Biology-Based Matched Signal Processing and Physics-Based Modeling For Improved Detection

by

Brian O'Donnell

A Dissertation Presented in Partial Fulfillment
of the Requirement for the Degree
Doctor of Philosophy

Approved November 2014 by the
Graduate Supervisory Committee:

Antonia Papandreou-Suppappola, Chair
Daniel Bliss
Stephen A. Johnston
Narayan Kovvali
Cihan Tepedelenlioglu

ARIZONA STATE UNIVERSITY

December 2014

ABSTRACT

Peptide microarrays have been used in molecular biology to profile immune responses and develop diagnostic tools. When the microarrays are printed with random peptide sequences, they can be used to identify antigen antibody binding patterns or immunosignatures. In this thesis, an advanced signal processing method is proposed to estimate epitope antigen subsequences as well as identify mimotope antigen subsequences that mimic the structure of epitopes from random-sequence peptide microarrays. The method first maps peptide sequences to linear expansions of highly-localized one-dimensional (1-D) time-varying signals and uses a time-frequency processing technique to detect recurring patterns in subsequences. This technique is matched to the aforementioned mapping scheme, and it allows for an inherent analysis on how substitutions in the subsequences can affect antibody binding strength. The performance of the proposed method is demonstrated by estimating epitopes and identifying potential mimotopes for eight monoclonal antibody samples.

The proposed mapping is generalized to express information on a protein's sequence location, structure and function onto a highly localized three-dimensional (3-D) Gaussian waveform. In particular, as analysis of protein homology has shown that incorporating different kinds of information into an alignment process can yield more robust alignment results, a pairwise protein structure alignment method is proposed based on a joint similarity measure of multiple mapped protein attributes. The 3-D mapping allocates protein properties into distinct regions in the time-frequency plane in order to simplify the alignment process by including all relevant information into a single, highly customizable waveform. Simulations demonstrate the improved performance of the joint alignment approach to infer relationships between proteins, and they provide information on mutations that cause changes to both the sequence and structure of a protein.

In addition to the biology-based signal processing methods, a statistical method is considered that uses a physics-based model to improve processing performance. In particular, an externally developed physics-based model for sea clutter is examined when detecting a low radar cross-section target in heavy sea clutter. This novel model includes a process that generates random dynamic sea clutter based on the governing physics of water gravity and capillary waves and a finite-difference time-domain electromagnetics simulation process based on Maxwell's equations propagating the radar signal. A subspace clutter suppression detector is applied to remove dominant clutter eigenmodes, and its improved performance over matched filtering is demonstrated using simulations.

To my parents Robert and Janice,

and my wife Kayley

ACKNOWLEDGEMENTS

My deepest thanks go to my advisor, Prof. Antonia Papandrou-Suppapola for her guidance, support, and mentoring. I would like to express special thanks to Prof. Dan Bliss for going above and beyond as a committee member and mentor. Thank you to the other members of my committee, Dr. Narayan Kovvali, Prof. Cihan Tepedelenlioglu, and Dr. Stephen Johnston for their guidance and useful feedback.

I would like to thank Dr. Phillip Stafford and Josh Richer for all of their assistance and patience with my questions about random sequence peptide arrays.

I would like to thank my colleagues in the Signal Processing and Adaptive Sensing Lab for their insightful discussions and friendship.

I am extremely thankful for my time at the University of New Hampshire's Center for Coastal and Ocean Mapping, where I worked with Dr. Brian Calder, and learned an immense amount about performing engineering research.

I would also like to acknowledge the University of New Hampshire's Electrical Engineering department, without whom none of this would have been possible.

This work was supported in part by Air Force Research Laboratory contract FA8650-08-D-1303, Defense Threat Agency Reduction Contract HDTRA1-12-C-0058, the Office of Naval Research, the Ira A. Fulton Schools of Engineering Dean's Fellowship, the Ira A. Fulton Fellowship, and a Doctoral Fellowship from the ASU Security and Defense Systems Initiative (SDSI) Institute.

TABLE OF CONTENTS

Page

CHAPTER

APPENDIX

LIST OF FIGURES

Figure Page

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Introduction and Motivation

Detection methods determine whether or not an observed noisy signal consists of useful information. Once information is detected, then information parameters need to be estimated, or specific information patterns need to be identified. One approach to improving detection performance is to first process an observed signal before applying statistical methods designed to determine the presence of a signal in noise. Signal processing techniques are most useful when they are designed to match the properties of the observed signal or when they can apply a physical-based model that describes the observation medium.

In this dissertation, we consider two detection applications that benefit from matched signal processing techniques, or techniques that rely on physics-based models. The first application is in molecular biology, and it involves mapping one-dimensional (1-D) protein sequences or three-dimensional (3-D) protein structures onto signals with highly-localized representations in the time-frequency plane. Using processing techniques that are matched to these mapped signals can provide important information for identifying diseases or for drug discovery. The second application involves detecting small radar targets in heavy sea clutter. Using a dynamic model for generating sea clutter based on the governing physics of water gravity and capillary waves provides useful information for improving the target detector design.

1.2 Signal Processing Methods in Molecular Biology

The area of bioinformatics is mainly involved with the management of biological information using computer technology and statistics. Signal processing for molecular biology, on the other hand, encompasses the development of algorithms and methodologies for extracting, processing and interpreting information from biological sequences [3–8]. Intelligent use of signal processing algorithms can provide invaluable insight into the structure, function and evolution of biological systems. For example, complex assays that determine the functional activities of analytes or peptide chips that manifest key residues for protein binding can provide a wealth of information on underlying biological systems. However, in each of these cases, appropriately designed processing is required to robustly extract the most relevant information. Images of array fluorescence are enhanced to improve the estimation of gene reactivity, while gene expression classification performance is increased by including biological and experimental variability in the algorithm design [6].

Genomics and proteomics, in general terms, study the functions and structures of genomes and proteomes, respectively. Genomes, which are genetic material of organisms encoded in deoxyribonucleic acid (DNA) or ribonucleic acid (RNA), and proteomes, which are expressed proteins in given organisms, provide discrete information, represented in sequences of unique molecules [9, 10]. More specifically, DNA are bio-molecules that are represented as letter sequences of precise orderings of four nucleobases; the different orderings correspond to patterns that influence the formation and development of different organisms. Similarly, proteins are bio-molecules represented as sequences of unique orderings of twenty linked amino acids, with each amino acid represented by a letter of the alphabet.

DNA and protein sequence analysis requires significant processing of the discrete gene orderings in order to identify intrinsic common features, or find gene variations such as mutations [11, 12]. One important application in genome analysis is the identification of gene sequence periodicity; this periodicity selects regions of genetic repetition that have been shown to correlate with functionally important genes [13, 14]. Gene periodicity has been analyzed using spectral methods [15–18]; such methods have also been used to estimate variations in base pair frequencies between organisms as they can indicate phylogenic origin from the species genome. Time-frequency (TF) signal processing methods such as wavelet transforms have also been used in gene sequencing to characterize long range correlations or identify irregularities in DNA sequences [19, 16, 20].

Signal processing methods have also been used for sequence alignment. This is a method for ordering sequences to identify regions of similarity due to functional, structural, or evolutionary relationships between the sequences [21, 22]. As thousands of organisms have been sequenced completely, and many more have been partially sequenced, searching for these similarities requires a vast number of computations. There are many algorithms designed to perform these searches including dynamic programming algorithms such as Smith-Waterman, basic local alignment search tool (BLAST), correlation based methods, Bayesian approaches, and TF based methods [23–27, 12, 28–30]. Computational alignment tools based on dynamic programming such as the Smith-Waterman algorithm are guaranteed to find all similarity matches, but they are slow and inefficient [23]. Other tools, such as BLAST [24, 25], are widely made available for database similarity searching as they were developed to provide a fast approach of approximating the complete alignment found by dynamic programming algorithms. BLAST runs very quickly, around an order of magnitude faster than the complete alignment algorithms, and finds most significant alignments under most circumstances. However, it tends to miss alignments for queries with repetitive segments. Correlation based methods map DNA or amino acid sequences to sequences of real or complex numbers and use correlation between sequences of numbers to compute similarity [28]. Correlation algorithms can be implemented efficiently using the fast Fourier transform; however, errors increase when aligning sequences of longer lengths. We have recently developed a TF based method that first uniquely maps DNA or amino acid sequences to highly-localized Gaussian waveforms in the TF plane and then uses the matching pursuit decomposition (MPD) algorithm to perform alignment [30–32]. The TF-based alignment approach was compared to other approaches and was shown to perform well with repetitive segments in real time without pre-processing.

In addition to gene sequencing, microarray analysis has also played a significant role in the extraction and interpretation of genomic information. Microarrays can provide measurements of expression levels of large numbers of genes. For example, peptide microarrays have been used to study binding properties and functionality of different types of protein-protein interactions and to provide insight into specific pathogens [33–37]. Peptide microarrays are a relatively new application in biological signal processing. The technology to create assays using single peptide chains has been around for a while in the form of the enzyme linked immunosorbent assay (ELISA) [38]. In recent years, as the cost of printing many peptide clusters onto a single substrate has been dropping, tens or hundreds of thousands of peptide clusters can be reasonably printed on a single array. In addition to being able to construct large scale peptide arrays to detect specific diseases, another important aspect are robust analysis methodologies used to interpret and analyze the extracted peptide data in order to establish relationships between peptide sequences and binding strengths. Some of these methodologies include support vector machine (SVM) modeling methods

[39], computational alignment approaches [40] and statistical tools such as t-test and analysis of variance linear regression [41–43].

1.3 Epitope and Mimotope Estimation Methods

The human body's response to a foreign pathogen is a complex process of creating cells to identify, inhibit, and eventually destroy the pathogen. Antibodies play an integral role in this response, as their primary function is to recognize and neutralize pathogens and alert the immune system on the pathogen's presence. Antibodies, acting as pathogen biomarkers, can be detected by locating the pathogen binding site or epitope [44, 45]. As antibodies can recognize epitopes of multiple amino acids (AAs) in length, each peptide may have several such binding sites. As a result, mapping antibody epitopes on a target pathogen is very critical in diagnosing diseases. Mimotopes are peptides that mimic antibody binding sites and have binding characteristics similar to epitopes [46]. In particular, an antibody for a given epitope antigen responds similarly to both the epitope and its derivative mimotope. Mimotopes have been shown to induce an epitope antigen response with vaccination [47, 48] and thus have the not yet realized potential to be used in developing new vaccines and diagnostics. It is also possible to design methods for mapping mimotopes to a source antigen in order to find the interacting epitope on the antigen [49].

There are a number of existing research tools that can be used to provide information about a pathogen's epitope as well as the binding strength of the antibody-to-pathogen interface and the effect of changing the epitope sequence on the antibody binding strength. Some of the more frequently used tools to study pathogens by proxy of the antibodies include ELISA [50], phage display combined with peptide panning [51–54], and peptide microarrays. Phage display results in a single epitope estimate which is likely to be the strongest binder to the targeted antibody. Phage display can also be combined with peptide panning to further refine the epitope estimate by measuring the binding strength of a set of peptide sequences which are single AA substitutions from the phage display estimate of the epitope. Peptide microarrays are printed with either a set of peptides from a sequenced proteome [37] or from adaptations on known epitopes [36]. They are commonly used to profile the humoral immune response by finding the antigenic regions of a proteome. Highly binding peptides are antigen candidate sequences that are further verified by checking which of the sequences are on the folded protein surface and are physically available to an antibody for binding. Epitopes are efficiently mapped using peptide microarrays acting as screening tools for profiling antibody signatures and discovering diagnostic signatures. Peptide microarrays are designed to diagnose a specific infectious disease pathogen, assuming that the pathogen has already been identified and sequenced before the microarray is designed. Mimotopes have thus far been discovered by phage display technology [51], and mimotope databases have been developed based on information acquired from phage display [55, 56].

The recently developed random-sequence peptide microarrays provide platforms for identifying antigen antibody binding patterns or immunosignatures [44, 57, 58]. These microarrays have a major advantage over tests designed for one specific antibody as they adopt an unbiased sampling of hundreds of thousands of random, but known, peptide sequences. These random-sequence peptides are important for recognizing multiple antibodies from pathogens present in the testing blood sample, without any *a priori* knowledge of a specific disease. As a result, they can be used to classify different diseases based on identifying patterns of peptides that exhibit high image pixel median fluorescence intensity (MFI) in the microarray. From the MFI of the aggregate set of peptides, it is possible to estimate antibody epitope sequences and obtain their corresponding binding strength. Specifically, peptides with the highest

MFI are composed of specific binding to sub-sequences which match antigen epitopes or potential antigen mimotopes. The fluorescence patterns for pathogens have been shown to be consistent across patients and thus can act as biomarkers to classify patients into disease groups [59–63].

Epitope and mimotope estimation has been attempted before on smaller arrays with 5,520 and 10,000 peptides [34, 44]. The peptides with the largest binding strength tended to resemble the true epitopes, but with so few peptides, exact epitope sub-sequences of more than a few AAs did not exist on these smaller arrays. Increasing the number of peptides on the array increases the number of unique sub-sequences on the array as well as the number of times that those sub-sequences are repeated. Arrays with 100,000 or more peptides contain enough unique sequences to estimate exact epitopes with a high degree of reliability and robustness, and thus are adequate for diagnosing diseases from sequenced proteomes [64, 65]. Furthermore, given that random sequence peptides can yield mimotopes for many different antigens [66], many disease associated antigen could, in theory, be detected.

A challenge of analyzing random-sequence peptide microarrays is how to integrate peptide sequences and MFI measurements to estimate epitopes or identify mimotopes. NNAlign is an algorithm which attempts to solve this problem by generating neural network models from subsets of the peptide array data and then combining those multiple models into a single motif [67]. This algorithm provides a representation of AA probabilities at each position in the estimated motif. Another method for motif/epitope estimation uses regular expressions, a formally defined sequence of characters which forms a search pattern, to estimate epitopes. This method includes a dependence on the sub-sequence position within the peptide sequence [2].

Random sequence peptide arrays are in contrast to panned peptide arrays which start from epitope candidate sequences either derived from phage display, the proteome of the pathogen of interest. Proteome sub-sequences are printed on arrays as individual peptides [68], and epitope candidates from phage display and random-sequence peptides are printed on the array with exhaustive substitutions and truncations to find minimal length, exact epitope sub-sequences [69, 70]. However, these methods require and initial step that can be computationally intensive, can prohibit comprehensive specificity analysis, and can limit the biological search space. For example, if the search is limited to only linear sequences of a pathogens proteome, the methods may not be able to identify conformal epitopes that can be detected in large random sequence peptide arrays. Additionally, not all antibodies are linear sequences to proteins, or are even necessarily in response to proteins, making it difficult to determine a suitable peptide panning set.

1.4 Processing of Protein Structures

Protein alignment methods are used to arrange protein sequences or structures to identify regions of similarity or homology between proteins with common function, structure or evolutionary relationships. These methods are important not only for drug discovery but also for providing associations between gene mutations and diseases. Early pairwise protein alignment methods were based on the protein's primary structure or 1-D amino acid sequence [71–73]. Sequence protein alignment provides some degree of similarity accuracy but can have a difficult time matching dissimilar sequences which result in similar 3-D folded structures. Protein structures demonstrate different shapes due to the hydrogen bonds, ionic bonds, and van der Waals attractions between the molecules that make up the amino acids. The invention of x-ray crystallography provided atomic coordinate information. As a result, alignments are also performed using protein secondary and tertiary structures to improve matching algorithm performance [74–76]. Recent research into protein alignment focuses on developing methods that integrate multiple protein attributes in order to improve the evolutionary relevance between the identified matches [77–80]. Finding protein similarity in sequence, function and structure can lead to accurately inferencing distantly related homologs for which only some of the protein's functions may be conserved and to decreasing the number of structurally similar matches that have no evolutionary relationship.

The integration of multiple protein information, when available, can aid in determining homology and contributing to protein functionality [81]. Such information includes amino acid sequences [77], hydrophobicity and structural factors [82–85], hydrogen bonding potential and structural substitution matrices [80, 86], proteins geometrical location [78, 79], and protein domains [87]. Note, however, that integrating the different information is not a simple problem. There is no known optimal method to perform the integration and different integration methods yield different matching results [78, 75]. In [74], a method is provided that takes into consideration both protein sequence and structure information. The method models the evolutionary cost of protein mutations, insertions and deletions that occurred on the structure level during transformation as implied by changes in the protein sequences. Structure alignment can be taken into consideration for amino acid substitution matrices such as local substructure mutation matrices and hydrogen-bonding similarity. The alignment accuracy for different types of information is a function of how distantly related the proteins are. Amino acid sequence information is more useful for closely related proteins, while more distantly related proteins require supplemental structure information [80].

Traditional bioinformatic methods of protein sequence and structure pattern matching can be computationally intensive due to the large amounts of proteomic data available in reference databases. An approach toward this problem is the use of signal processing techniques for protein alignment after appropriately mapping the protein element representations [6, 30, 88, 89]. In [30], a robust querying algorithm is used based on a time-frequency signal expansion matched to the waveforms in the mapped protein sequences. This approach is shown to outperform current-used sequence alignment methods such as BLAST for queries with repetitive sequence segments [90]. A 3-D Gaussian waveform is considered in [88] for protein structure alignment, with each waveform mapping an individual amino acid. Graph theory and additional properties of the protein are utilized in [91]. In [89], a 3-D waveform matching algorithm is provided for local and global alignments between multiple protein structures. In this approach, linearly separable, highly-localized Gaussian waveforms are used to map links between amino acids with inherent directionality information.

1.5 Sea Clutter Radar Signal Processing

The detection and tracking of small targets on the sea surface is difficult, as strong scattering from the sea can mask weaker target reflections. In particular, at low grazing angles and high sea states, transmitted signals with bandwidths large enough to observe reflections from breaking waves and sea spikes can result in a low, or even negative signal-to-clutter ratio (SCR) [92]. In such heavy sea clutter scenarios, the detection performance deteriorates and the targets cannot be realistically tracked. Increasing the received signal power through antenna gain, transmitter power, and pulse Doppler processing may not improve detection as sea clutter returns consist of the transmitted signal undergoing small Doppler shifts relative to the target.

One approach to improving target detection performance at low radar cross section (RCS) is by accurately modeling the sea clutter statistics. This was demonstrated in prior work using the compound Gaussian model that relates back to the physical sea clutter phenomenology [93]. The model assumes that the sea clutter return consists of speckle and texture components. The speckle return is primarily a function of small-scale capillary waves forming a large number of independent scattering from the incident signal. The texture is a function of the large-scale gravity waves; it is assumed to modulate the local mean power of the speckle return, while exhibiting spatial correlation based on the range resolution, sea state, and wind speed [94]. The compound Gaussian model has been validated using real sea clutter data, and has been used to construct improved detectors and configure waveforms [95–97].

1.6 Dissertation Contributions

In this dissertation, we have three main contributions, summarized as follows.

1.6.1 Epitope Estimation and Mimotope Identification For Random Sequence Peptide Microarrays

We propose a signal processing based method for epitope estimation and mimotope identification using random-sequence peptide microarrays. In particular, we select an appropriate time-domain basis signal to map the AAs in a peptide sequence. For a unique mapping, we transform the highly-localized Gaussian signal in the TF plane; we then map transformation parameters, such as time and frequency shifts to AA characteristics, such as type, position in the sequence, and rate of change of type over time. Signal processing is an established area of research in electrical engineering for processing time-domain signals. Furthermore, a multitude of algorithms have already been developed and evaluated for analyzing, detecting, estimating, identifying and classifying signals. As a result, once peptide sequences are mapped to time-domain signals, the problem is to select the appropriate algorithms for finding exact and single-substitution matches between peptide sequences and and peptide sub-sequences. We use the number of times a sub-sequence is found to occur in a random-sequence for epitope estimation and mimotope identification. The number of times a single substitution sub-sequence occurs is used to determine the effects of single AA substitutions on epitope and mimotope binding strength. Note that we have successfully applied the TF mapping approach for DNA alignment and showed that it outperformed BLAST in some alignment cases [30].

1.6.2 Generalized Mapping for Protein Multi-Alignment

We propose a signal processing based protein alignment approach to search for similarities and establish homology by integrating information from multiple protein attributes. Specifically, we perform alignment by integrating information from primary sequences to model the process of sequence evolution by mutations, insertions and deletions; from geometric structures to provide similarity between 3-D shapes; and from amino acid physical-chemical features, such as hydrophobicity, since the pattern of hydrophobic residues in substitution matrices can provide alignment information of distantly related proteins. Variation in physical-chemical properties can lead to substitution patterns represented by matrices including the codon substitution matrix, Dayhoff evolutionary mutation matrix, hydrophobicity amino acid substitution matrix and block substitution matrix (BLOSUM) [84]. An amino acid substitution matrix can, for example, show that hydrophilic amino acids are more frequently substituted by hydrophobic amino acids than the vice versa scenario [86]. Also, predicting the effects of amino acid substitutions can lead to information on protein function [98].

As proteins have distinct 3-D geometrical shapes, our signal processing based approach first models the protein secondary or tertiary *structure* as a linear combination of 3-D Gaussian waveforms, following our initial work in [89]. The Gaussian waveform provides a compact time-frequency representation while encoding the 3-D position of a protein's α-carbons. The covariance matrix of the Gaussian waveform is designed using pairwise angles between two neighboring α-carbons. Time-shifting the Gaussian waveform provides information on the 3-D coordinates of the two amino acids; frequency-shifting the Gaussian waveform provides directionality by pointing between neighboring α-carbons. Additional transformations of the Gaussian waveform in the higher-order time-frequency plane can be used to map different protein attributes. These attributes include the 1-D protein sequence, that characterizes the *location* of covalently linked amino acids, and numeric amino acid physical-chemical features, such as substitution matrix entries that could lead to protein *function* information. Note that multiple features, such as entries from different substitution matrices, could be represented by additional unique waveform transformations. The three protein attribute mappings, location, structure, and function (LoStrFn) provide unique representation methods for performing multi-alignment (or alignment based on multiple attributes) not only based on denotation and geometric similarities but also on property similarity leading to different protein functions.

1.6.3 Physics-based Sea Clutter Model For Improved Target Detection of Low Radar Cross-Section Targets

In this dissertation we present the detection results using an externally developed physics-based sea clutter generation model based on an electromagnetic simulation of gravity and capillary waves evolving through time. By computing radar returns from the simulated sea surface and low RCS target scattering, we utilize the statistical variation of the returns to separate the target from the clutter and thus improve target detection performance. We specifically compare the performance of a matched filter detector to that of a subspace clutter suppression detector [97]. The subspace clutter suppression detector is an eigenmode analysis algorithm that exploits the statistical independence of clutter compared to the target of interest [99, 100]. As we demonstrate, this detector can separate and suppress clutter from the radar returns, significantly improving SCR and detection performance.

1.7 Dissertation Organization

This dissertation is organized as follows.

In Chapter 2, we first discuss the immunosignature data collection procedure and explain how we form the peptide subsequences. We describe the amino acid to signal mapping technique and provide details on our proposed peptide subsequence estimation algorithm. We provide and discuss our findings on epitope estimation and mimotope identification using data from monoclonal antibody (mAb) array samples. We also discuss the different factors that affect the algorithm performance.

In Chapter 3, we first provide the protein-to-waveform mapping model for the location, structure, and functional attributes of a protein. We then describe the 3-D alignment in terms of the structure, sequence, and hydrophobicity alignment. We demonstrate our results using two human mutant ferrochelatase proteins, showing how the structure and functional attributes improve alignment.

In Chapter 4 we first provide the physics-based finite-difference time-domain sea clutter simulation model based on Maxwell's equations and define the environmental and target parameters that define the strong sea clutter, and electromagnetically weak targets under study. We consider the problem of detecting these low radar cross section targets and define the generalized matched filter detector. We then present the better performing subspace cutter suppression detector, and we provide simulations demonstrating the detection performance of the two detectors under different modeled sea clutter scenarios.

Finally, Chapter 5 summarizes the work presented in this dissertation, provides some concluding remarks, and comments on future directions for these lines of research.

In Appendix A we provide a comprehensive list of all of the acronyms used in this dissertation. In Appendix B we list the full results of our subsequence estimation algorithm for all of the monoclonal antibody samples, and in Appendix C we provide some preliminary epitope estimates for random sequence peptide array data collected using blood samples of patients infected with different diseases.

Chapter 2

EPITOPE ESTIMATION AND MIMOTOPE IDENTIFICATION USING RANDOM SEQUENCE PEPTIDE MICROARRAYS

2.1 Immunosignature Random-Sequence Peptide Microarrays

The random-sequence peptide microarrays are designed by the Center for Innovations in Medicine at the Biodesign Institute at Arizona State University [101]. The data corresponds to immunosignatures from eight different monoclonal antibody (mAb) samples. The immunosignature assay is performed by incubating diluted blood antibodies on a microarray of random-sequence peptides. The peptides are printed on standard glass slides or synthesized onto silicon dioxide wafers and diced into standard slides [101]. The 330k random-sequence peptide microarrays have 330,034 probes manufactured by HealthTell, Inc. in Chandler, AZ. The sequences are sufficiently long such that binding occurs between an antibody and a subsequence of the peptide, but not the entire peptide sequence. The average length of the peptide sequences on the 330k microarray is 11.2 amino acids (AAs), with a standard deviation of 1.3 AAs. More specifically, 95% of the peptides are between 5 and 14 AAs long; the minimum and maximum lengths are 1 and 22 AAs, respectively. From the 20 AAs, the AAs cysteine, isoleucine, methionine, and threonine are not included in the selection. Note that these lengths do not include the constant AA linker sequence `GSG` (glycine-serine-glycine), which attaches the AA chain to the array substrate.

The arrays are first washed in dimethylformamide for an hour. The solvent phase is transitioned to an aqueous phase over a six hour period using a phosphate-buffered saline incubation buffer before incubating in the presence of antibodies or serum. In order to bind the antibodies to the arrays, the arrays are washed in distilled water and then loaded into a multi-well 24-up gasket. Each well receives an incubation buffer and diluted sera solution containing antibodies. A secondary fluorescing antibody is added to bind to the appropriate primary antibody. After incubation for an hour, the arrays are washed using a plate washer. The removed arrays are scanned and the resulting images are processed to provide raw microarray image data. The amount of antibody binding to a feature measured remotely by fluorescence; more signal results when more primary antibodies bind to the peptide and thus more secondary antibodies bind to the primary antibodies. A calibrated picture is taken of the fluorescing array, where pixels in the image have been associated with specific peptide clusters. The median fluorescence intensity (MFI) is calculated as the median value of the pixels associated with a given peptide cluster.

2.2 Forming Peptide Subsequences

Our objective is to detect and identify subsequences or their single AA substitutions from a microarray peptide sequence. The subsequences could correspond to epitopes or mimotopes of a specific pathogen. We consider an immunosignature microarray consisting of M peptide sequences; we denote the mth peptide sequence of length L_m as $\mathcal{V}_m$,$m=1,\ldots,M$. As the maximum number of AAs in a peptide sequence is 22 using the 330k microarray, the maximum value of $L_m=22$. By shifting one AA at a time in the mth peptide sequence, we obtain at most $N_m \leq (L_m-\mathcal{L}+1)$ unique, length $\mathcal{L}$, subsequences of $\mathcal{V}_m$. In particular, the ℓth shifting operation, $\ell=1,\ldots,N_m$, generates the ℓth subsequence, whose first and last AAs correspond to the ℓth and $(\ell+\mathcal{L})$th AAs of the peptide, respectively. We denote the aforementioned shifting function by $h_\ell(\mathcal{V}_m;\mathcal{L})$, $\ell=1,\ldots,N_m$, $m=1,\ldots,M$. This function generates the length-$\mathcal{L}$ ℓth subsequence of the mth peptide $\mathcal{V}_m$ in the array by shifting the starting position of the subsequence from the first AA position of the peptide to the ℓth AA position of the peptide. Using this function, we represent the ℓth unique subsequence of $\mathcal{V}_m$ as

$$\chi(\ell; d_m, \mathcal{L}) = h_\ell(\mathcal{V}_m; \mathcal{L})\,. \tag{2.1}$$

Here, d_m is the MFI of the mth peptide sequence $\mathcal{V}_m$; it is the same value for all subsequences of peptide $\mathcal{V}_m$. For example, considering the $L_m = 10$ AAs long peptide $\mathcal{V}_m$=`ARVYHKHKHE`, we can generate at most $(L_m - \mathcal{L} + 1) = 8$ unique subsequences of length $\mathcal{L} = 3$. The subsequences are $\chi(1; d_m, 3)$=`ARV`, $\chi(2; d_m, 3)$=`RVY`, $\chi(3; d_m, 3)$=`VYH`, $\chi(4; d_m, 3)$=`YHK`, $\chi(5; d_m, 3)$=`HKH`, $\chi(6; d_m, 3)$=`KHK`, $\chi(7; d_m, 3)$=`HKH`, $\chi(8; d_m, 3)$=`KHE`. Since two of the subsequences are identical, $\chi(5; d_m, 3) = \chi(7; d_m, 3)$ = `HKH`, then the number of unique sequences is $N_m = 7$.

To achieve our objective, we find the number of times each unique subsequence of length $\mathcal{L}$ is repeated on the microarray. We form all possible unique subsequences as the union of all subsequences from the M microarray peptides. Specifically, there are at most $\mathcal{J} \le \sum_{m=1}^{M} N_m$ unique subsequences, χ_j, $j = 1, \ldots, \mathcal{J}$, in the set

$$\mathcal{S}_\mathcal{L} = \bigcup_{m=1}^{M} \bigcup_{\ell=1}^{N_m} \chi(\ell; d_m, \mathcal{L})\,. \tag{2.2}$$

Note that, in practice, it is uncommon for a single peptide to contain repeated subsequences; even when this occurs, it is only for the smaller length subsequences of $\mathcal{L} = 4$ or $\mathcal{L} = 5$ AAs. It is much more common that different peptides share the same subsequences.

2.3 Time-Frequency Mapping of Peptide Subsequences

The proposed peptide subsequence estimation algorithm is based on first mapping the peptide AAs to unique signals and then using time-frequency (TF) signal processing techniques to detect recurring patterns. The mapping uses the basic Gaussian signal, $g_b(t) = \pi^{-1/4} \exp(-0.5\,t^2),\ \ t \in (-T_g, T_g)$, as it is the most localized signal in the TF plane. The effective duration $2\,T_g$ is normally chosen to ensure minimum computational processing complexity. The basic Gaussian signal has unit energy and is centered at the TF origin. We design the AA-to-signal mapping as follows. Considering N_m subsequences of length $\mathcal{L}$ formed from the mth peptide $\mathcal{V}_m$ of length L_m, we map each AA to the time-shifted and frequency-shifted Gaussian signal

$$g(t; l, k) = g_b(t - lT)\ \exp(j2\pi kFt), \ \ t \in (lT - T_g,\ lT + T_g)\,. \tag{2.3}$$

The time shift parameter lT is used to represent the lth AA in the peptide subsequence, $l = 1, \ldots, \mathcal{L}$. The frequency shift parameter, kF, $k = 1, \ldots, 20$, is used to map the 20 existing AAs, as shown in Figure2.1. Using this mapping, the $\mathcal{L}$ AAs long ℓth subsequence $\chi(\ell; d_m, \mathcal{L})$, $\ell = 1, \ldots, N_m$, in Equation (2.1) can be represented by the linear combination of $\mathcal{L}$ TF shifted Gaussian signals as

$$x_{\ell,m}(t) = \sum_{l=1}^{\mathcal{L}} g(t; l, u[\{\alpha_l\}]) = \sum_{l=1}^{\mathcal{L}} g_b(t - lT)\ \exp(j2\pi\, u[\{\alpha_l\}]\, F\, t), \tag{2.4}$$

on the domain $t \in (\ell\, T - T_g,\ (\ell + \mathcal{L})T + T_g)$.

Note that we denote $x_{\ell,m}(t)$ to be dependent on m to clarify that the mapped signal originated from the mth peptide. This dependence is required for the estimation algorithm since we need to track the MFI of the subsequence; both the peptide and any of its generated subsequences have the same MFI. The function $u[\{\alpha_l\}]$ in (2.4), that replaced k in (2.3), is the integer-valued frequency shift that is used to map the type of the lth AA. Figure 2.2 provides an example of the mapping for the subsequence `EEDFRV` of length $\mathcal{L} = 6$ AAs. Note, for example, that time shifts $l = 1, 2$, share the same frequency shift, $u[\{\alpha_1\}] = u[\{\alpha_2\}] = 14$, since the type of AA (glutamic acid) is the same for both positions in the subsequence. Using the mapping, the weighted

Gaussian signal representation for the mth peptide $\mathcal{V}_m$ is given by

$$v_m(t) = \sum_{i=1}^{L_m} g(t; i, u[\{\alpha_i\}]) = \sum_{i=1}^{L_m} g_b(t - iT)\, \exp\left(j2\pi\, u[\{\alpha_i\}]\, F\, t\right), \qquad (2.5)$$

where $t \in (T - T_g,\ L_m T + T_g)$ and L_m is the length of the peptide AA sequence, $m = 1, \ldots, M$.

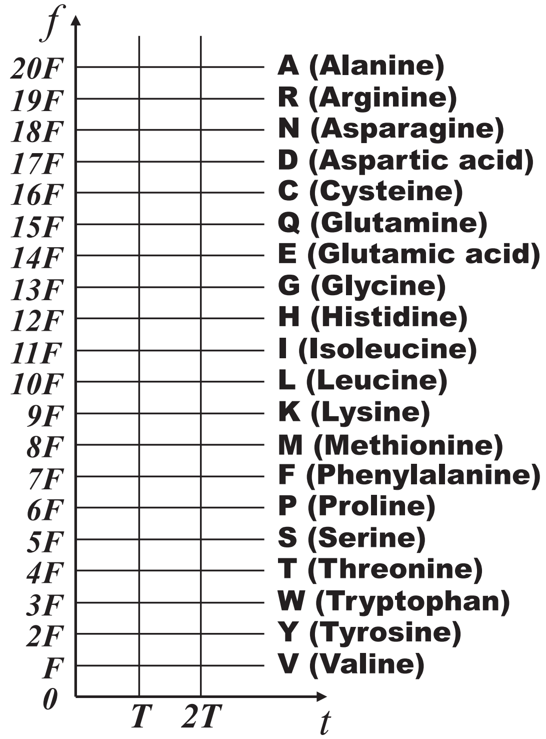


Figure 2.1: TF Representation of Mapping AA Type to Frequency Shifts

2.4 Peptide Subsequence Estimation Algorithm

Once the set $\mathcal{S}_{\mathcal{L}}$ of all unique subsequences of length $\mathcal{L}$ on a microarray are formed as in (2.2), we need to find the OCRC of each subsequence; we use occurrence count (OCRC) as a metric for the number of times each subsequence appears in the array.

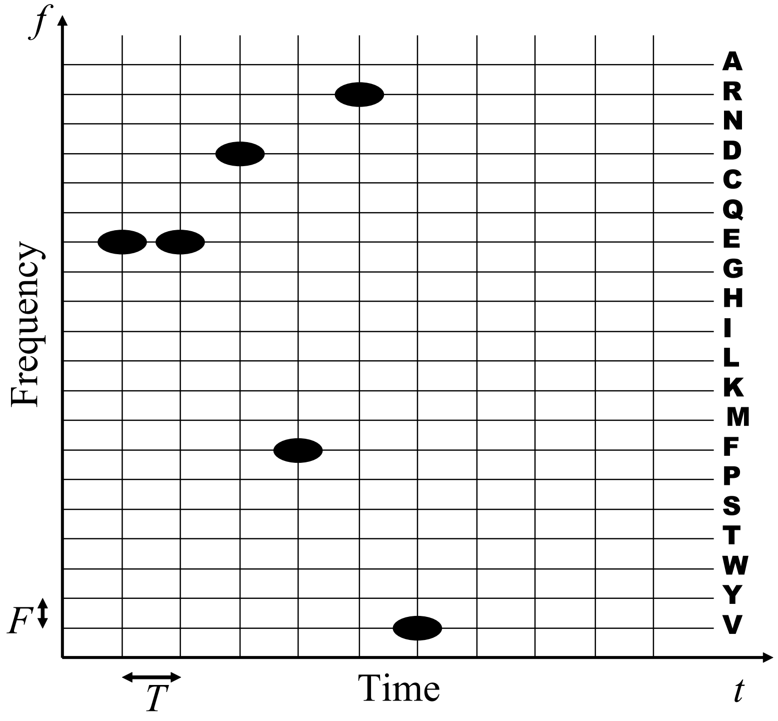


Figure 2.2: TF Representation of the Mapped AA Subsequence `EEDFRV`.

In addition to OCRC, we sometimes down-select the set of peptides used to find an occurrence count, and denote this as the down-selected occurrence count (DS-OCRC).

In particular, we want to detect the signal $x_{\ell,m}(t)$ in (2.4) that represents the ℓth subsequence $\chi(\ell; d_m, \mathcal{L})$ of length $\mathcal{L}$, $\ell = 1, \ldots, N_m$, of the mth peptide within all possible signals $v_m(t)$, $m = 1, \ldots M$, that represent the M peptides. This process is analogous to searching for similarity between a given subsequence and all the peptide sequences on the microarray. Essentially, we use this approach to estimate pathogen epitopes and identify candidate pathogen mimotopes.

We perform the subsequence estimation and identification method in TF using the matching pursuit decomposition algorithm [102]. The matching pursuit decomposition (MPD) is an iterative signal expansion technique that can be used to represent a signal with time-varying spectral characteristics as a linear combination of basis functions. Normally, the basis functions are selected from a dictionary that consists of a basic Gaussian signal that is centered at the TF origin as well as time-shifted, frequency-shifted and scaled transformed versions of this basic signal. Transformed Gaussian signals form the dictionary as they highly-localized in the TF plane; however, based on the application, the MPD can give a sparse representation if the dictionary is formed using real signals [103].

If the signal under processing is well-matched in TF to the Gaussian basis functions, then the algorithm converges after only a few iterations; otherwise, the MPD can be computationally intensive. For our application, the processing signals are perfectly matched to the Gaussian basis functions as we map the AAs in the peptide sequences directly to Gaussian signals. We thus expect the MPD to converge fast when used to identify subsequences, provided that the time shift and frequency shift transformations of the MPD dictionary are selected to be integer multiples of the time and frequency shift parameters T and F in Equation (2.3), respectively.

Algorithm 1 provides the steps of our proposed approach to determine the DS-OCRC / OCRC of each unique subsequence χ_j, $j = 1, \ldots, \mathcal{J}$, of length $\mathcal{L}$, in a microarray. In order to compute both the DS-OCRC/OCRC of each subsequence as well as keep track of the MFIs of the peptides that contributed to the count, we compute the DS-OCRC/OCRC of the length-$\mathcal{L}$ ℓth unique subsequence $\chi(\ell; d_m, \mathcal{L})$ of the mth peptide, $m = 1, \ldots, M$. The subsequence is represented by the signal $x_{\ell,m}(t)$ with duration $(\mathcal{L}T + 2T_g)$ and MFI d_m. To reduce computational cost, we need to ensure that we do not unnecessarily process two or more subsequences when their corresponding mapped signals $x_{\ell,m}(t)$ and $x_{\ell',m'}(t)$, $m \neq m'$ and any ℓ or ℓ', are identical; each subsequence to be processed is generated only once, because of how the subsequences are defined in Equation (2.2). The algorithm computes inner products between the linear combination of Gaussian signals in $x_{\ell,m}(t)$ that represent the ℓth subsequence and the linear combination of Gaussian signals $v_m(t)$ that represent the mth peptide. A perfect match is determined only when the sum of the inner product outputs is exactly equal to $\mathcal{L}$. The DS-OCRC/OCRC of the ℓth subsequence is the total number of perfect matches after processing all microarray peptides.

Algorithm 1 runs with the following considerations:

- Consider a sample microarray consisting of M random peptide sequences
- Either calculate the OCRC using $m = 1, \ldots, M$ peptides, or calculate the DS-OCRC using a down-selected peptide set of $m = 1, \ldots, Q$, where $Q \leq M$
- Use the approach in Section 2.2 to generate the length-$\mathcal{L}$ unique subsequence χ_j, $j = 1, \ldots, \mathcal{J}$, from the set $\mathcal{S}_{\mathcal{L}}$ in Equation (2.2); equivalently, by ensuring that each subsequence is not generated more than once when considering all peptides, generate the length-$\mathcal{L}$ unique subsequence $\chi(\ell; d_m, \mathcal{L})$, $\ell = 1, \ldots, N_m$, from the mth, length-L_m, peptide, $m = 1, \ldots, M$; note that $N_m \leq (L_m - \mathcal{L} + 1)$ and $\mathcal{J} \leq \sum_{m=1}^{M} N_m$
- Form a one-to-one correspondence between the indexing of the unique subsequences: χ_j is equivalent to $\chi(\ell; d_m, \mathcal{L})$, with $j = 1, \ldots, \mathcal{J}$, $\ell = 1, \ldots, N_m$, and $m = 1, \ldots, M$
- Use the TF mapping in Section 2.3 to represent: (i) the mth peptide of length L_m, $m = 1, \ldots, M$, by the signal $v_m(t)$ in Equation (2.5); and (ii) the ℓth sub-subsequence $\chi(\ell; d_m, \mathcal{L})$, $\ell = 1, \ldots, N_m$, of length $\mathcal{L}$ and MFI d_m, by the signal $x_{\ell,m}(t)$ in Equation (2.4)

Algorithm 1 Computation of OCRC Υ_j and Mean MFI $\bar{d}_j$ of Unique Subsequence χ_j in a Peptide Microarray

```
for $m = 1, \ldots, M$ do
  $\star$ Set count $= 0$ and $\bar{d} = 0$ to initialize the OCRC/DS-OCRC and the mean MFI respectively of the $\ell$th unique subsequence of the $m$th peptide
  $\star$ Form the MPD dictionary $\mathcal{D}_m = \{g(t; 1, u[\{\alpha_1\}]), \ldots, g(t; L_m, u[\{\alpha_{L_m}\}])\}$ using the signals in (2.3)
  $\star$ Denote any signal in the dictionary $\mathcal{D}_m$ by $y_m(t)$
  for $n = 0, 1, \ldots, (N_m - \mathcal{L} + 1)$ {shift the subsequence by one AA position at a time} do
    $\bullet$ Initialize the MPD iterations by setting $r_\ell^{(0,n)}(t) = x_{\ell,m}(t - nT)$
    for $\zeta = 0, \ldots, \mathcal{L} - 1$ {perform $\mathcal{L}$ MPD iterations} do
      $\diamond$ Compute $\mathcal{C}_{\ell,\mathcal{D}_m}^{(\zeta,n)} = \int r_\ell^{(\zeta,n)}(t)\ y_m^*(t)\, dt$, the correlation of $r_\ell^{(\zeta,n)}(t)$ with every dictionary signal
      $\diamond$ Select the dictionary signal with the maximum correlation
          $y_\ell^{(\zeta,n)}(t) = \underset{y_m(t) \in \mathcal{D}_m}{\operatorname{argmax}}\ \mathcal{C}_{\ell,\mathcal{D}_m}^{(\zeta,n)}, \quad t \in \Big((n+1)T - T_g,\ (n+\mathcal{L})T + T_g\Big)$   (2.6)
      $\diamond$ Compute the MPD coefficient $\lambda_{\zeta,n} = \int r_\ell^{(\zeta,n)}(t)\ \ y_\ell^{*\,(\zeta,n)}(t)\, dt$ {if the two AAs match, $\lambda_{\zeta,n} = 1$}
      $\diamond$ Compute the residue $r_\ell^{(\zeta+1,n)}(t) = r_\ell^{(\zeta,n)}(t) - \lambda_{\zeta,n}\, y_\ell^{(\zeta,n)}(t)$
    end for
    $\bullet$ Evaluate the sum of the MPD coefficients, $\Lambda_{m,n} = \sum_{\zeta=0}^{\mathcal{L}-1} \lambda_{\zeta,n}$
    if $\Lambda_{m,n} = \mathcal{L}$ then
      - Subsequence $x_{\ell,m}(t - nT)$, with fluorescence value $d_m$, is a perfect match in peptide $v_m(t)$
      - Update the mean MFI of as $\bar{d} = \big(\bar{d} \cdot \text{count} + d_m\big)\ /\ (\text{count} + 1)$
      - Increase OCRC/DS-OCRC of subsequence $x_{\ell,m}(t)$ by one, count $=$ count$+1$
    end if
  end for
end for
$\triangleright$ Obtain the final OCRC/DS-OCRC as $\Upsilon_j = \text{count}$
$\triangleright$ Algorithm output: The OCRC/DS-OCRC and mean MFI of subsequence $\chi_j$ are $\Upsilon_j$ and $\bar{d}$, respectively
```

2.5 Estimation of Subsequences with Single AA substitutions

Subsequences formed by replacing a single AA with another AA are called point mutations or single AA substitutions. Although substituting one AA can significantly change the peptide structure and binding characteristics, sometimes the effect is unimportant to structure or binding. Silent mutations occur when the substitution is by an AA with similar properties as the original AA, resulting in no significant change in functionality [104]. As a result, single substitutions of AAs with similar properties are important to consider for estimating specific types of subsequences such as epitopes and mimotopes, or substitutions are in the epitope, but do not form critical contacts with the antibody.

Algorithm 1 can be modified to estimate subsequences with single AA substitutions at a time. In particular, the design of the proposed algorithm is inherently matched to handle substitutions with computational ease. This is because the algorithm only needs to find subsequence matches with identical mapped time shifts, as they represent the position of an AA in the sequence; all frequency shifts are allowable as they represent the AA type. Note, however, that we need to keep track of the exact AA substitution in order to determine the OCRC of a silent mutation. The resulting approach for estimating silent mutations is described in Algorithm 2.

Algorithm 2 runs with the following considerations:

- Consider a sample microarray consisting of M random peptide sequences
- Either calculate the OCRC using $m = 1, \ldots, M$ peptides, or calculate the DS-OCRC using a down-selected peptide set of $m = 1, \ldots, Q$,where $Q \leq M$
- Following Algorithm 1, generate the length-$\mathcal{L}$ unique subsequence χ_j, $j = 1, \ldots, \mathcal{J}$, or equivalently $\chi(\ell; d_m, \mathcal{L})$, $\ell = 1, \ldots, N_m$, from the mth, length-L_m, peptide,

$m = 1, \ldots, M$; form a one-to-one correspondence between the subsequence indexing

- Following Algorithm 1, use TF mapping to represent the mth peptide sequence by $v_m(t)$ and the ℓth subsequence $\chi(\ell; d_m, \mathcal{L})$ by $x_{\ell,m}(t)$ in (2.4)

2.6 Peptide Sequence Down-Selection and Bias-Normalization

Although the microarrays consist of a very large number of peptides, not all peptides are applicable for detecting antibody subsequences that bind to specific antigens. In order to avoid unnecessary processing, we down-select the peptides using two different schemes. The first scheme involves down-selecting peptides with high MFIs; this is because only a small fraction of the peptides bind strongly and specifically to the monoclonal antibody samples. The remaining peptides bind weakly and non-specifically, and thus do not provide sufficient information on the sample antibodies. Antibody peptides that bind specifically, but only somewhat strongly to antigens are also not down-selected. To include these peptides, we use a second scheme which involves the calculation of Pearson's (PCC) that can be used to down-select peptides that bind strongly on only one of the monoclonal antibody samples. The PCC is calculated between a vector of MFIs and a reference vector, and it measures the similarity between the two vectors. PCC of -1, 0, and 1 imply negative correlation, no correlation, and positive correlation, respectively. For each of the M peptides in the ϱth microarray sample, $\varrho = 1, \ldots, \mathcal{P}$, the PCC is calculated as

$$r_{\varrho,m} = \Big(\mathbf{s}_m - \bar{s}_m \, \mathbf{1}_{\mathcal{P}}\Big)^{\mathsf{T}} \Big(\mathbf{b}_\varrho - \frac{1}{\mathcal{P}} \, \mathbf{1}_{\mathcal{P}}\Big) \tag{2.8}$$

for $m = 1, \ldots, M$. Here, $\mathbf{s}_m = [s_{1,m} \;\; \ldots \;\; s_{\mathcal{P},m}]^{\mathsf{T}}$, $s_{\varrho,m}$ is the mean MFI of the mth peptide in the ϱth microarray sample, $\bar{s}_m = (1/\mathcal{P}) \sum_{\varrho=1}^{\mathcal{P}} s_{\varrho,m}$ is the MFI of all the mth peptides in the $\mathcal{P}$ microarray samples, $\mathbf{1}_{\mathcal{P}}$ is a $\mathcal{P} \times 1$ column vector of ones, $\mathbf{b}_\varrho$ is

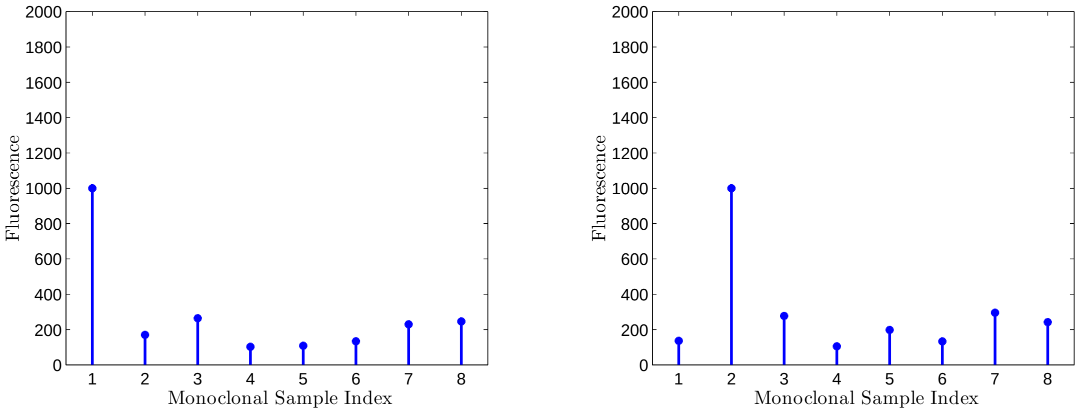
Algorithm 2 Computation of OCRC and Mean MFI of Subsequences With Single AA Substitutions.

```
for $m = 1, \dots, M$ do
  $\star$ Form the MPD dictionary using the signals in Equation (2.3)
  {Ensure that any position on the sequence can be substituted at a time by any of 16 possible AAs}
  {Exclude AAs threonine, methionine, soleucine, and cysteine that are not used in the 330k microarray; these AAs correspond to frequency shifts $k = 4, 8, 11, 16$, respectively, in Equation (2.3)}
      $\mathcal{D}_m = \{g(t;i,1), \dots, g(t;i,3), g(t;i,5), \dots, g(t;i,7), g(t;i,9), g(t;i,10), g(t;i,12), \dots, g(t;i,15) \mid i = 1, \dots, L_m\}$
  $\star$ Denote any signal in the dictionary $\mathcal{D}_m$ by $y_m(t)$
  for $n = 0, 1, \dots, (N_m - \mathcal{L} + 1)$ {shift the subsequence by one AA position at a time} do
    for $l = 1, \dots, \mathcal{L}$ {consider AA at the $l$th position of $\chi(\ell; d_m, \mathcal{L})$} do
      for $k = 1, 2, 3, 5, 6, 7, 9, 10, 12, 13, 14, 15$ {substitute AA at the $l$th position by the $k$th AA} do
        • Generate the $(l,k)$th new subsequence $q_{l,k,m}(t - nT)$ of $x_{\ell,m}(t - nT)$ by substituting the $l$th position of $x_{\ell,m}(t - nT)$ by the $k$th AA
            $q_{l,k,\ell,m}(t - nT) = g(t - nT; l, k) + \sum_{\substack{l'=1 \\ l' \neq l}}^{\mathcal{L}} g(t - nT; l', u[\{\alpha_{l'}\}])$        (2.7)
        • Set $\text{count}_{l,k} = 0$ and $\bar{d}_{l,k} = 0$ to initialize the OCRC/DS-OCRC and mean MFI of the $(l,k)$th subsequence
        • Initialize the MPD iterations by setting $r_\ell^{(0,n)}(t) = q_{l,k,\ell,m}(t - nT)$
        • Perform $\mathcal{L}$ iterations as in Algorithm 1 to obtain the MPD coefficients $\lambda_{\zeta,n}$, $\zeta = 0, \dots, \mathcal{L}$
        • Compute the sum of the MPD coefficients, $\Lambda_{m,n} = \sum_{\zeta=0}^{\mathcal{L}-1} \lambda_{\zeta,n}$
        if $\Lambda_{m,n} = \mathcal{L}$ then
          - Mapped subsequence $q_{l,k,\ell,m}(t - nT)$ that was derived from the mapped subsequence $x_{\ell,m}(t - nT)$, with MFI $d_m$, is a perfect match in the mapped peptide $v_m(t)$
          - Update the mean MFI of subsequence $q_{l,k,\ell,m}(t - nT)$:
                $\bar{d}_{l,k} = \left(\bar{d}_{l,k}\, \text{count}_{l,k} + d_m\right) / (\text{count}_{l,k} + 1)$
          - Increase by one the OCRC of subsequence $q_{l,k,\ell,m}(t - nT)$:
                $\text{count}_{l,k} = \text{count}_{l,k} + 1$
        end if
        $\triangleright$ Obtain the final OCRC as $\Upsilon_{l,k,j} = \text{count}_{l,k}$ and the final mean MFI as $\bar{d}_{l,k,j} = \bar{d}_{l,k}$
        $\triangleright$ Algorithm output: The OCRC and mean MFI of the subsequence formed by substituting the $l$th position of $\chi_j$ by the $k$th AA are $\Upsilon_{l,k,j}$ and $\bar{d}_{l,k,j}$, respectively
      end for{$k$th for loop}
    end for{$l$th for loop}
  end for{$n$th for loop}
end for{$m$th for loop}
```

a $\mathcal{P} \times 1$ reference vector that is defined as the ϱth column of a $\mathcal{P} \times \mathcal{P}$ identity matrix, and $\mathtt{T}$ denotes vector transpose. The reference vector indicates the correlation pattern needed to match the ϱth array.

Down-selecting based on the PCC provides an effective ranking metric for various cases, as illustrated in the following three examples. The first example assumes that all $\mathcal{P} = 8$ samples have approximately the same MFI. Such a situation can occur when all samples are either binding non-specifically to something in the antibody or not binding to anything. Using the reference vector $\mathbf{b}_1 = [1\ \mathbf{0}_7]$ for the sample $\varrho = 1$, the PCC is computed as $r_{1,m} = 0.01$ in (2.8), and $\mathbf{0}_\varrho$ is a $\varrho \times 1$ vector of zeros. The second example assumes a specific binding at the microarray for which the PCC is computed. Specifically, as shown in Figure 2.3a, the MFI of the specific binding in the $\varrho = 1$ sample is higher than the values of the non-specific binding in the $\varrho = 2, \ldots, 8$ samples. Using reference vector $\mathbf{b}_1$, the PCC is $r_{1,m} = 0.98$ for the $\varrho = 1$ sample. In the last example, the specific binding is for the $\varrho = 2$ sample, as shown in Figure 2.3b; using $\mathbf{b}_1$, the PCC for the $d = 1$ sample is $r_{1,m} = -0.22$. Thus, the correlation for the MFI in Figure 2.3a is very large as the binary vector matches the MFI pattern, whereas the correlation for the MFI in Figure 2.3b is negative as the binary vector does not match the pattern.

The PCC provides a better metric than MFI for ranking peptides with antigen binding subsequences. The binding to an epitope different than the original epitope can be equal to the the original binding. If that occurs, peptides with larger MFI on the sample of interest, relative to the same peptide on other samples, will be kept because of that specific binding. This is demonstrated for the monoclonal `Ab8` in Figure 2.4. Using the PCC instead of MFI to rank peptides resulted in a larger fraction of peptides with epitopes. This behavior was typical for most of the monoclonal samples. In the few cases where MFI ranking resulted in a higher percentage of the



(a) (b)

Figure 2.3: The MFIs in (a) and (b) Are Due to a Specific Binding for the First and Second Monoclonal Antibody Samples, Respectively, and Non-Specific Binding for All Other Samples.

selected peptides containing epitopes, the PCC also performed well in estimating the epitope. Note that when we used the MFI as the ranking metric for monoclonal `Ab8`, the epitope was not correctly estimated.

In some cases, it was found that the subsequence estimation performance increased when the MFIs of the down-selected peptides were normalized. The normalization tends to remove biases in the data resulting from inter-experimental variation (wafer-to-wafer-synthesis variation, temperature, duration, mechanical forces) or intra-experimental variation (sub-wafer variation, peptide location effects). The normalization approaches used include logarithmic (log10) normalization (resulting in Gaussian-like characteristics), median normalization, and linear model normalization [40, 105]. The effect of normalization is demonstrated in Figure 2.4 for monoclonal `mAb8`. For example, logarithmic normalization of the MFIs before computing the PCC resulted in more peptides with subsequences than combined logarithmic and median normalizations. Note, however, that the best estimation results were obtained when the MFIs were not normalized, indicating that the data are of consistent quality.

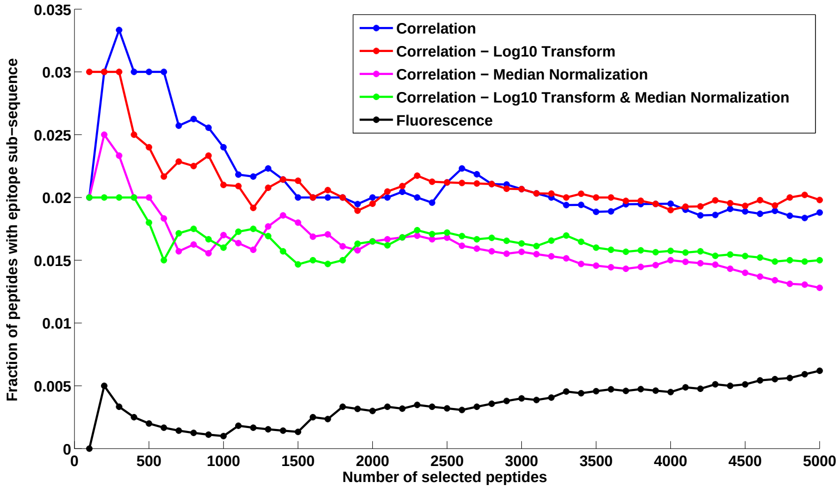


Figure 2.4: Fraction of Peptides With Epitopes for Different Numbers of Down-Selected Peptides for Monoclonal Antibody `Ab8`.

2.7 Subsequence Estimation Results

The analysis data consisted of 330k peptide microarrays for eight monoclonal antibody samples and a list of the synthesized peptides. The peptides are the same for all eight samples, allowing for comparison calculations across different samples for the same peptide. Algorithms 1 and 2 provide the steps for estimating epitopes and identifying mimotopes based on finding unique subsequences and their DS-OCRC. The most frequently occurring subsequences in the down-selected peptides are selected as the estimated epitopes. The algorithms also provide a list of additional subsequences that, although they do not occur as frequently as the epitope estimates, they still occur a sufficiently large number of times to warrant further investigation. These subsequences are proposed as potential antigen mimotopes as they appear to have readily permissible substitutions of the true epitopes.

2.7.1 Epitope Estimation Performance Results

We used the algorithms to estimate epitopes for the eight monoclonal samples as the most frequently occurring subsequences. The resulting estimated epitopes are listed in Table 2.1, together with their OCRC, DS-OCRC and mean MFI. As demonstrated in the table, the algorithms estimated exact subsequences for the full epitopes of the monoclonal antibodies `2C11`, `A10`, `Ab1`, `Ab8`, and `DM1A`; close matches were obtained for `4C1`, `Flag`, and `HA`. These results demonstrate both the diversity of the peptides on the microarray, spanning enough of the possible sequence space to bind all eight monoclonal antibodies, as well as the high performance of the epitope estimation algorithm in finding relevant epitopes.

Sample	**Full**	**Estimated**			**Mean**
mAb	**Epitope**	**Epitope**	**OCRC**	**DS-OCRC**	**MFI**
2C11	NAHYYVFFEEQE	VFFEEQE	22	7	805
4C1	LQAFDSHYDY	GYDSR	21	13	8,731
A10	EEDFRV	EDFRV	34	20	65,535
Ab1	NTFFRHSVVV	RHSVV	209	186	65,535
Ab8	TFSDLWKLLPE	DLWKL	63	6	1,174
DM1A	AALEKDYEEVGV	AALEKD	2,053	5	2,368
Flag	DYKDDDDK	AALEKDG	2,001	1,323	44,567
HA	YPYDVPDYA	YDAPE	16	14	61,414

Table 2.1: Epitope Estimates With OCRC, DS-OCRC, and the Mean MFI for Those Estimates.

The performance of the epitope estimation algorithm is tightly coupled to the frequentness and diversity of the subsequences in a microarray. By frequentness we mean how often a specific subsequence (of fixed length) occurs in the whole microarray; this is important because it affects the number of peptides the antibodies bind on and, as a result of that, the number of down-selected peptides that contain an epitope subsequence increases, and those subsequences are at the top of the DS-OCRC those peptides; and by diversity we mean the variety of peptide subsequences included in the whole microarray. We present next the processing of specific subsequences for four of the eight monoclonal antibody epitopes. For `Ab1`, we estimated the exact epitope whereas for `4C1`, `Flag`, and `HA`, we obtained comparable (not exact but similar) epitopes. Comparable and not exact epitopes are estimated because the true epitopes have low OCRC on the microarray and also the subsequences estimated have only moderately strong binding strength.

2.7.2 Epitope Estimation Analysis

As the microarray peptides are typically much longer than the estimated epitopes, the monoclonal antibodies only bind to a fractional portion of a peptide. It is thus only possible to infer that a particular subsequence contributed to the binding if that subsequence is present on multiple peptides with large MFIs. The success of the estimation algorithm also depends on the diversity of the microarray peptides; this is achieved using the sufficiently large 330k random sequence peptide microarray. In particular, many of the shorter length subsequences were found to repeat numerous times. As a result, this increased the robustness of the estimation algorithm and also allowed for an analysis of single AA substitutions based on binding strength.

In order to determine how well subsequences of different lengths are represented, we list the number of potential subsequences on the microarray in Table 2.2. On

Subsequence length	**# of unique subsequences**	**# of possible subsequences**	**% of unique subsequences**
4	58,700	65,500	89.5%
5	550,000	1,050,000	48.1%
6	1,490,000	1,680,000	9%
7	1,880,000	2,680,000	0.7%

Table 2.2: Number of Possible and Unique Subsequences of Varying Lengths on the Microarray.

the 330k peptide microarray, approximately 90% of length-4 (and 50% of length-5) subsequences occur on the array. Also, many of these subsequences are repeated multiple times, as shown in Table 2.3. As it can be observed, most of the length-4 and length-5 subsequences of the monoclonal epitopes are present on the array and are also repeated multiple times. This occurs for the epitopes of monoclonal antibody samples `2C11`, `A10`, `Ab1`, and `Ab8` and `DM1A`, for which we obtain exact epitope estimates. The results for the remaining three monoclonal antibody samples, `4C1`, `Flag`, and `HA` did not provide exact matches to the full epitopes. It is important to emphasize that the performance of the proposed estimation algorithm depends on the design of the random peptides on the microarray. More specifically, the performance depends on how frequently subsequences of the full epitope occur, whether the actual subsequences are present, and how strongly the antibodies bind to the peptides with these subsequences. As it is not possible to provide the details of every selected epitope, we illustrate next some specific examples which show trends in the data.

Our analysis demonstrated that it is possible that the full epitope does not correspond to the subsequence with the highest binding strength. This is demonstrated

Subsequence	**% of subsequences repeated at least $\mathcal{G}$ times**					
Length	$\mathcal{G}=5$	$\mathcal{G}=10$	$\mathcal{G}=50$	$\mathcal{G}=100$	$\mathcal{G}=500$	$\mathcal{G}=1,000$
4	99.8%	99.5%	95.2%	90%	69.1%	46.1%
5	94.2%	89.2%	61.5%	38.6%	1.2%	0.3%
6	57.8%	37%	2.6%	0.4%	0.2%	0.2%
7	5.9%	1.2%	0.2%	0.2%	0.1%	0.1%

Table 2.3: Percentage of Subsequences of Varying Lengths That Are Repeated in the Microarray at Least $\mathcal{G}$ Times.

with the monoclonal antibody sample `Ab1`, with full epitope `NTFFRHSVVV`. Table 2.4 lists the matched subsequences, their OCRC and corresponding mean MFIs for `Ab1`. Although the AA `T` occurs in the full epitope, we do not consider this AA in our estimation as it was not used to generate the peptides [101]. Also, when computing the OCRC of a short subsequence whose identical AA pattern appears in a longer subsequence, we do not include the OCRC of the longer subsequences. For example, when computing the OCRC of `HSVV`, we did not include the peptides which contain `RHSVV`, `RHSVVV` or any other higher-length subsequences of `NTFFRHSVVV`. This is because we wanted to ensure that the OCRC metric for `HSVV` is not influenced by the binding strength of longer subsequences. From Table 2.4, we can conclude that while `RHSVV` has the highest binding strength, the smaller length `HSVV` also has a high binding strength when compared to other subsequences. No conclusions can be made from the single occurrence of `RHSVVV` because some variability exists in the MFI measurements, and because multiple subsequence occurrences are required to disambiguate which subsequence on a peptide caused the antibody binding. Also, longer subsequences such as `FFRHS`, `FRHSV`, and `HSVVV` have very low binding strength.

The estimation results for Ab1 are typical for other samples in that not all subsubsequences of the epitope bind strongly to the antibody. Typically, the longest subsequence was estimated and listed in Table 2.1, and this often corresponded to the most dominant subsequence, that is, the subsequence with the highest binding strength. For Ab1, the dominant subsequence is RHSVV (shaded in Table 2.4). Note, however, that not only RHSVV but also HSVV occurred more frequently than the other length-4 and length-5 epitopes. However, RHSVV has comparatively larger binding strength.

Ab1 **subsequence**	**OCRC**	**Mean** **MFI**
$\mathcal{L} = 4$		
FFRH	44	1,394
FRHS	28	2,711
RHSV	87	3,119
HSVV	402	11,455
SVVV	5	1,087
$\mathcal{L} = 5$		
FFRHS	4	2,250
FRHSV	2	1,308
RHSVV	208	65,535
HSVVV	7	2,062
$\mathcal{L} = 6$		
RHSVVV	1	10,502

Table 2.4: Subsequences of Varying Lengths $\mathcal{L}$ for Ab1 Where the Shaded Row Corresponds to the Estimated Epitope.

The exact epitope was not estimated for the monoclonal antibody HA. The full epitope of this monoclonal is YPYDVPDYA; however, the estimated epitope YDAPE appears to be a substitution (at positions 3 and 5) of the exact epitope YDVPD. We thus selected this non-exact epitope as our estimate since the exact subsequence occurred very infrequently on the array. Tables 2.5a and 2.5b show the occurrences of different epitope subsequences and the mean MFIs for the antibody epitope subsequence YDVPD and the estimated epitope sequence YDAPE, respectively. While the antibody epitope sequence YDVPD occurred on the array with a high binding strength, the estimated epitope subsequence YDAPE occurred more frequently, and with almost as high binding strength. The exact epitope was also not estimated for the monoclonal antibody Flag. The non-exact estimate for Flag was AALEKDG, which is interesting because it is a close match to the true epitope for monoclonal antibody DM1A of AALEKD. The similarity of the estimated epitope for Flag, and the true epitope of DM1A is due to the similarities between their true epitopes, and the sparsity of sufficiently long true epitope subsequence for Flag. The important overlap between these two epitopes is the KD AA pair, and the permissive binding of Flag antibodies.

The sparsity of true epitope subsequences of Flag on the array is seen in Table 2.17, which lists the median MFI and the OCRC for each of the subsequences for this monoclonal antibody. The only true epitope subsequence with high binding strength was DYKDD; however, this subsequence only occurred twice on the array, which is not very frequently for a 5-mer, and therefore it is hard to identify it as an important subsequence. The overlap between the epitopes of the monoclonal antibodies Flag and DM1A is the AA pair KD. The MFI effects of this overlap can be seen by comparing the MFIs of peptides which contain subsequences similar to the epitopes. Figures 2.5 and 2.6) provide scatter plots of the MFIs for all the peptides on the array that contain a 4-mer or longer subsequence of peptide AALEKD, the estimated epitope for DM1A. In

Figure 2.5, the MFIs of HA are plotted with respect to the MFIs of Flag. As expected, the MFIs for HA are small as this sample has the unrelated true epitope YPYDVPDYA. This is in contrast to the scatter plot of Figure 2.6 which plots the MFIs of DM1A versus Flag, which have related epitopes and therefore the peptides containing these subsequences are bound strongly.

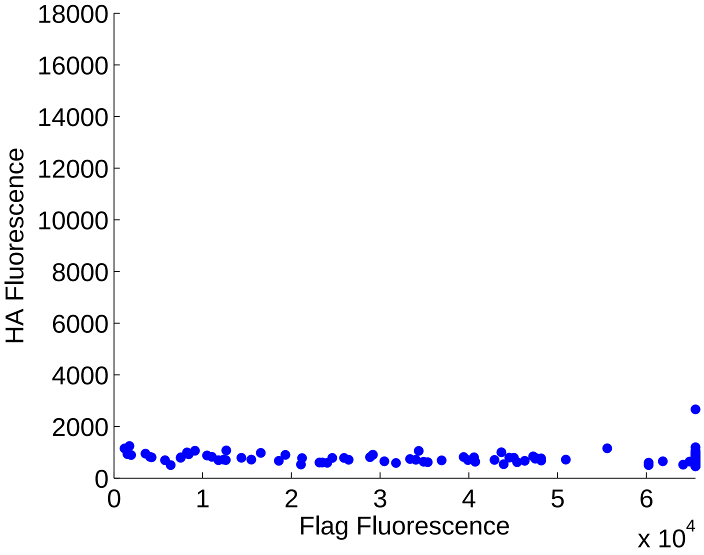


Figure 2.5: Scatter Plots of the Fluorescence of Flag Compared to the MFI of HA.

Here, we list out the top DS-OCRC results for monoclonal antibody samples 2C11, A10, and HA in Tables 2.6 - 2.14 to show what subsequence occur beyond the top estimate, what happens across multiple estimate lengths, and the characteristics of the potential mimotopes which occur in these lists.

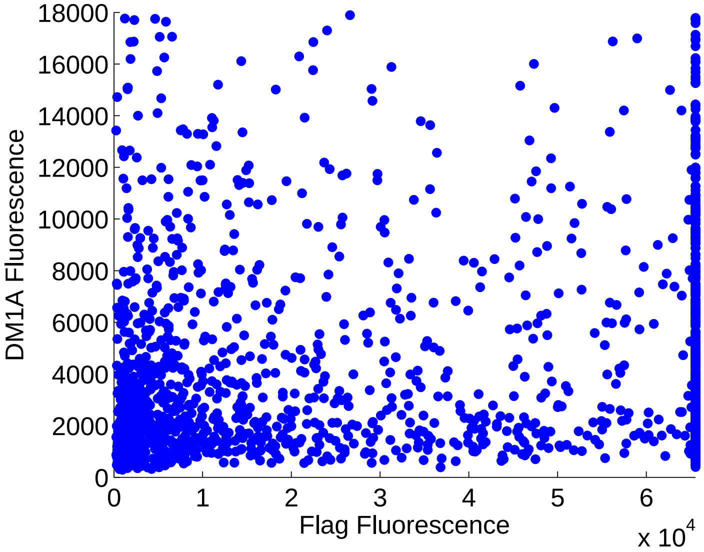


Figure 2.6: Scatter Plots of the Fluorescence of `Flag` Compared to the MFI of `DM1A`.

HA		Mean
subsequence	OCRC	MFI
$\mathcal{L}=4$		
YDAP	75	5,028
DAPE	98	884
$\mathcal{L}=5$		
YDAPE	16	61,414

(a)

HA		Mean
subsequence	OCRC	MFI
$\mathcal{L}=4$		
YPYD	22	813
PYDV	18	688
YDVP	42	3,377
DVPD	28	21,429
VPDY	19	746
PDYA	462	757
$\mathcal{L}=5$		
YPYDV	0	-
PYDVP	1	31,435
YDVPD	3	65,535
DVPDY	1	65,535
VPDYA	0	-
$\mathcal{L}=6$		
YPYDVP	0	-
PYDVPD	1	65,535
YDVPDY	0	-
DVPDYA	0	-

(b)

Table 2.5: Subsequences of Varying Lengths $\mathcal{L}$ for (a) the Estimated Epitope of HA, and (b) the True Epitope of HA. Note That in (a) the Shaded Row Corresponds to the Estimated Epitope, and That in (b) the $\mathcal{L}=5$ and $\mathcal{L}=6$ True Epitope Subsequence Do Not Occur Often on the Array.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
FEEQE	168	7	586	5,826
FFEEQ	117	7	636	5,826
VFFEE	87	7	676	5,826
ARWFN	54	6	931	65,535
AVNWF	64	6	760	187
PWFNK	139	6	848	2,144
WFNRL	30	6	1,010	1,704
ARLRP	120	5	1,098	4,613
ARRVR	30	5	1,980	4,142
DARWF	37	5	834	65,535

Table 2.6: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to mAb 2C11, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
FFEEQE	116	7	636	5,826
VFFEEQ	86	7	685	5,826
DARWFN	10	4	1,197	65,535
AWRGFN	7	3	997	1,692
FARLRE	9	3	1,183	3,327
FKYARL	24	3	1,208	2,414
HFFKAL	6	3	954	1,693
KARLRP	6	3	1,652	4,613
WFARLL	6	3	1,050	1,769
WFNGYA	12	3	938	1,470

Table 2.7: Estimated Top 10 Subsequences of Length $\mathcal{L} = 6$ Obtained When Algorithm 1 Was Applied to mAb 2C11, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
VFFEEQE	85	7	694	5,826
YVFFEEQ	22	3	805	2,089
AALEKDG	2,000	2	630	16,310
ALEKDGY	111	2	701	16,310
AVARPFQ	2	2	1,849	2,182
AVGWQAR	3	2	1,922	16,130
AWRGFNY	3	2	997	1,616
FARLREY	2	2	1,415	1,647
FEEQERY	13	2	656	1,559
FFEEQER	23	2	759	1,559

Table 2.8: Estimated Top 10 Subsequences of Length $\mathcal{L} = 7$ Obtained When Algorithm 1 Was Applied to mAb 2C11, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
WDVA	272	55	17,534	65,535
DVAW	473	52	8,790	65,535
DSAW	442	46	8,763	65,535
WQEA	135	46	65,535	65,535
DAAW	385	40	11,101	65,535
DVSW	239	36	19,765	65,535
QEYA	323	35	37,316	65,535
EDVA	242	34	20,428	65,535
WFEA	267	34	8,875	65,535
EWDA	346	32	10,617	65,535

Table 2.9: Estimated Top 10 Subsequences of Length $\mathcal{L} = 4$ Obtained When Algorithm 1 Was Applied to mAb A10, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
EDFRV	34	20	65,535	65,535
EWDVA	41	15	65,535	65,535
EDVAW	35	14	65,535	65,535
WFEGA	53	14	32,589	65,535
WDVAP	33	13	65,535	65,535
DAAWP	52	11	16,042	65,535
DVAWG	57	11	10,288	65,535
EWDAA	44	11	31,044	65,535
PWFEA	69	11	10,370	65,535
WDVAW	42	11	19,322	65,535

Table 2.10: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to mAb A10, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

In Tables 2.6 - 2.14 the most frequently occurring down-selected subsequences for lengths $\mathcal{L} = 7$, $\mathcal{L} = 5$, and $\mathcal{L} = 5$ respectively are the epitope estimates. While Algorithm 1 can be run for different length subsequences, we choose the longest, consistent top subsequence as the epitope estimate. For example, the top subsequence in Table B.4 is VFFEEQE. We see that the top few estimates in Tables 2.7 and 2.6 are all subsequence of VFFEEQE, e.g. FEEQE, FFEEQ, VFFEE, FFEEQE, and VFFEEQ. These top subsequence infer that the there is a longer length epitope subsequence. Similarly for HA and the top subsequence in Table 2.13 is YDAPE, and the top two subsequence in Table 2.12 are YDAP and DAPE.

While this sort of trend is seen in many of the monoclonal samples, sometimes the binding strength appears to be dependent on a more complete epitope. An example of this is seen for A10 in Table 2.10, where the top $\mathcal{L} = 5$ subsequence is the epitope estimate. Neither EDFR or DFRV are seen in the Table 2.9 $\mathcal{L} = 4$ subsequences, however

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
DFRVDW	22	8	35,188	65,535
FRVDWK	40	8	5,252	65,535
EDFRVD	6	5	65,535	65,535
EDVRPF	10	5	39,784	65,535
PWQEAS	7	5	65,535	65,535
AVWFEG	11	4	7,222	65,535
DVAWPF	12	4	22,508	65,535
EDARSG	6	4	34,672	65,535
EDVAPN	9	4	60,074	65,535
EDVAWP	6	4	65,535	65,535

Table 2.11: Estimated Top 10 Subsequences of Length $\mathcal{L} = 6$ Obtained When Algorithm 1 Was Applied to mAb A10, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

DFRV, DFRV, and FRV are present in the top three $\mathcal{L} = 6$ subsequences in Table 2.11. One of these two methods is used to determine which length subsequence should be the epitope estimate for each of the eight monoclonal antibodies.

Tables 2.6 - 2.14 also contain information about the potential mimotopes for those three monoclonal antibodies. The potential mimotope for 2C11 is DARWFN. It meets the four criteria listed for mimotopes, and some of its subsequences, ARWFN, ARWF, and WFN are seen in the $\mathcal{L} = 5$ subsequence in Table 2.6. Similarly, the potential mimotopes for A10 and HA meet the criteria for mimotopes, and subsequences of these mimotopes are seen in the top DS-OCRC lists of smaller lengths.

In addition to the results listed in Tables 2.6 - 2.14, we include a full list of Algorithm 1 results in Appendix B. We list the top 10 epitope estimates, ranked in descending order by DS-OCRC, for all eight mAbs, and $\mathcal{L} = 4, 5, 6, \text{and } 7$.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
YDAP	91	44	6,400	65,535
DAPE	114	31	1,537	65,535
ADAP	285	27	864	65,535
DVPE	93	25	1,008	65,535
DAPG	168	24	1,122	65,535
DVPD	33	24	31,506	65,535
DAPV	112	23	1,027	65,535
YDVP	47	23	4,846	65,535
LDVP	153	20	823	65,535
FDAP	47	18	2,071	65,535

Table 2.12: Estimated Top 10 Subsequences of Length $\mathcal{L} = 4$ Obtained When Algorithm 1 Was Applied to mAb HA, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
YDAPE	16	14	61,414	65,535
PYDAP	11	10	44,289	65,535
YDSPE	13	9	12,542	65,535
FDAPV	12	8	9,961	56,901
PFDAP	8	8	47,053	65,535
QYDAP	10	8	31,196	65,535
YDVPE	9	8	51,759	65,535
ADAPE	18	7	10,457	65,535
EDLPD	15	7	1,706	11,385
FYDAP	11	7	5,583	65,535

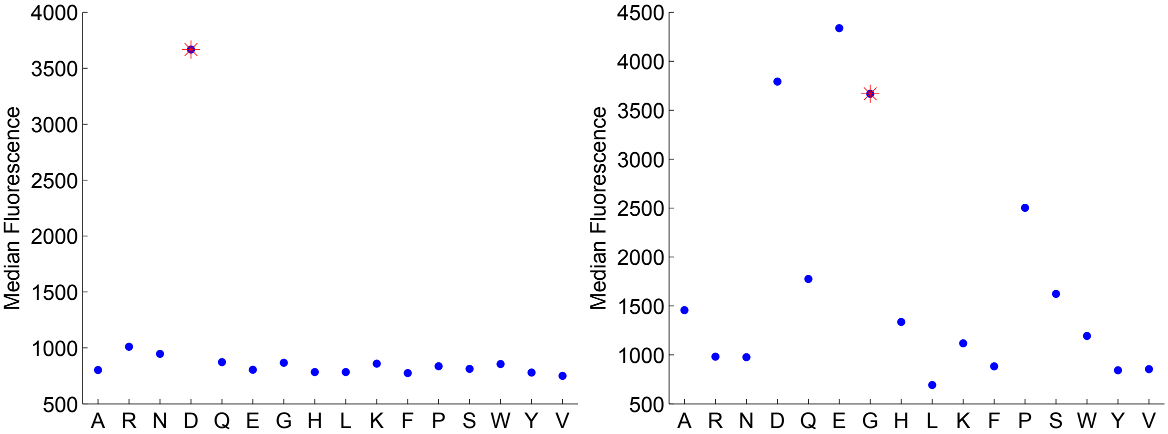
Table 2.13: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to mAb HA, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
FNYDSP	6	4	2,146	65,535
GYDAPE	4	4	59,422	65,535
NQYDAP	4	4	47,437	65,535
NYDSPE	4	4	11,997	65,535
AALEKD	2,053	3	694	11,285
ALEKDG	2,002	3	697	11,285
APYDAP	3	3	44,289	65,535
EDHPDG	3	3	4,984	40,563
EDLPDS	4	3	6,698	11,385
FFYDAP	3	3	6,135	65,535

Table 2.14: Estimated Top 10 Subsequences of Length $\mathcal{L} = 6$ Obtained When Algorithm 1 Was Applied to mAb HA, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

2.8 Substitution Analysis

The epitope estimates are derived from the peptides on the array which contain that epitope subsequence. In addition to that specific subsequence, there are other peptides on the array which contain that same subsequence, but with a single AA substitution. Our proposed algorithm for detecting subsequences using single AA substitutions is provided in Algorithm 2. Using this algorithm, we can analyze how these single AA substitutions affect the binding strength. In doing this, we see that the binding is not exact, but that some of the AAs in the epitopes can be substituted without much of a loss in binding strength; in some cases, these substitutions increase the binding strength. However, there are also specific AAs in the epitope subsequence which are required for the binding, and substituting them with different AAs can dramatically decrease the binding strength. One example of this is seen in Tables 2.15a and 2.15b which show AA substitutions at positions which are tolerant of substitutions and intolerant of substitutions, respectively. Figures 2.7a and 2.7a are plots of the MFI listed in the tables; the plots clearly show how much more tolerant of substitutions `4C1` is for epitopes in the first AA of the subsequence `_YDS` than it is for substitutions in the third AA of the subsequence `GY_S`. The tolerance for AA substitutions is particularly helpful when trying to estimate an epitope whose exact subsequences do not appear frequently on the array. This is true for `Flag`, where the third AA of the exact subsequence `KDDD` is substituted to form subsequence `KDGD`; this subsequence appears more frequently on the microarray.



(a) MFIs for _YDS substitutions

(b) MFIs for GY_S substitutions

Figure 2.7: The Mean MFIs for Two Different Substitutions of GYDS: (a) Substitution _YDS is Tolerant of Substitutions; and (b) Substitution GY_S is Not Tolerant of Substitutions.

AA	OCRC	Mean MFI
A	11	1,457
R	19	982
N	100	977
D	6	3,792
Q	85	1,776
E	129	4,337
G	158	3,667
H	201	1,337
L	5	693
K	67	1,119
F	107	883
P	55	2,503
S	9	1,624
W	50	1,194
Y	6	844
V	16	855

(a)

AA	OCRC	Mean MFI
A	267	803
R	195	1,011
N	26	947
D	158	3,667
Q	13	873
E	21	805
G	6	867
H	20	784
L	165	784
K	42	859
F	13	775
P	47	837
S	21	813
W	16	856
Y	9	780
V	37	751

(b)

Table 2.15: AA Substitutions With OCRC and Mean MFI for (a) GY_S and (b) _YDS.

2.8.1 Mimotope Identification

In addition to showing the top epitope results for all eight monoclonal antibodies, we show expanded results for monoclonal antibodies 2C11, A10, and HA in Tables 2.6 - 2.14. These results in these tables are listed in descending order by DS-OCRC. The most frequently occurring subsequence according to the DS-OCRC, and the subsequences can be used to find mimotope sequences.

Sample mAb	**Full Epitope**	**Potential Mimotope**
2C11	NAHYYVFFEEQE	DARWFN
4C1	LQAFDSHYDY	ADSWP
A10	EEDFRV	EWDVA
Ab1	NTFFRHSVVV	-
Ab8	TFSDLWKLLPE	-
DM1A	AALEKDYEEVGV	-
Flag	DYKDDDDK	ALEKDGD
HA	YPYDVPDYA	EDLPD

Table 2.16: Potential Mimotopes for the Monoclonal Antibody Samples.

In addition to finding the monoclonal antibody epitopes, we used the algorithms to identify potential mimotopes for the monoclonal antibody samples, as listed in Table 2.16. While these mimotopes do not match the AA subsequences of the full epitopes, they can potentially act as subsequences that uniquely bind to the monoclonal antibodies, without matching the AA composition of the epitope. We deduced the following considerations for potential mimotopes when analyzing random-sequence peptide microarrays: mimotopes are (i) distinctively different from the epitope of a specific monoclonal antibody sample; (ii) distinct across all eight monoclonal antibody samples; (iii) notably different from other peptide subsequences when comparing binding strength and/or occurrence count. From these considerations, we developed the following criteria to identify potential mimotopes. A potential mimotope of a monoclonal antibody sample is a subsequence that:

1. is not an exact or a single substitution match to a full or an estimated monoclonal epitope
2. is not sufficiently similar to high occurring peptide subsequences of other monoclonal antibody samples
3. has a sufficiently large MFI
4. has a large DS-OCRC, obtained using the down-selected monoclonal peptides.

Following these criteria led to potential mimotopes for the monoclonal antibodies samples `2C11`, `4C1`, `A10`, `Flag`, and `HA`. Subsequences of the remaining three monoclonal antibody samples did not meet all of the aforementioned criteria, and thus they were not identified as potential mimotopes.

The proposed approach identified some potential mimotopes, as listed in Table 2.18, for five of the monoclonal antibody samples we analyzed. As discussed in Section 2.7, we provide some criteria we developed on mAb mimotope identification. Although our mimotope analysis is only theoretical, we found that our criteria seem to match mimotope identification approaches in recent publications. More specifically, in [48], mimotopes were identified from peptide sequences by T cells with common receptors as they resulted in increased antigen immunity. As the authors discuss, optimizing the identification of mimotopes can lead to improvements in antigen-specific vaccines. Mimotopes were identified for a monoclonal cancer antibody using phage display efficient screening of random peptide libraries [46]. Similar to our findings, the mimotopes were selected based on their strong binding to the original peptides; it was also noted that stronger binding was obtained with AA substitutions. In [106], mimotopes displaying phages for monoclonal antibodies were investigated for biomarker assay development; it was found that the diversity of mimotope displaying phages of selected peptides is inversely correlated with binding strength.

Table 2.18 provides additional information on how we identified the mimotopes for the five monoclonal antibodies in Table 2.18. For each monoclonal antibody, the four criteria in Section 2.7 are met. In particular, all these mimotope subsequences have very large median or maximum fluorescence intensities.

`Flag` **sample**	**OCRC**	**Mean FMI**
$\mathcal{L} = 4$		
`DYKD`	16	947
`YKDD`	9	799
`KDDD`	2	523
`DDDK`	90	391
$\mathcal{L} = 5$		
`DYKDD`	2	23,744
`DDDDK`	22	376

Table 2.17: Subsequences of Varying Length $\mathcal{L}$ for `Flag`.

Sample	**Full**	**Potential**			**Mean**	**Max.**
mAb	**Epitope**	**Mimotope**	**OCRC**	**DS-OCRC**	**MFI**	**MFI**
2C11	NAHYYVFFEEQE	DARWFN	10	4	1,197	65,535
4C1	LQAFDSHYDY	ADSWP	20	10	12,769	65,535
A10	EEDFRV	EWDVA	41	15	65,535	65,535
Flag	DYKDDDDK	ALEKDGD	267	250	65,535	65,535
HA	YPYDVPDYA	EDLPD	15	7	1,706	11,385

Table 2.18: Identified Mimotopes for Five Monoclonal Antibody Samples With Corresponding OCRC, DS-OCRC Mean and Maximum Fluorescence Intensity.

2.9 Peptide Array Data Consistency Analysis

In the previous sections of this chapter, we have shown how the exact and single substitution matching algorithms can be used to successfully estimate epitopes, discover mimotopes, and analyze the effects of amino acid substitutions on binding strength. The epitope estimate results match closely with exact epitope subsequences, showing that the algorithms are performing well, and that the data are consistent enough to use PCC as a down-selection method. Using PCC instead of fluorescence as the ranking metric improved the epitope estimation accuracy, and further performance improvements could be expected with an improvement in consistency between the data. This section compares the peptide array datasets for similarity using statistical characterizations of the fluorescing peptides and the adjacent background substrate.

2.10 Data Collection

At the device level, after preparing and incubating the primary and secondary antibodies onto the array, the array is washed and dried, and then illuminated with fluorescent light. A 16-bit tagged image file format (TIFF) image of the fluorescing array is recorded. This image is calibrated to be able to spatially locate pixels either within the bounds of the peptide cluster, our outside of it. 12 pixels from within the peptide cluster are median averaged to obtain the MFI, while 104 pixels from outside of the peptide cluster are median averaged to obtain the median background intensity (MBI). In addition to the MFI and MBI, the the standard deviation of the fluorescence intensity (SFI) is calculated. These three quantities are the basis for analyzing the quality of individual peptide array sample datasets, as well as the consistency between them. Figure 2.8 is an illustration of the extent of the pixles (black squares) and the circular extent of the peptide cluster (blue circle). Pixels that lie entirely within the circular extent of the peptide cluster are used for calculating the MFI and SFI, while pixels entirely outside of the peptide cluster are used for calcuating MBI. Note that the source image is 16-bit, and therefore the individual pixel intensity values are between 0 and 65535, as are the MFI and MBI.

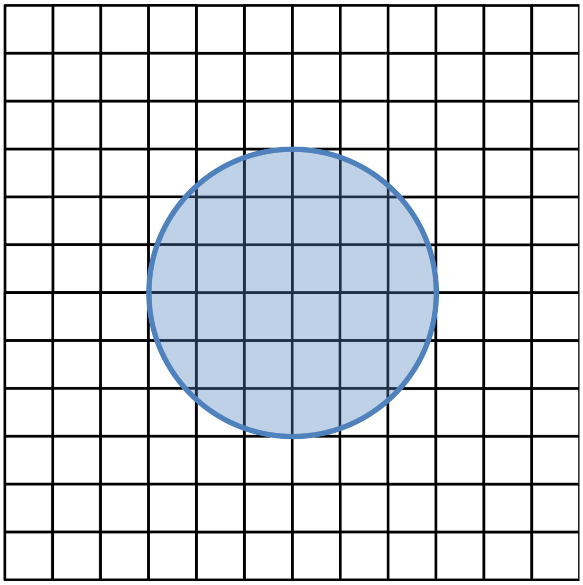


Figure 2.8: Image Pixels Inside of the Peptide Cluster Are Used to Estimate Statistics Related to the Fluorescence, and Pixels Outside of the Peptide Cluster Are Used to Estimate Statistics Related to the Background.

The eight monoclonal antibodies were collected at two different times. Five of the eight monoclonal antibody array samples `2C11`, `4C1`, `A10`, `Flag`, and `HA` were collected in January 2013, while the other three were samples, `Ab1`, `Ab8`, and `DM1A` were collected in June 2012.

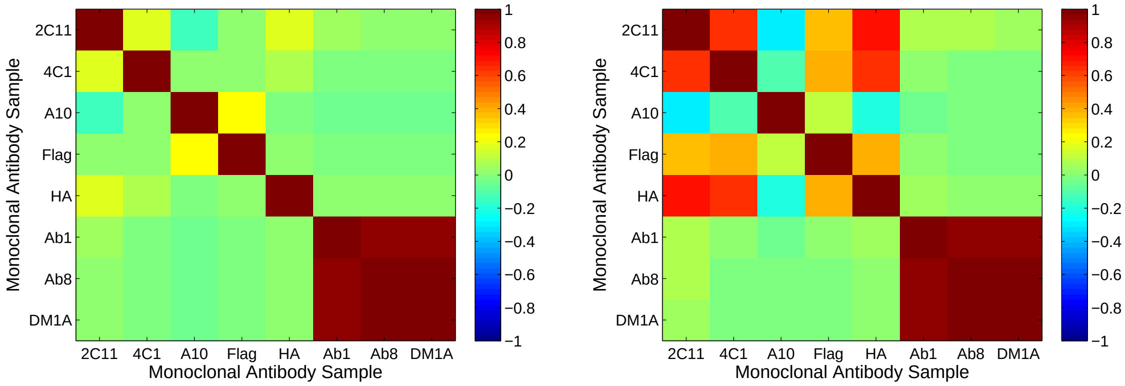
2.11 Dataset Analysis

One common method for comparing peptide array datasets is to compute the Pearson's correlation coefficient between the fluorescence values of two entire arrays.

This scalar value is a good first order approximation of similarity, because we expect that most of the peptides on the array will not be specifically bound to, that these background peptides will for the most part be the same when comparing two different monoclonal antibody samples.

Data normalization techniques seek to undo some of the biases and scalings that can occur when data are not collected using identical procedures. As was discussed previously, datasets can either be normalized through dividing the fluorescence values by the median fluorescence value so that all datasets have a median fluorescence value of 1, or by base 10 logarithmically transforming the datasets so that the distributions are more "Gaussian-like". When computing the Pearson's correlation coefficient between array samples, median normalization does not affect the correlation coefficient because part of that computation is to subtract out the mean array fluorescence. A color-scaled matrix of the correlation coefficients between array samples with no normalization and median normalization are shown in Figures 2.9a and 2.9b respectively. Note that the data set order has been rearranged such that samples from the same year are grouped together, i.e. the first five samples were collected in 2013, while the last three samples were collected in 2012. Note that the correlation between two samples is associative; the correlation between any two samples is the same regardless of order.

From the correlation analysis, we can see that the data from 2012 is well correlated independent of the normalization used, while the data from 2013 is a far less correlated. A correlation value greater than 0.9 would indicate that the data were collected using the same procedure, and that the non-specific binding between arrays is consistent. The logarithmic transformation significantly improves the correlation between most of the samples from 2013, with the exception of monoclonal antibody



(a) Unnormalized

(b) Logarithmic transformation

Figure 2.9: Array Sample Correlations for (a) Unnormalized Data, and (b) Logarithmically Transformed Data

`A10`. While additional normalization techniques do exist, they would require more calibration information than was available for this dataset.

Some of the array samples are very much uncorrelated (0.2-0.4), in apparent contradiction to the results shown earlier in this chapter. However, lower array-to-array correlation values and good estimation results are possible because the array-to-array correlation is an aggregate metric for all 330k peptides, while only 2.5k peptides, selected using the PCC ranking metric, are used to estimate the epitope. Improving the array-to-array correlation is likely to result in a set of peptides which is more likely to contain epitope subsequences.

As the individual pixel intensities are not provided in the peptide array data files, we work with the statistical estimates which are provided such as the MFI, MBI, and SFI. Mean estimates of the pixel intensity are also given, and from comparing the mean to the median fluorescence intensity we see only small differences, and therefore are able to conclude that the pixel intensity distributions are symmetric, and without additional information assume that they are mostly Gaussian. Thus, we look at the ratio of the MFI to the SFI as a measure of data quality, and also look at histograms of the MFI and MBI to see how they vary across peptide array sample.

Figure 2.10 is a plot of the MFI divided by the SFI. Because we assume an underlying Gaussian distribution, this quantity is an approximation for a quality factor. The median values of this distribution range between 5 and 9, showing that the measurements themselves are much larger than the statistical variation of that measurement. Additionally, we can see that the data again separates itself into two groups, where the data collected in 2012 and the mAb `A10` are in one group, and the remaining data collected in 2013 are in the second group. Based on this data quality metric, the data collected in 2013 (excluding `A10`) is higher quality data.

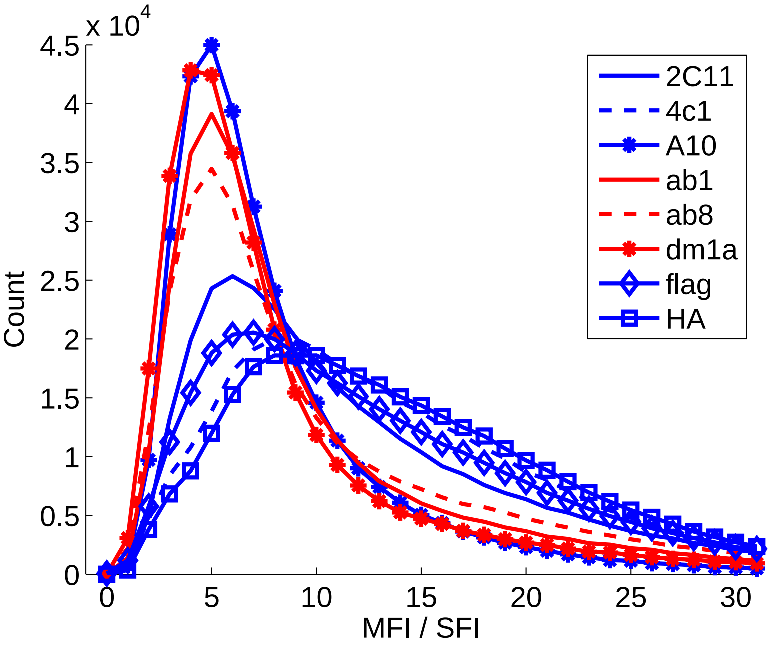


Figure 2.10: Histograms of the Median Fluorescence Intensity Divided by the Standard Deviation of the Fluorescence Intensity Pixels.

Figures 2.11 and 2.12 are histograms of the MFI and MBI respectively, for each of the eight monoclonal antibody samples. From these two plots we see the same groupings from previous figures. From these figures we can infer that the higher quality factors for the 2013 data group (excluding A10) seen in Figure 2.10 was a result of smaller SFI, as well as the relative size of MFI to SFI.

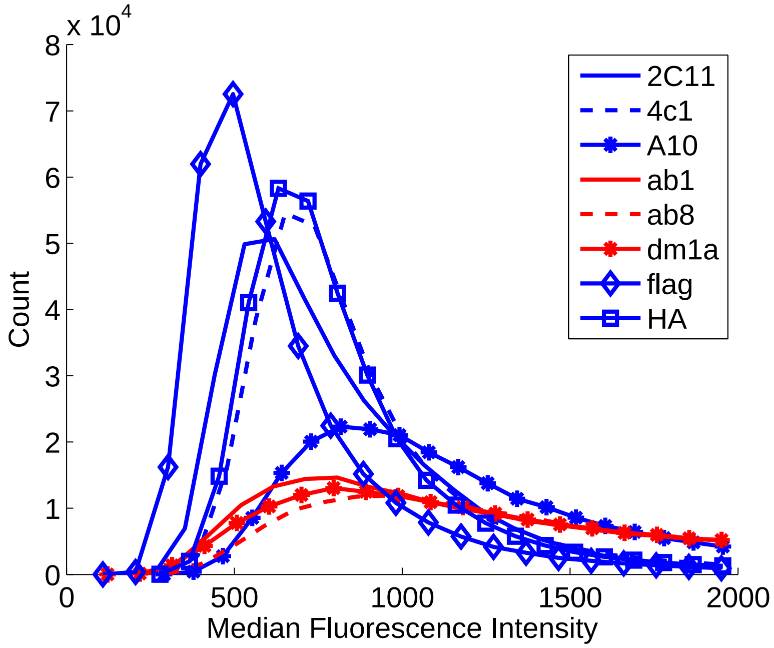


Figure 2.11: Histograms of the Median Fluorescence Intensity for All Eight Monoclonal Antibodies.

The analysis in this section has shown that the monoclonal antibody data collected in 2013 is, for the most part, of higher quality than the 2012 data. This is, according to the scientists collecting the data a result of improved data collection practices, including a higher level of consistency between samples. We can assume that if the mAb data were collected again, that the epitope estimates would be at least as accurate as they were with this data set.

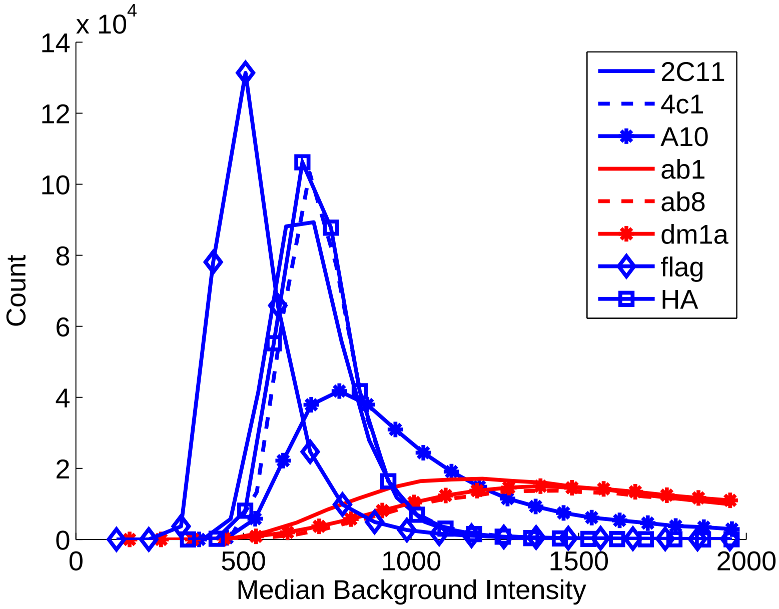


Figure 2.12: Histograms of the Median Background Intensity for All Eight Monoclonal Antibodies.

2.12 Comparison Methods

Random sequence peptide arrays containing hundreds of thousands of peptides are very new, and as a result, there are very few algorithms to which our epitope estimation algorithms can be compared. The most directly comparable method is a motif estimation algorithm [2] developed by the Center for Innovations in Medicine [101]. This method analyzes the peptide array data and searches for subsequences of length 3-7 which occur more than three times with MFI values that are statistically significant. They assess statistical significance by comparing the fluorescence of peptides which contain the subsequence of interest to the fluorescence of peptides selected at random from the array. Table 2.19 is a comparison of our results and the results from Figure 1 of [2].

The two methods find approximately the same epitope sequences for five of the eight monoclonal antibodies, however our method finds epitope subsequences for `2C11`, A10, and Ab8 at the top of our epitope estimate lists, while their method does not find an epitope subsequence for 2C11, and only finds epitope subsequences for A10, and Ab8 in their 11th and 13th spots in their epitope estimate list. As they state in their paper, these are difficult antibodies to estimate epitopes for because the antibodies bind non-specifically to a range of peptide sequences with the same strength as the specific binding. The reason we are able to estimate these epitopes is because we use the PCC instead of fluorescence. As we showed in Figure 2.4 for Ab8, using correlation as a ranking metric results in more peptides containing the epitope subsequences, and this is true for 2C11, and A10 as well.

Sample mAb	**Full Epitope**	**Estimated Epitope**	**Comparison Epitope**	**Motif Rank**
2C11	NAHYYVFFEEQE	VFFEEQE	-	-
4C1	LQAFDSHYDY	GYDSR	DSFDS	1
A10	EEDFRV	EDFRV	EDF	11
Ab1	NTFFRHSVVV	RHSVV	RHSVV	1
Ab8	TFSDLWKLLPE	DLWKL	SDLKW	13
DM1A	AALEKDYEEVGV	AALEKD	LEKD	1
Flag	DYKDDDDK	AALEKDG	DY_D	1
HA	YPYDVPDYA	YDAPE	DVPD, YDAPD	1

Table 2.19: A comparison of Our Epitope Estimates to the Motif Estimates in [2].

Chapter 3

ALIGNMENT OF MULTIPLE PROTEIN ATTRIBUTES USING WAVEFORM MAPPING

3.1 Protein Alignment and Generalization of Amino Acid to Waveform Mapping

Protein alignment methods find similarities in the structure and functional or evolutionary attributes of the proteins being compared. Aligning proteins is important for the discovery of evolutionary relationships, as well as discovery of important drug target locations, and for finding the effects of gene mutations. When aligning, the three dimensional structure of the protein is the most important important target to align on, however additional attributes are important to integrate as well, especially for more distantly related proteins, or more extreme mutations which may have significantly affected the structure. Examples of these additional attributes include the amino acid sequence, hydrophobicity, and sub-groupings of three consecutive amino acids.

As we have had previous success in incorporating additional information from one dimensional (1-D) signals into time-frequency transformations, in this chapter we propose a generalization to three dimensions (3-D). These additional dimensions allow for a concise representation of the structure, as well as time-frequency modulation to represent additional attributes.

3.2 Protein-to-Waveform Mapping Model

3.2.1 Selection of Highly-Localized Waveform

Current methods of protein structure alignment using waveform mapping do not consider all possible multiple global and local conformations [88]. Including directionality between linked amino acids as part of the waveform mapping leads to more successful structural alignment [89]. In particular, a protein-to-waveform mapping model needs to allow for translations and rotations in the three dimensional (3-D) plane to better represent the protein structure. As a result, there is a need for a parametric waveform representation with a unique 3-D shape and parameters that can be selected to identify changes in 3-D conformations. An important motivation for a 3-D basis representation of protein structures is the fact that distantly related proteins need not have similarity over the entire structure. Similarities can be localized, and if the representation is linearly separable, it can be used to analyze similar segments over shorter structure lengths.

For multiple attribute mapping, we use a 3-D Gaussian waveform that is highly-localized in the higher-order six dimensional time-frequency plane. The multivariate Gaussian waveform, $g(\mathbf{t})$, is defined across a 3-D time-domain, (t_x, t_y, t_z), with covariance matrix $\mathbf{\Sigma}$ as

$$g(\mathbf{t};\mathbf{\Sigma}) = A_g\,(2\pi)^{-1.5}|\mathbf{\Sigma}|^{-0.5}\,e^{-0.5\,\mathbf{t}^{\mathtt{T}}\mathbf{\Sigma}^{-1}\mathbf{t}}\,, \tag{3.1}$$

where $\mathbf{t} = [t_x\ t_y\ t_z]^{\mathtt{T}}$, A_g is a normalization constant so that the Gaussian waveform has unit energy, $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$, and $\mathtt{T}$ denotes vector transpose. The Gaussian waveform can also be transformed by changing its amplitude $a(\mathbf{t})$ and its phase function $\boldsymbol{\psi}(\mathbf{t})$ to obtain

$$s(\mathbf{t};\mathbf{\Sigma},\boldsymbol{\psi}) = a(\mathbf{t})\,g(\mathbf{t};\mathbf{\Sigma})\,e^{-j2\pi\boldsymbol{\psi}(\mathbf{t})}\,. \tag{3.2}$$

For example, the Gaussian signal can be time-shifted and frequency-shifted

$$s(\mathbf{t}; \boldsymbol{\Sigma}, \boldsymbol{\tau}, \boldsymbol{\nu}) = g(\mathbf{t} - \boldsymbol{\tau}; \boldsymbol{\Sigma}) e^{j2\pi\,(\mathbf{t}-\boldsymbol{\tau})^{\mathrm{T}}\boldsymbol{\nu}}\,, \tag{3.3}$$

where $\boldsymbol{\tau} = [\tau_x\ \tau_y\ \tau_z]^{\mathrm{T}}$ and $\boldsymbol{\nu} = [\nu_x\ \nu_y\ \nu_z]^{\mathrm{T}}$ are the time shift and frequency shift transformations along each axis, respectively.

3.2.2 Mapping Geometric Structure Attributes

We consider two neighboring amino acids, A_i and A_{i+1}, where each amino acid is centralized about a single α-carbon [107]. The 3-D geometric atomic coordinates of the ith amino acid are given by $\mathbf{x}_i = [x_i\ y_i\ z_i]^{\mathrm{T}}$; these 3-D structure coordinates are specified in protein data bank (PDB) files [1]. For the 3-D protein-to-waveform mapping, we use the time-shifted and frequency-shifted Gaussian waveform $s(\mathbf{t}; \boldsymbol{\Sigma}_i, \boldsymbol{\tau}_i, \boldsymbol{\nu}_i)$ in (3.3). We then select the time-shift parameter $\boldsymbol{\tau}_i$, frequency-shift parameter $\boldsymbol{\nu}_i$, and covariance matrix $\boldsymbol{\Sigma}_i$ to represent the neighboring A_i and A_{i+1} amino acids.

The time-shift parameter $\boldsymbol{\tau}_i = [\tau_{x,i}\ \tau_{y,i}\ \tau_{z,i}]^{\mathrm{T}}$ is selected such that the waveform in Equation (3.3) is centered between adjacent α-carbons. Thus,

$$\begin{aligned} \tau_{x,i} &= (x_i - x_{i+1})/2, \\ \tau_{y,i} &= (y_i - y_{i+1})/2, \\ \tau_{z,i} &= (z_i - z_{i+1})/2, \end{aligned}$$

corresponding to the mid point between A_i and A_{i+1}.

The covariance of the 3-D Gaussian is selected to model the amino acids 3-D orientation. Specifically, we want the energy of the Gaussian to be localized in the region between the two α-carbons such that $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$ appear as the two outermost points on the 3-D Gaussian function in the mapped (t_x, t_y, t_z) plane. Using eigendecomposition of the Gaussian covariance matrix, we can obtain the eigenvector matrix that can be shown to correspond to the orientation or rotation matrix from 3-D point $\mathbf{x}_i$ to $\mathbf{x}_{i+1}$. The geometric design of the rotation matrix is thus based on the spherical angles description, (ϕ_i, θ_i), of the vector between these two points. Using the coordinates for A_i and A_{i+1}, the angles are obtained as

$$\phi_i = \arctan\left(\frac{y_i - y_{i+1}}{x_i - x_{i+1}}\right), \theta_i = \arccos\left(\frac{z_i - z_{i+1}}{\sqrt{d_i}}\right) \tag{3.4}$$

where

$$d_i \ = \ (x_i - x_{i+1})^2 + (y_i - y_{i+1})^2 + (z_i - z_{i+1})^2. \tag{3.5}$$

Using the angles in (3.4), the rotation matrices [108] are given by

$$R_{\phi_i} = \begin{bmatrix} \cos(\phi_i) & -\sin(\phi_i) & 0 \\ \sin(\phi_i) & \cos(\phi_i) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$R_{\theta_i} = \begin{bmatrix} \cos(\theta_i) & 0 & -\sin(\theta_i) \\ 0 & 1 & 0 \\ \sin(\theta_i) & 0 & \cos(\theta_i) \end{bmatrix}.$$

The designed covariance matrix $\mathbf{\Sigma}_i$ is then calculated as

$$\mathbf{\Sigma}_i = R_{\phi_i}\, R_{\theta_i}\, \Lambda_i\, R_{\theta_i}^{\mathtt{T}}\, R_{\phi_i}^{\mathtt{T}} \tag{3.6}$$

where $\Lambda_i \ = \ \mathrm{diag}([d_i/36\ 0.1\ 0.1])$. The scaling on each axis by Λ is chosen to concentrate the energy between the amino acid locations, and to limit the amount of overlap between adjacent Gaussian waveforms.

The frequency-shift parameter $\boldsymbol{\nu}_i = [\nu_{x,i}\ \nu_{y,i}\ \nu_{z,i}]^{\mathtt{T}}$ is selected to provide directionality information. Fixing the frequency-shift along each plane demonstrates pointing from $\mathbf{x}_i$ to $\mathbf{x}_{i+1}$; the negative of the same frequency demonstrates pointing from

$\mathbf{x}_{i+1}$ to $\mathbf{x}_i$. Note that the time and frequency-shifted Gaussian waveform is sampled compactly so that the correlation between two mapped amino acids with different parameters is almost zero. An entire protein consisting of N amino acids is then modeled by the sum of $N-1$ time and frequency-shifted Gaussian waveforms

$$\mathbf{s}(\mathbf{t}) = \sum_{i=1}^{N-1} s(\mathbf{t}; \boldsymbol{\tau}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\nu}_i)\,. \tag{3.7}$$

3.2.3 Mapping Sequence and Function Property Attributes

The sequence and function property attributes can be mapped using other waveform transformations. For example, the time-varying phase function in (3.2) can be chosen to be a quadratic function along each axis with an amplitude modulation selected to preserve waveform orthogonality.

3.3 Protein Multi-Alignment Metric

The 3-D structure protein alignment requires rotating and shifting a query protein and a database protein to find their globally maximum alignment. We define the set of α-carbons describing the 3-D geometric atomic coordinates of N amino acids as $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_N]$. As we are using the location of amino acid α-carbons to model the location of the proteins, we only need to shift and rotate those point sets.

Structure Alignment Searching the entire 3-D space of feasible shifts and rotations is computationally prohibitive, so we utilize a search procedure where the 3-D protein structures are always in a position of at least partial alignment. The procedure focuses on a set of three consecutive amino acids A_{i-1}, A_i, A_{i+1}, with corresponding 3-D coordinates $\mathbf{x}_{i-1}$, $\mathbf{x}_i$, $\mathbf{x}_{i+1}$. The coordinates are shifted to place $\mathbf{x}_i$ at the origin $[0\ \ 0\ \ 0]$. The origin-shifted amino acid structure, with coordinates $(\mathbf{X} - \mathbf{x}_i)$, is then rotated using rotation matrix $\mathbf{Q}_i$. This matrix is obtained based on the following conditions:

(a) the vector from $\mathbf{x}_i$ to $\mathbf{x}_{i+1}$ lies along the x-axis;

(b) the vector from $\mathbf{x}_i$ to $\mathbf{x}_{i-1}$ lies in the x-y plane.

The 3-D coordinates of the origin-shifted and rotated structure are given by

$$\mathbf{X}_R = \mathbf{Q}_i(\mathbf{X} - \mathbf{x}_i)\,. \tag{3.8}$$

Figures 3.1 and 3.2 illustrate a short protein sequence before, and after shifting and rotating respectively.

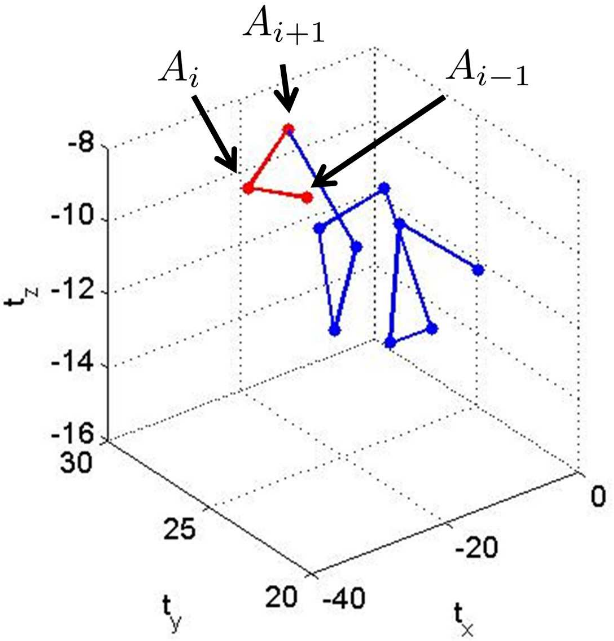


Figure 3.1: A Short Protein Sequence in an Arbitrary Coordinate System.

The structure is then mapped using the protein-to-waveform Gaussian mapping described in Equations (3.3) and (3.7). The mapping is applied to both the query and database structures. Considering a query amino acid $A_i^{(q)}$ and a database amino

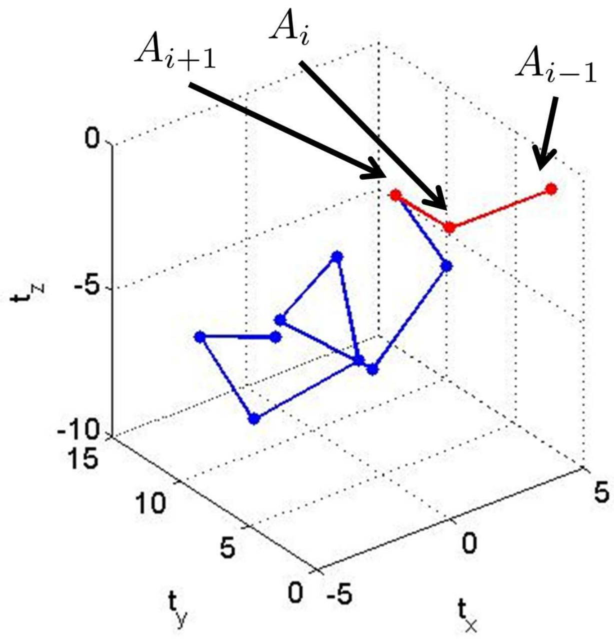


Figure 3.2: A Short Protein Sequence Shifted and Rotated to the 3-D Coordinate Origin.

acid $A_j^{(d)}$, the structure mapping for each amino acid results in the 3-D Gaussian waveforms $s^{(q)}(\mathbf{t};\boldsymbol{\tau}_i,\boldsymbol{\Sigma}_i,\boldsymbol{\nu}_i$ and $s^{(d)}(\mathbf{t};\boldsymbol{\tau}_j,\boldsymbol{\Sigma}_j,\boldsymbol{\nu}_j)$, respectively. We obtain a structure alignment score (Str) by calculating the inner product between these 3-D Gaussian waveforms as

$$\text{Str}_{i,j} = \int s^{(q)}(\mathbf{t};\boldsymbol{\tau}_i,\boldsymbol{\Sigma}_i,\boldsymbol{\nu}_i)\, s^{(d)}(\mathbf{t};\boldsymbol{\tau}_j,\boldsymbol{\Sigma}_j,\boldsymbol{\nu}_j)d\mathbf{t}\,. \tag{3.9}$$

Sequence Alignment The sequence alignment score (Seq) is calculated by looking at the corresponding rows and columns of the relevant block substitution matrix (BLOSUM), in this case, BLOSUM62 as

$$\text{Seq}_{i,j} = \frac{1}{11}\text{BLOSUM62}(A_i^{(q)}, A_j^{(d)})\,. \tag{3.10}$$

Note that this score is normalized to have a maximum match score of 1.

Hydrophobicity Alignment The hydrophobicity alignment score (Hydr) is calculated using

$$\mathrm{Hydr}_{i,j} = \frac{1}{13.64}(13.64 - |h_i - h_j|) \tag{3.11}$$

where h_i and h_j are the hydrophobicity values of amino acids $A_i^{(q)}$ and $A_j^{(d)}$, respectfully, in units of kJ/mol. The normalization constant 13.64 results in a maximum score of 1 when the amino acid sequences are identical and a score of 0 when the hydrophobicity values of the two amino acids are maximally different.

The overall location, structure, and function (LoStrFn) alignment score is obtained by combining the three scores in Equations (3.9)-(3.11). In particular, it is given by

$$\mathrm{SC}_{i,j} = \mathrm{Str}_{i,j} + \mathrm{Seq}_{i,j} + \mathrm{Hydr}_{i,j}\,. \tag{3.12}$$

Note that each of the scores can be weighted differently based on the available information on the importance of each attribute. In additional to computing separate scores and adding them together, as is done above in Equation 3.12, it is possible, but not implemented here, to integrate all of the information into a single time-frequency waveform where the inner product evaluated in Equation 3.9 would compute the entire score, and not just the structure score.

When all LoStrFn attributes are combined, we can use a single overall metric for protein alignment. An alternative method of obtaining a LoStrFn alignment metric is by performing each attribute mapping and corresponding alignment separately and then combining the three resulting metrics together to obtain a single metric [80]. This approach is less computationally intensive since the alignment algorithm becomes more complex as the signal transformation parameters increase. One possible way to perform each attribute mapping and alignment separately is by matching the one dimensional sequence information using the BLOSUM62 matrix and the matching pursuit decomposition method from [30]; and by matching the property amino acid or substitution matrix information directly from tabulated available results.

3.4 Alignment Results

A local LoStrFn alignment is performed to identify regions of similarity within long protein sequences that could be widely divergent. The metric score in (3.12) can be used for local alignment. A global LoStrFn alignment involves finding regions of similarity for the entire query protein. This alignment is computed using a structured search across all possible time and frequency-shifts and rotations of the query and database proteins. The 3-D structural alignment is performed for all three amino acid combinations of the database and query sequences. This is done by cycling through all of the possible combinations of $\mathbf{Q}_i$ in (3.8) for both protein sequences.

To demonstrate the effectiveness of the LoStrFn alignment method, we compare two human mutant ferrochelatase proteins. The mutations cause changes in the amino acid sequence as well as the 3-D structure of the protein. Here, the 7 amino acid query sequence (TSDHIET) is a sub-sequence of protein 2po5 and the 19 amino acid database sequence (ILLVPIAFTSDCIETLYEL) is a sub-sequence of protein 2po7. The location, geometry and hydrophobicity values for the two proteins are obtained from the Protein Data Bank [1]. The 3-D protein structure waveform mapping in the time-frequency plane for global alignment is demonstrated in Figure 3.3. The structure of the two protein subsections are highly overlapping. The query protein, 2po5, is plotted in blue, while the database protein, 2po7, is plotted in green. The global alignment scores for each attribute and the overall alignment scores are listed in Table 3.1. As expected, the overall score improves protein similarity as it incorporates the matching of different attributes that contribute toward the protein mutations.

Score Type	Score Values					
3-D Structure	0.66	0.95	0.98	1.00	1.00	0.99
AA Sequence	0.45	0.36	0.54	-0.27	0.36	0.45
Hydrophobicity	1	1	1	0.57	1	1
Overall	2.11	2.31	2.52	2.3	1.36	2.44

Table 3.1: LoStrFn Matching Scores

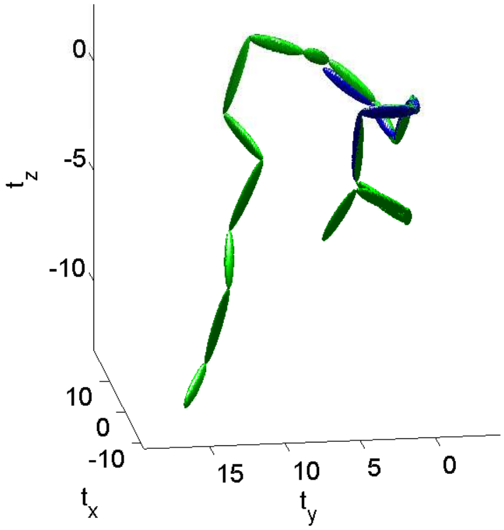


Figure 3.3: Structure Slignment of Two Human Mutant Ferrochelatase Proteins: Query 2po5 Protein (Blue) and Database 2po7 Protein (Green) From [1].

Chapter 4

PHYSICS-BASED SEA CLUTTER MODEL FOR IMPROVED DETECTION OF LOW RADAR CROSS SECTION TARGETS

4.1 Detection Problem in Rapidly Varying Sea Clutter

In highly cluttered environments, such as in heavy sea clutter, the problem of detecting a small target is very challenging. It is thus important to understand the statistical characteristics of the complex sea environment and obtain physics-based models that we can incorporate into our detector algorithm designs. This can lead to an increase in detection performance by using the model to minimize the impact of the environment.

Sea clutter is often characterized by the compound Gaussian sea clutter model. While described mathematically later in this chapter, the concept behind this model is that sea clutter is a summation of large amplitude and wavelength waves, called gravity waves, and smaller amplitude and wavelength waves called capillary waves. The capillary waves are are quick to decorrelate, and the strength of these random reflections is modulated by the size of the gravity waves. Thus, these large, quickly changing clutter reflections present a significant obstacle to the detection of electromagnetically small targets.

4.2 Physics-based Finite-Difference Time-Domain Sea Clutter Simulation

The sea clutter generation model includes two main processes. The first process is the generation of a three-dimensional (3-D) random dynamic sea surface that moves according to the governing physics of water waves as driven by the wind; the waves include gravity waves whose restoring force is gravity, and capillary waves whose restoring force is water surface tension. The second process includes two dimensional (2-D) finite-difference time-domain (FDTD) simulations. It is based on using Maxwell's equations to propagate radar pulses through the FDTD domain, where the incident electromagnetic (EM) field impinges on the sea surface and scatters. The implementation of a teleportation window [94] in the FDTD simulations separates the scattering field or clutter from the total field; it is then propagated to the far field and collected for processing. The 2-D FDTD simulations involve individual radar pulses incident on single cuts of the dynamic 3-D sea surface. The sea is modeled as a perfectly conducing surface (water cells are perfect electric conductors). As the radar pulse duration is on the order of nano seconds, the surface is a static snapshot during each pulse simulation but is propagated in between simulations of subsequent pulses according to the pulse repetition time. For a single radar pulse, multiple down-wind cuts of the 3-D sea surface are simulated. These cuts are strategically spaced in the cross-wind direction in an attempt to collect scattering samples of the 3-D sea and capture scattering across the radar footprint area on the sea surface. The superposition of radar backscatter collected from the multiple down-wind sea cuts constitutes quasi-3-D sea clutter. The features of the sea surface are developed in stages: incorporating 2-D static gravity waves and developing a capillary waves model, implementing a spreading function to expand into 3-D, and superimposing the capillary waves on the gravity waves while mathematically giving each random wave its respective phase velocity.

The gravity waves component of the 2-D sea surface are generated as described in [109]. The height $f(y_n)$ of the 2-D sea surface at points y_n along the surface is given by

$$f(y_n) = \frac{1}{L} \sum_{m=-N/2}^{N/2-1} F(K_m) \exp\left(jK_m y_n\right) \tag{4.1}$$

where

$$F(K_m) = \sqrt{2\pi L W(K_m)} R_m, \tag{4.2}$$

R_m is a zero-mean, unit-variance Gaussian random variable $r(0,1)$ for $m = 0, N/2$, and

$$R_m = ((r(0,1) + jr(0,1))/\sqrt{2} \tag{4.3}$$

for $m = 1, \ldots, N/2 - 1$. The function

$$W(K_m) = (\alpha/(4|K_m|^3) \exp\left(-\beta g^2/(K_m^2 U^4)\right) \tag{4.4}$$

is the P-M sea spectrum [110], where L is the sea surface length in meters, N is the number of surface sampling points, $K_n = 2\pi\Lambda_n$ is the wave number of the ocean wave, Λ_n is the ocean wave wavelength, $\beta = 0.74$, $\alpha = 0.0081$, $g = 9.81$ m/s^2 is the gravitational constant, and U is the wind speed in m/s.

The slope of the gravity waves versus time at any point on the sea surface is proportional to the amplitude modulation of the clutter (or the clutter texture) returning from that point of the sea, so it is vital that the instances of gravity waves are generated correctly. In order to verify that the gravity waves model is correct, we simulated 2-D FDTD incident EM plane wave illumination on our gravity wave surface $f(y_n)$. The parameters of the EM simulations were chosen based on information found in previous work. The proper discretization of the Pierson-Moscowitz (P-M) sea surface, to capture scattering from relevant details of the sea, is given in [111]. We chose an extreme case to reproduce for generating the far field scattering from large details of the sea (gravity waves), following an FDTD simulations study in [112]. The sea details of interest are relatively large gravity waves, with a sea state of 7 (wind speed of 20 m/s) and a significant wave height of roughly 6 meters. Thus, to capture details of this size, the free space EM illumination wavelength is chosen as $\lambda_0 = 7.49$ m. The sea surface is sampled at $\lambda_0/16$ intervals. The discretization cell in our FDTD space is $\lambda_0/16$, and the plane wave EM illumination angle of incidence is 20 degrees above the horizon. Using these settings, we simulated 40 independent, completely decorrelated, random instances of the gravity wave surface of total length $160\lambda_0$ m.

The scattering from the gravity waves is propagated to the far field over the horizon, where evidence of the changing slopes of larger gravity waves can be seen from one instance of sea to the next. The far field scattered intensity of 3 instances and the average of 40 instances (black curve) are shown in Figure 4.1. These results demonstrate that we have a valid simulation of large realistic sea surface details in the EM environment using the FDTD computational EM method. The results also show that the 2-D gravity wave surface is correctly generated.

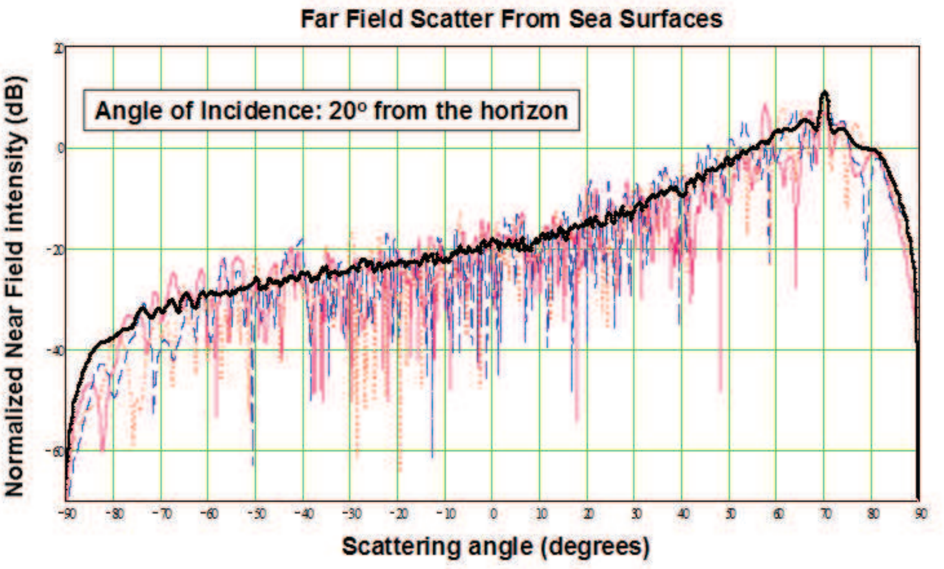


Figure 4.1: Far Field of 2-D Gravity Waves.

To include the smaller details of the sea surface, we develop an energy spectrum to generate capillary waves in the same manner that the P-M spectrum is used to generate gravity waves. The spectrum is obtained using results from other capillary wave studies. We first estimate an exponential function that relates the wind speed U and capillary wave wavelength Λ [113]. The total kinetic and potential energy of water waves is given by $E = \rho\pi A^2$, where A is the wave amplitude and ρ is the density of water [114], and the amplitude of the capillary wave of greatest height is $2A = 0.73\Lambda$ [115]. Using these relations, we estimate the energy spectrum for capillary waves using

$$w(K_n) = (4\alpha_c\rho\pi^3 U/|K_n|^2)\ \exp\left(-(K_B - \beta_c K_n)/(\beta_c^2 K_n)\right)^2, \tag{4.5}$$

where $K_B = 2\pi\Lambda_B$, Λ_B is the wavelength boundary between gravity and capillary waves, $\alpha_c = 0.0445$, and $\beta_c = 0.6$. This expression does not account for other phenomena such as the effect of the local gravity wave slope and the angle of incidence of the local wind and instantaneous wind speed. However, it yields roughly the correct capillary wave heights based on experimental data [113] and is suitable for our study. Note that, although capillary and gravity waves are generated by the same approach, we continuously change the random number sets to prevent repeating capillary wave patterns.

The waves are propagated using the phase velocity equation

$$v_p^2 = \frac{T}{\rho}K + \frac{g}{K}, \tag{4.6}$$

where T is the surface tension per length, ρ is the water mass density, g is gravity acceleration, $K = 2\pi/\Lambda$ is the water wave number, and Λ is the wavelength of the water wave [114]. The first and second terms in (4.6) correspond to the velocity of capillary waves and gravity waves, respectively. Each wave is given a phase velocity

$\phi_n = K_n y_n - \omega_n t$, following (4.6), where

$$\omega_n = (|K_n|^2(TK_n/\rho + g/K_n))^{0.5}. \tag{4.7}$$

To expand to 3-D surface, we implement a spreading function as in [116]. Superimposing the moving capillary waves on the moving gravity waves results in the full dynamic and random 3-D realistic sea surface.

The speckle component of sea clutter is backscatter from capillary waves. A simple test to demonstrate that our capillary weaves scatter clutter in as similar fashion as the real sea is to calculate the Pearson's correlation coefficient of the first returning clutter pulse with all returning pulses. If the capillary waves have the correct motion, the speckle component of the clutter decorrelates in the time that real sea speckle decorrelates, which is on the order of 10 ms. For this test, in order to best observe the capillary waves, we chose to use X-band radar, since the EM wavelength is on the order of the capillary wave wavelength and amplitude. We simulated 100 radar pulses incident on a single cut of time-varying 3-D sea, using 1 ms pulse repetition time, to capture the effect of the clutter. The correlation results are plotted in Figure 4.2, which shows that the speckle decorrelated in approximately 15 ms.

4.3 Detection Methods of Low Radar Cross Section Targets

We consider a radar system for detecting a target in heavy sea clutter. The target is assumed to have low radar cross-section due to its actual size and its relative size in relation to the wavelength of the illuminating radar. We assume that the radar system transmits a pulse train of K identical pulses, $s(t)$, which scatter off sea surface scatterers, and if present, the target. The two detection hypotheses describing this scenario for the kth transmit signal, $k = 1, 2, \ldots, K$, and the ith sea surface scatterer

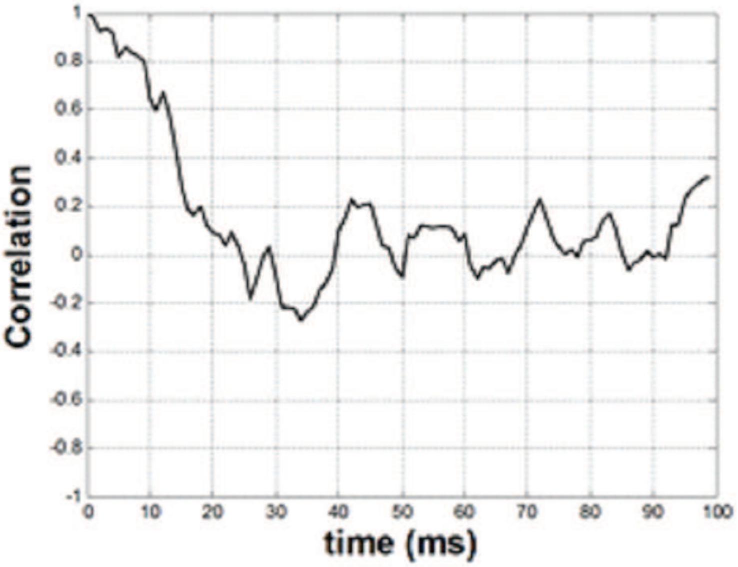


Figure 4.2: Decorrelations of the Radar Returns From the Simulated Sea Clutter Occur Over Approximately 15 ms.

are given by

$$H_0 : x_k(t) = \sum_i a_{k,i}\ s(t - t_i) + w(t)$$
$$H_1 : x_k(t) = b_k\ s(t - t_0) + \sum_i a_{k,i}\ s(t - t_i) + w(t)\,.$$

Under hypothesis H_0, we assume that that received signal consists of multiple scatterers with complex scattering coefficients $a_{k,i}$ and time delays t_i and white Gaussian noise $w(t)$. Under hypothesis H_1, we assume that, in addition to the scatterers and noise, the target is also present with a scattering coefficient b_k at time delay t_0. In both hypotheses, the signals are sampled using as sampling period T_s to yield the discrete time sequence $x_k[n] = x_k(tT_s)$, $n = 0, \ldots, N-1$. For the rest of the chapter, we assume that the clutter-to-noise ratio is very high and that the effects of noise on detection can be ignored. The pulse train radar waveform transmission scheme described above, is illustrated in Figure 4.3.

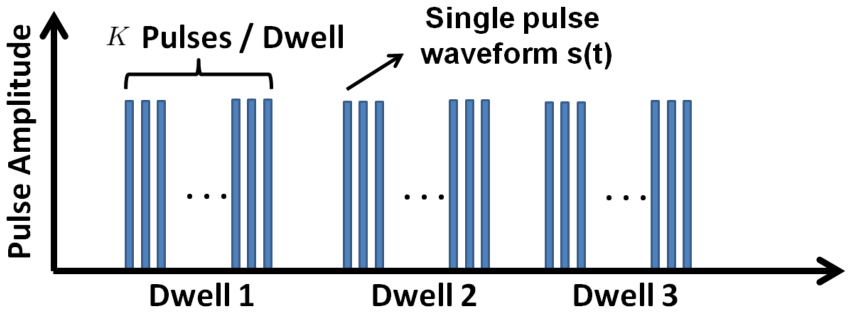


Figure 4.3: Illustration of the Radar Waveform Pulse Model.

4.3.1 Generalized Matched Filter Detector

We derive the generalize matched filter (GMF) detector, that under hypothesis H_1, assumes that the discrete-time incident signal $s[n]$ is known and deterministic but the target time-delay is unknown. After first estimating the time-delay using maximum likelihood estimation, the GMF detector is obtained by maximizing the probability of detection for a fixed false alarm rate. For our signal model, the discrete-time matched filter output corresponding to the kth pulse at the ℓth lag, $\ell = 0, \ldots, N-1$, with estimated n_0, is given by

$$r_k[\ell] = b_k z_s[\ell - n_0] + \sum_{n=-(N-1)}^{N-1} z_s[n] \; d_k[n+\ell] \tag{4.8}$$

where the autocorrelation function of the transmit signal $s[n]$ at lag ℓ is defined as

$$z_s[\ell] = \sum_{n=0}^{N-1} s[n] s^*[n-\ell], \tag{4.9}$$

and

$$d_k[\ell] = \sum_i a_{k,i} \tag{4.10}$$

is the aggregate scattering coefficient from all of the clutter scatterers that fall within the ℓth range bin. The decision threshold γ is set based on the distribution of $r_k[m]$ and by fixing either a desired value of false alarm rate $\mathbf{P_{FA}}$ or probability of detection $\mathbf{P_D}$.

4.3.2 Subspace Clutter Suppression Detector

The GMF detector is not expected to perform well for low radar cross section (RCS) targets in heavy sea clutter. In such cases, the clutter is much stronger than the signal, and for reasonable values of $\mathbf{P_D}$, the number of false alarms is large. This is expected as matched filtering does not involve clutter mitigation. The subspace clutter suppression (SCS) detector decomposes the signal into subspaces consisting of mostly clutter or mostly target energy. The detection performance is improved when only the subspaces that are orthogonal to the clutter are processed.

We assume a Swerling I point target so that the complex reflectivity of the target

$$\mathbf{b} = [b_1 \; b_2 \; \ldots \; b_K]^{\mathcal{H}}$$

for all K transmit pulses has a zero-mean complex Gaussian distribution with covariance matrix $\sigma^2 \mathbf{I}_K$, where $\mathbf{I}_K$ is the $(K \times K)$ identity matrix and $\mathcal{H}$ denotes complex transpose. For each of the K transmit pulses, the matched filter output at the ℓth lag or range bin can be written in vector form as

$$\mathbf{r}_\ell = \mathbf{b} \; z_s[\ell - n_0] + \sum_{n=-(N-1)}^{N-1} \mathbf{d}_{n+\ell} \; z_s[n] \tag{4.11}$$

where

$$\begin{aligned} \mathbf{r}_\ell &= [r_1[\ell] \; r_2[\ell] \; \ldots \; r_K[\ell]]^{\mathcal{H}}, \\ \mathbf{d}_{n+\ell} &= [a_{1,n+\ell} \; a_{2,n+\ell} \; \ldots \; a_{K,n+\ell}]^{\mathcal{H}}. \end{aligned}$$

The covariance matrix of the matched filter output depends on both the target and clutter characteristics, and it is given by

$$\begin{aligned} \mathbf{R}_\ell = E[\mathbf{r}_\ell \mathbf{r}_\ell^{\mathcal{H}}] = E[\mathbf{b}\mathbf{b}^{\mathcal{H}}] |z_s[\ell - n_0]|^2 \\ + \sum_{n=-(N-1)}^{N-1} \sum_{l=-(N-1)}^{N-1} E[\mathbf{d}_{n+\ell} \; \mathbf{d}_{l+\ell}^{\mathcal{H}}] \; z_s[\ell] \; z_s^{\mathcal{H}}[l] \, . \end{aligned} \tag{4.12}$$

The matrix can be re-written in the form of the compound Gaussian sea clutter model as

$$\mathbf{R}_\ell = \sigma^2 \mathbf{I}_K \; |z_s[\ell - n_0]|^2 + \sum_{n=-(N-1)}^{N-1} \Phi \; L_{n+\ell} \; |z_s[n]|^2 \tag{4.13}$$

where Φ is the speckle covariance matrix and L_ℓ is the sea clutter texture component.

Some existing detection methods use the above formulation to estimate the texture and speckle clutter components for use in a generalized likelihood ratio test. While a reasonable approach, estimating the texture and speckle clutter components is computationally intensive, and it is often performed using expectation maximization or another iterative method. A less computationally intensive approach, that also yields reasonably good results, estimates sample covariance matrix from the data in all the range bins in one coherent processing interval as

$$\mathbf{R} = \frac{1}{N} \sum_{\ell=0}^{N-1} (\mathbf{r}_\ell - \bar{\mathbf{r}}_\ell)(\mathbf{r}_\ell - \bar{\mathbf{r}}_\ell)^{\mathcal{H}} \tag{4.14}$$

where $\bar{\mathbf{r}}_\ell$ is the mean value of $\mathbf{r}_\ell$ at the ℓth lag. To suppress the clutter from the received signal, we decompose $\mathbf{R}$ into the eigenvector matrix $\mathbf{Q}$ and diagonal eigenvalue matrix $\mathbf{D}$ to obtain

$$\mathbf{R} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\mathcal{H}}, \tag{4.15}$$

where we assume that the eigenvalues along the diagonal of $\mathbf{D}$ are sorted in descending order. The eigenvector matrix $\mathbf{Q}$ is also sorted according to the ordered eigenvalue matrix.

Negative SCR values imply that the larger eigenvalues and associated eigenvectors are due to sea clutter and define the eigenvectors that we want to suppress. We form a matrix $\mathbf{Q}_c$ from the $J < K$ eigenvectors of $\mathbf{Q}$ whose columns are associated with the smallest J eigenvalues of $\mathbf{R}$. The projected signal onto the signal subspace is given by $\mathbf{Q}_c\mathbf{Q}_c^{\mathcal{H}}\mathbf{r}_\ell$; this is the clutter suppressed signal that results in a larger SCR than

$\mathbf{r}_\ell$. Using the clutter suppressed signal for target detection results in an improved detection performance when compared to that of the GMF detector.

4.4 Simulation Results

In order to evaluate the performance characteristics of the GMF and SCS detectors across a range of SCRs, we need to vary the strength of the clutter and target reflections. Varying these parameters is straightforward when clutter realizations are simulated using the compound Gaussian model or any other statistical model. However, this is not the case when using our proposed sea clutter generation model. The data generated from the physics-based FDTD model is controlled by physical properties of the sea surface (such as the size and shape of the waves), the target (such as the size of an object), and the radar (such as the radar beamwidth). While we have direct control over the strength of the clutter and target reflections, we do not know the exact numerical value of the SCR. This is because the reflected signal is a combination of both direct reflection from the target as well as delayed reflections from the sea surface; this makes it difficult to calculate just the target component or just the clutter component of the received signal.

As we cannot control the numerical SCR values, we cannot specify the exact detector performance, such as the probability of detection versus the probability of false alarm, for a given SCR. We can, however, evaluate the detector performance for relative ranges of SCR values, that is simulate scenarios for relatively larger or smaller SCRs. In order to accomplish this, we keep all parameters but one constant, and then we vary that one parameter to affect the SCR. In the following simulations, we hold the target size and radar beamwidth constant but increase the size of the waves resulting in varying SCR values. We consider three such scenarios to illustrate the detector performance across a wide range of SCRs. The SCR in these scenarios is ordered as these scenarios, $\mathrm{SCR}_1 > \mathrm{SCR}_2 > \mathrm{SCR}_3$.

The first scenario shows detection performance for a large SCR value. The receiver operating characteristic (ROC) curves, demonstrating probability of detection $\mathbf{P_D}$ as a function of the probability of false alarm $\mathbf{P_{FA}}$, for the GMF and SCS detectors at SCR_1 is shown in Figure 4.4. As expected, GMF detector outperforms SCS because of the positive SCR. When operating on a positive SCR the SCS detector removes the dominant mode, which in this case happens to be the signal of interest, and as a result its performance is poor.

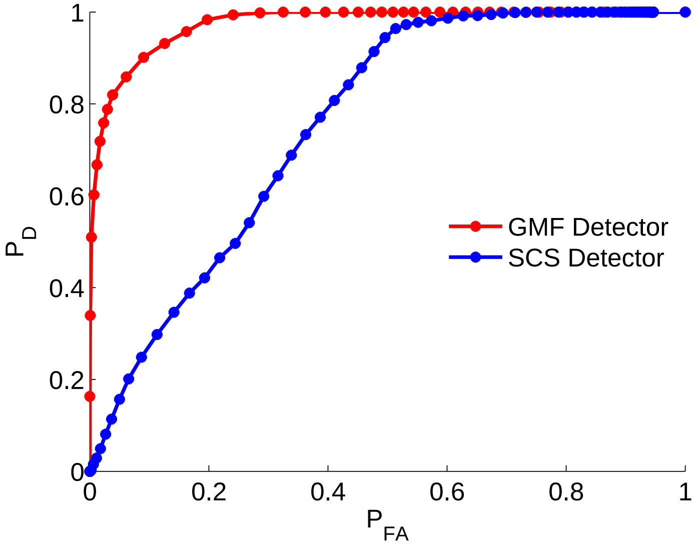


Figure 4.4: ROC Curves Comparing the Performance of the SCS and GMF Detectors Using SCR_1 Values.

In the second scenario, SCR_2, the SCR is now low enough that the SCS detector is removing clutter power from the radar returns, and is improving the SCS detector ROC curve. At this SCR, the GMF detector is performing much worse than before, and does not perform as well as the SCS detector. The ROC curves for both the GMF and SCS detectors are plotted in Figure 4.5.

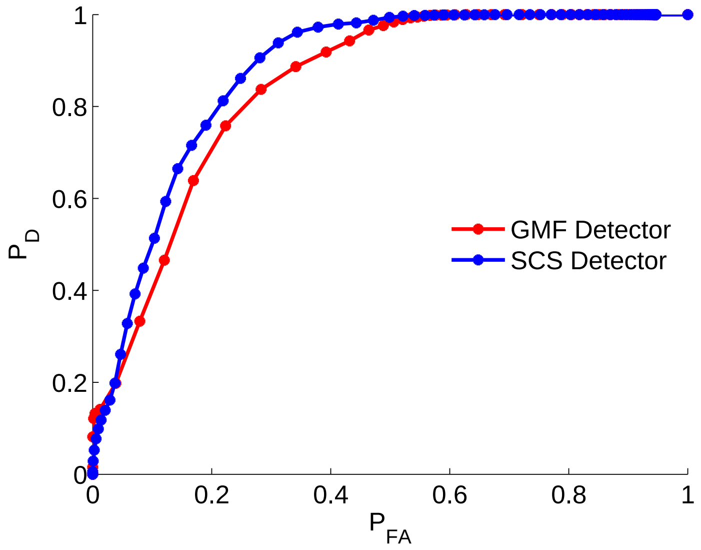


Figure 4.5: ROC Curves Comparing the Performance of the SCS and GMF Detectors Using SCR_2 Values.

The ROC curves for the third scenario, SCR_3 are shown in Figure 4.6. Here, the SCR is even more negative than in Figure SCR_2, and the processing gain of the SCS detector over the GMF detector has increased.

The simulated results shown in Figures 4.4, 4.5 and 4.6 use the sea clutter generation model with K=31 pulses; for the SCS detector, only the first eigenmode was suppressed (J=30). The number of clutter eigenvectors is data dependent and is often chosen by looking for an abrupt drop-off value in the eigenvalue amplitude from the ordered list of eigenvalues. For K=31, this drop-off value occurred after only one eigenvalue. We expect that for larger values of K, more than one clutter eigenmode would need to be suppressed. However, as in this set of simulations we considered a constant pulse repetition frequency, larger values of K resulted in poorer detection statistics because the sample covariance matrix was then computed from decorrelated sea clutter data.

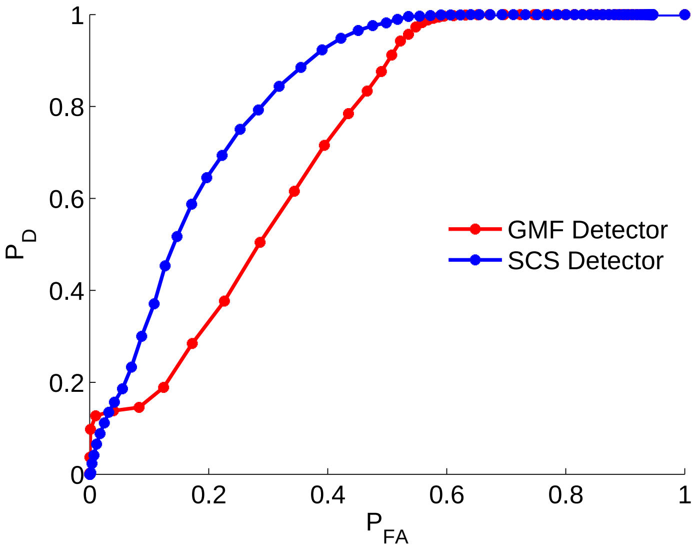


Figure 4.6: ROC Curves Comparing the Performance of the SCS and GMF Detectors Using SCR_3 Values.

Decreasing the SCR beyond what it is in the SCR_3 results in poorer performance for both the SCS and GMF detectors. While the clutter removed from the first eigenmode using the SCS detector improves the SCR, it does not remove enough clutter to continue improving the performance. A more complex algorithm would need to be developed to automatically select the number of eigenmodes which should be removed.

Chapter 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

The dissertation proposed signal processing methods for improving detection performance in molecular biology and radar applications. The signal processing algorithms included mapping biological sequences to signals and then using highly-localized time-frequency representations to estimate epitopes and identifying mimotopes from one-dimensional sequences and to perform alignment in protein structures. They also included an improved detector of a target in heavy sea clutter based on a high-fidelity physics-based electromagnetics simulation.

Random sequence peptide microarray analysis requires the detection and identification of antibody epitopes from microarray peptide sequences to discriminate between pathogens and diagnose diseases. This was achieved by first mapping characteristics of peptide and epitope sequences to parameters of highly-localized Gaussian waveforms in the time-frequency plane. After down-selecting the large number of sequences from a microarray, time-frequency based matching methods were used to estimate epitope candidates corresponding to specific pathogens. The performance of the novel epitope estimation and identification method was demonstrated using eight monoclonal antibodies. The candidate sequences that resulted in a stronger response for one antibody over the others corresponded well with the actual epitope sequences that generated the monoclonal antibodies. Using this method, we estimated exact epitope subsequences for five of the eight monoclonal antibodies, and we estimated epitope subsequences which closely resembled the exact subsequences for the other three monoclonal antibodies.

Initial results for the 330k array and a comprehensive description of the signal processing algorithms on subsequence estimation were published in a book chapter [117]. In [118] we present a more in-depth analysis which includes both epitope estimation and mimotope identification. We have also performed some related work on the adaptive learning of peptide features, using a smaller random sequence peptide array with 10k peptides [119]. This work was supported in part by DTRA, and by the Ira A. Fulton Schools of Engineering Dean's Fellowship.

We demonstrated a novel method of protein alignment incorporating multiple attributes by mapping them onto three dimensional (3-D) Gaussian waveforms and multiple waveform transformations between neighboring amino acids in the protein. In particular, we map information about a protein's sequence location, structure and characteristic properties, and use a combined matching score to obtain protein multi-alignment. As demonstrated by an example with query and database proteins from the Protein Data Bank, when multiple attributes are incorporated in the alignment, the overall alignment score improves and can lead to information on mutations that cause changes not only in the protein structure but also in the protein [32].

We have also considered a detection problem that made use of physics-based modeling but in a different application area. Specifically, we considered the detection of a low radar cross section (RCS) target in heavy sea clutter. The modeled sea clutter was generated using a 3-D random dynamic sea surface with gravity and capillary wave models. The model included gravity waves with gravity as the restoring force as well as capillary waves where the restoring force is the water surface tension. The generation model includes two dimensional (2-D) finite-difference time-domain (FDTD) sea clutter simulations. We investigated a simple matched filter detector and a subspace clutter suppression detector, and we used the FDTD simulations to compare the performance of the two detectors [120]. This work was supported in part by ONR, and a Doctoral Fellowship from the ASU Security and Defense Systems Initiative (SDSI) Institute. Note that some additional radar work was performed related to waveform design [121].

5.2 Future Work

5.2.1 Random Sequence Peptide Microarray Epitope and Mimotope Estimation

Future research into processing random sequence peptide array data will focus on improving the robustness of the algorithms presented in this dissertation. One of the major conclusions of the random sequence peptide array epitope estimation was that Pearson's correlation coefficient is an excellent ranking metric. It selects the peptides which do not fluoresce strongly relative to the median array response, yet are fluorescing strongly relative to how those peptides fluoresce in other antibody samples. It would be useful to confirm the results of this work with a new set of monoclonal antibody samples where the data are collected using exactly the same procedures such that the background fluorescence of all the samples are approximately constant. Furthermore, sometimes there is cross-reactivity between monoclonal antibody samples which can lead to worse performance of the Pearson's correlation coefficient as a ranking metric. The algorithm could be modified to be tolerant of this cross-reactivity in order, for example, for the epitope of `Flag` to be better estimated.

5.2.2 Protein Alignment Using Time-Frequency Encoded Waveforms

The work published in [32] included structural information in the 3-D Gaussian time-frequency modulation. While the locational and functional information was included in the alignment score, it was not part of the time-frequency modulation in Equation 3.12. Future work would involve incorporating that information into the time-frequency transform, and doing so in a computationally efficient manner. Additionally, this method of 3-D alignment could be tested using evolutionarily related proteins from the Protein Data Bank.

When incorporating additional information into the time-frequency transformation, the computational costs of additional information must be considered. Three dimensional alignment can be computationally intensive, because of the number of samples required to numerically evaluate Equation 3.9. When incorporating time-frequency modulation, the Nyquist sampling rate increases according to the bandwidth of the modulation in each of the three dimensions. While the sampling rate and resulting number of computations increases in a single dimension as $O(N)$, where N is the number of samples, the computational burden of evaluating Equation 3.9 increases as $O(N^3)$. This quickly becomes a significant computational burden, even on modern computers. One way to mitigate this is to consider time-frequency modulations which are a function of one dimension at a time, e.g. a sinusoidal modulation in the t_x dimension. This class of modulations allows for encoding additional information without an overwhelming increase in computational costs.

5.2.3 Sea Clutter Mitigation For Electromagnetically Small Targets

The radar sea clutter processing results showed improved performance for specific target and clutter geometries, and the problem should be extended to include more complex targets, and additional orientations between the sea clutter, target, and radar to establish the circumstances under which the target subspace and clutter subspace are meaningfully separable. Future work could also include dynamically estimating the number of clutter eigenmodes to cancel, and quantifying how dynamic estimation would affect the detection statistics.

REFERENCES

[1] [Online]. Available: http://www.wwpdb.org/

[2] J. Richer, S. A. Johnston, and P. Stafford, "Epitope identification from fixed-complexity random-sequence peptide microarrays," *under review Molecular And Cellular Proteomics*, 2014.

[3] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, pp. 8–20, July 2001.

[4] X.-Y. Zhang, F. Chen, Y.-T. Zhang, S. C. Agner, M. Akay, Z.-H. Lu, M. M. Y. Waye, and S. K.-W. Tsui, "Signal processing techniques in genomic engineering," *Proceedings of the IEEE*, vol. 90, no. 12, pp. 1822–1833, December 2002.

[5] P. P. Vaidyanathan, "Genomics and proteomics: A signal processor's tour," *IEEE Circuits and Systems Magazine*, pp. 6–29, 2004.

[6] E. R. Dougherty, A. Datta, and C. Sima, "Research issues in genomic signal processing," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 46–68, November 2005.

[7] Z. Aydin and Y. Altunbasak, "A signal processing application in genomic research: Protein secondary structure prediction," *IEEE Signal Processing Magazine*, pp. 128–131, July 2006.

[8] D. Schonfeld, J. Goutsias, I. Shmulevich, I. Tabus, and A. H. Tewfik, "Introduction to the issue on genomic and proteomic signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, 2008.

[9] L. Rowen, G. Mahairas, and L. Hood, "Sequencing the human genome," *Science*, vol. 278, no. 5338, pp. 605–607, 1997.

[10] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nature Methods*, vol. 5, pp. 16–18, 2008.

[11] R. M. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, 1998.

[12] G. D'Avenio, M. Grigioni, G. Orefici, and R. Creti, "SWIFT (sequence-wide investigation with Fourier transform): A software tool for identifying proteins of a given class from the unannotated genome sequence," *Bioinformatics*, vol. 21, no. 13, pp. 2943–2949, July 2005.

[13] H. Herzel and I. Große, "Correlations in DNA sequences: The role of protein coding segments," *Physical Review E*, vol. 55, no. 1, pp. 800–810, January 1997.

[14] H. Herzel, E. N. Trifonov, O. Weiss, and I. Große, "Interpreting correlations in biosequences," *Physica A: Statistical Mechanics and its Applications*, vol. 249, no. 1, pp. 449–459, 1998.

[15] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," *Bioinformatics*, vol. 16, no. 12, pp. 1073–1081, 2000.

[16] G. Dodin, P. Vandergheynst, P. Levoir, C. Cordier, and L. Marcourt, "Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences," *Journal on Theoretical Biology*, pp. 323–326, 2000.

[17] J. A. Berger, S. K. Mitra, and J. Astola, "Power spectrum analysis for DNA sequences," *IEEE Transactions on Signal Processing*, vol. 2, pp. 29–32, 2003.

[18] D. Sussillo, A. Kundaje, and D. Anastassiou, "Spectrogram analysis of genomes," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 1, pp. 29–42, January 2004.

[19] M. Altaiski, O. Mornev, and R. Polozov, "Wavelet analysis of DNA sequences," *Genetic Analysis: Biomolecular Engineering*, pp. 165–168, December 1996.

[20] T. Meng, A. T. Soliman, M. Shyu, Y. Yang, S. Chen, S. S. Iyengar, J. S. Yordy, and P. Iyengar, "Wavelet analysis in current cancer genome research: A survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 6, pp. 1442–14 359, 2013.

[21] A. L. Rockwood, D. K. Crockett, J. R. Oliphant, and K. S. J. Elenitoba-Johnson, "Sequence alignment by cross-correlation," *Journal of Biomolecular Techniques*, vol. 16, pp. 453–458, 2005.

[22] M. S. Rosenberg, Ed., *Sequence Alignment: Methods, Models, Concepts, and Strategies.* University of California Press, 2009.

[23] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, March 1981.

[24] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, October 1990.

[25] W. J. Kent, "BLAT–The BLAST-like alignment tool," *Genome Research*, vol. 12, no. 4, pp. 656–664, April 2002.

[26] J. Zhu, J. S. Liu, and C. E. Lawrence, "Bayesian adaptive sequence alignment algorithms," *Bioinformatics*, vol. 14, no. 1, pp. 25–39, 1998.

[27] W. Wang and D. H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Transactions on Signal Processing*, vol. 50, no. 3, pp. 628–634, March 2002.

[28] A. K. Brodzik, "A comparative study of cross-correlation methods for alignment of DNA sequences containing repetitive patterns," in *European Signal Processing Conference*, 2005.

[29] ——, "Phase-only filtering for the masses (of DNA data): A new approach to DNA sequence alignment," *IEEE Transactions on Signal Processing*, June 2006.

[30] L. Ravichandran, A. Papandreou-Suppappola, A. Spanias, Z. Lacroix, and C. Legendre, "Waveform mapping and time-frequency processing of DNA and protein sequences," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4210–4224, September 2011.

[31] L. Ravichandran, "Waveform mapping and time-frequency processing of biological sequences and structures," Ph.D., Arizona State University, Tempe, Arizona, 2011.

[32] B. O'Donnell, A. Maurer, and A. Papandreou-Suppappola, "Waveform processing for protein multi-alignment by mapping locational, structural and functional attributes," in *Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, November 2013.

[33] J. B. Delehanty and F. S. Ligler, "A microarray immunoassay for simultaneous detection of proteins and bacteria," *Analytical Chemistry*, vol. 74, no. 21, pp. 5681–5687, 2002.

[34] U. Reineke, C. Ivascu, M. Schlief, C. Landgraf, S. Gericke, G. Zahn, H. Herzel, R. Volkmer-Engert, and J. Schneider-Mergener, "Identification of distinct antibody epitopes and mimotopes from a peptide array of 5520 randomly generated sequences," *Journal of Immunological Methods*, vol. 267, no. 1, pp. 37–51, September 2002.

[35] X. Duburcq, C. Olivier, F. Malingue, R. Desmet, A. Bouzidi, F. Zhou, C. Auriault, H. Gras-Masse, and O. Melnyk, "Peptide-protein microarrays for the simultaneous detection of pathogen infections," *Bioconjugate Chemistry*, vol. 15, no. 2, pp. 307–316, 2004.

[36] F. Breitling, A. Nesterov, V. Stadler, T. Felgenhauer, and F. R. Bischoff, "High-density peptide arrays," *Molecular BioSystems*, vol. 5, no. 3, pp. 224–234, February 2009.

[37] J. V. Price, S. Tangsombatvisit, G. Xu, J. Yu, D. Levy, E. C. Baechler, O. Gozani, M. Varma, P. J. Utz, and C. L. Liu, "On silico peptide microarrays for high-resolution mapping of antibody epitopes and diverse protein-protein interactions," *Nature Medicine*, vol. 18, no. 9, pp. 1434–1440, September 2012.

[38] R. M. Lequin, "Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA)," *Clinical Chemistry*, vol. 51, no. 12, pp. 2415–2418, December 2005.

[39] G. Chen, Z. Zuo, Q. Zhu, A. Hong, X. Zhou, X. Gao, and T. Li, "Qualitative and quantitative analysis of peptide microarray binding experiments using SVM-PEPARRAY," *Methods in Molecular Biology*, vol. 570, pp. 403–411, 2009.

[40] B. Y. Renard, M. Lower, Y. Kuhne, U. Reimer, A. Rothermel, O. Tureci, J. C. Castle, and U. Sahin, "rapmad: robust analysis of peptide microarray data," *BMC Bioinformatics*, vol. 12, no. 324, pp. 1–10, 2011.

[41] R. D. Wolfinger, G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. S. Paules, "Assessing gene significance from cDNA microarray expression data via mixed models," *Journal of Computational Biology*, vol. 8, no. 6, pp. 625–637, 2001.

[42] J. R. Haan, S. Bauerschmidt, R. van Schaik, E. Piek, L. Buydens, and R. Wehrens, "Robust ANOVA for microarray data," *Chemometrics and Intelligent Laboratory Systems*, vol. 98, pp. 38–44, 2009.

[43] A. Hirakawa, C. Hamada, and I. Yoshimura, "Sample size calculation for a regularized t-statistic in microarray experiments," *Statistics and Probability Letters*, vol. 81, pp. 870–875, July 2011.

[44] R. F. Halperin, P. Stafford, and S. A. Johnston, "Exploring antibody recognition of sequence space through random-sequence peptide microarrays," *Molecular & Cellular Proteomics*, vol. 10, no. 3, March 2011.

[45] J. T. Ballew, J. A. Murray, P. Collin, M. Mäki, M. F. Kagnoff, K. Kaukinen, and P. S. Daugherty, "Antibody biomarker discovery through in vitro directed evolution of consensus recognition epitopes," *Proceedings of the National Academy of Science*, vol. 110, no. 48, pp. 19 330–19 335, November 2013.

[46] K. Schnatbaum, H.-U. Schmoldt, M. Daneschdar, L. M. Plum, J. Jansong, J. Zerweck, Y. Kühne, A. Masch, H. Wenschuh, M. Fiedler, O. Türeci, U. Sahin, and U. Reimer, "Peptide microarrays enable rapid mimotope optimization for pharmacokinetic analysis of the novel therapeutic antibody IMAB362," *Biotechnology Journal*, vol. 9, pp. 545–554, 2014.

[47] R. Knittelfelder, A. B. Riemer, and E. Jensen-Jarolim, "Mimotope vaccination from allergy to cancer," *Expert Opinion on Biological Therapy*, vol. 9, no. 4, pp. 493–506, April 2009.

[48] J. D. Buhrman, K. R. Jordan, D. J. Munson, B. L. Moore, J. W. Kappler, and J. E. Slansky, "Improving antigenic peptide vaccines for cancer immunotherapy using a dominant tumor-specific T cell receptor," *The Journal of Biological Chemistry*, vol. 288, no. 46, pp. 33 213–33 225, November 2013.

[49] W. H. Chen, P. P. Sun, Y. Lu, W. W. Guo, Y. X. Huang, and Z. Q. Ma, "MimoPro: A more efficient web-based tool for epitope prediction using phage display libraries," *BMC Bioinformatics*, vol. 12, pp. 1–13, May 2011.

[50] E. Engvall and P. Perlmann, "Enzyme-linked immunosorbent assay (ELISA): Quantitative assay of immunoglobulin G," *Immunochemistry*, vol. 8, pp. 871–874, 1971.

[51] G. P. Smith and V. A. Petrenko, "Phage display," *Chemical Reviews*, vol. 97, no. 2, pp. 391–410, 1997.

[52] F. Fack, B. Hügle-Dörr, D. Song, I. Queitsch, G. Petersen, and E. K. Bautz, "Epitope mapping by phage display: Random versus gene-fragment libraries," *J. of Immunological Methods*, vol. 206, pp. 43–52, 1997.

[53] C. Landgraf, S. Panni, L. Montecchi-Palazzi, L. Castagnoli, J. Schneider-Mergener, R. Volkmer-Engert, and G. Cesareni, "Protein interaction networks by proteome peptide scanning," *PLoS Biology*, vol. 2, 2004.

[54] Y. L. Yip and R. L. Ward, "Epitope discovery using monoclonal antibodies and phage peptide libraries," *Combinatorial Chemistry & High Throughput Screening*, vol. 2, no. 3, pp. 125–138, 1999.

[55] B. Ru, J. Huang, P. Dai, S. Li, Z. Xia, H. Ding, H. Lin, F. Guo, and X. Wang, "MimoDB: a new repository for mimotope data derived from phage display technology," *Molecules*, vol. 15, pp. 8279–8288, Nov. 2010.

[56] J. Huang, B. Ru, P. Zhu, F. Nie, J. Yang, X. Wang, P. Dai, H. Lin, F.-B. Guo, and N. Rao, " MimoDB 2.0: a mimotope database and beyond," *Nucleic Acids Research*, vol. 40, pp. 271–277, January 2012.

[57] P. Stafford and S. A. Johnston, "Microarray technology displays the complexities of the humoral immune response," *Expert Reviews in Molecular Diagnostics*, vol. 11, no. 1, pp. 5–8, 2011.

[58] P. Stafford, Z. Cichacz, N. W. Woodbury, and S. A. Johnston, "Immunosignature system for diagnosis of cancer," *Proceedings of the National Academy of Sciences*, vol. 111, no. 30, pp. E3072–E3080, July 2014.

[59] J. R. Brown, P. Stafford, S. A. Johnston, and V. Dinu, "Statistical methods for analyzing immunosignatures," *BMC Bioinformatics*, vol. 12, no. 349, pp. 1–15, August 2011.

[60] P. Stafford, R. Halperin, J. B. Legutki, D. M. Magee, J. Galgiani, and S. A. Johnston, "Physical characterization of the immunosignaturing effect," *Molecular & Cellular Proteomics*, vol. 11, no. 4, pp. 011 593–1–14, April 2012.

[61] A. Malin, N. Kovvali, J. J. Zhang, B. Chakraborty, A. Papandreou-Suppappola, S. A. Johnston, and P. Stafford, "Adaptive learning of immunosignaturing peptide array features for biothreat detection and classification," in *Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, November 2011, pp. 1883–1887.

[62] M. Kukreja, S. A. Johnston, and P. Stafford, "Comparative study of classification algorithms for immunosignaturing data," *BMC Bioinformatics*, vol. 13, no. 1, pp. 139–152, January 2012.

[63] J. B. Legutki, D. M. Magee, P. Stafford, and S. A. Johnston, "A general method for characterization of humoral immunity induced by a vaccine or infection," *Vaccine*, vol. 28, no. 28, pp. 4529–4537, June 2010.

[64] R. F. Halperin, P. Stafford, J. S. Emery, K. A. Navalkar, and S. A. Johnston, "GuiTope: an application for mapping random-sequence peptides to protein sequences," *BMC Bioinformatics*, vol. 13, no. 1, p. 1, January 2012.

[65] G. Bastas, S. R. Sompuram, B. Pierce, K. Vani, and S. A. Bogen, "Bioinformatic requirements for protein database searching using predicted epitopes from disease-associated antibodies," *Molecular & Cellular Proteomics*, vol. 7, no. 2, pp. 247–256, February 2008.

[66] C. G. Adda, R. F. Anders, L. Tilley, and M. Foley, "Random sequence libraries displayed on phage: Identification of biologically important molecules," *Combinatorial Chemistry & High Throughput Screening*, vol. 5, pp. 1–14, February 2002.

[67] M. Andreatta, C. Schafer-Nielsen, O. Lund, S. Buus, and M. Nielsen, "NNAlign: a web-based prediction method allowing non-expert end-user discovery of sequence motifs in quantitative peptide data," *PLoS ONE*, vol. 6, no. 11, pp. 1–11, November 2011.

[68] Y. Chen, Y. Pan, Y. Guo, L. Qiu, X. Ding, and X. Che, "Comprehensive mapping of immunodominant and conserved serotype- and group-specific B-cell epitopes of nonstructural protein 1 from dengue virus type 1," *Virology*, vol. 398, no. 2, pp. 290–298, March 2010.

[69] S. Buus, J. Rockberg, B. Forsstrom, P. Nilsson, M. Uhlen, and C. Schafer-Nielsen, "High-resolution mapping of linear antibody epitopes using ultrahigh-density peptide microarrays," *Molecular & Cellular Proteomics*, vol. 11, pp. 1790–1800, 2012.

[70] U. Reineke, "Antibody epitope mapping using arrays of synthetic peptides," *Methods in Molecular Biology, Antibody Engineering: Methods and Protocols*, no. 248, pp. 443–463, 2004.

[71] W. R. Pearson and M. L. Sierk, "The limits of protein sequence comparison?" *Current Opinion in Structural Biology*, vol. 15, pp. 254–260, 2005.

[72] P. Daras, D. Zarpalas, A. Axenopoulos, D. Tzovaras, and M. G. Strintzis, "Three-dimensional shape-structure comparison method for protein classification," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 3, pp. 193–207, 2006.

[73] R.Yan, D. Xu, J. Yang, S. Walker, and Y. Zhang, "A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction," *Scientific Reports*, vol. 3, no. 2619, 2013.

[74] G. Csaba, F. Birzele, and R. Zimme, "Protein structure alignment considering phenotypic plasticity," *Bioinformatics*, vol. 24, pp. 98–104, 2008.

[75] H. Hasegawa and L. Holm, "Advances and pitfalls of protein structural alignment," *Current Opinion in Structural Biology*, vol. 19, pp. 341–348, 2009.

[76] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," *Journal of Molecular Biology*, vol. 233, no. 1, pp. 123–138, 1993.

[77] O. O'Sullivan, K. Suhre, C. Abergel, D. G. Higgins, and C. Notredame, "3DCoffee: Combining protein sequences and structures within multiple sequence alignments," *Journal of Molecular Biology*, vol. 340, pp. 385–295, 2004.

[78] K. Chou, "Automated prediction of protein attributes and its impact on biomedicine and drug discovery," in *Automation in Proteomics and Genomics: An Engineering Case-Based Approach*, G. Alterovitz, R. Benson, and M. Ramoni, Eds. Wiley, 2009, ch. 5.

[79] P. H. Guzzi, M. Mina, C. Guerra, and M. Cannataro, "Semantic similarity analysis of protein data: Assessment with biological features and issues," *Briefings in Bioinformatics*, vol. 5, pp. 569–585, 2012.

[80] S. Wang, J. Ma, J. Peng, and J. Xu, "Protein structure alignment beyond spatial proximity," *Scientific Reports*, vol. 3:1448, pp. 1–7, 2013.

[81] G. A. Petsko and D. Ringe, *Protein Structure and Function.* New Science Press, 2004.

[82] L. Young, R. L. Jernigan, and D. G. Covell, "A role for surface hydrophobicity in protein-protein recognition," *Protein Science*, vol. 5, no. 3, pp. 717–729, 1994.

[83] B. Matthews, *Hydrophobic Interactions in Proteins.* Chichester: John Wiley & Sons Ltd, 1999. [Online]. Available: http://www.els.net

[84] M. J. Betts and R. B. Russell, "Amino acid properties and consequences of substitutions," in *Bioinformatics for Geneticists*, M. R. Barnes and I. C. Gray, Eds. Wiley, 2003.

[85] M. O. Jensen and O. G. Mouritsen, "Lipids do influence protein function-the hydrophobic matching hypothesis revisited," *Biochimica et Biophysica Acta*, pp. 205–226, 2004.

[86] I. Ladunga and R. F. Smith, "Amino acid substitutions preserve protein folding by conserving steric and hydrophobicity properties," *Protein Engineering*, vol. 10, pp. 187–196, 1997.

[87] C. Vogel, M. Bashton, N. D. Kerrison, C. Chothia, and S. A. Teichmann, "Structure, function and evolution of multidomain proteins," *Current Opinion in Structural Biology*, vol. 14, pp. 208–216, 2004.

[88] G. M. Maggiora, D. C. Rohrer, and J. Mestres, "Comparing protein structures: A Gaussian-based approach to the three-dimensional structural similarity of proteins," *Journal of Molecular Graphics and Modelling*, vol. 19, pp. 168–178, 2001.

[89] L. Ravichandran, A. Papandreou-Suppappola, A. Spanias, and Z. Lacroix, "Multiple protein structure alignment using time-frequency processing techniques," in *IEEE Biomedical Circuits and Systems Conference*, November 2010, pp. 94–97.

[90] S. F. Altschul, T. L. Madden, A. A. Schaffera, J. Zhang, Z. Zhang, and W. M. D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389–3402, 1997.

[91] A. L. Hu and K. C. C. Chan, "Utilizing both topological and attribute information for protein complex identification in ppi networks," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 10, no. 3, pp. 780–792, May 2013.

[92] M. Skolnik, *Radar Handbook*. McGraw Hill, 2008.

[93] S. Haykin, R. Bakker, and B. W. Currie, "Uncovering nonlinear dynamics - The case study of sea clutter," *Proceedings of the IEEE*, vol. 90, no. 5, pp. 860–881, 2002.

[94] M. E. Watts and R. E. Diaz, "Perfect plane-wave injection into a finite FDTD domain through teleportation of fields," Arizona State University, Tech. Rep., 2003.

[95] P. Stinco, M. Greco, and F. Gini, "Adaptive detection in compound-Gaussian clutter with inverse-gamma texture," in *IEEE Int. Conf. on Radar*, vol. 1, 2011, pp. 434–437.

[96] F. Gini and A. Farina, "Vector subspace detection in compound-gaussian clutter. Part I: Survey and new results," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, no. 4, pp. 1295–1311, 2002.

[97] S. P. Sira, D. Cochran, A. Papandreou-Suppappola, D. Morrell, W. Moran, S. Howard, and R. Calderbank, "Adaptive waveform design for improved detection of low-RCS targets in heavy sea clutter," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 1, pp. 56–66, 2007.

[98] P. C. Ng and S. Henikoff, "Predicting the effects of amino acid substitutions on protein function," *The Annual Review of Genomics and Human Genetics*, vol. 7, pp. 61–80, 2006.

[99] A. Suvorova, B. Moran, and M. Viola, "Adaptive modelling of sea clutter and detection of small targets in heavy clutter," in *International Radar Conference*, 2003, pp. 614–618.

[100] A. Haimovich, "The eigencanceler: Adaptive radar by eigenanalysis methods," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, no. 2, pp. 532–542, 1996.

[101] [Online]. Available: http://www.biodesign.asu.edu/research/research-centers/innovations-in-medicine/

[102] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, December 1993.

[103] A. Papandreou-Suppappola and S. B. Suppappola, "Analysis and classification of time-varying signals with multiple time-frequency structures," *IEEE Signal Processing Letters*, vol. 9, p. 9295, 2002.

[104] R. S. Hawley and C. A. Mori, *The human genome: A user's guide*, 3rd ed. Academic Press, 2010.

[105] B.-J. M. Webb-Robertson, M. M. Matzke, J. M. Jacobs, J. G. Pounds, and K. M. Waters, "A statistical selection strategy for normalization procedures in LC-MS proteomics experiments through dataset dependent ranking of normalization scaling factors," *Proteomics*, vol. 11, no. 24, pp. 4736–4741, December 2011.

[106] I. Hajdú, B. Flachner, M. Bognár, B. M. Végh, K. Dobi, Z. Lőrincz, J. Lázár, S. Cseh, L. Takács, and I. Kurucz, "Monoclonal antibody proteomics: Use of antibody mimotope displaying phages and the relevant synthetic peptides for mAb scouting," *Immunology Letters*, vol. 160, pp. 172–177, 2014.

[107] D. L. Nelson and M. M. Cox, *Principles of Biochemistry*, 5th ed. New York: W. H. Freeman, Feb. 2008.

[108] G. Strang, *Introduction to Linear Algebra*, 4th ed. Wellesley, MA: Wellesley Cambridge Press, Feb. 2009.

[109] E. Thorsos, "The validity of the kirchhoff approximation for rough surface scattering using a gaussian roughness spectrum," *J. Acoust. Soc. Amer*, vol. 83, p. 7892, 1988.

[110] W. J. Pierson and L. Moskowitz, "A proposed spectral form for fully developed wind seas based on the similarity theory of S. A. Kitaigorodskii," *Journal of Geophysical Research*, vol. 69, no. 24, pp. 5181–5190, 1964.

[111] J. V. Toporkov, R. T. Marchand, and G. S. Brown, "On the discretization of the integral equation describing scattering by rough conducting surfaces," *IEEE Transactions on Antennas and Propagation*, vol. 46, no. 1, pp. 150–161, 1998.

[112] S. L. B. Frank D. Hastings, John B. Schneider, "A monte-carlo fdtd technique for rough surface scattering," *IEEE Trans. AP*, vol. 43, no. II, 1995.

[113] L. C. Bobb, G. Ferguson, , and M. Rankin, "Capillary wave measurements," *Applied Optics*, vol. 18, pp. 1167–1171, 1979.

[114] A. J. W. Sommerfeld, *Mechanics of Deformable Bodies.* Academic Press, 1950, vol. 2.

[115] G. D. Crapper, "An exact solution for progressive capillary waves of arbitrary amplitude," *Journal of Fluid Mechanics*, vol. 2, pp. 532–540, 1957.

[116] J. Frechot, “Realistic simulation of ocean surface using wave spectra,” in *International Conference on Computer Graphics Theory and Applications*, 2006.

[117] B. O’Donnell, A. Maurer, and A. Papandreou-Suppappola, “Biosequence time-frequency processing: Pathogen detection and identification,” in *Excursions in Harmonic Analysis*. Springer, 2014, vol. 3.

[118] B. O’Donnell, A. Maurer, P. Stafford, A. Papandreou-Suppappola, and S. A. Johnston, “Monoclonal antibody epitope estimation and mimotope identification for random sequence peptide microarrays,” *Cancer Informatics*, sumbitted Nov. 2014.

[119] A. Malin, N. Kovvali, A. Papandreou-Suppappola, B. O’Donnell, S. Johnston, and P. Stafford, “Adaptive learning of immunosignaturing features for multi-disease pathologies,” in *Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, November 2013.

[120] B. O’Donnell, R. LeBaron, R. Diaz, and A. Papandreou-Suppappola, “Physics-based sea clutter model for improved target detection of low radar cross-section targets,” to be submitted, 2014.

[121] B. O’Donnell, J. Zhang, A. Papandreou-Suppappola, and M. Rangaswamy, “Waveform-agile multiple target tracking using probability hypothesis density filtering,” in *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2012 IEEE 7th*, Jun. 2012, pp. 245–248.

APPENDIX A

LIST OF ACRONYMS

1-D one dimensional

2-D two dimensional

3-D three dimensional

AA amino acid

BLAST basic local alignment search tool

BLOSUM blocks substitution matrix

DNA deoxyribonucleic acid

DS-OCRC down-selected occurrence count

ELISA enzyme linked immunosorbent assay

EM electromagnetic

FDTD finite-difference time-domain

GMF generalized matched filter

LoStrFn location, structure and function

mAb monoclonal antibody

MBI median background intensity

MFI median fluorescence intensity

MPD matching pursuit decomposition

OCRC occurrence count

PCC pearson's correlation coefficient

PDB protein data bank

P-M pierson-moscowitz

RCS radar cross-section

RNA ribonucleic acid

ROC receiver operating characteristic

SCR signal-to-clutter ratio

SCS subspace clutter suppression

SFI standard deviation of the fluorescence intensity

SVM support vector machine

TF time-frequency

TIFF tagged image file format

APPENDIX B

RESULTS FROM THE ESTIMATION OF SUBSEQUENCES OF MONOCLONAL ANTIBODY SAMPLES

In Chapter 2, we proposed a time-frequency based signal processing algorithm for estimating epitopes and identifying mimotopes for eight different monoclonal antibody samples. Here we provide, in detail, our results for all eight samples.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
PWFK	682	19	883	2,798
PWFN	781	18	726	2,390
ARLR	659	16	1,220	5,556
ARPF	593	15	869	9,512
AVWF	634	15	720	2,529
PWFF	430	15	730	2,327
RPWF	515	15	902	3,277
RWFN	184	15	929	65,535
YSAW	503	15	731	3,613
ARWF	278	14	902	65,535

Table B.1: Estimated Top 10 Subsequences of Length $\mathcal{L} = 4$ Obtained When Algorithm 1 Was Applied to mAb 2C11, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
FEEQE	168	7	586	5,826
FFEEQ	117	7	636	5,826
VFFEE	87	7	676	5,826
ARWFN	54	6	931	65,535
AVNWF	64	6	760	187
PWFNK	139	6	848	2,144
WFNRL	30	6	1,010	1,704
ARLRP	120	5	1,098	4,613
ARRVR	30	5	1,980	4,142
DARWF	37	5	834	65,535

Table B.2: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to mAb 2C11, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
FFEEQE	116	7	636	5,826
VFFEEQ	86	7	685	5,826
DARWFN	10	4	1,197	65,535
AWRGFN	7	3	997	1,692
FARLRE	9	3	1,183	3,327
FKYARL	24	3	1,208	2,414
HFFKAL	6	3	954	1,693
KARLRP	6	3	1,652	4,613
WFARLL	6	3	1,050	1,769
WFNGYA	12	3	938	1,470

Table B.3: Estimated Top 10 Subsequences of Length $\mathcal{L} = 6$ Obtained When Algorithm 1 Was Applied to mAb 2C11, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
VFFEEQE	85	7	694	5,826
YVFFEEQ	22	3	805	2,089
AALEKDG	2,000	2	630	16,310
ALEKDGY	111	2	701	16,310
AVARPFQ	2	2	1,849	2,182
AVGWQAR	3	2	1,922	16,130
AWRGFNY	3	2	997	1,616
FARLREY	2	2	1,415	1,647
FEEQERY	13	2	656	1,559
FFEEQER	23	2	759	1,559

Table B.4: Estimated Top 10 Subsequences of Length $\mathcal{L} = 7$ Obtained When Algorithm 1 Was Applied to mAb 2C11, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
GYDS	158	58	3,667	65,535
EYDS	129	32	4,337	65,535
DSWP	169	28	1,345	65,535
YDSR	127	27	1,611	65,535
DSHP	272	24	877	65,535
DSRP	303	24	942	65,535
EADS	180	24	1,275	53,944
YDSH	59	24	7,006	65,535
DADS	100	23	2,197	65,535
YDSK	113	23	2,129	65,535

Table B.5: Estimated Top 10 Subsequences of Length $\mathcal{L} = 4$ Obtained When Algorithm 1 Was Applied to mAb 4C1, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
GYDSR	21	13	8,731	65,535
YDSRP	18	11	4,315	65,535
ADSWP	20	10	12,769	65,535
YDSHP	13	10	28,346	65,535
ADSVP	34	9	1,728	29,299
FEYDS	25	9	3,320	61,256
GYDSW	11	9	6,107	33,735
YDSKG	12	9	18,650	65,535
EDADV	22	8	3,751	30,830
GYDSH	11	8	15,211	65,535

Table B.6: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to mAb 4C1, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
FDSVGG	4	4	38,206	65,535
FKQYDS	6	4	3,840	7,562
NGYDSR	6	4	14,295	65,535
PADSWP	5	4	14,259	65,535
PFDSVG	4	4	65,535	65,535
ADSWPP	5	3	12,244	15,036
APNDSG	4	3	50,593	65,535
ARPGYL	15	3	1,463	5,797
DADSVP	4	3	8,811	24,246
DADSWP	3	3	13,294	19,770

Table B.7: Estimated Top 10 Subsequences of Length $\mathcal{L} = 6$ Obtained When Algorithm 1 Was Applied to mAb 4C1, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
AAWRFFK	2	2	5,601	9,043
AGPGYDS	2	2	8,704	11,930
APNDSGG	2	2	50,593	65,535
ARGPFAR	4	2	2,857	5,287
ARPFYAR	6	2	1,563	3,254
AVGPNWF	6	2	1,116	11,397
AWRHFNY	4	2	1,080	1,977
AYAFDSN	2	2	34,179	65,535
DADSWPW	2	2	7,805	13,294
DLAPKEY	2	2	3,538	5,732

Table B.8: Estimated Top 10 Subsequences of Length $\mathcal{L} = 7$ Obtained When Algorithm 1 Was Applied to mAb 4C1, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
WDVA	272	55	17,534	65,535
DVAW	473	52	8,790	65,535
DSAW	442	46	8,763	65,535
WQEA	135	46	65,535	65,535
DAAW	385	40	11,101	65,535
DVSW	239	36	19,765	65,535
QEYA	323	35	37,316	65,535
EDVA	242	34	20,428	65,535
WFEA	267	34	8,875	65,535
EWDA	346	32	10,617	65,535

Table B.9: Estimated Top 10 Subsequences of Length $\mathcal{L} = 4$ Obtained When Algorithm 1 Was Applied to mAb A10, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
EDFRV	34	20	65,535	65,535
EWDVA	41	15	65,535	65,535
EDVAW	35	14	65,535	65,535
WFEGA	53	14	32,589	65,535
WDVAP	33	13	65,535	65,535
DAAWP	52	11	16,042	65,535
DVAWG	57	11	10,288	65,535
EWDAA	44	11	31,044	65,535
PWFEA	69	11	10,370	65,535
WDVAW	42	11	19,322	65,535

Table B.10: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to mAb `A10`, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
DFRVDW	22	8	35,188	65,535
FRVDWK	40	8	5,252	65,535
EDFRVD	6	5	65,535	65,535
EDVRPF	10	5	39,784	65,535
PWQEAS	7	5	65,535	65,535
AVWFEG	11	4	7,222	65,535
DVAWPF	12	4	22,508	65,535
EDARSG	6	4	34,672	65,535
EDVAPN	9	4	60,074	65,535
EDVAWP	6	4	65,535	65,535

Table B.11: Estimated Top 10 Subsequences of Length $\mathcal{L} = 6$ Obtained When Algorithm 1 Was Applied to mAb `A10`, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
DFRVDWK	22	8	35,188	65,535
EDFRVDW	6	5	65,535	65,535
FRVDWKH	33	4	2,294	65,535
AGNEYAL	4	2	38,101	65,535
APEDPED	167	2	830	65,535
APWFEDS	3	2	39,460	65,535
APWKEDS	4	2	34,151	65,535
APWNEAR	3	2	65,535	65,535
AQEYRPE	2	2	65,535	65,535
AVGPWQE	3	2	65,535	65,535

Table B.12: Estimated Top 10 Subsequences of Length $\mathcal{L} = 7$ Obtained When Algorithm 1 Was Applied to mAb A10, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
HSVV	618	196	16,682	65,535
RHSV	298	188	61,191	65,535
RRHS	78	17	4,082	65,535
PWFN	781	15	952	41,733
HPWF	632	14	1,069	15,186
PWHF	908	13	1,553	31,088
AAVW	611	12	1,436	43,546
AVRG	610	12	2,488	65,535
AWPF	562	12	1,033	41,005
FGAR	384	12	2,029	56,590

Table B.13: Estimated Top 10 Subsequences of Length $\mathcal{L} = 4$ Obtained When Algorithm 1 Was Applied to mAb Ab1, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
RHSVV	209	186	65,535	65,535
RRHSV	26	16	19,861	65,535
VRHSV	11	8	23,648	65,535
ARHSV	27	5	558	65,535
AVRGF	62	5	2,013	52,630
AYAWF	45	5	828	6,867
AVWHF	168	4	1,647	25,080
AWHFS	19	4	2,254	19,262
FKEYL	37	4	1,645	12,231
FQYAL	67	4	1,095	23,112

Table B.14: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to mAb `Ab1`, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
RRHSVV	21	16	21,999	65,535
VRHSVV	10	8	26,646	65,535
ARHSVV	13	5	33,836	65,535
RHSVVW	5	4	20,292	61,580
FFEEQE	116	3	760	3,217
RGHSVV	18	3	13,371	53,733
RHSVVD	3	3	65,535	65,535
RWHSVV	21	3	3,146	51,425
AARPFA	12	2	2,229	27,035
AARWFF	6	2	764	4,349

Table B.15: Estimated Top 10 Subsequences of Length $\mathcal{L} = 6$ Obtained When Algorithm 1 Was Applied to mAb `Ab1`, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
AFQYALV	3	2	1,485	2,275
APFKGRL	4	2	1,977	3,080
ARHSVVD	2	2	65,535	65,535
AVNWFLK	2	2	2,380	2,396
AVRHSVV	3	2	29,643	30,839
FFEEQEK	34	2	948	3,217
FSLKEWY	2	2	3,281	4,092
HVVLEEV	2	2	2,183	2,529
KYARNKR	3	2	2,099	3,054
LEEVLNL	167	2	1,121	15,534

Table B.16: Estimated Top 10 Subsequences of Length $\mathcal{L} = 7$ Obtained When Algorithm 1 Was Applied to mAb Ab1, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
PWHF	14	908	1,717	43,903
AWHF	13	872	2,055	32,972
VWHF	13	632	1,713	31,684
HPWF	12	632	1,264	15,576
PWFH	12	342	1,609	29,293
PWFK	12	682	1,306	39,635
WHFN	12	736	2,103	32,972
AARL	11	437	2,862	65,535
AVWG	11	662	1,420	64,190
AVWN	11	575	1,417	63,356

Table B.17: Estimated Top 10 Subsequences of Length $\mathcal{L} = 4$ Obtained When Algorithm 1 Was Applied to mAb Ab8, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
APEDP	7	172	1,107	33,036
AVGPW	7	97	1,428	23,273
DLWKL	6	63	1,174	10,504
EDPED	6	170	1,098	9,271
PEDPE	6	169	1,093	6,309
PWFAR	5	89	1,089	33,684
WFKYA	5	61	999	24,968
YALRV	5	55	1,500	12,950
AARLP	4	61	3,228	47,610
AAWHF	4	120	2,292	13,586

Table B.18: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to mAb Ab8, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
APEDPE	6	167	1,093	6,309
PEDPED	6	167	1,093	6,309
LAPEDP	4	107	1,045	5,434
APWFKY	3	14	1,443	3,086
AVGPWF	3	30	1,312	23,273
EDPEDK	3	14	1,197	3,392
EDPEDS	3	39	1,031	6,309
FNYALR	3	14	1,685	5,422
PWFARP	3	11	1,467	3,281
WHFFKY	3	9	1,037	4,614

Table B.19: Estimated Top 10 Subsequences of Length $\mathcal{L} = 6$ Obtained When Algorithm 1 Was Applied to mAb Ab8, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
APEDPED	6	167	1,093	6,309
LAPEDPE	4	107	1,045	5,434
PEDPEDK	3	14	1,197	3,392
PEDPEDS	3	39	1,031	6,309
APWFKYA	2	3	2,591	3,086
ARPWFHP	2	2	4,410	4,809
DHPADAW	2	2	1,962	2,479
DLDSDLW	2	2	3,713	5,483
DSDLWKL	2	2	3,713	5,483
DSWFKQG	2	2	1,458	1,534

Table B.20: Estimated Top 10 Subsequences of Length $\mathcal{L} = 7$ Obtained When Algorithm 1 Was Applied to mAb Ab8, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
AVWF	634	15	944	17,499
WHFN	736	14	2,034	32,980
PWFN	781	13	1,053	41,133
AVPW	545	12	1,419	50,578
AWPF	562	12	1,098	46,024
GPWF	611	12	1,082	41,133
PFFN	381	12	1,293	31,135
PWFF	430	12	769	10,266
PWFK	682	12	1,206	41,829
VPWN	186	12	1,313	29,619

Table B.21: Estimated Top 10 Subsequences of Length $\mathcal{L} = 4$ Obtained When Algorithm 1 Was Applied to mAb DM1A, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
ALNRP	109	6	2,314	59,404
AALEK	2,058	5	2,364	43,563
ALEKD	2,100	5	2,369	43,563
DLWKL	63	5	1,075	7,130
LAWHF	62	5	1,534	7,388
AAPWF	82	4	1,298	31,956
ALLRP	62	4	1,776	28,850
ARHPW	104	4	2,002	10,599
AVRNK	45	4	2,205	23,019
AVWFF	60	4	895	7,230

Table B.22: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to mAb DM1A, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
AALEKD	2,053	5	2,368	43,563
GPWFGY	11	3	1,137	4,580
LPAVFN	7	3	4,835	14,123
YALNRP	16	3	2,937	29,036
AAFYAL	6	2	1,591	2,287
AARWHF	10	2	2,850	12,048
AAVWFN	14	2	1,279	1,869
AAWARL	7	2	1,264	2,981
ADYRHF	3	2	3,295	3,474
AGPNWF	15	2	1,368	3,406

Table B.23: Estimated Top 10 Subsequences of Length $\mathcal{L} = 6$ Obtained When Algorithm 1 Was Applied to mAb DM1A, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
AALEKDG	2,000	2	2,416	43,563
AALEKDN	3	2	7,503	17,762
APEDPED	167	2	947	5,464
APFFNLS	2	2	3,227	3,874
AVWRGNF	3	2	2,189	2,903
DAVWRGN	4	2	1,839	2,903
FGALLGW	2	2	1,317	1,606
FYALNRP	4	2	2,420	3,599
GPFYAKR	2	2	1,913	2,404
HGPFYAK	2	2	1,913	2,404

Table B.24: Estimated Top 10 Subsequences of Length $\mathcal{L} = 7$ Obtained When Algorithm 1 Was Applied to mAb `DM1A`, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
ALEK	2,143	1,325	34,896	65,535
LEKD	2,156	1,325	34,296	65,535
EKDG	2,026	1,324	42,310	65,535
AALE	2,117	1,323	36,032	65,535
KDGD	284	250	65,535	65,535
KDGE	278	191	65,535	65,535
KDGA	250	162	28,590	65,535
KDGW	276	153	10,010	65,535
KDGP	206	128	33,848	65,535
KDGH	237	90	17,798	65,535

Table B.25: Estimated Top 10 Subsequences of Length $\mathcal{L} = 4$ Obtained When Algorithm 1 Was Applied to mAb `Flag`, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
ALEKD	2,100	1,324	37,480	65,535
AALEK	2,058	1,323	40,477	65,535
LEKDG	2,002	1,323	44,588	65,535
EKDGD	268	250	65,535	65,535
EKDGE	215	191	65,535	65,535
EKDGA	250	162	28,590	65,535
EKDGW	253	153	14,352	65,535
EKDGP	182	128	50,613	65,535
EKDGH	232	90	19,087	65,535
EKDGL	134	81	14,198	65,535

Table B.26: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to mAb `Flag`, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
AALEKD	2,053	1,323	40,576	65,535
ALEKDG	2,002	1,323	44,588	65,535
LEKDGD	267	250	65,535	65,535
LEKDGE	210	191	65,535	65,535
LEKDGA	250	162	28,590	65,535
LEKDGW	253	153	14,352	65,535
LEKDGP	181	128	51,217	65,535
LEKDGH	231	90	19,120	65,535
LEKDGL	134	81	14,198	65,535
LEKDGV	139	67	24,323	65,535

Table B.27: Estimated Top 10 Subsequences of Length $\mathcal{L} = 6$ Obtained When Algorithm 1 Was Applied to mAb `Flag`, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
AALEKDG	2,000	1,323	44,588	65,535
ALEKDGD	267	250	65,535	65,535
ALEKDGE	210	191	65,535	65,535
ALEKDGA	250	162	28,590	65,535
ALEKDGW	254	153	14,427	65,535
ALEKDGP	181	128	51,217	65,535
ALEKDGH	231	90	19,120	65,535
ALEKDGL	134	81	14,198	65,535
ALEKDGV	139	67	24,323	65,535
ALEKDGS	64	55	65,535	65,535

Table B.28: Estimated Top 10 Subsequences of Length $\mathcal{L} = 7$ Obtained When Algorithm 1 Was Applied to mAb `Flag`, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
YDAP	91	44	6,400	65,535
DAPE	114	31	1,537	65,535
ADAP	285	27	864	65,535
DVPE	93	25	1,008	65,535
DAPG	168	24	1,122	65,535
DVPD	33	24	31,506	65,535
DAPV	112	23	1,027	65,535
YDVP	47	23	4,846	65,535
LDVP	153	20	823	65,535
FDAP	47	18	2,071	65,535

Table B.29: Estimated Top 10 Subsequences of Length $\mathcal{L} = 4$ Obtained When Algorithm 1 Was Applied to mAb `HA`, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
YDAPE	16	14	61,414	65,535
PYDAP	11	10	44,289	65,535
YDSPE	13	9	12,542	65,535
FDAPV	12	8	9,961	56,901
PFDAP	8	8	47,053	65,535
QYDAP	10	8	31,196	65,535
YDVPE	9	8	51,759	65,535
ADAPE	18	7	10,457	65,535
EDLPD	15	7	1,706	11,385
FYDAP	11	7	5,583	65,535

Table B.30: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to mAb HA, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	OCRC	DS-OCRC	Mean MFI	Max. MFI
FNYDSP	6	4	2,146	65,535
GYDAPE	4	4	59,422	65,535
NQYDAP	4	4	47,437	65,535
NYDSPE	4	4	11,997	65,535
AALEKD	2,053	3	694	11,285
ALEKDG	2,002	3	697	11,285
APYDAP	3	3	44,289	65,535
EDHPDG	3	3	4,984	40,563
EDLPDS	4	3	6,698	11,385
FFYDAP	3	3	6,135	65,535

Table B.31: Estimated Top 10 Subsequences of Length $\mathcal{L} = 6$ Obtained When Algorithm 1 Was Applied to mAb HA, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

Subsequence	**OCRC**	**DS-OCRC**	**Mean MFI**	**Max. MFI**
AALEKDG	2,000	3	697	11,285
FNYDSPE	3	3	19,861	65,535
PFNYDSP	3	3	19,861	65,535
AAWRNWQ	2	2	3,196	4,288
AGPYDAP	2	2	31,213	60,226
ANQYDAP	2	2	42,126	65,535
ARFDAPV	2	2	40,791	56,901
ARPFYAR	6	2	1,556	2,171
AVWFKSL	3	2	1,107	1,393
AVWRNQR	3	2	3,705	9,178

Table B.32: Estimated Top 10 Subsequences of Length $\mathcal{L} = 7$ Obtained When Algorithm 1 Was Applied to mAb `HA`, With Their Corresponding OCRC, DS-OCRC, Mean MFI and Maximum MFI; the Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

APPENDIX C

RESULTS FROM THE ESTIMATION OF SUBSEQUENCES OF DISEASE SAMPLES

In addition to the monoclonal antibody data that was analyzed in Chapter 2, we also analyzed random sequence peptide array data which were collected using blood samples of patients infected with four different diseases. The samples analyzed were, Borrelia, Dengue, West Nile Virus, and Bordetella. Figures C.1 - C.4 are plots of the down selected occurrence count for the top 10 estimated subsequences for these four diseases. Two of the analyzed disease samples have literature-reported epitopes; Borrelia and Dengle, whose epitopes are `EDAK` and `AVHAD` respectively. The epitope for Borrelia was the first epitope estimate, while the epitope for Dengue was the fifth epitope estimate.

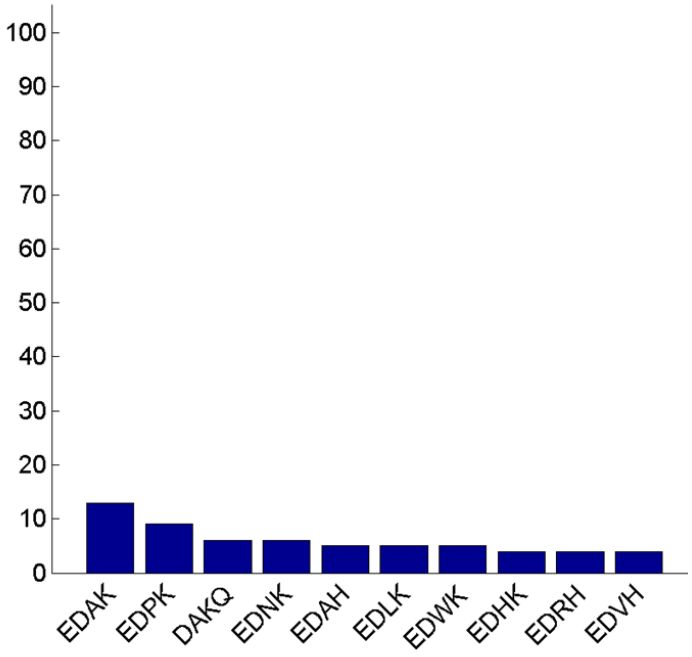


Figure C.1: Estimated Top 10 Subsequences of Length $\mathcal{L} = 4$ Obtained When Algorithm 1 Was Applied to the Borrelia Disease Sample. The Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

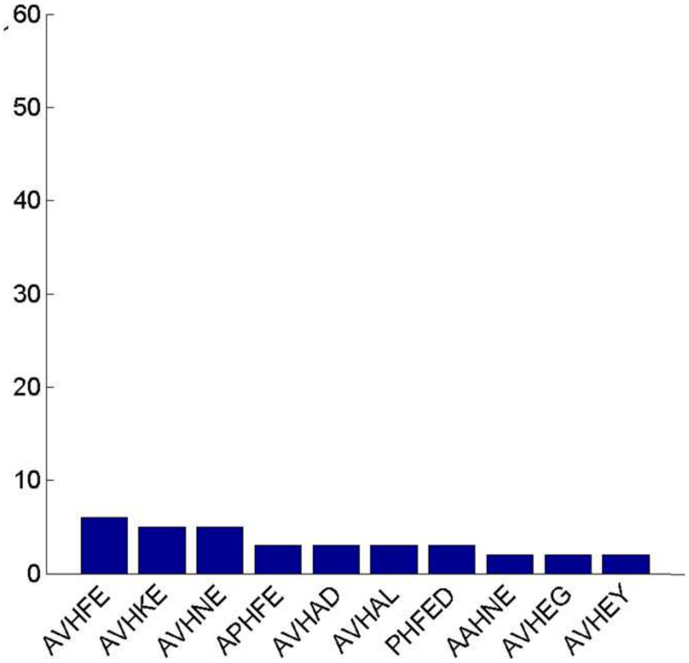


Figure C.2: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to the Dengue Disease Sample. The Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

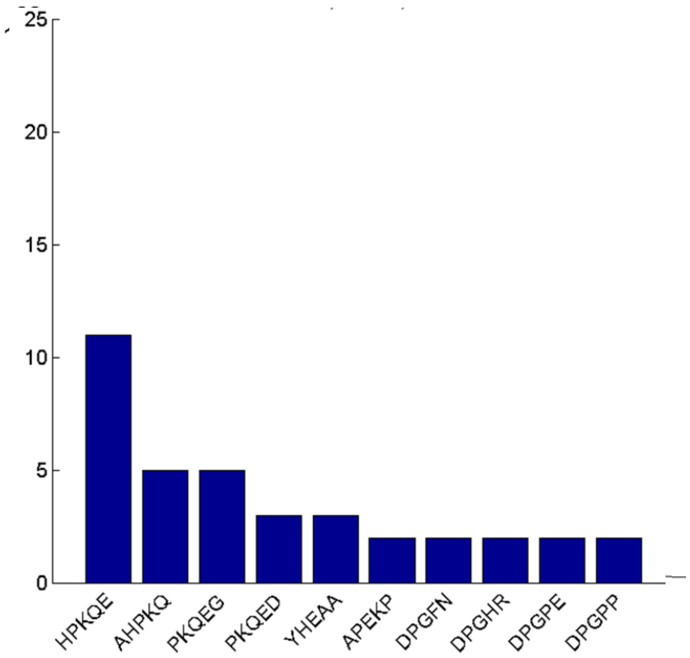


Figure C.3: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to the West Nile Virus Disease Sample. The Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.

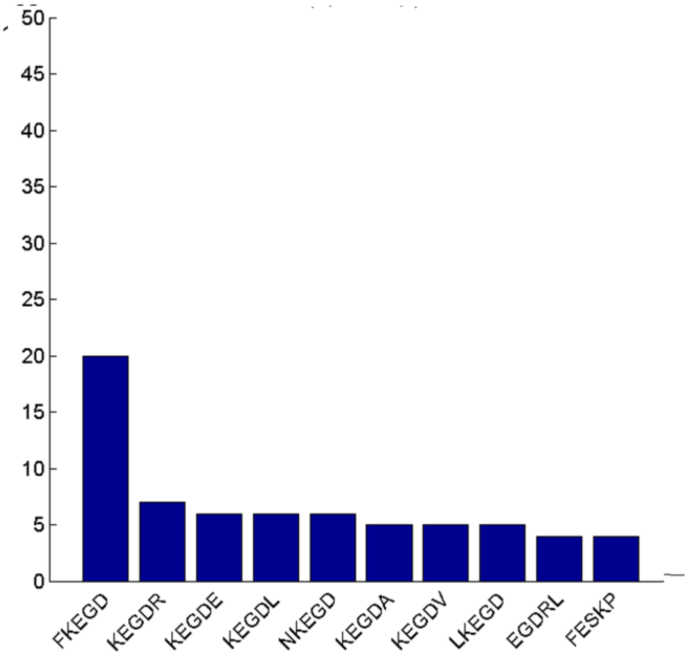


Figure C.4: Estimated Top 10 Subsequences of Length $\mathcal{L} = 5$ Obtained When Algorithm 1 Was Applied to the Bordetella Disease Sample. The Estimated Subsequences Are Sorted in Descending Order According to the Values of Their Corresponding DS-OCRC.