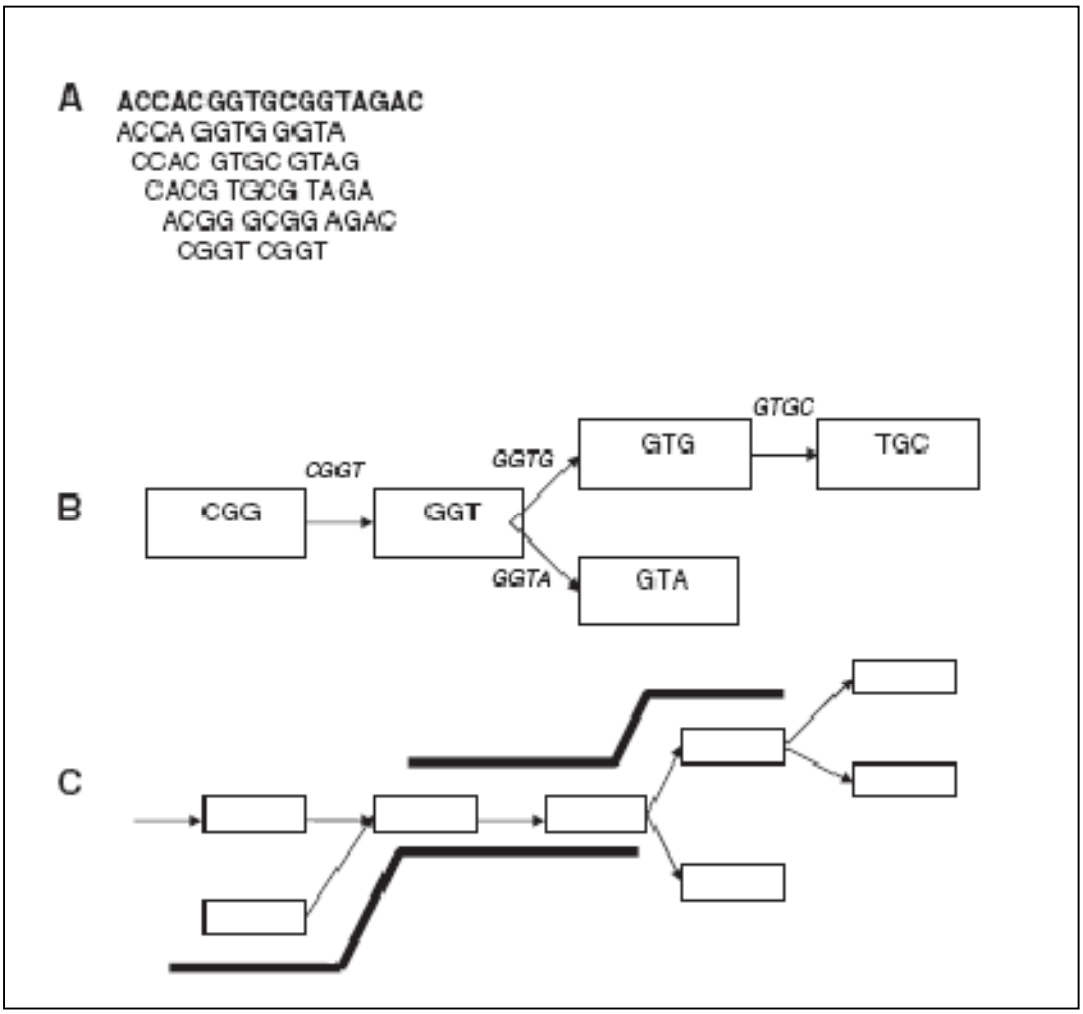


Figure 35: (A) kmer spectrum of a DNA string (bold) for k=4; (B) Section of the corresponding deBruijn graph. The edges are labeled with the corresponding kmer and (C) Overlap between two reads (bold) that can be inferred from the corresponding paths through the deBruijn graph[Pop, M . 2009].

In this approach the idea is to break each sequence reads into overlapping *k*mers, where a *k*mer is a substring of length *k* from the original sequence. In the next step a deBruijn graph is constructed using the *k*mer spectrum. Each edge in the graph corresponds to a *k*mer from one of the original sequence reads. The source and destination nodes correspond respectively to *k*-1 prefix and *k*-1 suffix of the corresponding *k*mer. The assembly problem then reduces to finding the correct eulerian path; there are usually a large number of eulerian paths. An eulerian path is a path that uses every edge exactly once.

Survey of Different Assembler Protocols

The following table 5 compares different currently available next generation sequence data assembly software.

Table 8: Various Assembly Algorithms

Assembler Protocols	Type	Group	Algorithm	Sequence Type	Test Data	Read Type	Coverage
VELVET	De Novo Assembly	EBI	Eulerian Path Approach	Solexa, 454,SOLiD, Sanger	5Mbp DNA Sequences E.Coli, S.cerevisiae, C.elegans, H.Sapiens	Single Ended Reads, Paired End Reads, Long Reads	50X
EULER - SR (Short Read)		UCSD	Eulerian Path Approach	454, Solexa	E.Coli, S.pneumoniae, Human BAC		
SHORTY	De Novo Assembly	Stony Brook	Eulerian Path Approach	Solexa, 454,SOLiD, Sanger	Streptococcuss Suis(simulated)	Paired end short reads	50X
SSAKE	De Novo Assembly	BCGSC	Extension with Prefix Tree Overlaps	Solexa	Viral, Bacterial and Fungal Genomes PhiX174, SARS, TOR2, H.influenzae	Single-ended, Paired end reads	400X for viral genomes and 100X for H.Influenzae

Table 8 (cont'd): Various Assembly Algorithms (cont'd)

Assembler Protocols	Type	Group	Algorithm	Sequence Type	Test Data	Read Type	Coverage
SHARCGS	De Novo Assembly	Max Planck Institute, Univ. of Regensburg, Regensburg, Germany	Extension with Prefix Tree Overlaps	Solexa	H.Acinonychis	Single Ended Reads	
Newbler	De Novo Assembly	454 Life Sciences	Overlap-Layout-consensus	454, Sanger	Humans, Plants, Yeast, Bacteria, Fungi, Viruses, YACs, BACs, Fosmids	Single Ended Reads, Paired End Reads, Long Reads	
VCAKE	De Novo Assembly	UNC, Wahington Univ.	Extension with Prefix Tree Overlaps	Solexa	Viral and Bacerial Genomes		50X
SHRIMP	Alignment	Univ. of Toronto	Alignment				
ALLPATHS	De Novo Assembly	Broad Institute	Eulerian Path Approach	Solexa		Paired End Reads	80X
FORGE	de novo assembly	JGI	restriction mapping				
Seqman	Assembly by Alignment		alignment plus de novo				
Pcap454	De Novo Assembly						
ABYSS	De Novo Assembly	BCGSC					

Genome Assembly Computational Challenges

The recent development of high throughput short read DNA sequencing has revolutionized genome sequencing. There are two approaches in genome assembly, de novo approaches and comparative approaches. De novo approaches must be used to reconstruct genomes that are not similar to any organisms previously sequenced but are often used for organisms with a potential reference as well. In computational complexity theory the de novo assembly problem falls within a class of difficult problems (NP – Hard, Non Deterministic Polynomial Time Hard), so no efficient computational solution

is known [Bodlaender et al 1995, Medvedev et al 2009]. Comparative approaches use the sequence of a closely related organism as a guide during the assembly process, this approach is much easier - essentially assemble a newly sequenced genome by aligning the set of reads onto a reference genome. The second generation sequencing data (discussed in this chapter, Solexa short read sequences generated using Illumina Genome Analyzer (GA) II machine) have several features impacting on assembly software and forces many challenges. The short reads generated forge difficulty in assembling repeats. Illumina GA II have mate-pairs protocols, pairs are about twice as expensive but gives twice as much coverage, in practice most data being generated are mated reads. With these short reads and variable read length and the large amounts of data, the existing assembly software has to be modified, new specific features has to be incorporated and require parallel implementations or specialized hardware when applied to large genomes. In this study, *S aureus* strains were assembled by the de novo approach, followed by a comparative approach for assembly validation.

Genome Assembly Metrics

The genome assembler result is a set of contigs. A complete genome cannot be constructed completely from the contigs alone. Contigs are contiguous assembled pieces of sequence reads. The mate pair information can be used to determine the relative placement of the assembled contigs along a genome. This process is called scaffolding. The output of this scaffolding process is a series of scaffolds. Two contigs can be inferred to be adjacent in the genome if one end of a mate-pair is assembled within the first contig, and the other end is assembled within the second contig [Pop, M. 2008]. The

scaffolding problem like the denovo assembly program is shown to be computationally difficult [Huson, D. H. et al 2001]. Most assemblers do contain a scaffolding module, but still they lack true scaffolding to this point. There are several metrics involved in validating the quality of the assembled contigs.

Number of Contigs Assembled

The number of contigs generated should be to a minimum. If the number of contigs rises to a maximum then the assembly generated is considered to be fragmented.

Genome Coverage/Number of Nucleotides Assembled

This metric looks at the percentage of base pairs in the original reference that was covered by the contigs assembled by the assemblers. This can be computed only if a reference (or closely related reference) genome is available or if the approximate genome size is known.

Maximum/Average Contig Length

The biggest contig and the average contig length are computed. The bigger contigs generated generally indicates good assembly.

N50

This is a standard measure for de novo assembly. It is a way of measuring the length of the contigs. N50 is the contig length such that 50% of the assembled genome lies in blocks of this size or larger. N90, N80, etc. are also used.

B2000

We define B2000 as the percentage of assembled nucleotide base pairs that are in contigs 2000 base pairs or longer.

The above mentioned metrics are very straight forward and convenient to test the quality of the contigs generated by the latest de novo assemblers. One of the important assessment of the quality of the assembly is by aligning it to the reference genome or the reference genome of the closely related species if it available, but this wouldn't be an option for de novo assemblies for which there is no reference available.

Sequence Parameters Analysis

This section briefly describes the sequencing projects of E coli and five different strains of staphylococcus aureus (MM25, MM61, MM66, MM66-4, MV8), the assemblies of these sequence data using VELVET and ABySS assemblers and discusses in detail how different sequencing parameters influence assembly.

Sequencing Projects

For both the genome sequencing projects, solexa short read sequences of variable read length and different coverage were generated using Illumina Genome Analyzer (GA) II machine. Sequence data from sequenced genomes of *Escherichia coli* and *Staphylococcus aureus* were utilized to study the preciseness of genome level assembly and the information thus obtained was used to guide future sequencing projects.

Sequence Data Information

Escherichia coli (E. coli). E.coli a gram negative rod-shaped bacterium with an approximately 4.6 Mbp genome was sequenced yeilding 36 base pairs, paired-end sequences of approximately 225-fold coverage.

Table 9: Sequence Read Information *Escherichia coli*

	Paired End Reads	**Sequence (Gbp)**	**Coverage**
E. coli	29,871,930	1,045,517,550	225X

Staphylococcus aureus (S. aureus). S. aureus is a gram positive spherical bacterium that occurs in microscopic clusters resembling grapes. They are present in nose and skin of healthy humans and belong to the bacterial family *Staphylococcaceae*. It has a 3 Mbp genome and we sequenced five strains. Each strain was sequenced on a single lane of a flowcell generating in excess of 1 Gb and 300X coverage (table 7). Each strain was sequenced with 90 nt paired reads sequenced from either end of a 380nt fragment.

Table 10: Sequence Read Information For All Five Strains

Strain	**Pairs**	**Sequence (Gbp)**	**Coverage**
MM25	10,365,826	1.87	643X
MM61	10,249,743	1.84	634X
MM66-4	10,204,136	1.84	634X
MM66	5,809,956	1.05	361X
MV8	11,438,281	2.06	710X

Assembly Hardware

Genome Assembly is a complex and computationally intensive task and requires large amounts of memory, especially when using second generation sequencing technology. For our assembly we used our local hardware resources and also ran our assemblies on Encanto, a New Mexico Super Computer which is an SGI Altix system, with 1,792 nodes each with two intel Xeon X5335 quad-core processors operating at 3.0 GHz for a total of 14,336 cores with 28.7 Terra bytes of RAM.

Assembly Software

For assembling the E. coli and strains of S. aureus ABySS and Velvet assemblers were used. **A**ssembly **By S**hort **S**equences (ABySS) [Simpson, J.T et al 2009] is a de novo sequence assembler designed for short reads developed by Canada's Michael Smith Genome Sciences Centre. It is a parallel assembler implemented using Message Passing Interface (MPI) capable of assembling larger genomes. Velvet [Zerbino, D. R et al. 2008] also a de novo genomic sequence assembler designed for short reads was developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute (EMBL – EBI).

De novo sequence assemblies

E. coli Assembly. Assembly was performed using the VELVET assembler with different kmer sizes which consumed memory ranging from 1.5 – 6 Gb. *E. coli* paired end assembly results using the VELVET assembler are presented below in table 11.

Table 11: E Coli Assembly Statistics

	I	II	III	IV	V	VI
Read Type	Pairs	Pairs	Pairs	Pairs	Single	Single
Coverage	76X	76X	225X	225X	225X	225X
kmer Length	21	31	21	31	21	31
kmer Coverage (35 nt's)	28X	7X	84X	19X	84X	19X
Reads	10,096,536	10,096,536	29,871,930	29,871,930	29,871,930	29,871,930
Contigs	684	4,964	1,636	405	6,094	855
Largest Contig	77,863	9,653	40,349	131,523	13,897	73,062
Contig N50	17,377	1,539	10,878	33,463	1,770	17,894
Total Length	4,701,905	4,640,491	4,631,853	4,651,087	4,654,435	4,567,569
Contigs >= 10kb	158	0	140	136	3	149
% nt's in Contigs >= 10kb	73%	0%	53%	89%	1%	73%
% Reads in Contigs	96%	88%	95%	88%	97%	89%

S. aureus Assembly. Here results of the genome assembly of five strains (MM25, MM61, MM64, MM66-4, and MV8) of *S aureus* using ABySS assembler with kmer size of 70 are presented.

Table 12: *S. aureus* Assembly Statistics

	MM25	MM66	MM66-4	MM61	MV8
Contigs	1,041	700	638	683	1,501
Bases Assembled	3,009,314	2,967,552	2,981,179	3,034,238	3,725,212
Max. Contig Length	69,478	95,793	68,016	83,874	192,790
N50	17,803	14,391	19,272	16,851	79,009
N90	43,049	45,582	45,529	63,318	183,701
B2000	94%	93%	94%	95%	93%

Parametric Intricacies in de novo Genome Assembly Process

In this section, we study in detail the connections between read length, read type, repeat complexity, quality score and coverage and how these parameters help in improving or diminishing the capability of the assembly programs to assemble the sequence data.

I. Influence of Read Type in Assembly. To study how single-end versus paired-end reads help in the assembly process we look in to the assembly results of *E. coli* (Table 11) using VELVET assembler and a strain of *S. aureus*, MM66 (Table 10) using ABySS assembler.

Table 13: *S. aureus* (MM66) – Read Type - Assembly Using ABySS

Reads	As Single ends	As Paired ends
Contigs	1,142	700
Bases Assembled	2,888,341	2,967,528
Max. Contig Length	44,778	95,793
N50	11,656	14,391
B2000	89.00%	92.86%

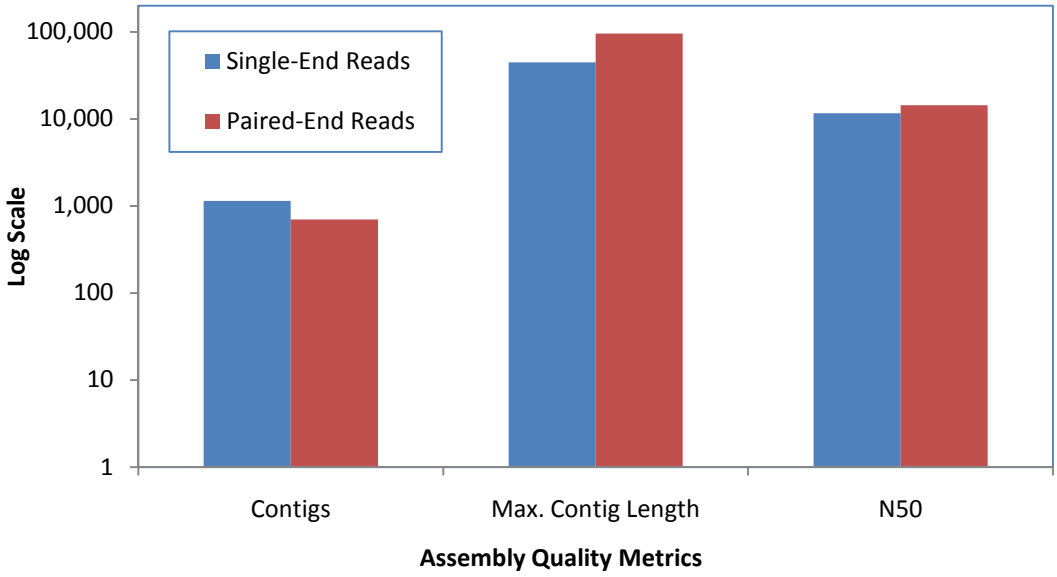


Figure 36: S aureus (MM66) ABySS assembly. Effect of Paired-End Read Types, the graph represents in log scale the number of contigs assembled, Maximum contig length, and N50 for single end reads vs. paired end reads.

The experimental results clearly indicate for both assemblies that paired-end reads help assemble more bases, maximum contig length has more than doubled in size and N50 and B2000 have increased. From the results it is clearly certain that paired end sequence reads help generate longer assemblies.

II. Influence of Read Length in Assembly. In this section the effect of read lengths on the quality if the assembled contigs has been studied. MM66 strain sequence data, which was originally 90 bases long, was trimmed to 75 bases long. Figure 37 shows quality graphs generated by Illumina machine for strain MM66 of s aurues. The quality scores drop down dramatically towards the end. This is very unsual in extent through quality does usually tends to drop down a bit towards the end. Sequences were trimmed from the 3' end. The reason for doing this is as per Illumina sequencing technology standard if the first 25 base pairs passes the quality score test then it generates the whole sequence, so the chances are high that the low quality bases are towards the end. So by doing this, the low quality bases were removed from the sequence. Two datasets were generated, one for 75 base pairs (~500-fold coverage after trimming the ends) and one for 90 base pairs (~600-fold coverage) and were assembled using ABySS. The assembly result is as shown in table below (Table 11)

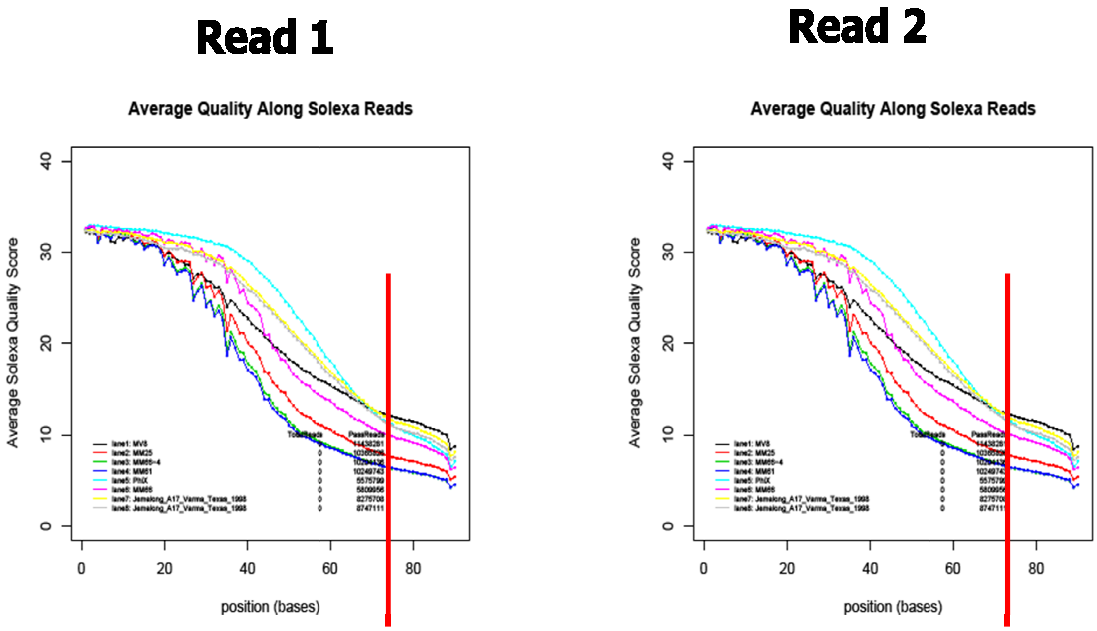


Figure 37: Average quality scores along the solexa reads generated by Illumina Sequencing Technology for s aureus (MM66 strain)

Table 14: *S. aureus* (MM66) - Read Length - Assembly Using ABySS

Read Length	75 mers (Trimmed)	90 mers (Original)
Contigs	1,186	700
Bases Assembled	2,935,514	2,967,528
Max. Contig Length	41,145	95,793
N50	6,696	14,391
B2000	85.96%	92.86%

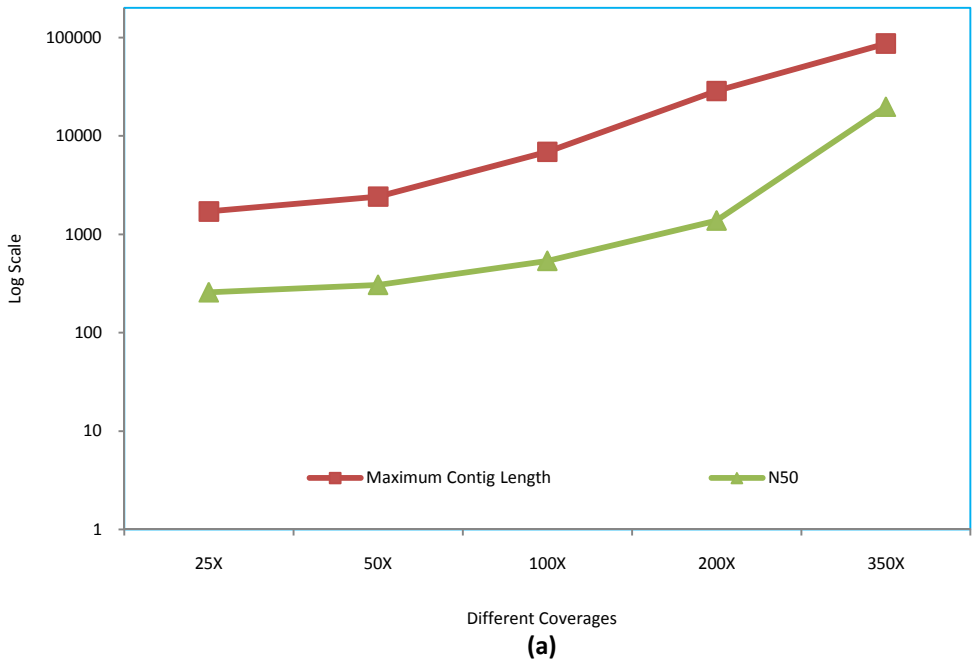
The results clearly indicate that by trimming the reads, a more fragmented assembly was obtained, the bases assembled and B2000 dropped slightly but the maximum contig length and the N50 were less than half.

III. Influence of Depth of Genome Coverage in Assembly. Here, the correlation between genome coverage and assembly was examined. Genome coverage is essentially the oversampling of the genome. To achieve this, the MM66 strain was considered and was sequenced at approximately 350X coverage. A sequence data sets that had 200X,

100X, 50X and 25X was created (Table 12). The same was done for *E. coli* (Table 11). These sequence data sets for different coverage are not independent, that is iterative subsets were taken ranging from 200X down to 25X. Assembly was done for each of the datasets, including the original sequences. The results clearly emphasize the advantage of having high coverage for genome assembly.

Table 15: *S. aureus* (MM66) - Varying Coverage - Assembly Using ABySS

Coverage	**25X**	**50X**	**100X**	**200X**	**350X**
Contigs	314	1,507	3,337	3,122	596
Max Contig Length	1,705	2,408	6,880	28,563	86,764
N50	257	306	537	1,379	19,705
B2000	0.00%	0.98%	10.64%	35.54%	94.69%



(a)

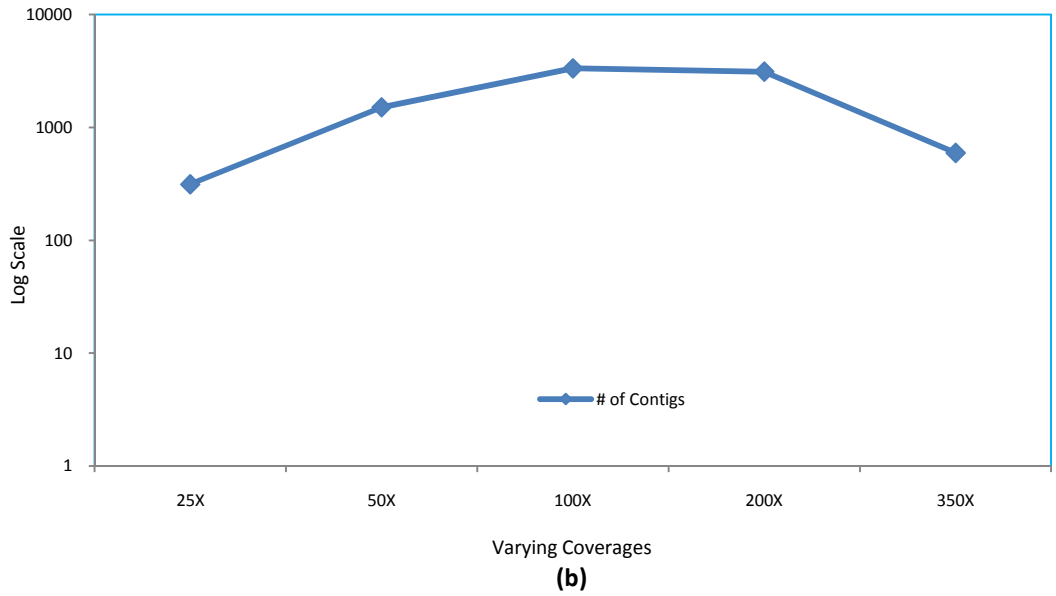


Figure 38: (a) With higher coverage the contig length and the N50 increase resulting in better assemblies. (b) With higher coverage most of the genome is assembled into a smaller number of contigs.

The consensus sequences statistics essentially tells us that at 350-fold coverage that most of the genome will be assembled into smaller number of contigs. As we go down in coverage the assembly looks more fragmented and the maximum contig length get smaller. Based on the above results it cannot be concluded that 350-fold coverage is required for good assembly, because sequencing at 350X coverage is not economically feasible and sheer computational requirements for larger genomes like plants and humans.

IV. Influence of High Quality and Low Quality Sequences in Assembly. In order to study how quality of sequences affect the assembly process, for MM66 strain all the sequences with coverage of 200X were considered and another data set with sequences

that only had an average quality score ≥ 20 was generated. Both data sets were run on the assembly program.

Table 16: *S. aureus* (MM66) - High Quality Sequences

Sequences	All Data	High Quality Data
Coverage	**approx. 200X**	**approx. 200X**
Contigs	3,122	3,302
Bases Assembled	2,545,008	2,517,292
Max. Contig Length	28,563	28,078
N50	1,379	1,248
B2000	35.54%	31.73%

The results from table 13 clearly indicate that removing low quality sequences did not improve assembly. But it is important to mention here that most assemblers do not take into consideration the quality score of the sequences. In this case ABySS assembler which was used for assembly does not take into the account quality scores but it performs some error correction based on bubbles in the graph that are not well supported [Simpson, J. T et al 2009]. It is suggested to do a quality check of the sequence data by aligning them to the reference genome using any available alignment program. But this wouldn't be possible if we are doing de novo assembly. In this case it would be a good experiment to try to apply pre filters on the quality scores, like removing all reads that contained ambiguous N characters (ABySS assembler do this automatically, while others simply replace the N with a random nucleotide A, C, G, or T. Also one could remove reads where the first 80% of the reads did not contain quality scores greater than Q30. Q30 refers to the Phred [Ewing, B et al 1998] score of that nucleotide base.

V. Influence of kmers on Assemblies. Several assemblies were performed on *E. coli* data (Table 11) with VELVET using different kmer values. Generally, the kmer value is limited on the upper side by the length of the reads and limited on the lower side by half the length of the reads. Technically, it is not limited on the lower end, and in practice one wouldn't want to go too low. Ideally a kmer value of 85% of the read length will allow for small amount of overlap, for example k=31 for 36 base pairs read length sequence data (our E coli data, Table 8). Smaller kmer values for example k=21 for 36 base pairs data would increase the connectivity of the graph. From the assembly results (Table 8) it is clearly determined that VELVET performs best with k = 31 on this data.

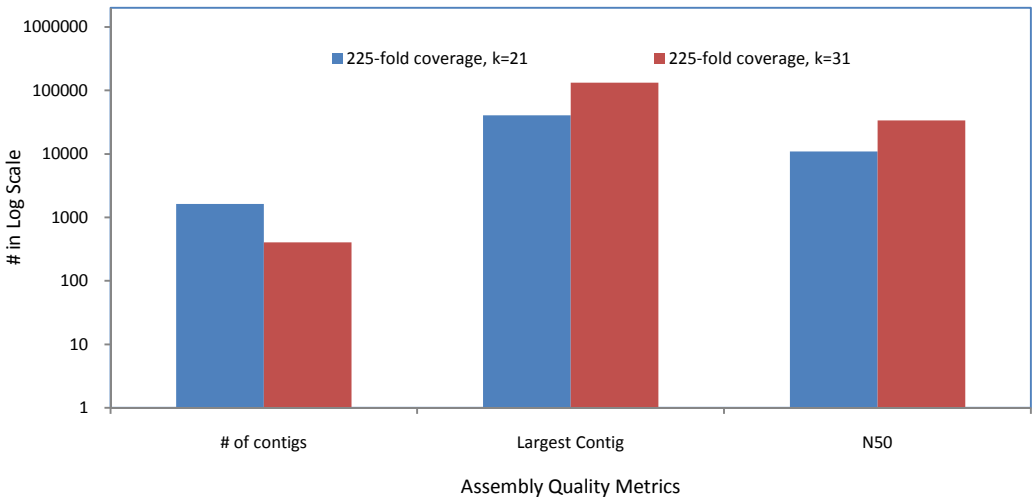


Figure 39: Represents number of contigs, largest contig, and N50 for E coli data with 225X coverage and kmer values 21(Blue) and 31(Red).

With more coverage and a high kmer value the number of contigs have been reduced by four times, the largest contig have been tripled the size, and finally the contig N50 also tripled in size. From this experiment we can conclude that with high coverage and high kmer value the assembler tends to generate decent assemblies.

Comparing Assemblers

In this section results from the ABySS and Velvet assemblers were compared. It is a non trivial problem to compare the results of different assembly programs. In this case, MM25 strain of *S._aureus* was run on ABySS and Velvet assembly programs under different kmer parameter. Both ABySS and Velvet are based on a de Bruijn graph theory. The sequence reads are divided into short kmers, all of the instances of a repeat collapse into single set of vertices. They then represent each read as a walk on the de bruijn graph, and search for a super walk that contains all the reads.

Table 17: *S. aureus* (MM25 strain)

Assembler	**Read Length**	**kmer size**	**contig N50**	**Max Contig Length**	**Bases Assembled**	**B2000**
ABySS	75	50	9k	39k	2.9Mb	88%
	90	60	21k	79k	3.4Mb	83%
Velvet	75	31	1k	9k	3.2Mb	31%
	90	N/A				

Using ABySS assemblers definitely generated better quality assemblies when compared to VELVET. Also couple of key points to be mentioned here is that, 1) it has to be pointed that ABySS is a parallelized and we ran it on Encanto (Super computer); 2) the VELVET assembly version used and it could accept only a maximum kmer size of 31 (newer versions can now do higher kmer values), but the reads assembled were 75 base pairs and 90 base pairs.

Inference

In the process of assembling the *S. aureus* genome it has been demonstrated how different read types, various read lengths, depth of genome coverage, and low and high quality sequences help in resolving the proper layout of the genome. This section has a summary of the analyses.

1. Paired end reads, also called mate pairs, come from opposite strands at an approximate known distance from the source genome. By allowing for spanning of repeats or regions of low coverage or high polymorphism, this mate pair information allows the assembler to join together longer genomic regions. From the assembly results it clearly indicates that having paired end reads definitely improved assembly and we were able to get good results.
2. Longer sequences in our case 90 base pair long reads helped improve assemblies compared to shorter sequences, 75 base pair long reads, despite the low quality of the nts from 76 to 90.
3. Higher coverage sequences tended to assemble better than lower coverage sequences.
4. ABySS assembler outperformed VELVET in terms of computing time, memory resources and consensus sequence statistics.
5. Removing low quality sequences did not lead to improvement in assembly. In this case we would like to use more effectual quality pre filters and see if it leads to improvement of assembly.

Assembly Parameter Optimization

To get better results out of the assembly programs the different assembly parameters have to be optimized.

Kmer selection

The size of the kmer which is used in the construction of the graphs in the entire de Bruijn graph based assembler plays a very crucial role. To predict an optimal kmer value to begin with is hard to determine because it depends on the read length, coverage and other factors. By selecting small and large kmer value there is a balance between sensitivity and specificity determined by k. The best approach is performing several assemblies over a range of kmers and selecting the one that generates the best contigs. Also sometimes assembling all the contigs generated by the assembler different kmer values tends to yield a good assembly.

Genome Coverage

From the above experiments it can be concluded that high coverage improved assembly. The caveat here is, for bacterial sequencing, it is possible to sequence higher than 100-fold, but sequencing deeper than 50-fold for higher organisms becomes very expensive.

Assembly Post Processing

Mutation Analysis of MM66 and MM66-4 Strains

Resultant assemblies of MM66 and MM66-4 strains were aligned to the COL strain (NCBI ID: NC_002951) using BLAST (Altschul et al, 1990) to identify MM66 and MM66-4 alleles at specific locations. The COL strain was reportedly isolated as a penicillinase-negative strain in the early 1960s from the operating theatre in a hospital in Colindale, England [Gill, S. R. et al, 2005]. The COL strain genome sequence was downloaded from NCBI and was blasted against the MM66-4 contigs generated by the assembler. Nucleotide blast (blastn) program was used to blast, which searches a nucleotide database (Reference) using a nucleotide query. The COL strain (NC_002951) was used as the reference and the MM66-4 contigs was used as the query. From the blast results known SNP positions where mutations occurred in the COL strain (reference) were identified and it is corresponding nucleotide position on the MM66-4 contigs (query) was checked to see if it was mutated. All the mutations were verified and confirmed in MM66-4 strain. Similarly the COL strain was blasted against the MM66 contigs to check if they were wild type and confirmed mutations in this strain as well.

Validation and Correction for High Quality Assembly

From the analysis of the assembly contigs from different projects it cannot be concluded that the assembly programs that were used are perfect and reconstructed good consensus sequences. The occurrences of assembly errors are common due to several reasons, incomplete or incorrect sequences provided to the assembly program, or due to

the limitations of the algorithm used in the assembly programs. So a downstream post assembly analysis of the assembly data will help in calibrating the accuracy of the assembly. Most of the assembly validation was primarily looking at the statistics related to the number of contigs, contig N50, bases assembled and comparing it to the whole genome, and maximum contig length which are very preliminary validation. In order to get a broader validation, the reconstructed assembly could be aligned to the reference sequences using alignment programs like BLAST, GMAP. This alignment could also be used for identifying structural variants. Also aligning reads to the assembled contigs will help determine insert sizes which are different from the expected one and to detect misassemblies.

Discussion

The assembly of genomes of different organisms using current Illumina sequence reads can be performed using a number of publicly available assembly programs such as VELVET, ABySS, or SOAPdenovo. This study presents a comparison and analysis of assembly results from different assemblers (ABySS, VELVET) on solexa sequence data for different organisms. The goal was not to identify the best assembly program but to try different assemblers that would produce good consensus sequences for our sequencing projects and help learn more about what data it takes and computational resources required for a successful assembly that would allow us to list some guidelines for generating more efficient assemblies. For all the aforementioned sequencing projects, de novo assembly was used. With an accelerated rise in the number of sequences, for many organisms the organism itself or their closely related species have been sequenced.

For these organisms, by using the comparative genome assembly programs (AMOS project) assembly can be reconstructed by mapping the sequence onto a reference genome. For *S. aureus*, a comparative genome assembly can be done and this can be compared to the de novo assembly and the comparison between the assemblies might provide an in depth understanding of the genome assembly results.

The assembly results presented in the above sections have several implications in terms of assessing the best read length to maximize accuracy and minimize sequencing cost, how many reads are needed before accuracy is no longer improved (represents the genome coverage required), and how mate-pair information can be used to determine the relative placement of the contigs along a genome (called scaffolding, to link two contigs in order to reduce the impact of experimental errors). Also it raises questions in terms of assembly validation.

Future Directions

The recent advances in genome sequencing technology have provided means to sequence increasingly large and complex genomes. Data generation is now no longer limiting complex genome sequencing. The ability to assemble this data is currently limited by a lack of dedicated bioinformatics tools that are designed to cope with the nature of the sequence reads (Short Reads, Short Insert paired End Reads (SIPE), and Long Insert Paired End Reads (LIPE)) and genomes especially for higher organisms.

With the release of each new and improved assembly algorithms there has been improvement in terms of assembly such as increase in contig sizes (longer contigs), a

decrease in the total number of contigs and a decrease in the number of assembly errors and improvement the overall quality of assemblies.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

Studying and understanding antibody – antigen interactions and epitope predictions has receiving much attention using both experimental as well as computational methods in porteomics. Also in addition, studying antibody – antigen interaction is of crucial importance for drug discovery and clinical diagnostics. Antibody epitope mapping allows for the revalation of structural information regarding the conformation of a protein that an antibody recognizes. Often structural information provides insight into the functional properties of a protein. We achieved this in our research work described in Chapter 2; a data set comprised of the 3D structures of 62 non-redundant Ag-Ab complexes, from the Protein Data Bank (PDB), was assembled and used to determine the general physical as well as biochemical features of the Ag-Ab interfaces.

Given the large number of proteins of unknown structure, developing non traditional methods for generating structural information regarding proteins that are not amenable to traditional structure determination methods such as NMR and x-ray crystallography is likely to have very broad impact in the area of structural genomics/proteomics. To address and overcome the short comings of traditional structural determination methods, we present in Chapter 3, MSA- EPIMAP an improved version of the epitope prediction program called EPIMAP, which facilitates antibody epitope mapping in Chapter 3. Our improved algorithmic approach uses the the substitution matrix derived in Chapter 2 and produces promising results. We would like

to apply this MSA – EPIMAP approach to number of additional antibody – antigen complexes whose 3D structures are known.

Lastly in Chapter 4, we looked into the problem of de novo genome assembly for short reads generated using Illumina Genome Analyzer. The assembly of genomes of different organisms are assembled using a number of publicly available assembly programs such as VELVET, and ABySS. We present results of several de novo assembly experiments. We explored the connections between the read length, read type, coverage, data quality, sort inserts paired end reads, long insert paired reads in genome assembly. This study helps us understand the parametric complexity of the genome assembly problem. Improvements in assembly parameters discussed, quality of the sequence data, quantity, and coverage will of course also assist in producing better assemblies.

REFERENCES CITED

Altschul, S.F.: Amino acid substitution matrix from an information theoretic perspective. J. Mol. Biol., 291, 555-565 1991

Altshcul, S. F., Gish. W., Miller, W., Myers, W. E., Lipman, D. F.: Basic Local Alignment Search Tool. Journal of Molecular Biology (1990) 215, 403-410

Angel, T.E., et al.: Structural waters define a functional channel mediating activation of the GPCR, rhodopsin. Proc Natl Acad Sci U S A, 2009. 106(34): p. 14367-72

Apraiz, I., Mi, J., Cristobal, S.: Identification of Proteomic Signatures of Exposure to Marine Pollutants in Mussels (Mytilus edulis). Molecular and Cellular Proteomics 5:1274-1285, 2006

Bahadur, R. P., Zacharias, M.: The interface of protein-protein complexes: Analysis of contacts and prediction of interactions. Cell Mol. Life Sci. 65 (2008) 1059-1072

Baker, D., Sali, A.: Protein structure prediction and structural genomics. Science 294(5540), (2001) 93-96

Bao, H., Guo, H., Wang, J., Zhou, R., Lu, X., Shi, S.: MapView: Visualization of short reads alignment on a desktop computer

Barlow, D. J., Edwards, M. S., Thornton, J. M.: Continuous and Discontinuous Protein Antigenic Determinants. Nature, 1986, August 21-27;322(6081):747-8

Bennett, K. L., Matthiesen, T., Roepstorff, P.: Probing Protein Surface Topology by Chemical Surface Labeling, Cross linking, and Mass Spectrometry. Methods in Molecular Biology, vol. 146: Protein and Peptide Analysis: New Mass Spectrometric Applications Edited by: J. R. Chapman © Humana Press Inc., Totowa, NJ. P.113-131

Bentley, D.R. et al.: Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456 (7218), 53-59 (2008)

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E.: The Protein Data Bank. Nucleic Acids Research 2000; 28:235-242

Blythe MJ Flower DR Benchmarking B cell epitope prediction: Underperformance of existing methods. Prot Sci 14:246-248, 2005

Bodlaender, HL, Downey, RG, Fellows, MR, Hallet, MT, Wareham, HT.: Parameterized Complexity Analysis in Computational Biology. CABIOS, 1995, Vol. 11, No. 1: 49-57

Bogan, A. A., Thorn, K.S.: Anatomy of hot spots in protein interfaces. J. Mol. Biol. 280:1-9, 1998

Branden, C., Tooze, C.: Introduction to protein structure, Garland Publishing, New York, NY. (1999)

Buyong Ma, Tal Elkayam, Haim Wolfsom, and Ruth Nussinov.: Protein-Protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. PNAS, May 13, 2003, vol. 100, no.10, 5772-5777

Bystroff, C., Baker D.: Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 281, 1998, 565-77

Bystroff, C., Shao, Y.: Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. Bioinformatics, 2002, 18 Suppl 1, S54-61

Chen J, Liu H, Yang J, Chou K-C. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33: 423-428, 2007

Chen, T, Skiena, SS.: A Case Study in Genome-Level Fragment Assembly. Bioinformatics 2000, Vol. 16 No. 6: 494-500

Chothia et al.: Conformations of immunoglobulin hypervariable regions. Nature, 1989, 342:877-883

Chou, KC.: A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Structure, Function, and Genetics 21:319-344 (1995)

Christopher Reynolds, David Damerell, and Susan Jones: ProtorP: a protein-protein interaction analysis server. Bioinformatics 2009 25(3):413-414

Conte, L. L., Chothia, C., Janin, J.: The Atomic Structure of Protein-Protein Recognition Sites. J. Mol. Biology (1999) 285, 2177-2198

Dandekar, T., Argos, P.: Applying experimental data to protein fold prediction with the genetic algorithm. Protein Eng. 1997 Aug, 10(8):877-93

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C.: A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure. National Biomedical research Foundation, Washington DC, Vol.5 Suppl. 3, pp. 345-352, 1978

Dhungana, S., Fessler, M. B., Tomer, K. B.: Epitope mapping by differential chemical modification of antigens. Methods Mol Biol, 2009. 524: p. 119-34

Dubchak, I., Holbrook, S. R. and Kim, S. H.: Prediction of protein folding class from amino acid composition. Proteins, 1993, 1679-91

Eisenhaber, F., Frommel, C. and Argos, P.: Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. Proteins, 1996, 25, 169-179

Ewing B, Hillier L, Wendl MC, Green P. : Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research. 1998 March, 8(3):175-85

Eyal, E., Frenkel_morgenstern, M., Sobolev, V., Pietrokovski, S.: A Pair-to-Pair Amino Acids Substitution Matrix and its Applications for Protein Structure Prediction. PROTEINS:Structure, Function, and Bioinformatics 67:142-153 (2007)

Garey, M., and Johnson, D. 1979. Computers and Interactability: A guide to the theory of NP – Completeness, W. H. Freeman and Co.

Gill, S. R. et al.: Insights on Evolution of Virulence and Resistance from the Complete Genome Analysis of an Early Methicillin-Resistant *Staphylococcus aureus* Strain and a Biofilm-Producing Methicillin-Resistant *Staphylococcus epidermidis* Strain. J Bacteriol. 2005 April; 187(7): 2426–2438

Glaser, F., Steinberg, D., Vakser, I., Ben-Tal, N.: Residue Frequencies and Pairing Preferences at Protein-Protein Interfaces. Proteins: Structure, Function, and Genetics 43:89-102 (2001)

Halperin, I., Wolfson, H., Nussinov, R.: SiteLight: Binding-site prediction using phage display libraries. Protein Science (2003), 12:1344-1359

Henikoff, S., and Henikoff, J.G.: Amino Acid Substitution matrix from protein blocks. Proc. Natl. Acad. Sci., USA, 89, 10915-10919, 1992

Hernandez, D., Francois, P., Farinelli, L., et al.: De novo bacterial genome sequence: Millions of very short reads assembled on a desktop computer. Genome Research, March 10 2008

Huang, W., Marth, G.: EagleView: A genome assembly viewer for next-generation sequencing technologies. Genome Research, June 11 2008

Huson DH, Reinert K, Myers E.: 'The Greedy Path- Merging Algorithm for Sequence Assembly'. In: Lengauer T, Sankoff D, Istrail S, et al. (eds). Proceedings of the Fifth Annual International Conference on Computational Biology (RECOMB), 2001, pp. 157–163. Association for Computing Machinery, Montreal, Canada

Jacobsen,R.B., Sale,K.L., Ayson,M.J., Novak,P., Hong,J., Lane,P., Wood,N.L., Kruppa,G.H., Young,M.M., Schoeniger,J.S.: Structure and dynamics of dark-state bovine rhodopsin revealed by chemical cross-linking and high-resolution mass spectrometry. Protein Sci., 2006, 15, 1303-1317

Janin J. Henrick K. Moult J., Sternberg M.J. Vajda S. Vakser I. and Wodak S.J.: CAPRI: Critical Assessment of Predicted Interactions. Proteins, Structure Function and Bioinformatics, 52, 2-9

Jesaitis, A., Gizachew, D., Dratz, E., Siemsen, D., Stone, K., Burritt, J.: Actin surface structure revealed by antibody imprints: Evaluation of phage-display analysis of anti-actin antibodies. Protein Surface, Vol. 8 (1999) 760-770

Jin,L.Y.: Mass spectrometric analysis of cross-linking sites for the structure of proteins and protein complexes. Mol. Biosystems, 2008, 4, 816-823

Jones, S., Thornton, J. M.: Analysis of Protein-Protein Interaction Sites using Surface Patches. J. Mol. Biol. (1997) 272, 121-132

Keskin, O., Tsai, Chung-Jung, Wolfson., H., Nussinov, R.: A new, structurally nonreduntant, diverse data set of protein-protein interfaces and its implications. Protein Science (2004), 13:1043-1055

Li, L., Khuri, S.: A comparison of DNA Fragment Assembly Algorithms. Department of Computer Science, San Jose State University

Lu, H., Lu, L., Skolnick, J.: Development of Unified Statistical Potentials Describing Protein – Protein Interactions. Biophysical Journal, Volume 84, March 2003, 1895 – 1901

Ma, B., Elkayman, T., Wolfson, H., Nussinov, R.:Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed surfaces., Proc Natl Acad Sci 100:5772-5777, 2003

MacCallum, R., Martin, A., Thornton, J.: Antibody-antigen Interaction: Contact Analysis and Binding Site Topography. J. Mol. Biol. (1996) 262, 732-745

Martin, O., Schomburg, D.: Efficient comprehensive scoring of docked protein complexes using probabilistic support vector machines. Protein 2008; 70:1367-1378

McConkey, B. J., Sobolev, V, Edelman, M.: Discrimination of Native Protein Structures using Atom-Atom Contact Scoring. Proc. Natl. Acad. Sci. U S A. 2003 March 18; 100(6): 3215–3220

Medvedev, P., Brudno, M.: Maximum Likelihood Genome Assembly. Journal of Computational Biology, Volume 16 Number 8, 2009

Medvedev, P., Georgiou K., Myers G., Brudno, M.: Computability of Models for Sequence Assembly. Workshop on Algorithms in Bioinformatics (WABI 2007). Philadelphia, PA: Springer, 2007

Mihel, J., Sikic, M., Tomic, S., Jeren, B., Vlahovicek, K.: PSAIA – Protein Structure and Interaction Analyzer. BMC Structural Biology 2008, 8:21

Miller, J. R., et al.: Aggressive assembly of pyrosequencing reads with mates. Bioinformatics, Vol 24, no. 24, 2008, pages 2818-2824

Miller, S, Janin, J, Lesk, M, Chothia, C.: Interior Surface of Monomeric Proteins. Journal of Molecular Biology, 196: 641-656

Miller, W.: Comparison of genomic DNA sequences: solved and unsolved problems. Bioinformatics Review, Vol. 17 no. 5, 2001, 391-397

Mumey, B., Bailey, B., Kirkpatrick, B., Angel, T., Jesaitis, A., Dratz, E. : A new method for mapping discontinuous antibody epitopes to reveal structural features of proteins. J. of Computational Biology, Vol. 10 Issue 3-4 (2003) 555-567

Mumey, B., Ohler, N., Angel, T., Jesaitis, A., Dratz, E.: Filtering Epitope Alignments to Improve Protein Surface Prediction. ISPA (2006) 648-657

Myers, EW. Toward Simplifying and Accurately Formulating Fragment Assembly. Journal of Computational Biology 1995;2:275-90

Nagarajan, N, Pop, M.: Parametric Complexity of Sequence: Theory and Applications to Next Generation Sequencing. Journal of Computational Biology 2009, Vol. 16, No. 7: 897-908

Orban, T., et al.: Visualizing water molecules in transmembrane proteins using radiolytic labeling methods. Biochemistry. 49(5): p. 827-34

Orfan, Y., Rost, B.: Analyzing Six Types of Protein – Protein Interfaces. Jounal of Molecular Biology (2003) 325, 377-387

Padlan, E.: X-ray crystallography of antibodies. Adv. Protein Chem., Vol. 49 (1996) 57-133

Pevzner, P. A., Tang, H., Waterman, M. S.: An Eulerian path approach to DNA fragment assembly. PNAS, Vol. 98, no. 17, August 14, 2001

Phillippy, A. M., Schatz, A, C., Pop, M.: Genome assembly forensics: finding the elusive mis-assembly. Genome Biology, March 14 2008

Pinak Chakrabarti, and Joel Janin: Dissecting Protein – Protein Recognition Sites. Proteins: Structure, Function and Genetics 47:334-343 (2002)

Pinilla C, Appel JR, Hougthen RA. Functional importance of amino acid residues making up peptide antigenic determininants Mol. Immunol 30: 577-585, 1993

Ponomarenko, J. V., Bourne, P. E.: Antibody-protein interactions:benchmark datasets and prediction tools evaluation. BMC Structural Biology 2007, 7:64

Pop, M, Phillipy A, Delcher, AL, Salzberg, SL.: Comparative Genome Assembly. Briefings in Bioinformatics 2004. Vol. 5. No. 3: 237-248

Pop, M., Salzberg, S. L., Shumway, M.: Genome Sequence Assembly:Algorithms and Issues. IEEE, 0018-9162, 2002

Pop, M.: Genome Assembly Reborn: Recent Computational Challenges. Briefings In Bioinformatics, March 2009. Vol. 10, No. 4: 354-366

Pop., M., Kosack, D. S., Salzberg, S. L.: Hierarchical Scaffolding with Bambus. Genome Research, 14: 149-159 2004

Regenmortel, Marc H. V.: Mapping Epitope Structure and Activity: From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity. METHODS: A Companion to Methods in Enzymology 9, 465-472 (1996)

Reynolds, C., Damerell, D., Jones, S.: ProtorP: a protein-protein interaction analysis server. Structural Bioinformatics, Vol. 25 no. 3 2009

Rubinstein, N.D., Mayrose, I., Halperin, D., Yekutieli, D., Gershoni, J. M., Pupko, T.: Computational characterization of B-cell epitopes. Mol. Immunology 45 (2008) 3477-3489

Rudra Prasad Saha, Pinak Chakrabarti.: Parity in the number of atoms in residue composition in proteins and contact preferences. Current Science, Vol. 90, No. 4, 25 February 2006

Sanner, M., Olson, A., Spehner, Jean-Claude.: Reduced Surface: An Efficient Way To Compute Molecular Surfaces. Biopolymers. 1996 March;38(3):305-20

Shulman-Peleg, A., Shatsky, M., Nussinov, R., Wolfson, H. J.: MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. Nucleic Acids Research, 2008, Vol 36, Web Server Issue

Simpson, JT, Wong, K, Jackman, SD.: ABySS: A Parallel Assembler for Short Read Sequence Data. Genome Research, February 27 2009. 19:1117-1123

Sommer, D., et al.: Minimus: a fast lightweight genome assembler. BMC Bioinformatics, February 26 2007

Suan Hafenstein, Valorie D. Bowman, Tao Sun, Christian D. S. Nelson, Laura M. Palermo, Paul R. Chipman, Anthony J. Battisti, Colin R. Parrish, and Micheal G. Rossmann.: Structural Comparoison of Different Antibodies Interacting with Parvovirus Capsids. Journal of Virology, June 2009, p. 5556-5566

Sun, J., Wu, D., Xu, T., Wang, X., Xu, X., Tao, L., Li, Y. X., Cao, Z. W.: SEPPA: A Computational Server for Spatial Epitope Prediction of Protein Antigens. Nucleic Acids Res. 2009 July 1; 37(Web Server issue): W612–W616

Sundberg, E. J., Mariuzza, R. A.: Molecular recognition in Ag-Ab complexes. Adv. Protein Chem. 61, 119-160 (2002)

Susan Jones, Janet M. Thornton: Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA, Vol. 93, pp. 13-20, January 1996

Wang L, Liu J, Zhu S,Gao YY. Prediction of linear B-cell epitopes using AAT scale. IEEE Xplore 978-1-4244-2902-8/09 2009

Wang, X., et al.: Probing rhodopsin-transducin interactions by surface modification and mass spectrometry. Biochemistry, 2004. 43(35): p. 11153-62

Wilson, I. A, Stanfeild, R. L.: Antibody – Antigen Interactions: New Structures and New Conformational Changes. Current Opinions in Structural Biology, 1994, Dec;4(6):857-67

Yan, C., Honavar, V., Dobbs, D.: Identification of interface residues in portease-inhibitor and antigen-antibody complexes: a support vector machine approach. Neural Comput. & Applic. (2004) 13: 123 – 129

Yu, Yi-Kuo, Altschul, S. F.: The Construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. Bioinformatics, Vol 21 no. 7 2005, 902-911

Yu, Yi-Kuo, Wootton, J. C., Altschul, S. F.: The Compositional adjustment of amino acid substitution matrices. PNAS, December 23, 2003, Vol. 100, no.26 15688-15693

Yuan, X., Shao, Y., Bystroff, C.: Ab initio Protein Structure Prediction Using Pathway Models. Comparative and Functional Genomics, 2003, 4(4):397-401

Zerbino, DR, Birney, E.: Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs. Genome Research 2008. 18: 221-829

Zheng, X., Wintrode, P. L., Chance, M. R. :Complementary structural mass spectrometry techniques reveal local dynamics in functionally important regions of a metastable serpin. Structure, 2008. 16(1): p. 38-51

APPENDIX A

EXTENDED TABLES AND MATRIX

Table A1: Data set of Antigen – Antibody Complexes

# of Complexes in this group (Group I- Peptides) = 15							
Code	Name	Antigen		Antibody			References
		Chain(s)	Size	Chain(s)	Size		
1CFT	anti-p24 (HIV-1)/ Protein (Antigen Bound Peptide)	C	5	A,B	214, 213	427	Keitel et al. (1999)
1E4W	ANTI-TGFALPHA ANTIBODY FAB-FRAGMENT / Cyclic Peptide	P	7	L,H	214, 213	427	Hahn et al. (2000)
1CE1	CAMPATH-1H / Peptide Antigen	P	8	L,H	211, 220	431	James et al. (1999)
1FRG	IGG2A 26/9 FAB / Influenza Hemagglutinin HA1 (STRAIN X47) (Residues 101-108)	P	8	L,H	217, 220	437	Ghurchill & Wilson (1994)
1HIN	IGG2A-KAPPA 17/9 FAB / Influenza Hemagglutinin HA1 (STRAIN X47) (RESIDUES 100-107)	P	8	L,H	217, 220	437	Rini & Wilson (1992)
1F90	FAB Fragment of Monoclonal / Antigenic Nonapeptide	E	9	L,H	219, 220	439	Afonin et al. (2000)
2FWO	H-2 Class I histocompatibility Antigen, K-D Alpha Chain / Beta-2-microglobulin / TYQRTRALV Peptide From Nucleoprotein	P	9	A,B	283, 100	383	Mitaksov & Fremont (2006)
1CFN	IGG1-KAPPA Antibody CB41 / Protein (Bound Peptide)	C	10	A,B	214, 213	427	Keitel et al. (1999)
2BRR	MN20B9.34 Anti-P1.4 Antibody, FAB Light & Heavy Chain / Class 1 Outer Membrane Protein Variable Region 2	P	11	H,Y,L,X	225, 215	880	Oomen et al. (2005)
2QHR	13F6-1-2 FAB Fragment Heavy Chain. V lambda x Light Chain / Envelope Glycoprotein Peptide	P	11	H,L	218, 222	440	Lee et al. (2007)
1P4B	Antibody Variable Light & Heavy Chain / GCN4(7p-14P) Peptide	P	12	L,H	135, 124	259	Zahnd et al. (2003)
2OTW	FV Light & Heavy Chain Variable Domain VL & VH / Poly-Gln Peptide Antigen	E,F	12	A,C,B,D	115, 118	466	Li (2007)
1A3R	IGG2A 8F5 FAB / Human Rhinovirus Capsid protein VP2	P	15	L,H	220, 218	438	Tormo et al. (1998)
1TJI	Anti-HIV Antibody 2F5 Light & Heavy Chain / Envelope Glycoprotein GP41	P	17	L,H	214, 237	451	Ofek et al. (2004)
1F58	IGG1 ANTIBODY 58.2 / Protein (Exterior Membrane Glycoprotein (GP 120)	P	23	L,H	216, 228	444	Stanfield et al. (1998)

Table A1: (continued)

# of Complexes in this group (Group II - Small Antigens) = 26						
Code	Name	Antigen		Antibody		References
		Chain(s)	Size	Chain(s)	Size	
1FCC	IGG1 MO61 FC / Steptococcal Protein G (C2 Fragment)	C,D	56	A,B	206	Sauer-Eriksson et al. (1995)
1HEZ	KAPPA LIGHT CHAIN OF IG / PROTEIN L	E	61	A,C,B,D	214, 224	Graille et al. (2000)
1MHH	FAB / PROTEIN L DOMAIN C	E,F	63	A,C,B,D	220, 217	Graille & Stura (2002)
2BDN	Smaal Inducible Cytokine A2 / Antibody Light & Heavy Chain 11K2	A	76	L,H	214, 217	Boriack-Sjodin et al. (2005)
1E08	[FE]-Hydrogenase / Cytochrome	E	78	A,D	371, 88	Morelli et al. (2000)
2JEL	JEL42 FAB Fragment / Histidine-Containing Protein	P	85	L,H	217, 218	Parasad et al. (1998)
2R69	Major Envelope Protein E / Light & Heavy Chain of 1A1D-2	A	97	L,H	212, 214	Lok et al. (2007)
1BJ1	FAB Fragment / Vascular Endothelial Growth Factor	V,W	102	L,J,H,K	214, 231	Muller et al. (1998)
1JRH	Antibody A6 / INTERFERON-GAMMA RECEPTOR	I	108	L,H	213, 219	Winkler & Sogabe (1997)
1ZTX	Envelope Protein / Heavy & Light Chain of E16 Antibody	E	108	L,H	212, 219	Nybakken et al. (2005)
2P4A	Ribonuclease Pancreatic / Antibody CAB-RN05	A,C	124	B,D	121	Tereshko et al. (2007)
1BZQ	(RNASE A) / Antibody CAB-RN05)	A,B,C,D	124	K,L,M,N	124	Decanniere et al. (1998)
1BQL	HYHEL-5 FAB / BOBWHITE QUAIL LYSOZYME	Y	129	L,H	212, 215	Chacko & Davies (1995)
1C08	Anti-Hen Egg White Lysozyme Antibody (HYHEL - 10) / Lysozyme	C	129	A,B	107, 114	Shiroishi et al. (1999)
1DQJ	Anti-Lysozyme Antibody HYHEL-63 / Lysozyme	C	129	A,B	214, 210	Li & Mariuzza (2000)
1FDL	IGG1-KAPPA D1.3 FAB / Hen Egg White Lysozyme	Y	129	L,H	214, 218	Fischmann & Poljak (1990)
1JHL	IGG1-KAPPA D11.15 FV / PHEASANT EGG WHITE LYSOZYME	A	129	L,H	108, 116	Chitarra et al. (1993)
1RJC	Camelid Heavy Chain / Lysozyme	B	129	A	137	De Genst et al. (2003)
1YQV	HYHEL-5 Antibody Light & Heavy Chain / Hen Egg White Lysozyme	Y	129	L,H	211, 215	Cohen et al. (2005)
2DQJ	Lysozyme Binding Ig Kappa Chain V23-J2 Region / Ig VH, Anti-Lysozyme / Lysozyme C	Y	129	L,H	107, 114	Shiroishi et al. (2006)
1BVK	HULYS11 / LYSOZYME	C,F	129	A,D,B,E	108, 117	Holmes et al. (1998)
1FBI	IGG1 F9.13.7 FAB / Guinea Fowl Lysozyme	X,Y	129	L,P,H,Q	214, 221	Lescar & Alzari (1995)
1JTO	Vh Single-Domain Antibody / Lysozyme	L,M	129	A,B	148	Decanniere et al. (2001)
1MLC	IGG1-KAPPA D44.1 / HEN EGG WHITE LYSOZYME	E,F	129	A,C,B,D	214, 218	Braden et al. (1995)
1P2C	LIGHT CHAIN ANTI-LYSOZYME ANTIBODY F10.6.6 / HEAVY CHAIN VH+CH1 ANTI-LYSOZYME ANTIBODY F10.6.6 / LYSOZYME C	C,F	129	A,D,B,E	212, 218	Cauerhff et al. (2003)
1ZMY	Antibody cabbcll-10:lys3 / Lysozyme C	L,M	129	A	142	Saerens et al. (2005)

Table A1: (continued)

# of Complexes in this group (Group III - Large Antigens) = 21						
Code	Name	Antigen		Antibody		References
		Chain(s)	Size	Chain(s)	Size	
1LK3	Interlukin-10 / 9D7	A,B	160	L,M,H,I	210, 219	Josephson et al. (2002)
2UZI	ANTI-RAS FV Heavy & Light Chain / GTPASE HRAS	R	166	L,H	114, 104	Tanaka et al. (2007)
1XIW	T-Cell Surface Glycoprotein CD3 Epsilon Chain, Delta Chain / Immunoglobulin Light & Heavy Chain Variable Region	A,E,B,F	105,79	C,G, D,H	108, 122	Arnett et al. (2004)
2I9L	Antibody 7D11 Light & Heavy Chain / Virion Membrane Protein M25	I,J,K,L	184	A,C,E,G,B,D,F,H	219, 219	Su et al. (2006)
2DD8	IGG Heavy & Light Chain / Spike Glycoprotein	S	202	H,L	245, 213	Prabakaran et al. (2006)
1ADQ	IGG4 REA FC / IGM-LAMBDA RF-AN FAB	A	206	L,H	213, 225	Corper et al. (1997)
1E6J	Immunoglobulin / CAPSID PROTEIN P24	P	210	L,H	210, 219	Berthet_Colominas et al. (2000)
2B2X	Integrin Alpha-1 / Antibody AQC2 FAB Heavy & Light Chain	A,B	223	H.I,L,M	226, 213	Clark et al. (2005)
1G6V	Carbonic Anhydrase / Antibody Heavy Chain (CAB-CA05, Variable Domain)	A	260	K	126	Desmyter et al. (2000)
2R0K	Poly(A)-Specific Ribonuclease	A	283	L,H	214, 225	Nagata et al. (2008)
1YYM	Exterior Membrane Glycoprotein(GP120) / Antibody 17b Light & Heavy Chain / CD4M33, scorpion-toxin mimic of CD4	G,P	313	L,Q	214	Huang et al. (2005)
2NY7	Envelope Glycoprotein GP120 / T-Cell Surface Glycoprotein CD4 / Antibody 17B Light & heavy Chain	G	317	L,H	215, 230	Zhou et al. (2006)
2Q8B	Apical Membrane Antigen 1 / 1F9 Light & Heavy Chain	A	336	L,H	214, 210	Gupta et al. (2007)
1A14	Neuraminidase/NC 10 FV	N	388	L,H	104, 120	Malpy et al. (1997)
1NMC	NEURAMINIDASE / SINGLE CHAIN ANTIBODY	A,N	388	B,H,C,L	122, 190	Malpy et al. (1997)
2AEQ	Neuraminidase / FAB Light & Heavy Chain	A	395	L,H	214, 217	Venkatramani et al. (2005)
2J4W	Apical Membrane Antigen 1 / FAB Fragment of Monoclonal Antibody F8.12.19	D	445	L,H	213, 225	Igonet et al. (2006)
2NXY	Beta-Lactamase TEM	A,B	317,184	C,D	214,229	Wang et al. (2003)
1QFU	Protein (HEMAGGLUTININ (HA1 CHAIN)) / Protein (Immunoglublin IGG1-KAPPA Antibody Light & Heavy Chain)	A,B	328, 184	L,H	217, 223	Fleury et al. (1999)
2B4C	Envelope Glucoprotein / T-Cell Surface Glycoprotein CD4 / Anti-HIV-1 GP120 Immunoglobulin X5 Light & Heavy Chain	G, C	344. 181	L, H	215, 235	Huang et al. (2005)
2QAD	Envelope Glycoprotein GP120 / T-Cell Surface Glycoprotein CD4 / Anti-HIV-1 Antibody 412d Light & Heavy Chain	A,E,B,F	388, 181	C,G, D,H	214, 231	Huang et al. (2007)

Table A2

Protein Antigen Data - Group I (Peptides) # of Complexes = 15					
Amino Acid Residues	Epitope Surface		Entire Surface		Occurrence Propensity
	Raw Occurrence	Average Molar Fraction	Raw Occurrence	Average Molar Fraction	
ALA	8	0.057	10	0.068	0.847
ARG	8	0.054	8	0.052	1.036
ASN	8	0.058	8	0.055	1.050
ASP	10	0.071	10	0.070	1.011
CYS	0	0.000	0	0.000	0.000
GLN	15	0.091	15	0.091	1.009
GLU	12	0.082	12	0.074	1.107
GLY	4	0.034	5	0.043	0.804
HIS	6	0.044	6	0.042	1.044
ILE	2	0.015	2	0.015	1.000
LEU	13	0.099	16	0.111	0.887
LYS	6	0.046	7	0.048	0.959
MET	0	0.000	0	0.000	0.000
PHE	2	0.017	2	0.017	1.000
PRO	9	0.070	9	0.069	1.013
SER	6	0.051	7	0.056	0.915
THR	9	0.060	9	0.060	1.000
TRP	2	0.011	2	0.010	1.167
TYR	3	0.024	6	0.049	0.488
VAL	10	0.073	10	0.070	1.040

Table A2: (Continued)

Protein Antigen Data - Group II (Small Proteins) # of Complexes = 26					
Amino Acid Residues	Epitope Surface		Entire Surface		Occurrence Propensity
	Raw Occurrence	Average Molar Fraction	Raw Occurrence	Average Molar Fraction	
ALA	8	0.020	35	0.028	0.728
ARG	38	0.109	171	0.117	0.930
ASN	46	0.132	164	0.112	1.179
ASP	30	0.085	106	0.078	1.095
CYS	1	0.002	2	0.001	1.338
GLN	22	0.068	64	0.045	1.510
GLU	20	0.059	85	0.069	0.854
GLY	28	0.079	98	0.069	1.152
HIS	4	0.012	26	0.018	0.681
ILE	3	0.009	16	0.012	0.708
LEU	14	0.036	53	0.039	0.932
LYS	38	0.108	157	0.117	0.924
MET	4	0.010	11	0.008	1.298
PHE	3	0.008	24	0.018	0.450
PRO	8	0.025	51	0.034	0.737
SER	16	0.049	93	0.063	0.772
THR	28	0.085	103	0.076	1.106
TRP	10	0.079	32	0.023	3.506
TYR	18	0.050	42	0.031	1.612
VAL	8	0.024	58	0.041	0.576

Table A2: (Continued)

Protein Antigen Data - Group III (Large Proteins) # of Complexes = 21					
Amino Acid Residues	Epitope Surface		Entire Surface		Occurrence Propensity
	Raw Occurrence	Average Molar Fraction	Raw Occurrence	Average Molar Fraction	
ALA	10	0.037	64	0.021	1.744
ARG	22	0.075	220	0.073	1.017
ASN	22	0.072	307	0.099	0.733
ASP	22	0.076	231	0.077	0.987
CYS	2	0.087	23	0.012	0.359
GLN	15	0.052	280	0.066	0.785
GLU	26	0.093	321	0.108	0.866
GLY	7	0.022	85	0.030	0.736
HIS	11	0.037	77	0.026	1.403
ILE	20	0.066	102	0.035	1.884
LEU	9	0.034	128	0.039	0.859
LYS	31	0.114	437	0.124	0.923
MET	8	0.026	21	0.009	2.836
PHE	3	0.008	51	0.019	0.437
PRO	20	0.069	170	0.057	1.194
SER	10	0.035	182	0.057	0.618
THR	19	0.073	251	0.073	1.013
TRP	5	0.027	33	0.013	2.060
TYR	11	0.047	62	0.027	1.719
VAL	10	0.033	102	0.036	0.900

Table A2: (Continued)

Group II & III (Small & Large ProteinsCombined) # of Complexes = 47					
Amino Acid Residues	Epitope Surface		Entire Surface		Occurrence Propensity
	Raw Occurrence	Average Molar Fraction	Raw Occurrence	Average Molar Fraction	
ALA	18	0.028	99	0.025	1.162
ARG	60	0.092	391	0.095	0.964
ASN	68	0.102	471	0.105	0.970
ASP	52	0.081	337	0.077	1.041
CYS	3	0.045	25	0.007	0.468
GLN	37	0.060	344	0.056	1.080
GLU	46	0.076	406	0.088	0.861
GLY	35	0.051	183	0.049	1.026
HIS	15	0.025	103	0.022	1.111
ILE	23	0.037	118	0.024	1.575
LEU	23	0.035	181	0.039	0.895
LYS	69	0.111	594	0.120	0.924
MET	12	0.018	32	0.009	2.127
PHE	6	0.008	75	0.018	0.443
PRO	28	0.047	221	0.046	1.022
SER	26	0.042	275	0.060	0.698
THR	47	0.079	354	0.075	1.060
TRP	15	0.053	65	0.018	2.977
TYR	29	0.049	104	0.029	1.662
VAL	18	0.028	160	0.039	0.727

Table A2: (Continued)

Group I & II & III (Peptides & Small & Large ProteinsCombined) # of Complexes = 62					
Amino Acid Residues	Epitope Surface		Entire Surface		Occurrence Propensity
	Raw Occurrence	Average Molar Fraction	Raw Occurrence	Average Molar Fraction	
ALA	26	0.038	109	0.039	0.980
ARG	68	0.079	399	0.081	0.979
ASN	76	0.087	479	0.089	0.987
ASP	62	0.077	347	0.075	1.032
CYS	3	0.030	25	0.004	0.468
GLN	52	0.071	359	0.067	1.048
GLU	58	0.078	418	0.084	0.934
GLY	39	0.045	188	0.047	0.959
HIS	21	0.031	109	0.029	1.079
ILE	25	0.030	120	0.021	1.437
LEU	36	0.056	197	0.063	0.890
LYS	75	0.090	601	0.096	0.930
MET	12	0.012	32	0.006	2.127
PHE	8	0.011	77	0.018	0.619
PRO	37	0.055	230	0.054	1.019
SER	32	0.045	282	0.059	0.767
THR	56	0.073	363	0.070	1.043
TRP	17	0.039	67	0.015	2.596
TYR	32	0.040	110	0.036	1.125
VAL	28	0.043	170	0.049	0.875

Table A3

Antibody Data # of Complexes = 62					
Amino Acid Residues	Paratope Surface		Entire Surface		Occurrence Propensity
	Raw Occurrence	Average Molar Fraction	Raw Occurrence	Average Molar Fraction	
ALA	11	0.013	317	0.031	0.415
ARG	53	0.068	673	0.057	1.190
ASN	55	0.066	473	0.047	1.417
ASP	66	0.077	716	0.067	1.143
CYS	1	0.001	39	0.004	0.261
GLN	23	0.035	683	0.067	0.522
GLU	53	0.045	770	0.069	0.651
GLY	40	0.048	531	0.054	0.883
HIS	20	0.028	113	0.010	2.651
ILE	26	0.034	105	0.012	2.763
LEU	18	0.017	299	0.029	0.570
LYS	23	0.037	1209	0.107	0.347
MET	6	0.007	41	0.005	1.511
PHE	19	0.025	58	0.006	3.810
PRO	20	0.021	687	0.064	0.324
SER	69	0.086	1714	0.167	0.515
THR	66	0.085	1152	0.111	0.764
TRP	54	0.059	112	0.011	5.527
TYR	169	0.213	417	0.041	5.134
VAL	18	0.019	244	0.023	0.824

Table A5: Average distance in Angstroms (Å) from the epitope/paratope surface centers and the standard deviation of each distance (values listed in descending order).

Amino Acid Residues		Epitope Surface		Amino Acid Residues		Paratope Surface	
	Raw Freq.	Average	Std. Dev.		Raw Freq.	Average	Std. Dev.
GLY	40	9.99	3.44	GLN	23	12.31	4.47
TRP	17	9.81	2.95	GLU	53	11.85	2.81
THR	56	9.76	3.26	ILE	32	11.69	3.74
SER	32	9.74	3.91	PRO	21	11.61	4.63
ARG	68	9.67	3.5	ALA	11	11.57	5.28
VAL	28	9.47	4.31	SER	71	11.41	3.23
ASN	76	8.93	2.73	ASP	67	11.32	4.46
GLU	62	8.86	2.84	LYS	24	11.28	2.5
HIS	21	8.84	3.1	MET	7	10.78	4.15
LYS	75	8.79	3.28	GLY	43	10.67	3.49
GLN	52	8.77	3.45	CYS	1	10.56	10.56
LEU	37	8.76	2.77	THR	73	10.46	2.68
ASP	62	8.63	2.28	ASN	55	10.09	2.9
PRO	37	8.54	3.91	ARG	62	9.88	2.47
TYR	37	8.04	2.98	VAL	18	9.69	3.21
PHE	8	8.01	3.07	HIS	24	8.88	2.62
ALA	29	7.95	3.36	TYR	194	8.8	2.72
MET	12	7.36	3.28	LEU	19	8.66	2.3
ILE	25	6.61	3.55	TRP	54	8.41	4.55
CYS	2	4.95	1.73	PHE	19	8.3	2.55

Table A6: Average Epitope Occurrence Fraction

Amino Acid Residues	Average Epitope Occurrence Fraction				
	Group I	Group II	Group III	Group II & III Combined	Group I & II & III Combined
ALA	0.533	0.033	0.030	0.032	0.153
ARG	0.333	0.227	0.133	0.185	0.221
ASN	0.467	0.204	0.083	0.150	0.227
ASP	0.400	0.198	0.067	0.139	0.202
CYS	0.000	0.000	0.010	0.004	0.003
GLN	0.400	0.295	0.058	0.189	0.240
GLU	0.400	0.146	0.111	0.130	0.196
GLY	0.300	0.142	0.020	0.087	0.139
HIS	0.333	0.058	0.089	0.072	0.135
ILE	0.067	0.024	0.052	0.037	0.044
LEU	0.494	0.088	0.019	0.057	0.163
LYS	0.367	0.211	0.127	0.174	0.220
MET	0.000	0.042	0.057	0.049	0.037
PHE	0.133	0.032	0.012	0.023	0.050
PRO	0.533	0.134	0.101	0.119	0.219
SER	0.267	0.076	0.033	0.057	0.108
THR	0.400	0.150	0.039	0.100	0.173
TRP	0.067	0.141	0.030	0.091	0.085
TYR	0.333	0.179	0.062	0.127	0.177
VAL	0.533	0.056	0.042	0.050	0.167

Table A7: Comparison of amino acid composition in percent area contributions to the interface area of our data set to the Lo Conte et. al. data set

Amino Acid Residues	B	
	Our Data (All Interface regions)	Lo Conte et.al Data
ALA (A)	2.8	1.8
CYS (C)	0.4	0
ASP (D)	7.0	7.3
GLU (E)	5.5	4
PHE (F)	2.2	3
GLY (G)	5.3	5.9
HIS (H)	2.5	1.4
ILE (I)	3.2	3.1
LYS (K)	6.8	6.8
LEU (L)	5.0	3
MET (M)	1.2	0.8
ASN (N)	7.1	9.2
PRO (P)	2.9	2.6
GLN (Q)	5.2	3.8
ARG (R)	8.6	9.2
SER (S)	6.5	7.4
THR (T)	6.5	6.4
VAL (V)	4.0	1.5
TRP (W)	5.3	5.7
TYR (Y)	12.2	16.6

B - Percent Area Contributions to the Interface

Table A8: Substitution matrix before finding the Lograthmic Values

Antigen Epitope Surface(Ag)	Antibody Paratope Surface (Ab)																				
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
A	0.101	0.106	0.105	0.127	0.135	0.115	0.110	0.109	0.117	0.109	0.091	0.111	0.086	0.092	0.114	0.102	0.096	0.092	0.086	0.141	A
R	0.106	0.145	0.130	0.148	0.162	0.144	0.131	0.131	0.152	0.135	0.110	0.143	0.100	0.102	0.136	0.122	0.121	0.108	0.105	0.160	R
N	0.105	0.130	0.138	0.164	0.190	0.150	0.141	0.145	0.158	0.145	0.114	0.145	0.113	0.118	0.158	0.132	0.115	0.118	0.095	0.180	N
D	0.127	0.148	0.164	0.216	0.263	0.188	0.181	0.177	0.189	0.175	0.136	0.166	0.129	0.150	0.195	0.159	0.130	0.133	0.110	0.236	D
C	0.135	0.162	0.190	0.263	0.429	0.236	0.219	0.233	0.239	0.203	0.159	0.199	0.143	0.185	0.251	0.215	0.132	0.126	0.089	0.292	C
Q	0.115	0.144	0.150	0.188	0.236	0.178	0.166	0.168	0.179	0.161	0.126	0.164	0.115	0.125	0.178	0.152	0.122	0.113	0.095	0.211	Q
E	0.110	0.131	0.141	0.181	0.219	0.166	0.166	0.156	0.166	0.149	0.124	0.146	0.109	0.128	0.169	0.140	0.116	0.107	0.098	0.202	E
G	0.109	0.131	0.145	0.177	0.233	0.168	0.156	0.177	0.174	0.150	0.118	0.156	0.109	0.126	0.179	0.154	0.113	0.107	0.083	0.202	G
H	0.117	0.152	0.158	0.189	0.239	0.179	0.166	0.174	0.193	0.167	0.130	0.169	0.123	0.135	0.187	0.157	0.128	0.120	0.097	0.214	H
I	0.109	0.135	0.145	0.175	0.203	0.161	0.149	0.150	0.167	0.161	0.122	0.155	0.128	0.123	0.171	0.137	0.117	0.117	0.092	0.198	I
L	0.091	0.110	0.114	0.136	0.159	0.126	0.124	0.118	0.130	0.122	0.109	0.120	0.101	0.108	0.136	0.108	0.100	0.091	0.084	0.149	L
K	0.111	0.143	0.145	0.166	0.199	0.164	0.146	0.156	0.169	0.155	0.120	0.165	0.116	0.106	0.168	0.142	0.121	0.113	0.089	0.192	K
M	0.086	0.100	0.113	0.129	0.143	0.115	0.109	0.109	0.123	0.128	0.101	0.116	0.143	0.113	0.144	0.100	0.090	0.092	0.077	0.147	M
F	0.092	0.102	0.118	0.150	0.185	0.125	0.128	0.126	0.135	0.123	0.108	0.106	0.113	0.145	0.148	0.113	0.094	0.096	0.090	0.153	F
P	0.114	0.136	0.158	0.195	0.251	0.178	0.169	0.179	0.187	0.171	0.136	0.168	0.144	0.148	0.208	0.157	0.117	0.116	0.086	0.224	P
S	0.102	0.122	0.132	0.159	0.215	0.152	0.140	0.154	0.157	0.137	0.108	0.142	0.100	0.113	0.157	0.143	0.105	0.103	0.080	0.179	S
T	0.096	0.121	0.115	0.130	0.132	0.122	0.116	0.113	0.128	0.117	0.100	0.121	0.090	0.094	0.117	0.105	0.112	0.109	0.100	0.139	T
W	0.092	0.108	0.118	0.133	0.126	0.113	0.107	0.107	0.120	0.117	0.091	0.113	0.092	0.096	0.116	0.103	0.109	0.146	0.104	0.131	W
Y	0.086	0.105	0.095	0.110	0.089	0.095	0.098	0.083	0.097	0.092	0.084	0.089	0.077	0.090	0.086	0.080	0.100	0.104	0.115	0.103	Y
V	0.141	0.160	0.180	0.236	0.292	0.211	0.202	0.202	0.214	0.198	0.149	0.192	0.147	0.153	0.224	0.179	0.139	0.131	0.103	0.277	V
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	