1-1-2013

Computational Approaches To Anti-Toxin Therapies And Biomarker Identification

Rebecca Jane Swett
Wayne State University,

COMPUTATIONAL APPROACHES TO ANTI-TOXIN THERAPIES AND BIOMARKER IDENTIFICATION

by

REBECCA JANE SWETT

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2013

MAJOR: CHEMISTRY

Approved by:

_____

Advisor Date

_____

Co-Advisor Date

_____

_____

_____

_____

DEDICATION

To all the friends, family and loved ones who stood by me and supported me through everything. Thank you.

ACKNOWLEDGEMENTS

There have been so many important people that have helped and guided me through this incredible experience. I would not have made it to this point without the friendship, love, and support you all shared.

To my advisors Andrew and Andrés, thank you so much for all your help. I am forever grateful that I ended up with the best possible combination of mentors. I consider myself incredibly lucky to have been able to work with both of you.

Andrew, thanks for not booting me from the lab after the repeated disasters in the cell room. I hope I made up for those first few months. Thank you for consistently pushing me to work hard while still managing to give me a lot of freedom with my projects. You always encouraged me to be creative with my research; and it really paid off. I learned a lot more than I would have if I would have just rolled along using available techniques.

Andrés, I am so glad you let me join your lab as well. I really appreciate you giving me a chance, your insight and guidance made a world of difference with my work. Thank you for challenging me, getting me out of my comfort zone, and giving me a chance to really take a shot at creating something completely new. Our work really broadened the way I think about research.

Along the way a lot of people have encouraged me in science. Thanks to Mr. Colassaco at WHS who was my first introduction to chemistry and Dr. Wickenheiser at NMU who was so helpful when I was an undergrad. Most of all Dr. Mark Paulsen, my advisor through my bachelors and masters studies. Thank you for turning me on to computational chemistry; you changed the course of my future.

Thanks to all my family for supporting my choice to become a scientist. I love you guys. I promise next Christmas I won't say "I'm still in school" when you ask what I'm doing. Thank you Mom for being amazing. I couldn't have asked for a more loving, supporting and encouraging parent. You were my first and most important teacher. Dad, I know I'm a bit different than the rest of the kids, thanks for accepting that I like computers instead of hunting gear. Jeff, thank you for being there for me while you were able. I know being with me through this was hard, you made it easier. Marlyss, I know you're gone, but thank you as well. You taught me far more than you'll ever know.

For my friends, both old and new, thanks for being there when I needed to relax or let off some steam. Everybody in Marquette who always made themselves available when I visited, you're awesome. For all my grad student friends at WSU, I'm glad we all met. We really broke the stereotype of awkward introverted geeks with our karaoke parties! Thanks specifically to Adam and Angelo. Adam, even though most of the time you were a box on a screen, you were always there when I needed to talk. Angelo, you're amazing. Thanks for always being there with a coffee delivery when I was working too late and being willing to put up with editing papers over a couple beers.

Thank you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CDAD
Clostridium difficile associated diseases

CPD
Cysteine protease domain

CROP
Clostridial Repetitive Oligopeptide

FEP
Free Energy Perturbation

GMD
Generalized Masked Delaunay

GPU
Graphics Processing Unit

GT
Glucosyltransferase

HDAC
Histone Deacetylase

MCCE
multi-conformer continuum electrostatics

MD
Molecular Dynamics

MM/GBSA
Molecular Mechanics/Generalized born solvent approximation

NMR
Nuclear Magentic Resonance

P1
Peptide EGWHAHT

P2
Peptide HQSPWHH

PCA
principle component analysis

POLL
DNA Polymerase Lambda

QM/MM
Quantum Mechanics/Molecular Mechanics

QSAR
Quantitative structure-activity relationships

RMSF
root mean square fluctuation

TcdA

Clostridium difficile Toxin A

UDP
Uridine diphosphate

Chapter 1 Computational techniques at every level of biochemical discovery

1.1 Introduction

As modern biochemistry, computational chemistry, and computing power advance, there are ever expanding opportunities for computational methods to be developed and applied to the process of discovery. The rapid development of technology has changed our lives in unprecedented ways. More computing power is available in the average cell phone than was used to put a man on the moon in 1969 (2). We are able to store and search the whole human genome on an iPod (3). When combined with bright and creative minds, the ability to perform complex calculations has been truly transformative for science. In biochemical discovery computational theory is being applied everywhere. Quantum chemistry tools can accurately model the electronic structure of a ligand molecule, and bioinformaticians can compare hundreds of complete human genomes to look for the roots of genetic disease. In terms of molecular biology, the applications range from target identification and initial druggability testing, to final structural modifications on potential small molecule drugs and fundamental studies on the nature of biomolecules. The concept of a "rationally designed drug", one that has been created through a deep understanding of both the receptor structure and the salient small molecule interactions was an entirely new concept when it was suggested in 1986 (4). In 2012 alone there were more than 40 publications on successful rational drug design.

Historically the development of pharmaceuticals was based on "whatever works", rather than understanding how a bioactive substance acts. In the long years since herbal remedies were simply put into pill form, we have gained the ability to capably understand the "why"

associated with effective treatments. As a scientific community are now at the stage where we can look at a biological system as a whole, from genome to phenotype. This has led to a much broader perspective for examining a biological system.

The scope of our biological understanding has expanded, and so has the associated data. The rapid expansion of "omics" fields threatened to drown researchers under a deluge of "big data". It seemed initially that our ability to generate data had outpaced our ability to extract meaningful information from it. This situation drove the development of a whole range of computational and statistical analytics that up until recently, would have been unnecessary. Not only has the creative application of statistical modeling and data deconvolution necessary, it's illuminating. For the first time, we are able to not only think about, but model whole interaction networks and observe how they react when perturbed. The exponentially growing field of systems biology is a testament to the value of combining traditional molecular biology with methods originally developed for computer science.

Within the context of probing a disease for either cause or cure, computational chemistry has become invaluable at each stage of discovery. Genomic and bioinformatic investigations can assist in finding causal genes or potential drug targets. Applications of systems biology can determine how those genes or drug targets are interacting within a broader system. Network analysis is shifting the way we think about drug design. Should we hit one target hard? Or hit many associated targets weakly? Will off target hits render this a dead end for treatment? Structural biology literally allows us to look at the proteins we are working with. Rather than a line on a gel, we can hold up a three-dimensional structure and get a good look at what makes it tick. Once we know how it ticks, rational drug design techniques can allow us to build a molecule that stops the clock.

From start to finish, computational chemistry has changed the way we function as scientists and will continue to do so in the future. This work will focus on current applications of computational chemistry to biochemical discovery. The techniques discussed will include novel bioinformatics applied to both human and bacterial genomes, structural and dynamic analyses applied to both fundamental understanding of biomolecules and drug discovery, and small molecule development methods. The following chapters will illustrate productive application of these techniques to the design of inhibitors for *Clostridium difficile* toxin, and the development of bioinformatic software for both human and bacterial genomic data.

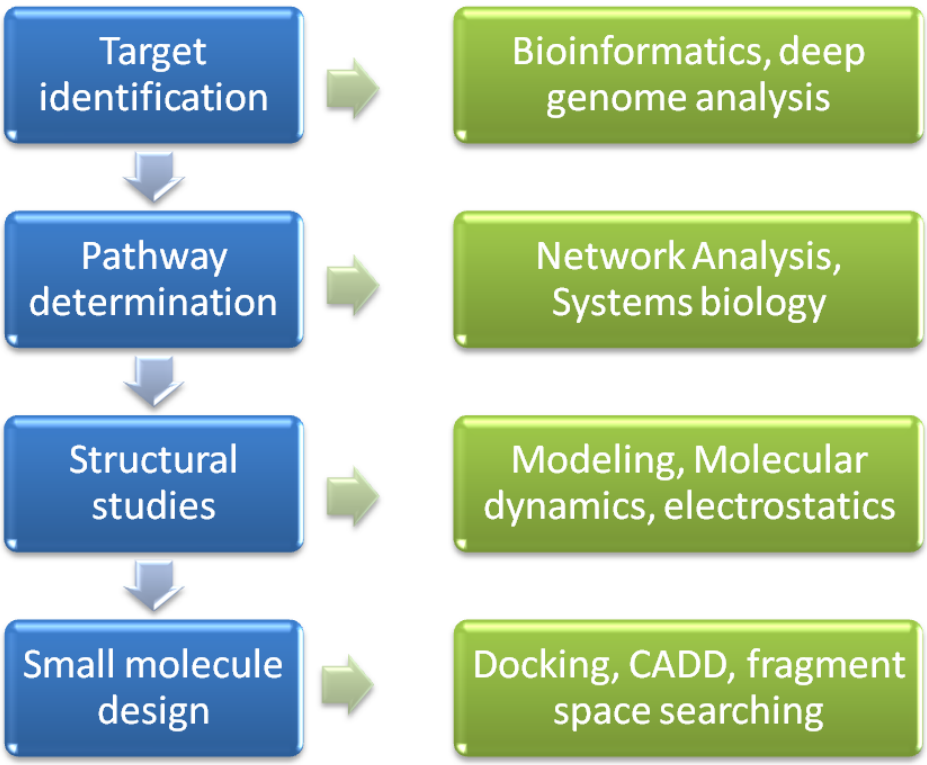


Figure 0-1 Opportunities to apply computational methods to biochemical discovery

The workflow here describes one possible pathway to biochemical investigation or drug design. Initially, a gene or protein of interest is identified as a target for further study. Biochemical interaction pathways are then pursued, traditionally by methods of molecular biology. Structural characterization is carried out where possible to deepen the understanding of the system, followed by small molecule design either for the purposes of drug development, or for developing biochemical tools to interrogate the system. At each step, computational intervention can assist the process and provide otherwise intractable information.

1.2 Target identification

Traditional methods of identifying targets for biochemical investigation involve direct biochemical methods, where a gene, RNA or protein is identified, purified, characterized and then further studies are performed based on the goal of the project. These methods often require a heroic amount of work, along with sophisticated technologies and years of training for skilled scientists in the field. With the advancement of high-throughput methods such as DNA microarrays and RNA sequencing data collection has greatly accelerated generating an overwhelming mountain of data. It's been well documented that the rate at which publically available sequence data has been expanding has far outpaced our ability to handle that data manually(5, 6). Making sense out of a massive dataset has required complex computational analysis techniques, as significance cannot be assessed by simple examination. Bioinformatics has exploded as a field, and its popularity is well warranted. By combining mathematical and statistical analysis with big "omics" data, not only are significant associations detected, but patterns of associations, and the patterns are where it gets interesting. The following sections will discuss applications of bioinformatics to genomic level data.

1.2.1 Genome-wide association studies

Following the completion of the Human Genome Project, the technologies for whole genome sequencing developed extremely rapidly. This has led to the practice of investigating diseases thought to have a genetic link with Genome Wide Association Studies (GWAS). In these experiments, large numbers of individuals with a specific disease are subjected to whole-genome genotyping, and the results are deposited in massive databases. While this work has led to a number of striking discoveries, most well known of which is the BRCA1 and BRCA2 breast cancer genes(7-9), many scientists have been disappointed in the low productivity of the project(10). The reigning paradigm at the time was that introduction of GWAS studies would

result in the deconvolution of many genetic diseases: one gene, one mutation, one disease- termed the "common-variant hypothesis". Simple statistical methods were employed to determine strongly predictive mutations in diseased individuals in the hope that this would result in a causal relationship between mutation and disease(11). Unfortunately, this paradigm became the exception rather than the rule. Rather than revealing simple links between genome and disease, frequently a huge number of associations were detected between diseases and genomes, or weak clinical correlation between strong statistical predictors and disease phenotypes(12). Manolio et al. in 2009 was among the first to openly discuss GWAS studies as not meeting the expectations of the scientists involved and suggest that the "disease-common variant" hypothesis was incorrect. The interplay between numerous mutations may be more biologically relevant, something that would not be detected by trying to apply extremely stringent statistical methods to find a single direct connection. Others suggested that the worth of GWAS studies lies in the detection of rare variants (13), that statistical analysis of GWAS are confounded by rare variants (14), or that the combination of several alleles into a haplotype may be a more accurate predictor (15, 16).

While many people were disheartened at the failure of the common-variant hypothesis, this development did spur investigations into pathway associations and combinations of low-significance mutations. The association of multiple loci with disease phenotypes has resulted in a reimagining of the function of GWAS studies, that rather than seeking a single mutation, GWAS experiments have the capability to expose novel biological pathways. Notably, the linkage between Crohn's disease and both the autophagy and interleukin-23 related pathways was completely unexpected (17). The illumination of unexpected biological pathways relevant to disease has resulted in an explosion in the realm of "big data analysis" applied to GWAS data. While the deposition of data still proceeds at a much faster rate than the growth of processing

power, novel analysis and data storage methods are making an attempt to keep pace with this ever expanding wealth of data(5).

The complexity of genetically linked diseases has spurred creativity in the computational algorithms used for analysis, far surpassing the simple statistical methods that were initially applied to the GWAS data. To date, over 35 major software suites have been developed specifically for the handling and analysis of GWAS data (5), ranging from methods for genome assembly (18-24), to "big data" storage and query methods(25-27), to novel data mining techniques(28-34). Furthermore, there has been some considerable success in identifying potentially "druggable" targets within the human genome using bioinformatics techniques (35-40). With the development of novel computational techniques, and the continued re-examining of GWAS data(41-45), there is much more potential for a deep understanding of genetically linked diseases as well as a future in genome-to-pharmacy drug design.

1.2.2 Data mining in single genome studies

While the term "bioinformatics" has been around since the 1970's(46), a turning point came in 1984 when a technique had only previously been applied to mathematical data, was turned on a biological system. A method was developed to align sequences from the Dayhoff Atlas of Protein Sequence and Structure, and very rapidly thereafter bioinformatics became associated with genetic and proteomic data(47). Largely, modern computational techniques in data mining are applied to organisms for which whole genome sequence data is available. Rather than analyzing a cohort of genomes for variation, these techniques focus on pattern determination and deep understanding of the function and interaction of the genome of a single organism. Frequently programs are designed to find patterns or comparisons in the genomes that would be undetectable without assistance. Bioinformatic software at this level ranges from the well established sequence alignment tools available such as ClustalX, BLAST, and TCoffee (48-50), to highly complex pattern and motif searching tools like PairMotif and DISCOVER (51-

55). A comprehensive overview of the breadth of methods available is beyond the scope of this document, but the categories of genomic analysis to which computational methods have been critical will be discussed.

While the utility of these kinds of bioinformatics tools varies widely, we will focus on application of two general categories of "small data" investigation to antimicrobials. Finding genes that code for bacterial pathogens or that may be exploitable with new antibiotics comprises a non-human form of "Target Identification".

Target identification within bacterial genomes can generally be divided into two fields: sequence-based, or composition-based. Sequence based methods rely on comparison of multiple bacterial sequences(56-59), while composition-based methods require only a single genome sequence (60-65). Recent application of genomic interrogation techniques has been quite successful in the study of antimicrobials.

One such success story describes the identification of 214 unique enzymes in *Pseudomonas aeruginosa*. By using comparative genomics based on information from the Kyoto encyclopedia of Genes and Genomes (KEGG) Perumal et al. were able to identify essential genes unique to *Pseudomonas aeruginosa.* A case study was pursued using a homology model of one of the identified proteins, targeted with *in silico* docking of several inhibitors. Results indicated that this protein would be a suitable target for antibiotic development and further *in vitro* studies.

Study of genomic islands has become quite popular in trying to understand the underlying mechanisms of antibiotic resistance and pathogenicity transfer. Genomic islands are clusters of genes, acquired by horizontal gene transfer, which frequently confer some evolutionary advantage and are frequently associated with pathogenesis or antibiotic resistance. While many of the studies focus on determining druggable targets in pathogenic organisms(66,

67), some focus on organisms with inherent antimicrobial activities. (68-74). Originally these mobile genetic elements were analyzed using gene cloning techniques(75), but with the availability of genomic data on many organisms several computational techniques have expedited the process(65, 76, 77). The goal of many of these studies is to identify genomic islands associated with pathogenicity, in hopes of identifying possible targets for future anti-microbial work.

A comparative analysis of three strains of antibiotic resistant *Acinetobacter baumanii* is one such application. Once fully sequenced, the genomes were examined using comparative genomics analysis and individual genomic island detection with IslandViewer (56, 57). Work by Zhu et al (78)revealed that while the three strains shared drug-resistant genomic elements, the context in which these elements were present was quite different. In the most genetically divergent strain, antibiotic resistance was conferred by a large resistance island, containing multiple drug-resistance genes and several transposons. The analysis indicated that one challenge that will be encountered by researchers targeting *Acinetobacter baumanii,* is the genomic flexibility of this organism.

Motif searching is also a broadly applied term in both bacterial genomics and proteomics. The ability to quickly identify genetic regions characteristic for protein-binding(79, 80), sRNA-controlled regulation(80), subcellular localization(81), determination of sequence families(82, 83) as well as many other genome-based inquiries has a wide range of applications (83-89). However, computational identification of motifs remains challenging and recent surveys of the field indicate that there is no "best" method (90, 91). Whether informed by experiment or built on probabilistic modeling and neural networks, there has been no clear success in the field. Perhaps the best example of fully utilizing the various methods of motif recognition is the work done by Hu and Rajasekaran (92, 93) in developing a consensus ensemble algorithm. This

method combines the results from multiple external motif searches, and results in a 6-45% improvement in accurate motif identification.

As completing a full genome for an organism has become increasingly fast and cost-effective, the tools facilitating thorough exploration have proliferated. In addition to the bioinformatic methods discussed above, there is a wide range of bioinformatic techniques that can be applied to the transcriptome and proteome. In brief, the development of computational tools for single-genome analysis has been both expansive and creative, facilitating the discovery of numerous genomic features that can be exploited for either fundamental learning, or antimicrobial development.

1.3 Pathway and Network analysis

As was discussed in section 1.2.1, the richness of GWAS studies lies not in directly connecting a single gene mutation to a disease, but with illuminating networks of interactions whose cumulative effect may be more salient. Recent work in graph theory and "big data" analysis has led to the growth in the field of network theory, a subset of graph theory, which seeks to represent complex sets of interactions in tractable ways. Following the construction of a network of various components, data mining and mathematical sub-analysis can be applied. While it's easy to think of networks in terms of genes or protein-regulation, network theory is being applied far more broadly in computational chemistry. Network theory has been applied to cheminformatics, drug design, repurposing of drugs, and target identification.

In order to mentally parse and bring coherent understanding to the sweeping datasets capable of being generated by both next-generation lab techniques and those generated computationally, a human-readable expression of the data is necessary. While visualization packages are rapidly developing, it still requires human intuition and interpretation to recognize

some of the more biologically relevant patterns. In combination with a researchers deep fundamental understanding of their research focus, network analysis can provide a fascinating way of stepping outside the system and looking at a biological problem in a broader context.

1.3.1 Network theory in visualization

Just as novel approaches have changed the way we approach the data that we're aggregating, so to do new or innovative visualization methods change the way we literally view the data. Representing biological pathways as complicated networks in textbooks has very nearly reached trope status. Complex, difficult-to-interpret diagrams of cellular interaction pathways are in nearly all biochemical textbooks. While this is technically an application of a network representation, there is distinct room for improvement. Visually, there are huge strides being made towards visual deconvolution of networks. Uninformative hairballs have been the norm for genetic network interpretation, which rather than aiding in interpretation, can convolute it.

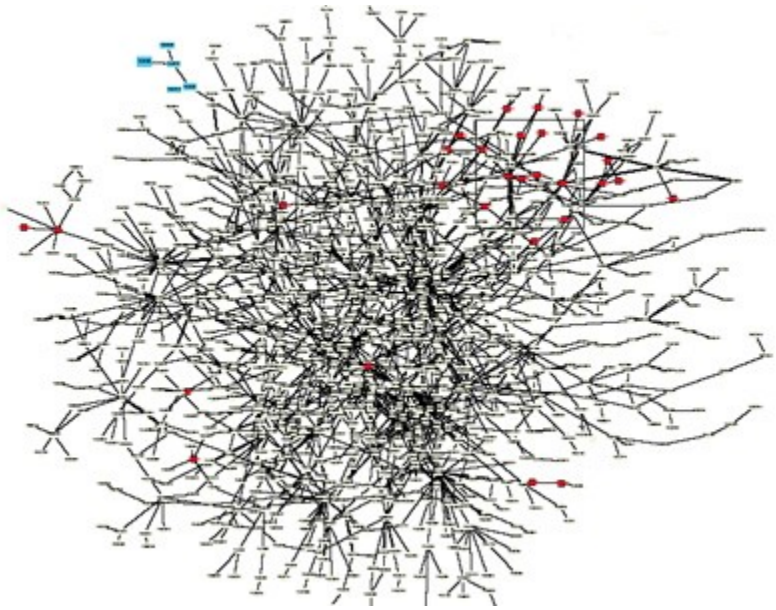


Figure 0-2 Yeast interaction network hairball

A visually uninformative, but typical interaction hairball reproduced from www.visualcomplexity.com

A thorough survey of modern network analysis algorithms has been published by Csermely et al.(94, 95) which includes an overview of available visualization and analysis programs as well as suggested methods for use. Appropriate use of network analysis must necessarily take into account the viewer, and the viewers paradigm for data interpretation. Effective application of network analysis can both accommodate and add to pre-existing paradigms for biological representation. We are familiar with the "alphabet soup" interaction diagrams in biochemistry textbooks. Chromosomes and sets of genes are represented as neatly organized straight lines. Plasmids are perfect circles with cleanly annotated features. Well heeled scientists have these visual organizational structures deeply embedded in how we think about our science and how we communicate effectively with others. The addition of network analysis on top of those pre-existing templates can help us clearly understand the complex system of connections and interactions that can be layered on top of that fundamental understanding. As the field of systems biology develops, we will likely become open to novel visual organizational structures, but as the "hairball" above illustrates, focus on understanding is key.

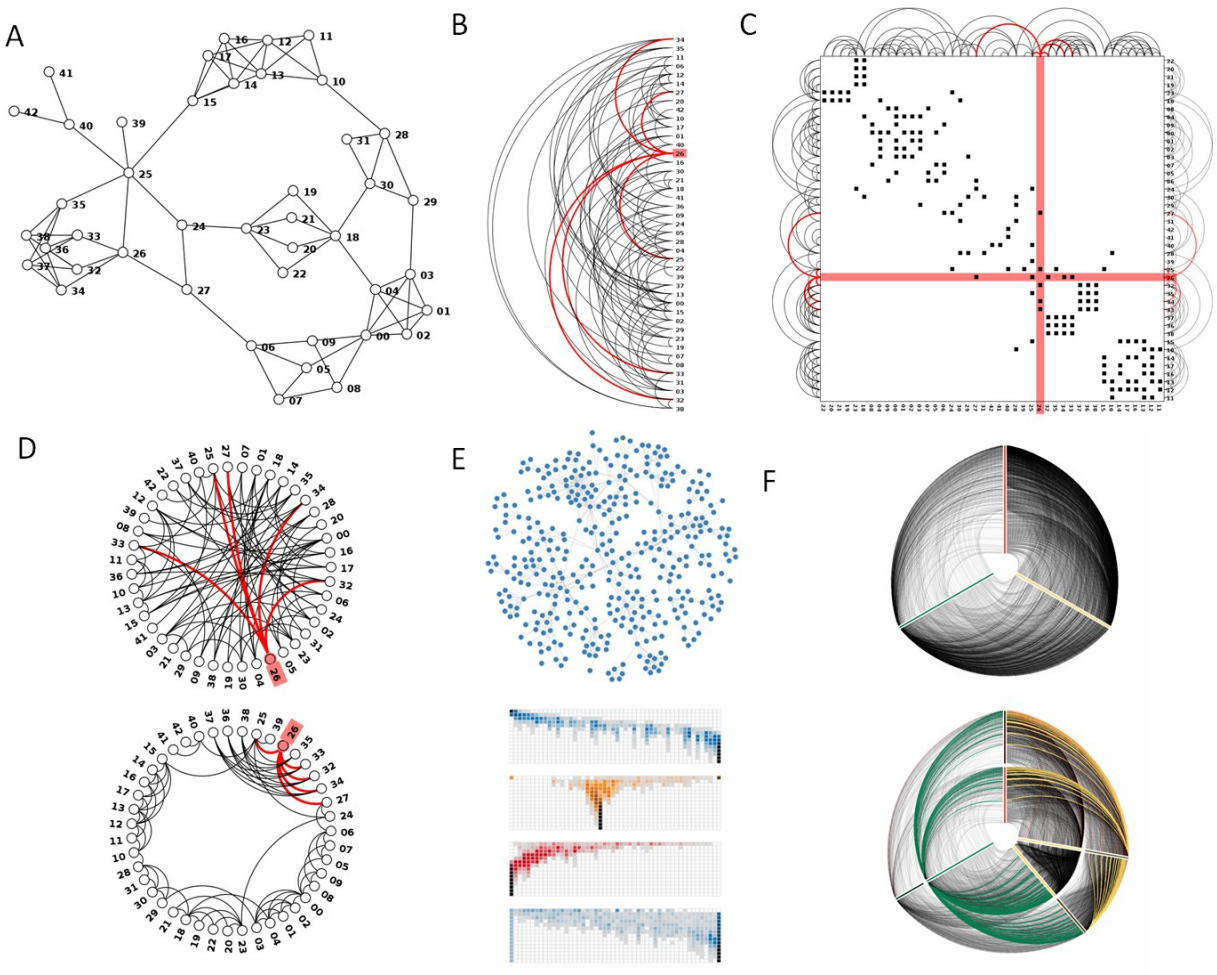


Figure 0-3 Various forms of network visualization

Example plots of each of several types of network visualization. A simple force directed graph is shown in (A), these are most similar to biological network schemes. An arc plot, useful for depicting relationships on gene regions or chromosomes is shown in (B). An adjacency matrix plot combined with an arc plot is shown in (C), these are frequently useful for connecting data that is spatially or temporally distant. Panel D shows a radial/circular plot commonly used in depicting bacterial genome interactions. A network representation of data interpreted by an attribute driven heatmapping method is shown in E. This method is excellent for decomposing complex sets of unrelated interactions. Panel F shows the relatively new "hive plots", where nodes are organized by characteristic leading to a dramatic deconvolution of hairy networks. Data reproduced from references 78-94.

Numerous algorithms have been developed to assist in the visual reinterpretation of networks, and they can be decomposed into several categories: Force-directed(96-98), arc-type(99-102), adjacency matrix (also known as correlation plots)(103-107), circular layouts(108-111), attribute driven and hive plots(95, 112). While each of these methods has applications in many fields, all of them can be useful in the analysis of biological data. Figure Figure 0-3 depicts some of the more common types of network layouts

In thinking about how to best represent data, it is critical to place emphasis what the viewer should take away. For example, with a network of interacting proteins, understanding which sets interact closely, and which are only peripherally involved in a processes can help us understand what role each protein plays. For this application, a force directed graph would be useful. Force directed graphs use mathematically applied attraction and repulsion to distribute nodes in a network. This visually recreates many of the standard characteristics of a biological interaction scheme and can illustrate which nodes are highly connected and integral to the system.

Chromosomes and gene loci are commonly represented as linear. Information on regulation or common protein-binding locations could be layered onto linear plots using an arc method without occluding the original data. By literally connecting the dots, it becomes visually tractable to determine if there are "hot spots" of activity on a given chromosome, or if a DNA binding protein is interacting with a specific family of proteins. Circular networks perform similarly for plasmid information, but have other applications as well. Circos (113, 114) released in 2009, has gained popularity as a way to visualize relationships in genomic data. Adjacency matrices are used to describe relationships between spatially or temporally distant data, in molecular dynamics these are used as "correlation plots". Representing data as a matrix has a range of advantages, including quick visual assessment of symmetry. One recent paper in Nature applied both circular and adjacency matrix plotting to yeast genomic inter-chromosomal interactions(115). By reducing the visual complexity of the interactions to colored diagrams, it becomes simple to quickly find the "take away" of the data.

There are times when the density of a dataset becomes intractable, even using appropriate network visualization techniques. With the generation of both ever larger, and higher dimensionality data sets on gene expression, regulation, and various other properties, it

becomes necessary to step away from traditional visualizations and move towards methods that may be able to handle such complex data.

Newer methods of attribute driven binning can reduce the complexity of a network to a heatplot, representing a single network characteristic. The recently available GraphPrism (116) software decomposes networks into sub-plots relevant to a single node or characteristic. This deconvolutes extremely dense data into visually interpretable pieces. Each heatmap is a statistical summary of the interactions of a node or metric. As this technique is extremely new, there are no published scientific articles that have effectively used this technique, though one could imagine its utility when applied to variable expression data, or large scale regulatory networks.

Hive plots are gaining traction for numerous applications, as these allow the grouping of nodes into axes by user-defined properties (112). Developed by Martin Krzwinski specifically for the interpretation of systems biology information, this allows the creator of the plot to decide which groupings of data points are most relevant when viewed together and reduces the visual clutter of networks. Typically color is required to bring out the salient characteristics of these plots, but they are ideal for displaying relationships between families of related nodes.

Regardless of which type of visualization algorithm is applied, appropriate use facilitates the recognition of patterns and connections in the data that would otherwise be hidden. As the size and scale of biological data grows, so too does our understanding of the interconnected nature of it all. Unfortunately, the working memory of the human mind is not linearly increasing as well. Effective graphical representations are necessary to assist in the comprehension and communication of the salient information an author wishes to express.

1.3.2 Network theory applied to genomics

Network analysis is frequently applied to large-scale genomic data to avoid deferring to reductionist tactics. The proposal of genome-wide regulatory networks as an alternative to the "disease-common variant" interpretation of GWAS data relies heavily on the treatment of genome level data with network theory and an analysis (117). Interrogation of regulatory networks rather than single genes has been illuminating in several fields, as discussed by Arujaho et al. (118). He suggested that as our ability to collect broad information on cellular function and subcellular interaction increases, we should increasingly implement that knowledge in our biochemical investigations. For example, high throughput methods allow us to identify genes that are overexpressed in tumors. Targeting those genes without considering the greater context of the genomic system may lead researchers down a dead end path. The detected associations may be peripheral, reactionary, or have a compensatory backup system in the cell. When using large-scale data for target identification, it is absolutely essential to consider the complexity of the system; network analysis is ideal as a solution.

In the application of network modeling to the human genome project, it has been suggested that the inclusion of genomic analysis methods may increase the amount of potential human drug targets by up to fourfold (119-121) depending on the method of estimation. While no new druggable protein families have been detected, bioinformatics methods have been applied to search for new members of those families. It should be made clear however that identifying a druggable protein does not necessarily result in a drug target, as many proteins, which can bind small molecules are not currently disease associated(122, 123).

Statistical analysis of genomic data is useful, however, novel insights have been obtained using network visualization. The concept of "bottleneck, hub and peripheral" nodes is

one entirely resultant from the interrogation of interaction networks of both genes and proteins, and provides an interesting perspective (124, 125). Hub nodes have numerous connections to other nodes, and typically manifest as a "starburst" when graphed. Peripheral nodes have a low number of connections, and may be only secondarily connected to a hub. Bottleneck nodes connect hub nodes, and usually have no redundant connections, i.e. information that needs to travel from one hub to another may pass through a bottleneck node if there are no other routes of communication available. Analysis of several known drug-target inclusive networks indicate that hub nodes are generally not targetable as they contribute significantly to the stability of the overall network(125). Many drug targets have turned out to be peripheral nodes, as removal of these nodes is less likely to destabilize the entire network system. Similarly to hub nodes, bottleneck nodes are unlikely to be good drug targets, and frequently knockouts of these genes are lethal as the removal of a bottleneck node destabilizes the overall network(124).

As network analysis increases our understanding of genomic and protein interactions, this opens the door to reimagining drug design. Excellent work by Agoston et al. indicated that weak perturbation of several nodes in a transcriptional regulatory network was more efficacious in altering the system than singling out a particular node (126). The transcriptional regulatory networks of *E. coli* and *S. cerevisiae* were modeled as weighted networks, and then subjected to a series of perturbations. Each node was "attacked" in several ways; complete knockout, partial removal of interactions, or all interactions were attenuated by decreasing the weight of the interaction. These "attacks" mirror the biological phenomena of strong inhibitors, weak inhibitors and broad-spectrum inhibitors. Combinatorial simulation of the networks was carried out to determine what combination of effects maximally perturbed the system. Surprisingly, weakly targeting multiple nodes in the network was more able to shut down the system than the complete removal of a node. This fundamental work on a simulated model system echoes failed "common-variant" hypothesis from GWAS work. The general consensus of the GWAS literature

indicates that multiple mutations or factors are more likely to be causal than a single mutation. Looking forward, Agoston's work suggests that with respect to pharmacology, the design of multiple ligands may be more efficacious than striving towards ever more potent inhibitors. The concept of polypharmacology has caught like wildfire and has caused a push towards designing multiple ligands capable of weakly interacting with many target structures.

1.3.3 Network theory applied to drug development

While no completely novel drug families have been found using network analysis, statistical network modeling has made some surprising discoveries. Frequently the overarching concept in drug design is to find one target, and bind that target as tightly as possible. Many branches of medicinal chemistry are devoted to substrate optimization and understanding ligand binding affinity. However, following the initial work by Agoston et al. mathematical modeling of perturbed interaction networks has repeatedly stated that there may be significant benefit in approaching multi-target drugs (127-129). Modest interference in an interaction network at several points has been shown to result in a more significant overall downstream effect(128, 130). So far this method has been successfully applied to the discovery of ligand-gated ion channels as potential drug targets(131), as well as novel methods of targeting kinases in anti-cancer treatments(132). This has led to the application of network theory to the field of drug development.

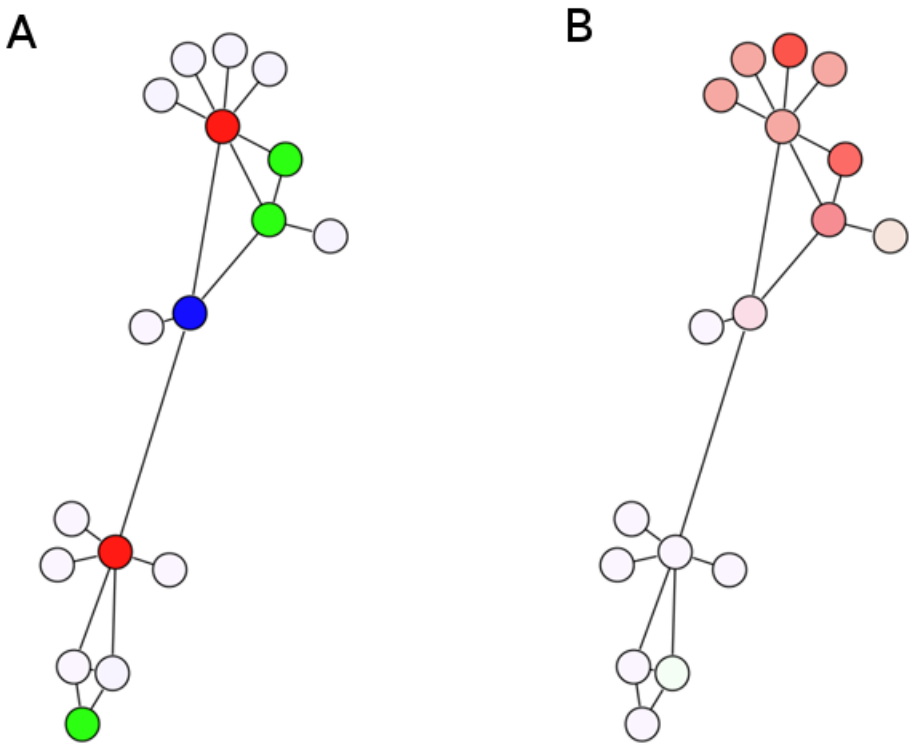


Figure 0-4 Example network illustrating standard and multiple ligand pharmacology

Both panels show the same randomly generated interaction network. We see two typical small regulatory networks connected by a bottleneck. Panel A illustrates the one drug, one target paradigm. Panel B illustrates the concept of multiple ligands.

These early successes in polypharmacology led to the pursuit of "designed multiple ligands"(133-135). Figure 0-4 describes the concepts behind both traditional drug design and polypharmacology. Both panels show the same randomly generated interaction network. We see two typical small sub-regulatory networks connected by a bottleneck. Panel A illustrates the one drug, one target paradigm, with several targets highlighted for inhibition. Shown in red are hub nodes, which would destabilize separate sections of the network. In blue is a bottleneck node, inhibition of which would decouple the two, but not alter their independent behavior. In green are peripheral nodes, which would affect their respective small networks without completely shutting them down. Panel B illustrates the concept of multiple ligands. If a substrate weakly targets numerous nodes in a regulatory network this may provide more complete

inhibition than strongly targeting a single node. Small molecules that affect multiple cellular targets within a single cellular interaction network have shown some promise in a range of diseases including cancer, HIV, neurodegenerative disorders and obesity (136-139). Many drugs whose mechanism of action was not known during their discovery and trial phases have been determined to exploit multiple targets within a system(140).

1.3.3 Integration of genomics, network analysis and pharmacology

In the discovery workflow described in Figure 0-1, network analysis is shown as being connected to pathway determination of known targets. While this is indeed a valid application for network analysis, it also provides an excellent bridge between the genomic data involved in target validation and the downstream steps involved in small-molecule design. Integration of genomic level data with interaction networks and current ligand databases have been used to "repurpose" a number of small molecule drugs. Several successes have been published connecting genome level data to disease treatment through network analysis(141, 142). Perhaps one of the most thorough is Sanseau et al. applying GWAS level information to drug repositioning. In this study, disease-associated SNPs were analyzed and through several iterations of database mining were linked to various disease therapies. This resulted in several small molecules being suggested for alternative usage in diseases such as Crohn's and smoking cessation(143).

Regardless of the application, integrating genome level data, network analysis, and pharmacology is promising for future work from target identification to pathway analysis. Once potential targets have been found and the systems they participate in elucidated, further studies are warranted. Whether the goal is novel small molecule design, or fundamental understanding, structural information is an irreplaceable contributor.

1.4 Structural analysis

Computational chemistry has long been applied to the structural analysis of biomolecules. The initial techniques of X-ray crystallography were applied to large biomolecules in the early 1950's, and have steadily progressed since then. Use of 3D structural information can greatly assist in the interpretation of *in vitro* experimental results, and frequently obtaining a structure quickly becomes a goal when an interesting target or system is discovered.

Static structures themselves can be informative, as much information can be gleaned from thorough examination of active sites and known substrate binding regions. However, queries into the nature of the biomolecular surface and the dynamic properties of the molecule are also quite valuable. Wet-lab technologies are making great strides towards more real-world structural data, with the development of single molecule techniques(144-147), atomic force microscopy(148-151), and the combination of individual-particle electron tomography and focused electron tomography reconstruction (152) in addition to traditional NMR and X-ray techniques. Structural information is a fantastic resource, but observing biomolecules in a cellular setting is still far beyond our current capabilities. Until those experiments are possible, computational techniques can attempt to fill the gap.

Given a static structure, surface characteristics and electrostatic analyses provide context for some of the behaviors of the molecules, both alone and in contact with protein partners and can be used to guide both protein and ligand design. Where available, multiple crystal structures can provide snapshots of the motions of a protein, but dynamic calculations such as normal mode, molecular dynamics, and advanced sampling techniques are required to turn the stills into movies, so to speak. The following sections will address some of the computational techniques that assist in structural investigation of a biomolecule once it has been assessed as a potential target for study.

1.4.1 Static analyses

The importance of static structural characteristics in structural analysis should not be understated. Innumerable mutagenesis studies have shown that altering just a single residue in an active site can abolish activity. Similarly substitution on the surface of a protein can dramatically alter protein-protein interaction. When the alteration is not dramatic enough to perturb the structural dynamics of the protein, the effects of substitution can be attributed to either catalytic relevance, or surface effects. Electrostatics are critical to binding events between proteins(153), as well as with substrates(154). Polarizability and electron movement have a distinct role in the behavior of proteins at the catalytic level, but assessment of all quantum computational techniques applied to biomolecules is outside the scope of this document. With respect to the workflow discussed in Figure 0-1, the interjection of computational chemistry to the analysis of biomolecular structures can aid in fundamental understanding, engineering of proteins, and ligand design. The surface property most commonly discussed is largely electrostatics(154-164), while hydrophobicity(165-168), ligand binding pocket analysis(169-171), and protein-protein interaction hotspot detection(172-177) are all also popular.

Electrostatic calculations applied to proteins suffer from a few assumptions that must be made in order to facilitate the calculations. As always, computational cost and accuracy are inversely related. The more accurate your calculations are, the more computationally demanding they become. There have been attempts to use combined QM/MM and full quantum mechanical calculations to obtain highly accurate electrostatic potentials of small proteins, however this is not the norm. By and large, most electrostatic calculations are done under continuum rather than explicit solvent, as solvation greatly increases the computational cost of the calculations. Some excellent comparisons have been made by Tan, Lee and Godschalk with respect to the differences in explicit and implicit electrostatic calculations (178-180) and will

not be discussed here, rather we will focus on the information that can be discerned from application of electrostatic calculations in implicit solvent.

Within the context of continuum electrostatics, there are two primary techniques: solution of the Poisson-Boltzmann equation for the system, or application of the generalized Born approach. A broad array of programs exists to facilitate these calculations, and several overviews are available in the recent literature (162, 181-183) . Two of the most common are DelPhi and APBS, the adaptive Poisson-Boltzmann solver(184, 185). While the ability to obtain an electrostatic representation of a protein is in itself a valuable thing, electrostatic surfaces can be used in a multitude of ways to aid in the analysis and prediction of structural properties. Electrostatic surface calculations can be applied to the detection of protein-protein binding(153, 157, 158), ligand interactions(154), or the predictions of ligand binding pockets, among many others.

Experimentalists are able to find protein-protein binding partners through various techniques such as pull down assays, but frequently investigating the protein-protein binding interface is difficult without extensive mutagenesis studies. Electrostatic calculations have been used to guide protein-protein docking(158, 186-189), protein-protein interface detection(177, 190-192), and have been determined to be strongly involved in protein-localization(193, 194). With respect to fundamental studies, analysis of protein-protein interfaces and docked conformations can assist in directing mutagenesis studies(195, 196) and has made great contributions in the field of designed proteins(156, 197).

1.4.2 Dynamic properties from Normal mode to molecular dynamics

When interrogating a static structure, we are excluding all the dynamic properties of a protein in a warm solvated system. It is not yet possible to observe biomolecules in their natural environment, but by simply incorporating information on the motion of a molecule, we can

greatly expand our understanding. Simple estimations of predicted motion such as Normal mode or hinge region decomposition provide computationally cheap methods for study of motion. Normal mode calculations applied to proteins give an estimate of the low energy vibrational modes by grouping sets of amino acids into dummy groups and then treating them like an oscillating system. Normal-mode theory involves a harmonic approximation of the potential energy around a global minimum, in most cases the crystal structure is used at a starting point and is assumed to be near an energetic minimum. This allows the solution of the equations of motion by diagonalization of the Hessian matrix. The eigenvectors of this matrix provide the normal modes and the eigenvalues are the square of the frequencies of the oscillations. Every atomic displacement vector is a linear superposition of normal modes weighted by each of its eigenvector coordinates for the atomic mass, an amplitude, phase, and frequency. This technique has been in use since the early 1980's and have proved highly reliable for predicting overall structural movement. Recent estimations suggest that in many cases where there are known changes in tertiary structure, normal mode calculations correctly represent them roughly 50% of the time(198), and given an EM map, application of normal mode analysis can aid in interpretation of the diffuse density observed in experiment up to 20Å(199). Furthermore, frequently representative motions are contained within one of the three lowest energy normal modes (200-202). While domain level movement is particularly interesting in enzymes with open-closed conformational transitions such as polymerases (203), certain hydrolases (204, 205) and immunoglobulins (206), these types of calculations are less effective when the goal of a project is to interrogate a catalytic site.

Methods that move beyond domain level include coarse grain, all atom molecular dynamics, and accelerated sampling techniques, where the equations of motion are approximated for a system. The objective of these methods is to simulate real world phenomena as accurately as our technological limits will allow. In these types of simulations, classical

physics is applied to a three dimensional system. Molecular dynamics propagates Newton's equations of motion using a force field to evaluate the potential energy. The most commonly used force fields describe bonded and non-bonded terms and are parameterized to reproduce experimental data. Bonds and angles are represented as springs, dihedrals are represented using a sine function. This reproduces the energy cost of rotating a group through an eclipsed position. Van der Waals and electrostatics are included as non-bonded terms and are typically modeled using a Lennard-Jones potential and Coulumb's law respectively. A range of methods are being developed to improve on the standard force fields, in hopes of better reproducing physical phenomena. The development of a force field including polarizability is expected to improve the accuracy of simulation, and these are being pursued by various methods including electronic structure theory, induced dipoles, point charges, distributed multipoles and density fitting.

While there are certainly concessions to the method, all-atom molecular dynamics has been broadly accepted as the gold standard for exploring macromolecular motion on a per-residue scale. As we cannot yet observe the sub-domain motions in a protein barring collecting crystallographic snapshots, dynamic methods give insight into systems that are beyond the reach of experiment and are likely to remain that way for some time.

Coarse grained methods involve representing clusters of atoms as larger "dummy atoms", assigned various properties such as size, weight, and charge based on various metrics. This reduction in complexity is particularly useful for simulating large or complex systems where discrete residue-based information isn't required, but where sub-domain detail is needed. Successful applications to membrane bending (207-211), lipid layer perturbation (212-216) and motor proteins (217, 218) have been among some of the most interesting in the literature. Frequently, coarse-grained simulations are capable of reproducing some gross characteristics of an experimentally measured system. For example, membrane tubule diameter was

successfully reproduced by coarse grained simulation of membrane-bending by BAR domain proteins (208, 209). While excellent in conjunction with biophysical experimental methods, coarse graining does not allow for analysis at the atomic level.

All-atom molecular dynamics (MD) are capable of illustrating the behavior of biomolecules on an atomistic scale, and can describe very fine networks of interactions when properly constructed. The capability in dynamics simulation has exploded in recent years, with studies surpassing the millisecond long milemarker in 2009 (219). While extremely long timescale dynamics are now within reach, frequently the resources to facilitate those studies are not. This has led to the development of accelerated dynamics techniques, designed to sample the conformational space present in an extremely long timescale simulation without explicit performance of those dynamics. Accelerated MD(220, 221), replica exchange MD(222-224), temperature accelerated dynamics(225-229) and other forms of enhanced sampling (230-237) are all methods that seek to decrease the computational cost associated with traditional molecular dynamics. Most traditional studies limit themselves to the nanosecond range and a plethora of analysis tools have been developed to facilitate the extraction of data from these simulations. Detection of correlated motions(238), solvent effects(239), per-residue rearrangement to facilitate salt bridges and hydrogen bonds are all common analysis techniques.

Detailed understanding of the motions of a biomolecular system are incredibly informative, but until recently, the computational cost, resources, and skill needed to perform robust dynamics simulations were prohibitive. As the methods develop to thoroughly and efficiently investigate biomolecular structures in a semi-native environment, the breadth of systems and downstream applications for which molecular dynamics has been used has greatly expanded. In terms of the workflow discussed in Figure 0-1, biomedical research has recently seen a huge push towards incorporating the data from dynamic techniques into drug design.

1.4.3 Molecular dynamics applied to drug design

As our ability to perform molecular dynamics has increased, there has been a continual press to increase the usage of these types of studies in drug design. There is a considerable body of evidence suggesting that drug design can be enhanced by studying the interactions of ligands with an array of protein conformations (240-244). The concept of a rigid receptor and rigid small molecule has lost favor over the years, as our understanding of protein conformational dynamics has evolved. Conformational selection and conformational capture are gaining favor as more realistic interpretations of protein-ligand interaction (245-251). To design drugs around a cadre of protein conformations, one must first have access to that conformational space. Molecular dynamics, as well as Markov state modeling and enhanced sampling have been applied recently to the generation of protein structures, which can then be used for ligand design.

Following the profound success in drug design on HIV-1 integrase leading to the now FDA-approved drug raltegravir, application of MD to drug design gained a considerable amount of notoriety(252, 253). A cryptic binding site on the surface of HIV-1 integrase was detected following molecular dynamics studies. Further work with molecular modeling and optimization was used to develop a set of small molecules which were then tested *in vitro* and found to have significant inhibition constants. This substrate optimization was directly based on the detection of a novel trench detected in the molecular dynamics work.

There have been a few other notable success stories, though none have been as popular in the news. Histone deacetylases have been long known to be valuable, but challenging drug targets. Work by Estiu et al. on the numerous crystal structures of HDAC 8 led to the differentiation of several critical interactions governing isoform selectivity in small molecules(254, 255). Molecular dynamics simulations indicated that the HDAC8 structure was highly mobile, with a few critical contacts in the interior of the protein. It was determined that the

alterations in HDAC8 conformation when bound to small molecules were essential for inhibitor function and isoform selectivity. This information was exploited to design isoform specific HDAC8 inhibitors (256-258). Further studies extended this hypothesis on isoform selectivity to HDAC10 and 11, where it was determined that small molecules which tightly bind HDAC8 have weaker contacts with the other two structures (256-259).

Inhibitor design targeting influenza A virus M2 (IA/M2) used molecular dynamics to investigate the binding dynamics of amantadine, the well known anti-viral. In this case, the dynamics studies were used to dissect the interactions in the active site and determine which contacts were stable. Reoptimization of the initial ligand structure resulted in a set of amantadine analogues that perform similarly in simulation and which are now being investigated as potential drugs(260, 261).

In application to protein-protein interface disruptors, molecular dynamics was used to detect the critical contacts in tubulin polymerization, which were then built into antimitotic peptides(262). Simulations of tubulin dimers allowed access to the conformational and energetic characteristics of the interaction. Simulation following computational alanine scanning was used to discern which components of the interface were most critical. Based on the essential sequences, a series of peptides were designed to mimic those tight interactions. Tubulin-peptide simulations were performed and in several cases the peptides remained tightly associated with the tubulin monomer. *In vitro* assays indicated that the peptides had a strong inhibitory effect on tubule formation. A control peptide that dissociated *in silico* was found to have no inhibitory properties *in vivo* suggesting that the efficacy of the design was not random.

While there has been some considerable success, the popular opinion is that there needs to be far more inclusion of computational methods in drug design. Historically most dynamics studies went after structures with interesting physical properties, or those that were

interesting to the fundamentals of biochemistry. The few attempts at combining MD and drug design have been quite productive, but as the concept of combining these techniques is still quite new it is difficult to ascertain extent of that productivity . A text-mining and citation tracking foray reveals that of 35 studies performing molecular dynamics on potential drug targets since 2001 (252, 253, 263-296), 21 of these resulted in the synthesis of an active inhibitory molecule (131, 263, 287, 291, 294, 297-312). That is not to say that studies without verified synthetic molecules have been unsuccessful; many computational investigators lack the facilities to synthesize and test potential drug-like molecules. When facilities or collaborations exist, it still may take time to translate computational results into wet-lab successes. All told, a 60% success rate in the last dozen years is still nothing to sneeze at.

The response from key figures in the drug design field has been unanimous, whatever can possibly be done to increase the amount of dynamic data related to drug design should be done. High-throughput GPU powered MD facilities, black box servers for the generation of Markov model structures, integration of data resources, and improvement in protein and small molecule forcefields have all been suggested as ways to facilitate the incorporation of dynamic information into drug design (313-319). Computational methods that give additional structural information about receptors are already becoming invaluable to the process.

1.5 Small molecule development

With a thorough understanding of the target, the pathway, and the structure in hand, the final step towards drug design or investigation is to determine the ligand possibilities available. When receptor structure is unknown, or where there are multiple substrates with binding data available, there are a range of methods possible to assist in the development of small molecules. The broad field of quantitative structure-activity relationships (QSAR), bootstrapping, bioisostere and rescaffolding methods are just a few(320-323). When a receptor structure is

available however, computational tools can provide a considerable advantage over either high-throughput screening or "spaghetti-to-wall" synthesis. It is generally agreed that the return on investment for chemical high-throughput screens is lower than expected, and they are limited by the contents of the screening library. Similarly, purely virtual compound screens suffer from a staggering number of algorithmic options, as well as a lack of consensus in scoring methods for potential drugs (319). An alternative to high-throughput screening that has performed well is chemical fragment screening. In these experiments low molecular weight molecules are screened for binding affinity against a target. Once small fragments are identified, these results are typically combined with crystallography and computational methods to develop a more drug-like compound. A review by Murray et al presents a list of the numerous successes in both industry and academic literature (324). Fragment-based drug design is an excellent example of how the combination of experimental and computational techniques can be more productive than either technique alone.

As structure building and conformational optimization for small molecules have become commonplace, these methods will not be discussed. In the last stage of the workflow discussed in Figure 0-1, design of small molecules for either clinical inhibition or fundamental inquiry is a common pursuit when studying a biomolecular system. Molecular modeling and docking, as well as novel structure reoptimization methods can considerably improve the pace at which a new molecule can be brought to bench. Additionally, this is the part of the workflow where computational chemistry has been most frequently applied. The following chapter sub-sections will focus on techniques relevant to the previous discussions, and new advancements in the field.

It is important to note that with respect to small molecule development, there are a spectrum of goals that can be worked towards. Small molecule design can include optimization of physiological transport properties, tuning of binding, increasing stability *in vivo*, and selecting

for various protein isoforms or multiple targets. As discussed in section 1.3.3, tight binding is not always the goal of a drug design project. Tuning a small molecule to target a range of proteins is becoming increasingly popular as polypharmacology catches on. Furthermore, the behavior of a small molecule inside a body needs to be considered. Many drug leads that perform spectacularly *in vitro* have little or no *in vivo* activity, alternatively off-target effects can result in considerable toxicity. Pharmacokinetics must be considered as well. The suite of biological processes known as ADME (absorption, distribution, metabolism and excretion) all affect the pharmacokinetic efficacy of a drug and there are numerous computational methods directed at tuning these properties, the well known "Rule of five" is the result of one such study.

1.5.1 Small molecule docking

In assessing the interaction between a ligand and receptor, a conformation for the ligand is required. Numerous algorithms have been developed since the 1980's to perform this task, and it has been largely established that docking is a good way to generate a putative ligand binding conformation. While the number of docking programs has proliferated over the years, the fundamental protocol remains the same. Generate a set of starting structure guesses, and score them to assess binding or interaction affinity.

Generating the docked structures can be done in a number of ways. Docking a rigid ligand to a rigid receptor has long since been abandoned as inefficient and ineffective for small molecules, so these historical techniques will not be discussed. The dominant paradigm places flexible ligands into protein active sites of variable flexibility. Many programs still use a rigid receptor, while some include side chain "softness", while some accommodate docking to a group of related structures. Of the techniques commonly in use today, programs can be sorted by algorithm: Monte-Carlo methods, grid-based methods, genetic algorithm methods, fragment-based methods, and "soft-docking".

Monte Carlo methods rely on the statistical theory of Metropolis et al. (325). In its simplest form, this method can be thought of as a repeated series of three steps. A ligand is moved in cartesian space, the new position is evaluated by some heuristic, and then the move is accepted if the new position is more energetically favorable, or rejected if it is above an energetic threshold and statistical probability based on a Boltzmann distribution. In the program AutoDock, the AMBER force field is used for energetic scoring(326). Frequently Monte Carlo steps are used as a cheap form of minimization, such as in ProDock(327). The stochastic nature of the search means that compared to many other docking algorithms, MC methods can be slow. There are some structural shortcuts that decrease the computational cost, such as representing the protein receptor as a grid rather than an explicit 3D structure. A particularly interesting new stochastic method is swarm optimization, where the search is conducted using a model of swarm intelligence. The method of particle swarm optimization was originally developed to model flocking behavior, but was observed to be a mathematically efficient method for optimization. SODOCK(328) and PSO@AUTODOCK(329, 330), are just two examples.

Systemic search methods use an algorithmic approach to explore all possible conformational space available to a ligand, within a designated interaction space. GLIDE(331, 332) performs an exhaustive brute-force search, as does FRED(333). Another way of performing a systemic search is to apply a genetic algorithm. In the programs GOLD(334) and AutoDock(335), ligand docking is treated like an evolving system. A population of ligand conformations are generated, moved, and ranked. Poorly scoring conformers are removed from the pool and the process is repeated with the newly culled ligand pool.

Fragment based methods explore the protein binding site by decomposition of the ligand into small pieces which are placed in the active site, providing seeds for reconstruction of the ligand. While there are a range of methods for fragmentation of the ligand and placement of the initial fragments from which building will proceed, nearly all of them use an interaction based

assessment. Fragment division is typically performed in a manner that will leave major functional groups intact. For example, FlexX severs the ligand at all acyclic single bonds. This would leave protein backbone fragments and Murcko cyclical drug fragments intact (336-340). The advantage of the fragment based methods is their relative speed. Pre-assessment of the protein binding site, and the incremental construction of the ligand minimizes the amount of of calculations performed per docking solution(340, 341).

While there are many methods used to generate the set of ligand conformation, evaluation of the conformations is also required. A plethora of scoring algorithms is available, and a thorough review of recent developments is available in Ramsland et al. (342). Comparative studies of docking and scoring algorithms for the past few years have universally reached the same conclusion. While current docking methods all are capable of generating excellent docking conformations, the scoring algorithms are insufficient and generally are difficult to assess. Inclusion of solvation estimates, entropy, knowledge-based scoring based on database information, and even machine learning has only moderately improved the situation, and in no system has one scoring method proven unequivocally superior. In most cases human assessment of the docking conformation is still required. Regardless, the ability to generate structures that are reliable, even when a human eye is required for the final assessment, is a significant achievement.

1.5.2 Lead optimization

It frequently happens that a small molecule is found to bind a target protein, just not with a particularly good dissociation constant. When a structure of the protein is available, and the ligand has been modeled, computational lead optimization can be performed. Before the development of explicit algorithms for optimization, much of the restructuring of the ligand was

done by eye. Newer methods however have increased our ability to explore the chemical space available for synthesis and more accurately assess active site interactions.

One popular computational technique for retooling small molecules is one that was borne out of the combinatorial docking studies. The generation of ligand libraries facilitated the generation of ligand fragment libraries. CAVEAT was the first program designed to mine fragment libraries in the hopes of retooling lead compounds(343). Fragment-space searching seeks to replace unfavorable pieces of bound ligands with small molecule fragments in the hopes of improving the binding. Various protocols incorporate shape complementarity, electrostatics, desolvation, and more recently explicit water binding (344-348). Upon obtaining a docked conformation, a researcher inspects the binding pose and either by eye or using a scoring method assesses which contacts or contact regions are negatively impacting binding. A region of ligand is selected to be replaced, or extended, and then a global search of the fragment library is used to do the substitution. Fragments are pre-screened before docking by assessing the region of the receptor in which they'll be binding and if possible a pharmacophore can be designated as a desired point of interaction. A library of derivatives is generated in this manner and they can be docked, scored and compared to the original ligand. There are several drawbacks to this method, one of which is that only fragments currently in the fragment library can be assessed. Second, it is highly dependent on the initial selection of regions for replacement and iterative investigation can be time consuming. However, the recent implementation of the fragment search software ReCore (349-352) has shown some promise in the literature (353).

Other methods of lead optimization are based on free energy perturbation theory (FEP), and thermodynamic integration (TI). While some computational groups have applied these methods to drug design, in very productive ways, there are no "easy" ways to perform these calculations. Application of these methods has been successful in the design of fructose 1,6-

bisphosphatase inhibitors (354-359), COX1-COX2 inhibitors(360-363) and fatty acid amide hydrolase inhibitors(364-367). Cheaper and easier methods based on molecular mechanics/generalized born solvent approximation (MM/GBSA) have not proved accurate enough for functional lead optimization(368).

1.6 Conclusions

As computational techniques advance and become both more sophisticated and less expensive, there is no denying that there is benefit to inclusion at every step on the path to biochemical understanding. From target identification and systems biology to structure interrogation and ligand design, computational chemistry can play an important role.

Application of bioinformatic and statistical techniques to genome level data has the ability to deconvolute the massive amount of data generated. High throughput sequencing is capable of generating a staggering amount of data very quickly, and it's only recently that computational techniques have been catching up to the pace. Working in conjunction with geneticists and clinicians, bioinformaticians have the unique ability to find novel targets and biomarkers hidden in the data. When data mining is applied to non-human genomes it opens the door to finding novel antimicrobial targets. Beyond that, the foundation of human genetics was laid by studying regulation in model organisms. We now have the ability to more deeply understand the underlying patterns of gene expression and regulation in simple organisms, which may in time lead to a greater understanding of human genetics.

With that in mind, the extrapolation of graph and network theory to biological data has unprecedented potential to change the way we think about biosystems. Not only does this type of modeling lead to a clearer understanding of the interactions in a system through deconvolution, mathematical modeling has revealed entirely new paradigms for drug design.

The prospect of generating as complete a network model as possible, and then probing the effects of perturbations on the system holds a lot of promise. The validation of the work of Csermely and Nussinov on network perturbation applied to multi-target drugs has opened the door for a completely new method of drug design. Multiple protein-binding ligands and network based repurposing of known molecules may help remedy the well-known dearth of novel drugs being released. Application of network analysis to both gene and transcript level data allows the formation of meaningful connections and novel perspectives in drug design.

At the level of atomic structure of biomolecules is where computational chemistry really shines. When you get down to the level of a 3D atomic structure, computational chemists are able to give insights on structure and function that are just out of the reach of experimentalists. It is at this point that the *in vitro* data from experiment can be at least partially interpreted by computational results at the atomic level. Static assessment of protein, RNA or DNA structures can help us understand the physical characteristics that are observed in experiment. Electrostatic and surface properties govern the subcellular interactions of biomolecules; computational investigation can help give context to those interactions. Protein-protein binding is one of the major areas where electrostatic and surface properties can assist experimentalists. Though determination of the protein-binding interface can be examined by mutagenesis, computational methods can help guide those studies and vice versa. With respect to dynamic information, such as normal mode or molecular dynamics studies, the level of understanding increases exponentially. As we begin to think about proteins as more than just rigid bodies with a catalytic function, dynamic structural information becomes a necessity. Molecular dynamics studies in particular are both highly valued, and rarely applied. There is incontrovertible evidence that MD investigations are highly productive when applied to drug design, but both computational cost and skill required are hurdles that need to be overcome.

Small molecule studies are the oldest application of computational chemistry, and also some of the most well developed. Algorithms for *in vacuo* structure and energy optimization became the gold standard long ago and we can only hope that much of the rest of computational chemistry goes that way as technology advances. Modeling of protein-ligand complexes has reached the point where the question is no longer "Did we find a binding orientation?" but is rather "Which one of these is best?". Scoring and binding energy assessment accuracy still lag behind the generation of docked conformations, but not for lack of effort. In this area, computational chemists and skilled modellers still require the input and "eye" of skilled wet-lab chemists. However, with respect to reoptimization of binding conformations, fragment space searching is a great assist in the generation of derivatives. A thorough fragment space search capably assesses the result of performing literally millions of substitutions on a molecule. While there is no software that will hand over a "magic molecule", these types of studies can provide inspiration, and are an excellent starting point for conversations between computational chemists and synthesists.

Computational chemistry can make critical contributions at all stages in the biochemical discovery workflow. From identifying targets and facilitating genome level analysis, to the development of systems and network biology, to structural assessment, drug design, and small molecule optimization, integration of computational techniques is becoming essential. The ability of computational methods to handle extremely large datasets, be it genomic or dynamic puts a wealth of information within reach. Whether for fundamental or goal-oriented studies, the combination of computational and experimental work is extremely powerful.

Looking forward, across all facets of biochemical discovery, communication and collaboration is key in solving the challenges we face. While presented as four separate steps in a process, target identification, network analysis, structural investigation and small molecule development each have the potential to affect other parts of the discovery pathway.

Collaborations between experimentalists and computational chemists are highly valued, but beyond that, to really make the most of the new developments in computational chemistry, we need to expand our reach.

The rapidly expanding amount of "omics" data available is presenting a huge challenge in data storage, analysis and interpretation. To help us develop creative and innovative ways to meet the challenges we face we need to "think outside our field". Network and graph theory have permanently changed the way we think about genomic data and small molecule inhibition, though the initial studies had nothing to do with biochemistry. Bioinformatics is assisting in building a bridge between computer science and biochemistry, but the application of purely statistical techniques to biology has not been entirely helpful.

The initial results from genome wide association studies were considered a let-down, but we as scientists should know better than to approach a disproven hypothesis as a failure. The common-variant hypothesis failed. Genome wide association experiments did not. In fact, they are changing our perception of human genetics is linked to disease. Now that we have access to large scale genomic data, we should move past reductionist techniques and attempt to understand the complex interactions underlying our complex diseases.

When we cannot communicate our experiments, results, or interpretation, science suffers. As we layer network theory on top of dense systems biology information to find understanding in a broader context, a major challenge will be finding a way to effectively present our findings. While current network visualization techniques can build off of pre-established ideas, we need to develop a visual language that can effectively communicate the depth of information in pathways networks and systems. In the future perhaps we can look to cognitive psychologists or graphic designers to help us find new methods to express our research in a way that doesn't overload our brain-processing power.

Novel drug targets and disease associated proteins are being identified regularly and in order to really get to the root of how these proteins tie into their network, we need to understand them. Our ability to quickly and accurately determine molecular structure has grown immensely, providing a wealth of information for modelers. Over 30% of the structures in the RCSB protein data bank have been deposited in just the last three years. As our ability to investigate structure grows, we need to propagate that information both backwards and forwards in our discovery process. Surface analysis methods have not dramatically changed in recent years, but novel applications of those methods should be something to strive towards particularly as inhibition at the protein-protein interface becomes popular.

There is a resounding chorus calling for more molecular dynamics in drug design, and one would hope that in the future, the dynamic properties of a potential drug target would be commonly investigated. Rationally designing a clinically valid therapy from a single crystal structure is similar to trying to figure out the ending of a movie from a still frame. You may be able to figure out some bits and pieces and make a good guess, but you're still missing the big picture.

While we should constantly strive for higher accuracy and better sampling methods, current methods allow for a basic investigation even into relatively large biomolecules. As both the methods and computational power improve, there is no doubt that in the future molecular dynamics should become a much more widely applied technique.

Challenges in the future of small molecule design seem to be largely human. Both the experimental and computational technologies are well advanced with respect to building new drugs, but a quick survey of the literature indicates a disconnect. In the future small molecule development needs to be enthusiastically interactive. The plethora of tools designed for small molecule building, docking, ranking, optimization, and physiochemical property determination

are staggering, and this proliferation of techniques and publications can be unwieldy to process. Similarly the work involved in synthesis, purification, assay development and testing can be gargantuan.

Purely computational methods lack the persuasiveness of *in vitro* experiments as the results are semiquantitative at best due to the approximations that are involved. Frequently organic chemists balk at synthesizing molecules from a pure computational screen. This is appropriate; frequently structural biologists and computational chemists are less familiar with the practical aspects of organic synthesis and lack the ability to quickly weed out structures that a specialist in pharmacokinetics or drug metabolism would easily spot as problematic. Synthesists overlook the capabilities of computational chemists when it comes to suggesting ore reoptimizing a structure. It is far faster to determine whether a substituent location will enhance or interfere with ligand binding computationally than it is to synthesize and test. Also, the ability to search millions of subtituent fragments may inject creativity into the drug design process.

Try as we might, no individual will ever be capable of working effectively on all facets of biochemical research. Rather than attempt to create polymaths of us all, collaboration and communication should necessarily connect experimentalists and theorists at all levels of biochemical discovery. Hopefully in the future our ability to generate and manage large datasets should go hand in hand with generating and managing profound discoveries.

Chapter 2 Conformational analysis of *Clostridium difficile* Toxin B and its implications for substrate recognition[1]

2.1 Background

One of the most common and serious hospital-acquired infections is *Clostridium difficile* (*C. difficile*), responsible for a suite of diseases collectively known as *Clostridium difficile* associated diseases (CDAD) (369, 370). *C. difficile* typically affects patients undergoing antibiotic treatment for other infections, as it leaves the GI tract susceptible to colonization by this highly virulent pathogen due to the reduced protection by the normal gut microbiota (371, 372). Currently, U.S. health care costs associated with treating CDAD are estimated to be between $750 million and $3.2 billion (372-376). With the emergence of an epidemic strain that is both hypervirulent and more resistant to current therapies (377-379), costs will surely continue to rise, so new approaches to treating CDAD are needed.

C. difficile is a spore forming bacillus, and thus is difficult to effectively sanitize against, as the spores are easily spread from person to person and can survive most normal sanitization methods (380). Figure 0-1 shows an electron micrograph of both a vegetative bacillus and a spore, red and blue respectively (381).

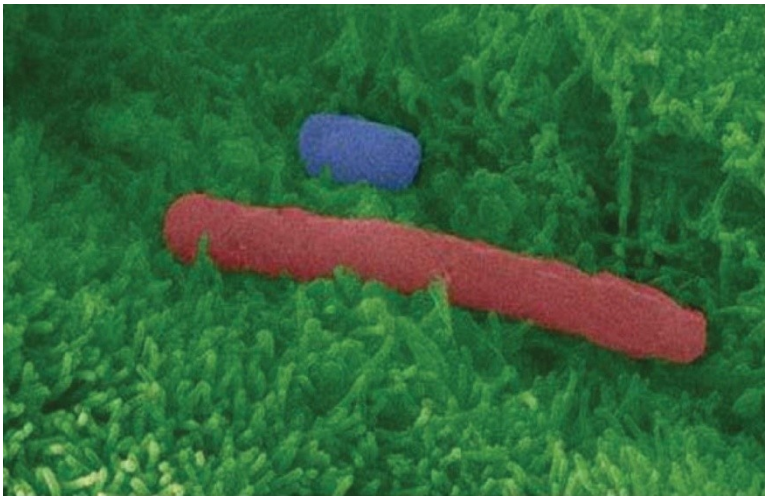


Figure 0-1: Electron micrograph of C. difficile spore (blue) and bacillus (red).

Although two thirds of patients infected with *C. difficile* are asymptomatic, those that are symptomatic can experience severe complications such as fulminative colitis which is potentially fatal (372). A recent upswing

[1] Sections of Chapter 2 have been previously published (449)

in symptomatic infection rates has spurred numerous hospital facilities to track and report outbreaks as they occur. In one instance, it was observed that that patients who acquired a *C. difficile* infection averaged a hospital stay of 3.6 days longer than unaffected patients, and increased their average hospital cost by $3669(375). *C. difficile* is also a rapidly changing bacterium that has acquired progressively more antibiotic resistance. Though this resistance varies by strain and geographic location, strains can be found that exhibit resistance towards most antibiotics.(377, 382-384) It also has been implied that the increase in *C. difficile* infection is due to the emergence of such widespread antibiotic resistance (382). As the *C. difficile* is an opportunistic pathogen, taking over when healthy intestinal flora are suppressed by broad-spectrum antibiotics, antibiotic resistance is prerequisite to infection. Currently the main form of treatment for this antibiotic-initiated disease, are stronger antibiotics. Treating a naturally antibiotic-resistant bacterium with ever-stronger doses of antibiotics makes no sense. It simply makes *C. difficile* harder to treat.

The complement of diseases known as CDAD is the result of cellular damage in the intestines due to exogenous virulence factors produced by the bacillus(385, 386). *C.difficile* produces two large cytotoxins known as Toxin A and Toxin B, referred to as TcdA and TcdB. These toxins are members of the type A glucosyltransferase family, and cause cellular damage by targeting small G-proteins such as RhoA, Rac and Cdc42 (387). Utilizing their UDP-Glucose as a substrate, the toxins glucosylate a conserved threonine residue on their target G-protein, inactivating it permanently and ultimately leading to cell death (387, 388). The toxins themselves are large, and consist of four domains. Both contain C-terminal repetitive oligopeptides referred to as the CROP domain, which acts as a receptor binding region, a translocation domain and the N-terminal catalytic domain responsible for glucosylation (387, 389-391).

2.2 *C. difficile* toxins A and B

C. difficile damages the intestines primarily through the action of TcdA and TcdB as described above (371). These are members of the lethal subclass of large clostridial toxins (392). The holotoxins are ~300 KD and are comprised of four domains, each having a specific function related to cellular uptake and toxicity (393). The CROP domain (Clostridial Repetitive Oligopeptide) helps to identify and bind to appropriate target cells by recognizing cell surface glycoproteins and inducing endocytosis (394-397). The translocation domain is responsible for forming a transmembrane pore capable of passing the two remaining domains from the endosome to the cytoplasm (397-400). A cysteine protease domain, activated by inositol hexakisphosphate in the cytoplasm, intramolecularly cleaves the cytotoxic glucosyltransferase (GT) domain from the holotoxin (401-403). This last step is critical since at this point the GT domain is released into the cytosol where it can act on the RhoA, glucosylating residue T37 in the switch I region (or its equivalent S/T residue in the case of other Rho family members) (388). Glucosylation of RhoA permanently inactivates it, causing defects in the cell-signal pathways that lead to cell rounding and ultimately apoptosis (388).

While one could develop new antibiotics to better target *C. difficile*, resistance is likely to be a major concern with any new agents. A potentially complementary approach to antibiotic therapy is to develop methods that target and neutralize the GT domain of the toxin (404). Several steps in the etiology pathway could be targeted for inhibition, however this work focuses solely on the glucosyltransferase domain.

Several approaches are currently being used to target TcdA and TcdB. Clinical studies are under way with humanized monoclonal antibodies that recognize and sequester the toxins, but this approach has some issues and will not be suitable for all patients (378, 405, 406). Peptides and small molecules that recognize and inhibit toxin function are also being studied

(407). By better understanding the domain structures of the holotoxin, it will be easier to design or select molecules that disrupt their activity.

2.3 Glucosyltransferase domain of TcdB

The GT domain from TcdB (PDBID: 2BVL) was crystallographically characterized several years ago (408). This domain was found to be a 543 amino acid domain that adopts a characteristic GT-A glucosyltransferase fold, and binds a catalytically-important Mn(II) ion. Previous studies comparing the *C. difficile* toxins to other glucosyltransferases, as well as extensive mutagenesis analysis on the toxins themselves, have identified a number of amino acid side chains critical for activity (409-411). Figure 0-2 illustrates some of the important structural elements of TcdB that will be discussed later in this chapter. A four helix amphipathic bundle comprising residues 1-87 (shown in blue) has been implicated in membrane association (412); we will show that it is a key component in the large scale molecular motions exhibited by TcdB. Residues 510-522, shown in yellow, are part of a mobile loop which supports the catalytic manganese and includes a standard DXD motif. The two regions shown in cyan will be referred to as "upper promontories". The function of these structural motifs is not yet understood, although they participate in a scissoring motion that will be described below. The beta hairpin shown in purple (residues 374-387) will be referred to as the active site flap and may have implications in catalysis and substrate recognition. The green region (residues 436-456) has been shown to be involved in recognition of RhoA by TcdB (413). Finally, the red region (residues 483-497) shows motions that are highly correlated to those of the recognition site (residues 436-456) in our analyses (411, 414, 415). Shown in transparent orange is RhoA, following docking.

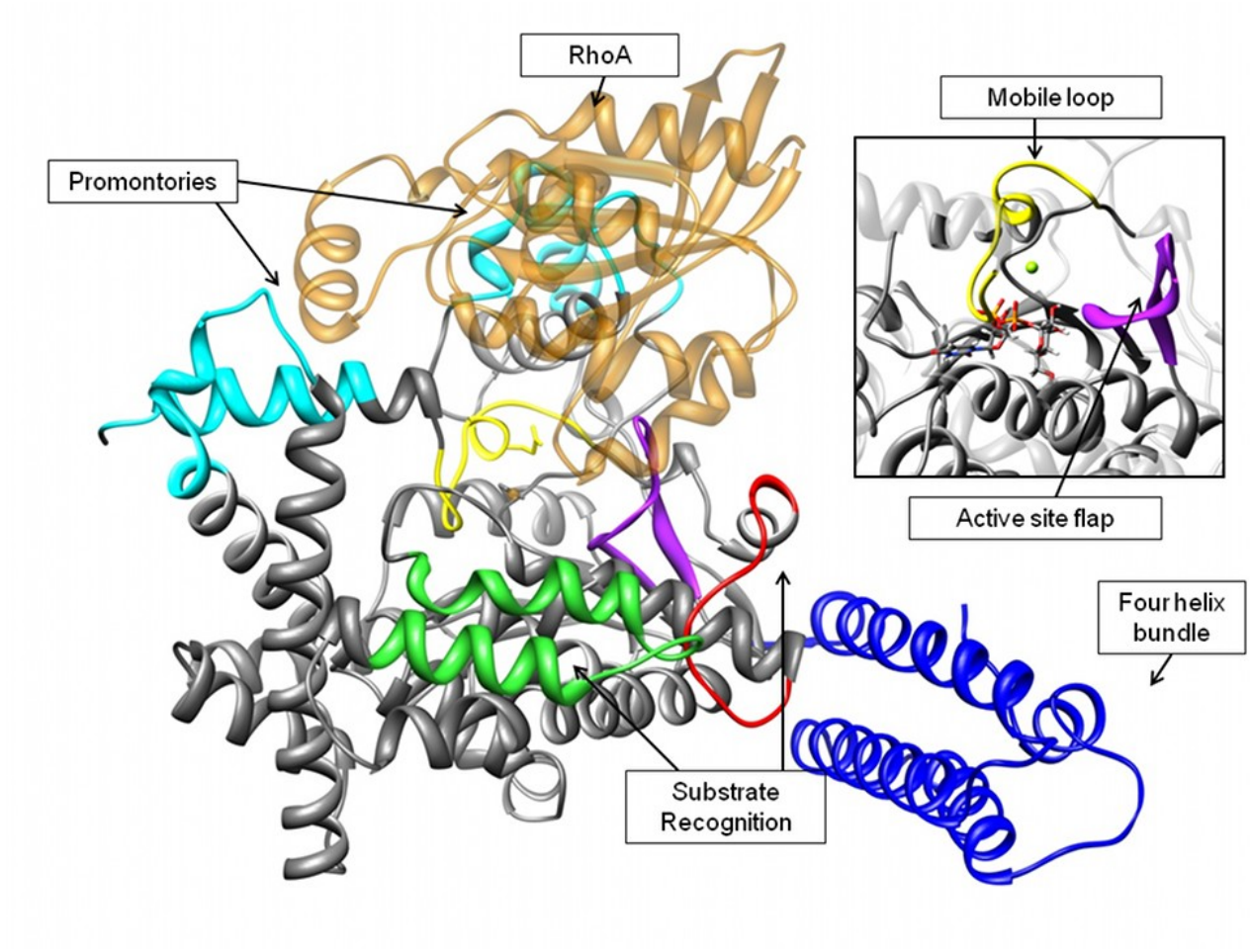


Figure 0-2: Structure of TcdB Glucosyltransferase domain

Relevant regions for discussion: The four helix bundle is shown in blue, the mobile loop containing the DXD motif is in yellow, the catalytic manganese is shown in black. Regions shown in green and red are involved in RhoA recognition. The B-hairpin shown in purple will be referred to as the active site flap. The upper regions in cyan are two flexible promontories unique to TcdB. RhoA is shown in transparent orange.

2.4 Experimental design and rationale

A comprehensive understanding of the conformational space that TcdB can occupy will better guide design of potential inhibitors. TcdB must pass through a pore to gain entry into the cell, therefore it is expected to have a flexible form to facilitate transient unfolding and refolding during translocation. Hinge region (416, 417) and normal mode analysis (418) were applied to determine the location and extent of the primary flexions. Both of these techniques have

previously proven useful in determining the major motions attributed to well-studied systems, and give a fundamental impression of the overall motions one should expect to see in a flexible protein.

Long timescale unbiased molecular dynamics (MD) simulations may give insight about the conformational space a protein occupies, as well as the mechanism of transition between those conformations. Additionally, the atomic scale detail in these simulations allows us to take a look at how large scale motions can have consequences in small regions, such as within an active site.

Understanding in a broad sense how TcdB moves and flexes both on its own and in contact with RhoA is expected to lead to better understanding of catalysis, substrate recognition and most importantly, drug design. The GT domain of TcdB has not yet been crystallized bound to substrates other than UDP-Glucose, and thus, nothing is known about the range of conformational space it can occupy, or what consequences binding to the RhoA protein might have. Recent evidence suggests that RhoA employs a conformational selection mechanism (419), rather than induced fit or lock and key. Thus, it is expected that a toxin targeting such a protein might have similar properties. Here we report normal mode and hinge region analysis, as well as long timescale molecular dynamics of TcdB. Additionally, macromolecular docking and long timescale simulation of the TcdB/RhoA complex was performed. Principle component analysis (PCA) and Generalized Masked Delaunay (GMD) analysis of the resulting conformations were used to help understand the conformational space TcdB occupies both alone and in complex with RhoA as well as the nature of the transitions between these conformational spaces.

2.5 Flexibility studies on TcdB

To quickly assess TcdB's flexibility range, initially two computationally inexpensive methods were employed. A hinge region analysis was performed to determine simply if there were regions that would be amenable to macroscopic motions. Following this analysis, normal mode calculations were carried out.

2.5.1 Hinge region analysis

Hinge regions are determined by finding the consensus between two methods, StoneHingeP and StoneHingeD. StonehingeP incorporates ProFlex, which decomposes the structure into rotatable and non-rotatable bonds. These are then used to analyze flexibility based on bond rotation constraints. StoneHingeD incorporates DomDecomp, which uses gaussian normal mode analysis. StoneHingeP identifies the two largest rigid domains in the protein, and designates hinge regions as residues between these domains. StoneHingeD identifies hinge residues as those that are between domains or at domain boundaries, following normal mode analysis. StoneHinge (420) determined three major hinge regions on TdcB. The total search returns any StoneHingeD residue within five residues of a StoneHingeP residue. The results for TcdB are shown in Figure 0-3. These hinge regions involve the residues connecting the four-helix bundle to the body of the protein, as well as the region on the underside of the active site, and between the two upper promontories to either side of the active site.

Figure 0-3: Hinge regions identified by StoneHinge shown in green.

Backbone is shown as a chain trace, hinge residues are represented as green spheres. Hinge regions are observed to occur between regions of flexibility in the normal mode analysis

2.5.2 Normal Mode analysis

Once it was determined that hinge regions were present that would allow for macroscopic conformational changes, a Normal Mode analysis was carried out. elNemo (421) was used to find the low energy modes of TcdB. Normal-mode theory involves a harmonic approximation of the potential energy around a global minimum, in this case the crystal structure is assumed to be near a minimum. This allows the solution of the equations of motion by diagonalization of the Hessian matrix. The eigenvectors of this matrix provide the normal modes and the eigenvalues are the square of the frequencies of the oscillations. Every atomic displacement vector is a linear superposition of normal modes weighted by each of its

eigenvector coordinates for the atomic mass, an amplitude, phase, and frequency. The normal modes were represented by a moving gif. Several of the modes appeared to have significant effect on the conformation of the active site, and this information was used to assist in generation of a putative RhoA-TcdB complex. The primary motions were a wagging of the four helix bundle, and scissoring of the upper promontories described in Figure 0-2. The intersection of these modes results in an opening and closing motion in the active site center.

2.6 Macromolecular docking

Currently no crystal structure of TcdB bound to RhoA exists, though individual crystal structures of both TcdB and RhoA are available. Studies performed by the Aktories group determined the conformation of RhoA preferred during attack by TcdB (388). It's also known that TcdB functions by glycosylating a key Threonine residue on RhoA (388, 422). The combination of these pieces of information was then used to begin study of the protein-protein interface during the glycosylation of RhoA.

The first step in understanding the protein-protein interface was macromolecular docking to determine whether or not the structures were compatible in their near crystallographic forms. To this end, a multi-step protocol was developed. Rough dockings using Hex 4.5 (423, 424)were performed, followed by a more refined search using RosettaDock (425).

2.6.1 Hex 4.5 docking

Hex 4.5 models rigid protein bodies using three dimensional expansions of real orthogonal spherical polar basis functions for both shape and electrostatics. By using the vector coefficients of these basis functions along with a surface representing the shape of the protein, it's possible to evaluate a docked conformation through the overlap of pairs of these functions. The docked conformations being evaluated are generated through a Monte Carlo search

allowing for translational and rotational steps. This relatively simple and fast search was used to find initial conformations that place Threonine 37 of RhoA near the active site of TcdB.

2.6.2 RosettaDock docking

To further refine the docking poses, RosettaDock was employed. Rosettadock allows for additional steps of minimization and side chain refinement. Beginning with a coarse grained Monte Carlo search, rigid body docking is carried out, following 500 Monte Carlo steps, explicit side chains are added using a backbone-dependent rotamer packing algorithm, and are optimized using a simulated annealing Monte Carlo search. A Davidson-Fletcher-Powell quasi-Newtonian minimization technique (425) is employed to find a local minimum of the structure with the newly added sidechains. This process of repacking and optimization is repeated through fifty cycles. Initially, side-chains are repacked in a sequential manner, but once every eight cycles, the side chains are repacked combinatorially. Following side chain repacking, the structure is randomly perturbed by rigid body translation and rotation, and the process is repeated.

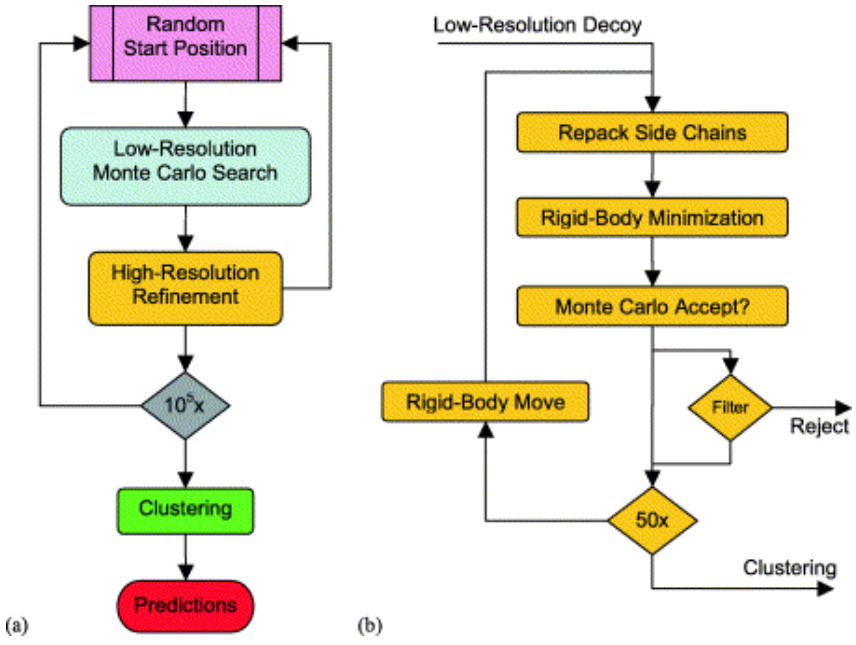


Figure 0-4: Flowchart of RosettaDock algorithm

Reproduced from Gray et al(1) Rosetta dock uses a series of low cost conformational searches followed by high resolution refinement steps to generate docked conformations

Docking results for the initial trials did not result in a “Docking Funnel” following cluster analysis indicating that no preferred low energy conformation was reached. Failure to find a binding conformation using RosettaDock is typically due to conformational flexibility in one or more of the binding partners. It is known that RhoA has a region of high flexibility, however these studies were performed with the TcdB crystal structure. To incorporate the information from the flexibility studies discussed in Section 2.5 Flexibility studies on TcdB, we repeated these experiments with the most open of the normal mode conformations.

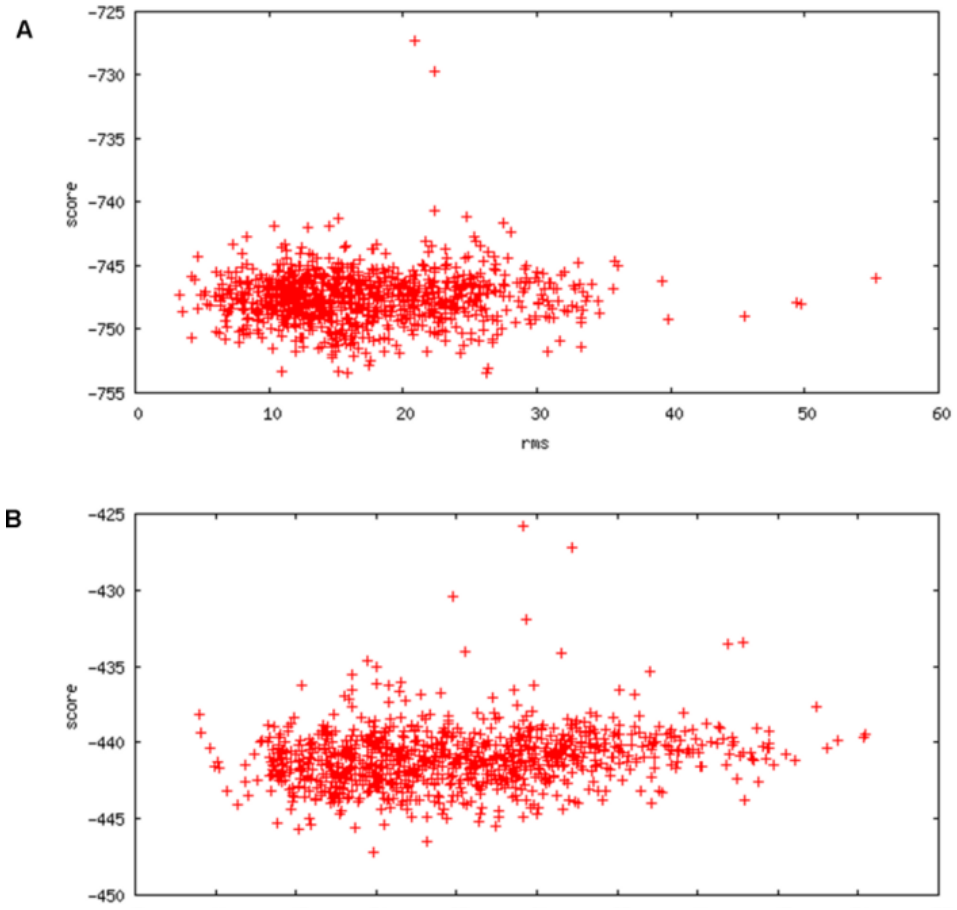


Figure 0-5: Structure Energy plots generated following RosettaDock

Panel A shows RhoA docking to the crystal structure of TcdB, and it can be observed that all energies are relatively high, and no cluster of low energy structures is observed. Panel B shows RhoA docked to the normal mode relaxed structure of TcdB. A reduction in docking energy is observed, and a few low energy regions are apparent. Of note is the improvement in docking when the normal mode structure of TcdB is utilized, indicating that flexibility in the face presented for docking may be a feature in TcdB’s target recognition process.

2.6.3 Normal Mode dockings

Following the failure of crystal-crystal conformation docking, The steps described in sections 2.6.1 Hex 4.5 docking and 2.6.2 RosettaDock docking were repeated using the crystal structure of RhoA and the most open normal mode conformation of TcdB.

The normal mode docked conformations showed improvement in binding over the crystal-crystal docked structures in proximity of the glucosylation site to the catalytic manganese. In the original docking, threonine 37 had a closest approach of 18 Å to the catalytic manganese. Subsequent docking to normal mode structures yielded a closest approach of 12.38 Å. A fully docked conformation might be expected to have a contact distance of between 7.1Å and 7.7Å based on comparison to several glycosyltransferases crystallographically characterized in the presence of UDP and an appropriate acceptor (426, 427). Additionally, improvements were noted in the structure/energy plots. Overall complex energy was lower, and docked solutions are more tightly clustered. However, while the use of a normal mode structure improved the docking, none of the structures that were obtained were catalytically valid. It was concluded from these results that while the normal mode calculation represented some measure of the flexibility of the toxin, it was insufficient to model a conformation capable of glucosyltransferase activity--particularly with respect to the regions in and around the active site.

2.7 Molecular dynamics and associated analyses

To fully elucidate the interaction between these partners, all atom molecular dynamics simulations were performed. Apo-TcdB and the structure of the normal mode conformation docked to RhoA (NM-RhoA) were simulated for a minimum of 150ns. Our purpose in performing a full all-atom simulation was to determine what conformational changes occur in the TcdB/RhoA pair to allow binding when compared to TcdB in the absence of substrate.

All simulations were performed using a parallel build of NAMD(428) employing the CHARMM (429) force field on the WSU grid. The simulations were run on eight nodes, of eight processors each. Structures were solvated, and appropriate counterions were added to reach 0.5mM NaCl. A timestep of one femtosecond was used, along with a 1-4 scaling factor of 0.4, recommended for protein simulations, to prevent additional contribution of bonded atoms to the calculated Van der Waals forces as this energetic contribution is partially included in the torsional terms of the force field. As in any molecular motion, there is energy associated with folding or unfolding of a protein. In a natural system, this energy is typically dissipated into or taken from the surrounding environment, however in a molecular dynamics simulation; the extensive surroundings that typically accommodate this thermodynamic process are absent. In order to maintain the simulation temperature in the form of a Boltzmann distribution around 300K, the velocities and trajectory of the atoms is regularly rescaled using a method based on the Langevin equations of fluid motion. In the context of a simulation, the technique of Langevin Dynamics (430) provides apparent viscosity and an element of randomness while maintaining the Boltzmann distribution around the selected temperature. As the simulations were run piecewise, they were joined, and all water and counterions were removed using CatDCD (431) to facilitate statistical analysis.

2.7.1 Principal Component Analysis

In order to more effectively compare the conformational space occupied by TcdB through the MD trajectories, PCA was applied. PCA is useful in that it decomposes the complex motions of the simulation into the major types of movements that are observed across the entire trajectory. These can be observed as series of conformations varying in a single dimension.

Analysis of the long MD simulations by PCA indicates that the principal component motions of the simulations echo the normal mode conformations as seen in Figure 0-6. Figure 0-6A shows a superposition of snapshots from the Apo molecular dynamics simulation. Figure

0-6B shows the results of the fundamental normal mode analysis. Figure 0-6C shows the first principal component extracted from the simulation of Apo-TcdB. Figure 0-6D displays the first principal component of the simulation of NM-RhoA. In normal mode analysis, MD, and PCA, the wagging motion of the four-helix bundle dominates, while the scissoring motions of the promontories is secondary. In each case, movement of these three regions affects the conformation of the highly flexible active site. The coupling of the motions of large peripheral structural elements of TcdB with highly specific rearrangements in the active site appears to be relevant to the process of substrate accommodation. Because normal mode analysis accurately predicts global protein movements in approximately 70% of cases (421, 432), agreement between these methods can be used as a measure of validation for the molecular dynamics simulations. In addition, it is apparent that in the NM-RhoA, the extent of flexibility is highly restricted (see Figure 0-6C and Figure 0-6D). Qualitatively the motions remain quite similar, with the exception of movement in regions near the active site which will be discussed below.

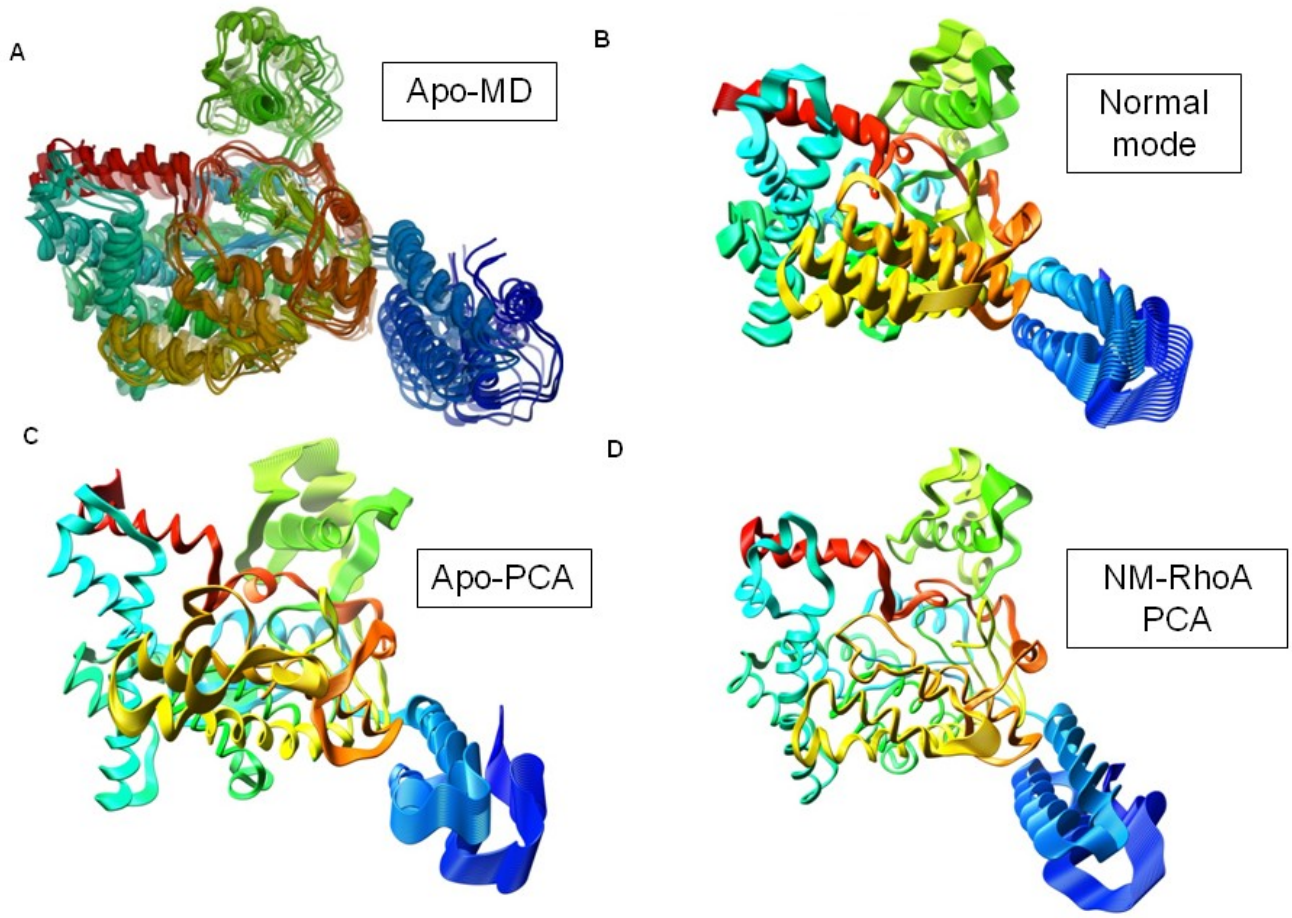


Figure 0-6: Comparison of general motile features of TcdB analyses and simulations

All structures are colored by rainbow per residue to allow better comparison between structures. A) Superposed frames representing various conformations in the Apo simulation, transparency indicates progression through the simulation. B) normal mode structures of TcdB in the apo form. C) First principal component of the Apo simulation. Degree and direction of displacement is shown by broadened ribbons. D) First principal component of NM-RhoA simulation.

Upon visual inspection the primary normal mode shows considerable similarity to the principal component motion of both the Apo-TcdB and NM-RhoA simulations throughout both trajectories, as can be observed by comparing panels B, C and D from Figure 0-6. It should be noted that the degree of motion is less pronounced when the protein is in contact with RhoA. This result is expected since there is a physical object impeding flexibility. Also, the second principal component, represented by the wagging of the upper promontories comprises a larger fraction of the variance in the Cartesian motions of both simulation

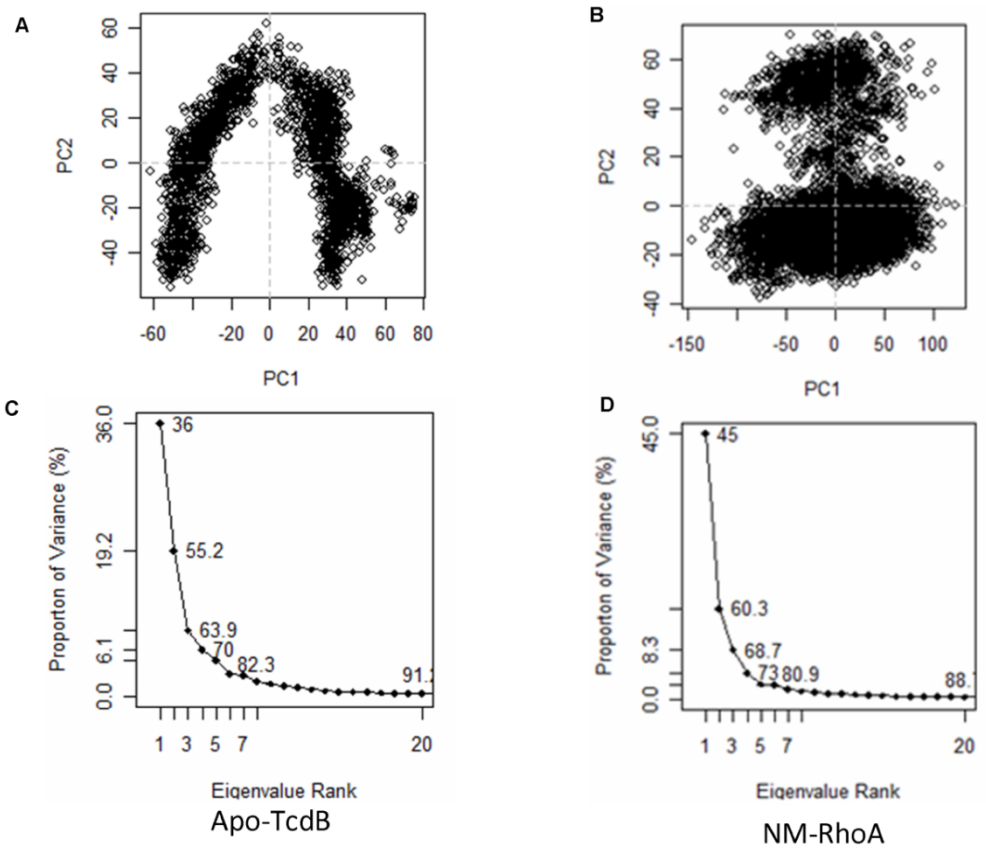


Figure 0-7: Crossplots and Breakdown of Variance for Apo-TcdB and NM-RhoA simulations

Panels A and B are the crossplots of the first and second principal components of the simulations. Each data point represents a single conformation from the MD simulations, and can be used to interpret the occupancy of the conformational space of a simulation. Panels C and D plots of the proportion of variance to Eigenvalue rank. These indicate relative contributions of the lower order principal components.

It is apparent in panel A that Apo-TcdB has a broad range of conformations available. Panel B shows three clusters of conformations observed during the NM-RhoA simulation, one of which is heavily populated. Plots of the proportion of variance to Eigenvalue rank indicate relative contributions of the lower order principal components. In the NM-RhoA simulation a slightly higher contribution from the primary normal mode is observed relative to the Apo structure. The slight decrease in the contribution from the second principal component in the NM-RhoA PCA analysis indicates that the scissoring motion of the upper promontories is less prevalent.

The primary difference between the Apo-TcdB simulation and the NM-RhoA simulation occurs upon approach of RhoA to the catalytic center of TcdB. In the Apo-TcdB simulation, the active site flap (Figure 0-2 shown in purple) performs a repetitive back and forth motion, never completely obstructing the active site. During the course of the NM-RhoA simulation, the active site flap folds down directly over the TcdB active site, completely precluding access to the catalytic manganese. We interpret this behavior as indicative of the order of binding required for catalysis. In the absence of UDP glucose, the TcdB conformation required for successful RhoA is not accessible, and folding of the active site flap precludes close association. In the presence of UDP-glucose, this folding would not be possible, as the sidechains of the active site flap would run into the bound UDP-glucose. However, the similarities between the simulations indicate that the majority of the large-scale motion of TcdB has been captured, and this may be of interest to those designing RhoA mimics.

2.7.2 Quantitation of RhoA-TcdB contacts

To assess improvements in the protein-protein interface following molecular dynamics, three structures were analyzed. One structure was selected as a representative frame from the

most populated cluster throughout the simulation. The structure of the closest approach between Threonine 37 of RhoA and the catalytic manganese of TcdB was selected, as was the original normal mode docked structure; NM-RhoA. Table 0-1: Quantitation of RhoA-TcdB contacts, lists the total number of interactions, number of hydrogen bonds, hydrophobic, ionic, aromatic-aromatic interactions, and cation-pi interactions. Hydrogen bonds are divided into main chain-main chain, side chain-main chain, and side chain-side chain interactions. The structures of both closest approach and most populated cluster both show improvement in the total number of interactions relative to NM-RhoA. Between the original docking and both MD structures, a shift from side chain-main chain interactions to side chain-side chain interactions occurs. No main chain-main chain hydrogen bonds were observed in any of the structures. A significant increase in ionic interactions is also observed relative to the original docked structures.

Table 0-1: Quantitation of RhoA-TcdB contacts

	NM-RhoA[a]	Closest[b]	Cluster[c]
Total interactions	33	45	42
H-bonds	20	24	20
MC-MC	0	0	0
SC-MC	19	4	4
SC-SC	1	20	16
Hydrophobic	8	5	7
Ionic	0	15	13
Aro-Aro	1	0	0
Cation-pi	4	1	2

[a] Structure of RhoA docked to the most open normal mode of TcdB.
[b] Structure of closest Thr37-Mn approach within NM-RhoA MD simulation.
[c] Structure of representative frame from the most populated cluster of the NM-RhoA MD simulation.

2.7.3 Normal Mode and molecular dynamics correlation

In the simulation of TcdB alone, movements were observed that mimicked the normal mode motions we anticipated. Moreover, in the presence of RhoA, these motions were also observed. To determine how well correlated the normal mode motions were to the simulation, a plot of backbone RMSD from the Apo-TcdB simulation to each frame of the normal mode was created. Periodicity across the simulation is observed, as are fluctuations indicating normal mode motions coming slightly in and out of phase with the simulation

Both a three dimensional landscape and binned heat plot were prepared to visualize the correlation between the normal mode and molecular dynamics trajectories. Figure 0-8 shows the RMSD from the normal mode structures across the dynamics trajectory. Panel A shows RMSD vs. Normal mode frame vs.MD frame, with coloring by RMSD. Panel B is a binned version of this plot where RMSD is plotted as a color scale while molecular dynamics trajectory frame and normal mode frame are on the y and x axes, respectively. This correlation results in a plot where the fluctuations in RMSD can be interpreted as the MD motions going in and out of phase with the normal mode conformations. For example, at roughly frames 25, 50 and 97 within the scaled trajectory, a low RMSD relative to the most open conformation of normal mode (Frame 41 on the x axis) is observed. This indicates that during the course of the molecular dynamics trajectory, Apo-TcdB exhibited a conformer similar to that of the normal mode structure, rebounded from that open conformation, and returned to the same open conformation later in the trajectory.

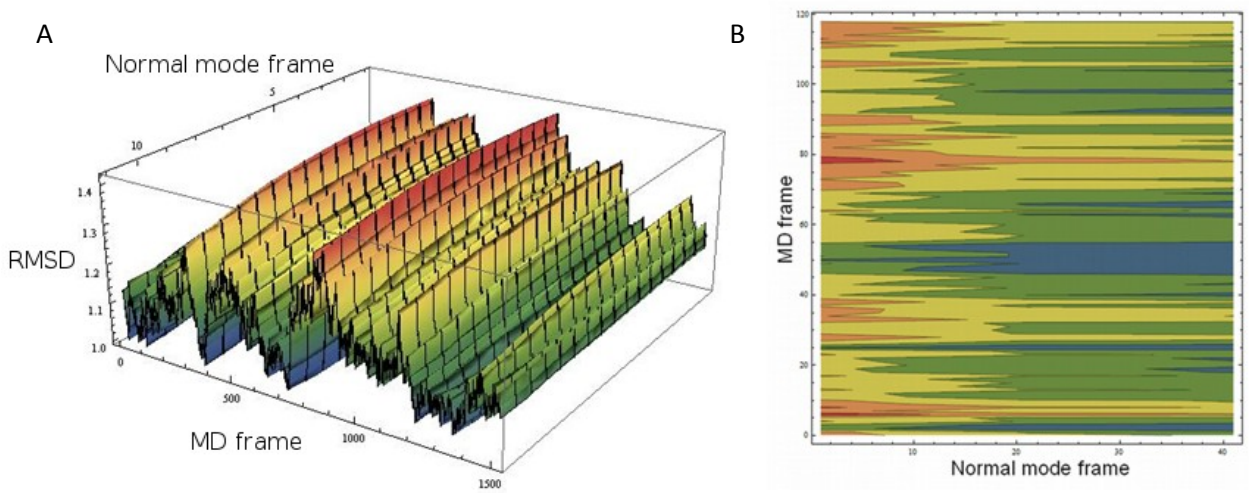


Figure 0-8: Correlation of MD structures to Normal Mode Structures

Normal mode frame is on the X-axis, MD frame is on the Y-axis, and RMSD is shown as gradient from blue (low) to red (high). This arrangement allows observation of the correlated motions between the normal mode and the simulation. As the RMSD becomes low between the various normal mode structures and the MD simulations, occupation of the extremes of the normal mode conformations are observed. The periodicity seen in the plot can be interpreted as Apo-TcdB flexing through the range of normal mode conformations.

2.7.4 Generalized Masked Delaunay analysis

GMD analysis shows the rate of occurrence of significant events over the course of a molecular dynamics simulation. To create a time-dependent contact graph sensitive to large-scale conformational changes the GMD analysis performed utilized a Delaunay tetrahedralization. In this technique, a recrossing filter is applied to remove transient local positional changes that are the result of thermal motion. In order to separate trivial from non-trivial motion, a protein structure is converted into a Voronoi graph in which a representative atom from the side chain of each residue is considered the center of the Voronoi cell. A Delaunay triangulation is the dual graph of the Voronoi tessellation, but as we are in three dimensions, in practicality it is a tetrahedralization. The Delaunay graph, after excluding all faces of significant distance from the representative sidechain atoms and in combination with a recrossing filter, is then used to separate significant persistent motions from trivial recrossings.

The recrossing filter gives several angstroms leeway for atoms to thermally fluctuate across a plane in the Delaunay graph without that motion being considered persistent and significant. When a residue crosses a plane in the Delaunay graph and remains there for the duration of the frame window selected, that motion is considered significant and is reported as an “event” for further analysis.

A plot of events per frame is generated following analysis, where the pattern of detected events in the context of contact making, breaking and total activity can be observed. In our analysis we observed no major folding events, and used the plots for comparative analysis of activity patterns.

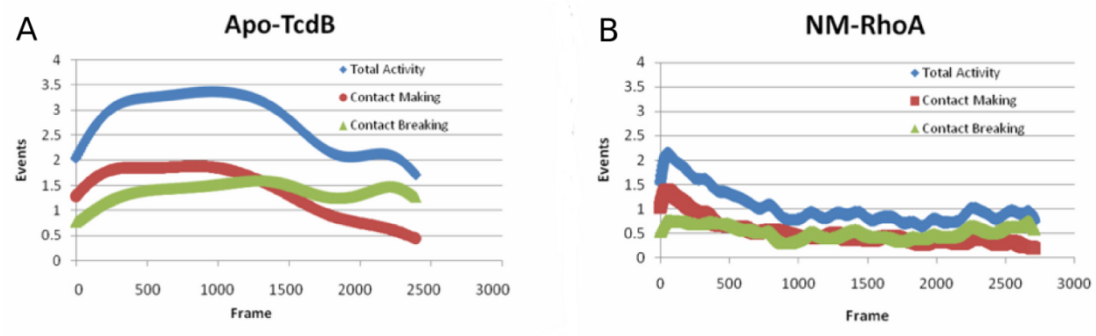


Figure 0-9: GMD plots of Apo-TcdB and NM-RhoA simulations

This analysis plots events per frame through the course of the simulation. Total activity is shown in blue, contact making shown in red, and contact breaking in green. The event pattern indicates that while Apo-TcdB is flexing through its conformational space at a relatively constant pace, the NM-RhoA simulation undergoes a brief period of conformational rearrangement and then persists at a low level of activity through the rest of the simulation.

Figure 0-9 panels A and B are the results of a Generalized Masked Delaunay analysis across the molecular dynamics trajectories of Apo-TcdB and NM-RhoA respectively. Activity is plotted as events per frame, and is decomposed from total activity, shown in blue, to contact making (red) and contact breaking (green). The patterns of activity for Apo-TcdB compared with that of NM-RhoA are markedly different, with Apo-TcdB showing a relatively high level of activity throughout the simulation, while NM-RhoA very rapidly settles down and then exhibits a much lower level of activity throughout the simulation. This can be interpreted as a rearrangement

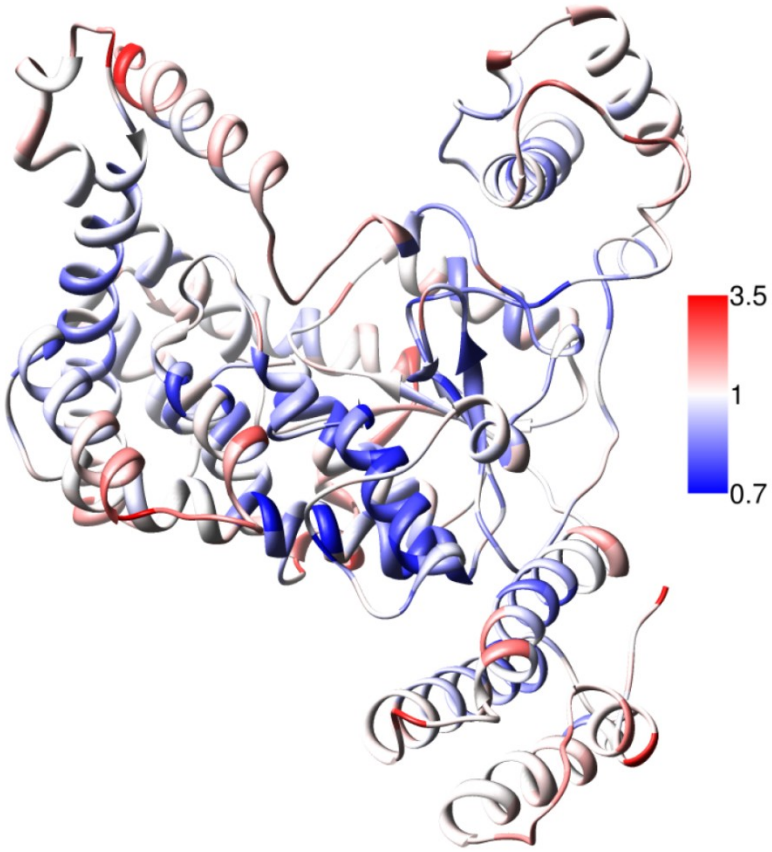
followed by reduction of the available conformational space, or alternatively, a slowing of the transit between available conformations.

Throughout the Apo-TcdB simulation, the number of events per frame as shown in Figure 0-9 does not change dramatically, indicating a steady fluctuation between conformations rather than defined transitions. This can be interpreted as smooth flexion, rather than spontaneous and rapid conformational switches, providing support for the argument that the GT domain of TcdB utilizes a conformational selection mechanism to find its targets. It is likely that TcdB with bound substrate will have access to an alternative range of conformations that affects the movement of the active site flap when in contact with RhoA. While there is some overlap in conformational space of the Apo and bound simulations, the absence of UDP-Glucose precludes formation of a catalytic complex.

2.7.5 RMSF analysis

Over the course of the Apo simulation, major rearrangements have been observed in and around the active site. Both the mobile loop supporting the catalytic center, and the regions responsible for recognition of RhoA appear to be highly flexible. This flexibility is illustrated by the relative RMSF (root mean square fluctuation) as shown in Figure 0-10, representing atomic freedom of motion over the time course of the simulation. It is expected that residues on a protein surface are quite flexible, while interior residues tend to be less mobile (433, 434). The RMSF of TcdB ranges between 0.7Å and 3.9Å. In our simulation both mobile loops near the active site reach RMSF values of near 2Å and thus undergo quite significant motions over time. The flexibility of the active site is unusual but understandable for this protein. Since the toxin must interact with a protein target well known for its conformational switch (435), flexibility near the active site would increase the ability to capture and glucosylate RhoA regardless of the

conformation in which the switch is presented. Detailed analysis of the active site motions from MD simulations of TcdB in complex with UDP-Glc will be discussed at length in Chapter 3.



.

Figure 0-10: RMSF of Apo-TcdB simulation

RMSF was calculated across the Apo-TcdB simulation, and mapped onto the TcdB structure. Ribbons are colored by average atomistic rmsf per residue, from 3.5Å (red) to 0.7Å (blue). High flexibility is observed at the periphery of the protein, while the core of the four helix bundle and RhoA recognition site are stable. The active site flap and mobile loop reach rmsf values near 2Å

2.8 Conclusions

Unbiased long timescale simulations of TcdB from *C. difficile* both Apo and in contact with RhoA were performed. Analysis on these trajectories included GMD, PCA, and comparison to motions observed in normal mode analysis. Large-scale flexibility was observed both in the presence and absence of a protein binding partner without a catalytic binding event being observed. The dramatic rearrangement of the TcdB active site and the consequences for substrate binding point to the possibility that TcdB utilizes a conformational selection mechanism rather than lock and key, or induced fit binding.

Application of normal mode analysis to the crystal structure of TcdB from *C. difficile* correctly captured the large-scale motions of this prototypical glucosyltransferase. The great degree of flexibility of TcdB is both expected and shown in evidence through normal mode analysis and molecular dynamics. A loose fold and considerable flexibility would be practical as the glucosyltransferase domain TcdB must, by necessity, thread through the membrane pore created by the translocation domain. The normal mode conformations bind RhoA moderately well while the crystal structure conformation of TcdB is completely incapable of forming a docked protein-protein complex. While the docking was unable to achieve a fully accommodated form where the toxin has Thr37 fully in the active site, this is a solid step towards determining the manner in which TcdB recognizes the Rho-family GTPases and excludes alternative G-proteins that might be structurally similar but which are not viable substrates.

In simulation, the conformations sampled between the Apo-TcdB and NM-RhoA bound structures are similar with respect to the primary normal modes. PCA plots in Figure 5, indicate that the NM-RhoA and Apo simulations are separately populated, with distinct conformational space occupancy. Taken together, this provides evidence for a conformational selection mechanism, which has been perturbed by Apo-Apo binding. In light of the dramatic alterations in

the active site landscape through the course of the simulations. it is possible that the presence of substrate may shift the conformation of TcdB towards a more suitable orientation for protein-protein binding.

Very recently, high resolution crystal structures for *C. difficile* Toxin A were reported, both alone and in complex with UDP-Glucose (436). These proteins are highly homologous and catalyze the same glucosylation chemistry. Superposition of the TcdA structures shows considerable rearrangement of the active site in both the mobile loop, and active site flap. This has implications for the RhoA binding we observed. During the course of the NM-RhoA simulation, RhoA approach and active site flap orientation were correlated. In the absence of UDP-Glucose, the active site flap motions precluded close approach of RhoA to the catalytic center. In light of the rearrangements observed in the TcdA crystal structures, it is likely that conformational changes initiated by UDP-glucose binding are required before RhoA can be fully accommodated.

It is logical that a protein that seeks out Rho GTP-ases would employ a conformational search mechanism, as Rho GTP-ases are known to employ conformational selection in their binding interactions both with small molecules and macromolecules

The exploration of this non-catalytic binding event has large implications for the kinetics of glucosyltransferase-substrate interactions. As anticipated, flexion in the active site alters substrate binding, and further study will elucidate the consequences of substrate binding on the conformational space available to TcdB. The combination of normal mode analysis, MD and GMD and PCA was shown to be a very effective method for study of protein-protein interactions.

2.9 Methods

Normal mode analysis of the toxin structures in question were performed via the El Nemo (418) web server and confirmed via hinge analysis using the StoneHinge (416, 417) hinge region prediction software. Docked conformations of the Apo-Toxin in contact with RhoA were generated using the RosettaDock (1) server using Hex 4.5 (424) for preliminary conformation generation, and systems were selected for simulation based on proximity to the catalytic binding site.

MD simulations were run using the CHARMM27 (437-440) force field with the NAMD (428) suite of programs on the WSU rocks cluster. The canonical ensemble was maintained via periodic boundaries, Langevin dynamics and thermostat (430). Simulation stability was verified by use of the trajectory analysis tools available with the VMD software (431). Stability was monitored by energy and RMSD. Two systems were prepared and subjected to MD: Apo-TcdB and NM-RhoA.

The Apo-TcdB simulation includes only the TcdB structure, while the NM-RhoA simulation contains TcdB and RhoA in a putative docked conformation following protein-protein docking as described above.

The systems were solvated with TIP3P water, neutralized with counter ions and subjected to 1000 steps of conjugate gradient minimization and temperature ramped to 300K. The Apo-TcdB simulation contains 543 residues, 28,330 water molecules, and a total of 94,013 atoms. The NM-RhoA simulation contains 719 residues, 30,780 water molecules and a total of 102,970 atoms.

Frames from the trajectories were written every 1 ps. Apo-TcdB was simulated for 300ns and NM-TcdB was simulated for 150ns post minimization. The solvation box includes a 15Å pad on each face of the box. Electrostatics were calculated using the particle mesh Ewald (441), and van der Waals were calculated with a nonbonded cutoff of 8Å and a switching function between

7-8Å. Results were analyzed by use of the GMD method, via the TimeScapes (442) software from the D.E. Shaw research group, as well correlation analysis manually handled by the Mathematica software (443). For the purposes of the correlation analysis, a corkscrew interpolation was applied to the eleven original normal mode structures, resulting in a total of 41 normal mode structures. MD frames were selected evenly throughout the simulation, and pairwise RMSDs were calculated.

Analysis of the protein-protein interface was carried out across three structures using PIC (436). Following clustering, a representative frame from the most populated cluster was selected, designated Cluster 1. The frame representing closest approach between Threonine 37 on RhoA and the catalytic manganese of TcdB, and the NM-RhoA structure described above. Hydrogen bond analysis was broken into two types, side chain-main chain interactions, and side chain-side chain interactions. Main chain-main chain interactions were looked for, but none occurred. Additionally, hydrophobic pairs, ionic, aromatic, and cation-pi interactions were tabulated.

Chapter 3 Development of peptide based inhibitors of TcdA and B[2].

3.1 Background

Clostridium difficile infection is increasingly becoming problematic to treat, due to both the intrinsic antibiotic resistance and the emergence of hypervirulent strains. An opportunistic pathogen, *C. difficile* primarily affects patients taking, or having recently completed, a course of broad-spectrum antibiotics (377). Development of anti-virulence therapies as opposed to antibiotics may be an effective way of mitigating the damage of an infection without inciting further antibiotic resistance(383, 384). Toxins A and B (TcdA and TcdB) are responsible for the bulk of the cellular damage that occurs upon infection, and thus are excellent targets for development of antitoxin therapies. While investigation into both immunotherapy (13), toxin-binding materials (444), and probiotics (445) are making progress, none have been approved for clinical use.

It has been proposed that inhibition of the glucosyltransferase activity of toxins A and B may provide some protection against CDAD(407, 446, 447). Considering the etiology of the toxins discussed in Chapter 2, preventing the final step in cellular intoxication may preclude apoptosis of intestinal cells while a *C. difficile* infection is present. As a proposed treatment, GT inhibitors would be co-administered at the start of broad-spectrum antibiotics to abrogate tissue damage if or when *C. difficile* is contracted. By preventing the cellular death and consequent structural damage, many of the symptoms of *C. difficile* infection will be avoided. From a clinical perspective, decreasing the symptoms of *C. difficile* infection would decrease the amount of patien-to-patient transmission. Additionally, preservation of tissue integrity may provide

[2] Sections of Chapter 3 have been previously published (407,484)

protection against recurrence as *C. difficile* spores will be less likely to remain in the digestive tract.

Early work by S. Abdeen on anti-toxin therapies included a phage display experiment designed to find heptapeptides with potential anti-TcdA/TcdB activity. This experiment was successful and a library of peptides with inhibitory potential was discovered. Due to the way the experiment was performed, it was unclear if the potential lead peptides were inhibiting glucosyltransferase activity, or protein-protein binding. Computational experiments were designed to determine the binding modes of these peptides and improve on their *in cellulo* activity. This chapter describes the successful development of a potent anti-toxin molecule.

3.2 Identification of a library of inhibitory peptides

TcdA and TcdB both target a wide range of Rho GTP-ases, and as such it is possible that they would recognize and bind to a broad array of peptides. To determine if any peptides have affinity for TcdA/B, a phage display experiment was performed. The PhD-7 phage library was queried with an affinity capture method. As shown in Figure 0-1 the phage display query was performed in several steps. The recombinant GT domain of TcdA rTcdA^{540} containing a histidine tag was generated, to allow collection on nickel coated magnetic beads. The phage pool was first pre-cleared against the Ni-NTA beads to preclude selection of phage that bind nickel rather than rTcdA^{540}. The pre-cleared phage were allowed to interact with rTcdA^{540} coated beads, which were then washed. Three rounds of selection were performed with different elution methods. Following each cycle of elution, the recovered phage were amplified in an E. coli host, and used as the phage pool for the next cycle of selection. The first round elution was carried out using EDTA. The following three rounds of elution were performed with RhoA, to select for phage that bind either the protein-binding face or active site of TcdA.

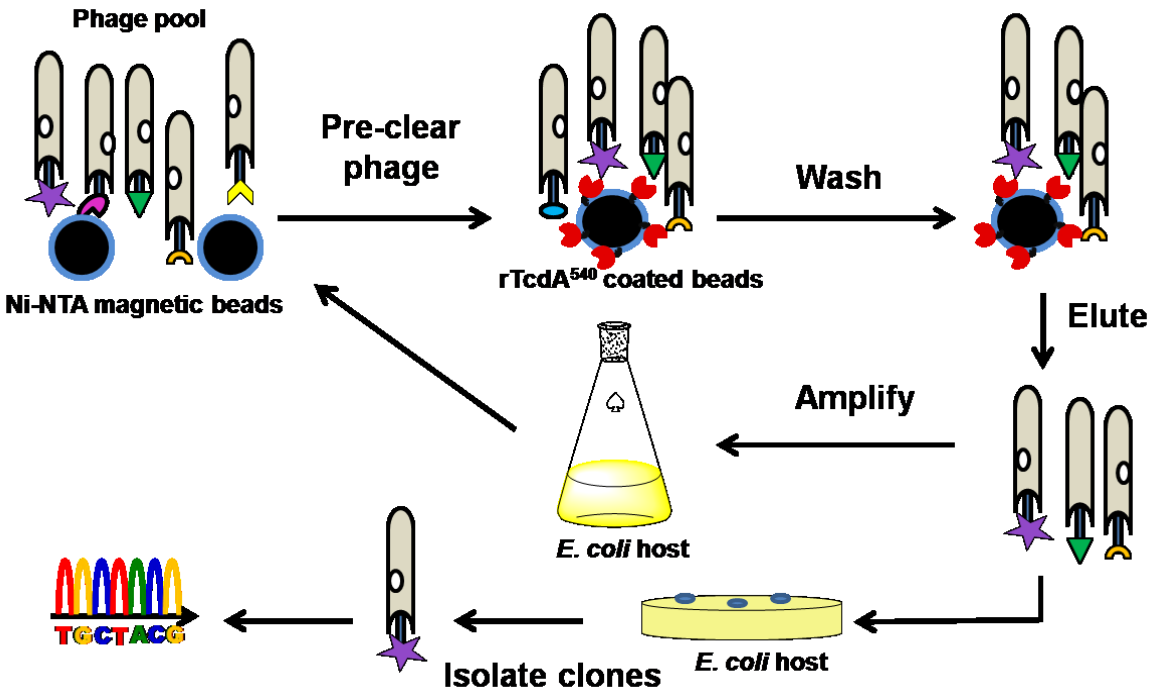


Figure 0-1 Flowchart of phage-display experiment.

A PhD 7 phage library was pre-screened against Nickel-coated magnetic beads. The pre-cleared phage pool was then exposed to beads coated in recombinant TcdA. Following a wash, the bound phage were eluted, amplified in an *E. Coli* host, and the cycle was repeated several times. Following the final round of selection, phage were grown, isolated and sequenced.

Phage were then sequenced and grouped into families by S. Abdeen. This provided an excellent starting point for the computational assessment of peptide binding. Figure 0-2 shows the breakdown of phage sequences into families and subfamilies. A phage-based ELISA assay was performed to determine the apparent Kd for all sequences. Marks to the right of each sequence indicate the range of binding affinity. Red marks indicate binding affinity less than 200 nM; blue, 200-1000 nM; and gray, >1 μM. It is of note that the affinity of TcdA for RhoA is poor, with a Km of over 300 μM(448). Relative to the natural substrate, several of the phage exhibited what would be considered "tight" binding.

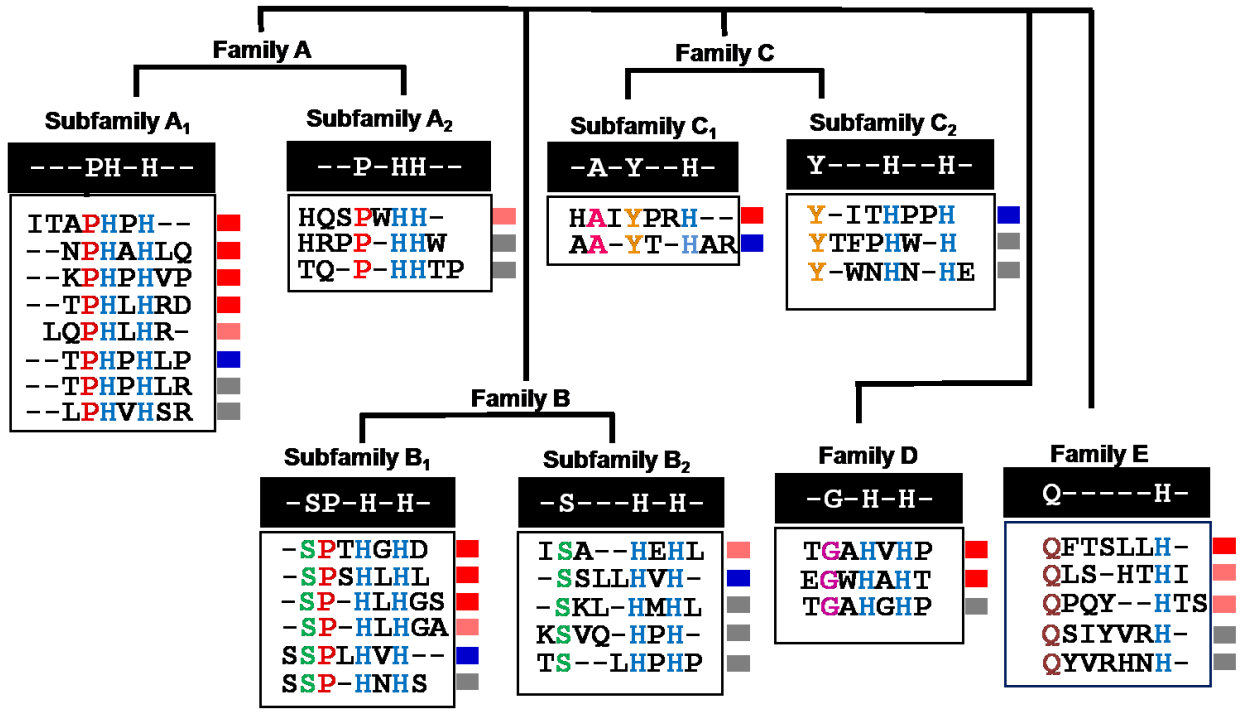


Figure 0-2: Phage sequences from biopanning experiment.

Sequences were sorted into families based on sequence similarity. Bars to the right of the sequence indicate Kd. Red indicates a Kd <200nm, Blue have Kd between 200-1000nm, and Gray bars indicate Kd > 1μM

However, as polyvalent phage were used for these assessments, chelate and avidity effects were a consideration. Before moving forward with peptide binding assays, computational modeling was used to suggest putative binding modes for the peptides.

3.3 Computational studies of inhibitory peptides

To better understand the possibilities of the inhibitory peptides found in the study, several computational investigations were launched. Initial dockings of the peptides to the TcdB crystal structure were carried out. As TcdB is highly flexible and MD relaxed structures were available from the work performed in Chapter 2, we also docked the peptides to several snapshots from the MD. Lastly molecular dynamics studies were performed.

Previous studies (449) showed that a conformational selection mechanism is likely at work in TcdB, and that the presence of a protein binding partner can dramatically alter the conformational space of the toxin. While all peptides in the scan were subjected to initial assessment, two were selected for more in depth analysis. A combination of computational and experimental evidence suggested that the peptides EGWHAHT and HQSPWHH were likely to have the most potential as inhibitors. These two peptides had good docking scores to both the crystal and MD relaxed structure, as well as good *in vitro* activity in RhoA glucosyltransfer assays. EGWHAHT was found to have the highest Kd of all inhibitory peptides in the phage based ELISA assay, and HQSPWHH showed up the most frequently in the sequencing of clones. To determine the effects of substrate binding on the conformational space of TcdB, the structure PDBID:2BVL was simulated in the presence of it's native substrate UDP-Glucose, and these two inhibitory peptides.

3.3.1 Peptide docking with LeadIT

FlexX was selected as the best candidate based on its inclusion of user-definable coordination for metals. Investigation of the peptide binding characteristics of TdcB began following a proof of method docking with UDP-Glucose,. The peptides obtained from the phage display experiments were built using Spartan, minimized at the AM1 level and concatenated into a database file for use with FlexX (338). Active site designation was carried out by selecting all residues within 20Å of any atom in the crystallographic substrates UDP and Glucose. Water visible in the crystal structure within the active site area were allowed full rotation, and were displaceable in the event that a fragment of the docking candidate bound more favorably in a location occupied by water.

The docking results ranked the peptides by favorability of binding using an arbitrary ranking scheme. It was determined that TcdB contains two primary grooves where peptides are capable of binding. The model peptide shown in blue has the sequence SPHLHGS; peptides

binding in this location tend to form contacts with two glutamine residues and an asparagine, thus this nitrogen heavy groove has been termed the N pocket. The peptide shown in green has the sequence EGWHAHT, and the groove in which it binds contains several charged residues, including two glutamic acid residues, an aspartic acid residue, it has been termed the A pocket due to the acidic nature of the residues contacting the peptides that bind in this location.

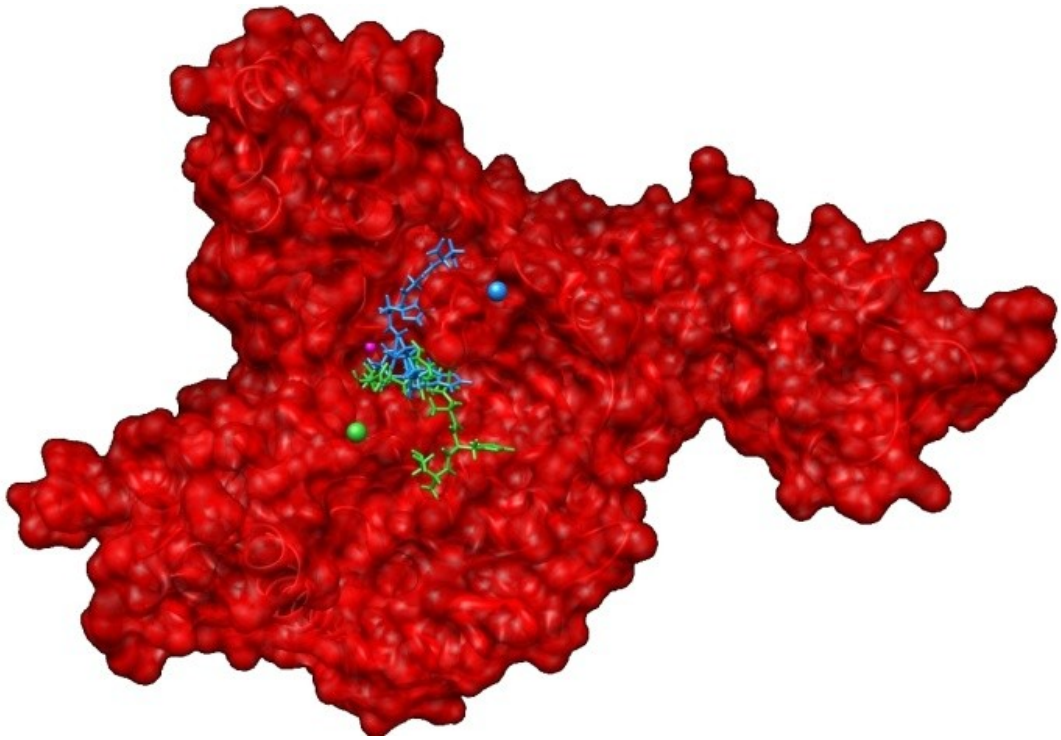


Figure 0-3 Depiction of two peptide binding pockets of TcdB

Representative peptides EGWHAHT (green) and SPHLHGS (blue) shown in their docked conformations to illustrate the two binding pockets of TcdB

EGWHAHT ranked higher than the rest of the peptides, this result is confirmed by the inhibition studies being performed concurrently. EGWHAHT has a Ki of approximately 10^5 pfu and is ranked first among the docked peptides.

Peptides were ranked using a modified Boehm's scoring algorithm, which includes terms for hydrophobic interactions, amphiphilic residue interactions, steric clashes and number of rotatable bonds. includes a comparison of the docking scores with Kd and apparent binding pocket.

Table 0-1: Comparison of docking scores to experimental data. Kd and Ki experiments performed with phage.

Ligand	Score	Kd	Ki (pfu)	Pocket
EGWHAHT	-57.3	100±5	10^5	Green
SPTHGHD	-53	900±400		Blue
QPQYHTS	-49.5	200±40		Blue
NPHAHLQ	-49	115±5		Blue
SPHLHGA	-48.9	105±10		Blue
TPHLHRD	-48.1	530±80		Blue
ITAPHPH	-47.7	840±200		Blue
QLSHTHI	-46.7	500±50		Blue
SPHLHGS	-45.6	770±130		Blue
HQSPWHH	-43.7	330±40	10^3	Both
KPHPHVP	-42.3	145±35		Blue
UDP-Glucose	-40.3	N/A		N/A
QFTSLLH	-36.6	210±10		Green
HAIYPRH	-35	170±10		Green
ISAHEHL	-18.1	480±60		Blue
LQPHLHR	-6.7	100±5		neither

3.3.2 Peptide docking to MD relaxed TcdB

To incorporate information from the molecular dynamics studies discussed in Chapter 2, the above docking protocol was applied to the MD relaxed structure of TcdB. The peptides were built using the Spartan (450) modeling program, minimized at the AM1 level of theory, subjected to fragmentation and docked. Base placement was performed using a triangle-matching scan, which gives preference to base placements that pair complementary functional groups such as Hydrogen bond donor-acceptor pairs. Docking results were ranked using an internal scoring function (451-453). Results from all techniques described here and below were visualized and all images generated using UCSF Chimera(454).

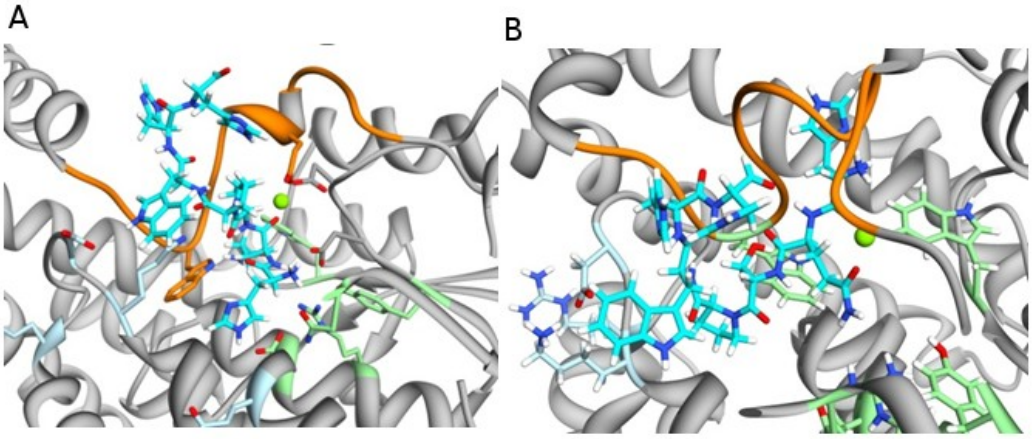


Figure 0-4 Alteration in peptide binding following MD.

Alteration in the conformation of peptide HQSPWHH docked to both the TcdB Crystal structure in Panel A, and the MD relaxed structure in panel B. Active site residues are shown in green. The peptide interacts with a mobile loop shown in orange to a greater degree in the MD relaxed structure.

It was determined that the peptides dock deeper into the active site, exhibit more hydrogen bonding and ring stacking interactions, and their ranking scores better agree with *in vitro* inhibition data relative to the crystal structure dockings.

Table 0-2 Comparison of Crystal and MD docking scores to experimental data

Ligand	Kd (nM)	Ki (nM)	Crystal rank	Crystal Score	MD rank	MD Score
EGWHAHT	100±5	500	1	-57.3	3	-23.6
SPHLHGA	105±10		4	-48.9	2	-35.3
NPHAHLQ	115±5		3	-49	6	-38.8
HAIYPRH	170±10		9	-35	8	-37
QFTSLLH	210±10		8	-36.6	7	-32.5
HQSPWHH	330±40	300	7	-43.7	1	-41.1
ISAHEHL	480±60		10	-18.1	4	-27.3
SPHLHGS	770±130		6	-45.6	9	-39.4
ITAPHPH	840±200		5	-47.7	5	-21.5
SPTHGHD	900±400		2	-53	10	-40.1

We determined from this study that peptide binding is improved when the structural constraints imposed by crystallization are alleviated. Table 0-2 synthesizes the experimental and computational docking information. Peptides are ordered by ascending Kd. For The two peptides selected for lead optimization, Ki values were calculated. The relative rank and score for the dockings to the crystal and MD relaxed structures are also shown. It is noteworthy that EGWHAHT outperformed the other peptides in the crystal docking, while HQSPWHH docked best to the MD relaxed structure. These results suggested that the flexible nature of TcdB may play a role in designing molecules with inhibitory properties.

3.4 Determination of mechanism by Molecular Dynamics and Analysis

While the studies discussed in Chapter 2 and Section 3.3 showed that conformational flexibility in TcdB affects both protein-protein binding and small molecule binding, it was unclear what effects the binding of small molecule substrates would have on the conformational flexibility of the toxin. Interesting behavior in inhibition assays indicated that while EGWHAHT had a tighter Kd, HQSPWHH had a better Ki for both TcdA and TcdB.

Table 0-3 Differential activity of two inhibitory peptides.

Sequence	K_d (nM)	K_i (peptide) nM	
		$TcdA^{540}$	TcdB
EGWHAHT	100 ± 5	500 ± 200	54 ± 20
HQSPWHH	330 ± 40	300 ± 200	18 ± 9

This suggests that even the peptides may inhibit the toxins in different ways, possibly through disruption of the intrinsic motions of the protein. By applying MD followed by GMD and PCA, it was determined that a conformational selection mechanism is likely at work in this system. The specific consequences of substrate presence in the active site were determined with respect to the subsets of conformational space contingent on substrate binding. To facilitate interpretation of the results from these experiments, an overview of the structure of TcdA and B and relevant domains for analysis is shown in Figure 0-5

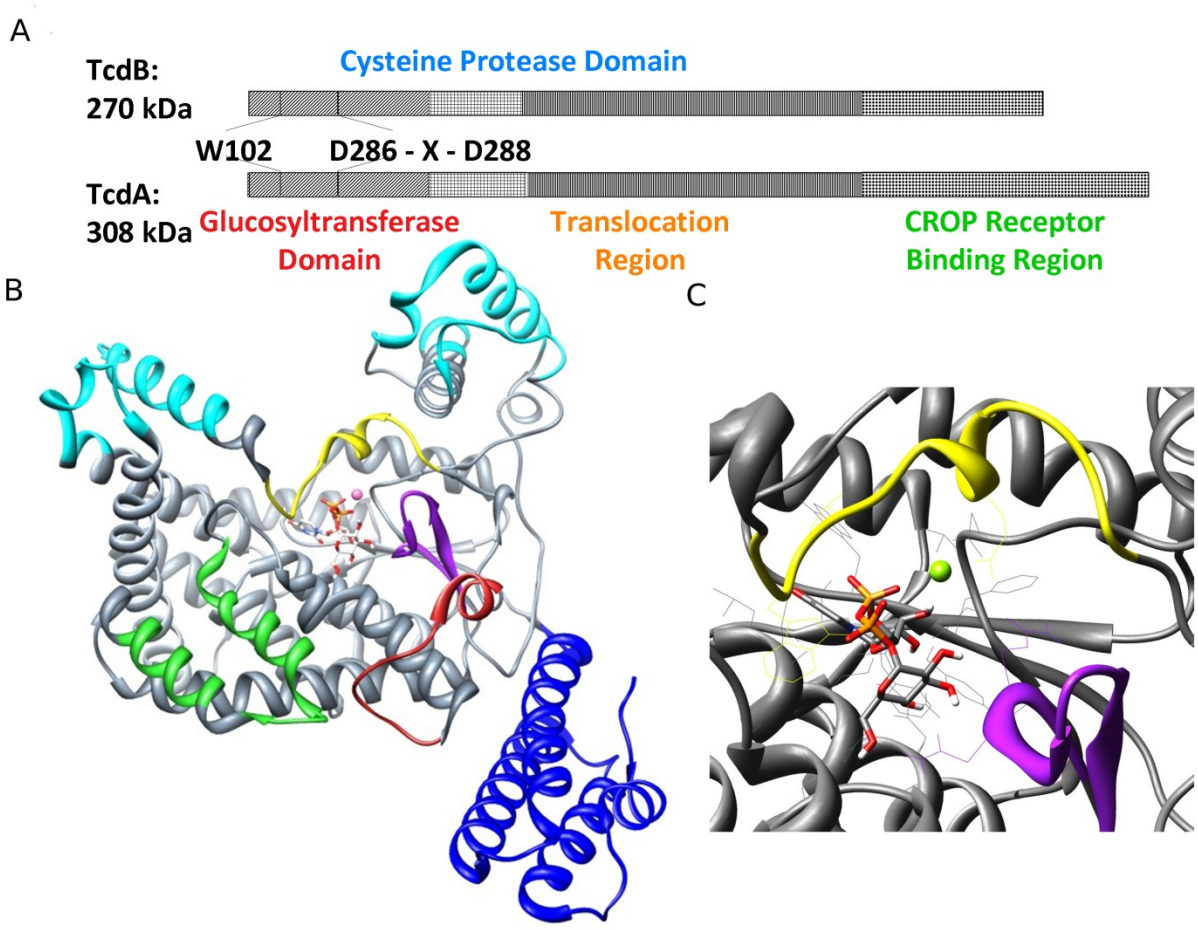


Figure 0-5 Domain organization of *C. difficile* toxins, structure of *C. difficile* Toxin B glucoysltransferase domain (TcdB).

Panel A:Toxins A and B share a common domain organization, differing in the size of the CROP receptor binding region. The glucosyltransferase domain is cleaved from the translocation and CROP domains by the cysteine protease domain upon endocytosis. Panel B: Structure of the glucosyltransferase domain. Toxin specific upper promontories shown in cyan, DXD supporting mobile loop shown in yellow, active site flap shown in purple, Protein-protein recognition loops shown in green and red, and N-terminal four helix bundle shown in blue. Panel C. Inset showing orientation of UDP-glucose in the active site, relative to the mobile loop and active site flap.

A protein that employs a conformational selection mechanism occupies a large conformational space, which is then restricted or modified by interactions with its substrates or binding partners (455). To understand known substrates and develop novel binding partners such as inhibitors, it is necessary to recognize the malleability of the active site and thoroughly understand the consequences of each internal motion and interaction. One avenue to evaluate the available conformational space and understand how it is affected by substrate binding

involves using a combination of molecular dynamics, General Masked Delaunay (GMD) analysis and principal component analysis (PCA). Long-timescale unbiased molecular dynamics allows us to sample a the conformational space available to a given protein, without biasing the population density. Utilization of GMD analysis allows us to pinpoint significant transitions between conformations or clusters of conformations, without relying solely on clustering or PCA (442).

3.4.1 Simulation of Apo, UDP-Glucose and peptide bound TcdB

Four simulations were carried out to study the inhibitory peptides EGWHAHT and HQSPWHH. For the purpose of the description below we will refer to EGWHAHT as peptide 1 (P1) and to HQSPWHH as peptide 2 (P2). TcdB was also simulated in the Apo conformation, and bound to its native substrate UDP-Glucose (UPG). Peptide-bound conformations were created by docking using LeadIT, and then performing MD simulation of the docked structures according to the protocol described in the Methods section. All simulations were carried out for 75 ns under unbiased conditions. Analysis was performed using PCA and GMD. All simulations completed normally, and observation of root mean square deviation (RMSD) and total energy indicated that they were continuously stable. The docked structures are shown in Figure 2, where EGWHAHT is shown in panel A in red and HQSPWHH is presented in panel B in green. Both peptides bind in the active site, interacting with the yellow mobile loop and purple active site flap. The active site conformation shown in the docking is consistent the mass spectrometric analysis of peptides crosslinked to TcdA (407). Following completion of the dynamics a comparison between docking clusters and dynamics peptide conformations was carried out, to verify agreement between both methods.

3.4.2 Clustering analysis

A complete clustering analysis workflow is shown in Figure 0-6. The aim of this comparison was to determine whether or not the docked peptide conformations generated by LeadIT were well represented within the simulation and vice versa. Agreement between the two experiments indicates that the docking succesfully generated a series of low energy conformations, and that the molecular dynamics was run long enough to explore a range of peptide conformers.

All docking results as well as the two molecular dynamics simulations were clustered. To assess the presence of peptide conformations in both the docking and MD simulated structures, a cluster comparison was performed. All docking conformations were superposed on selected structures from the four most populated clusters from the molecular dynamics. In all cases, following superposition, RMSDs were calculated and cluster membership assessed. As shown in Table 0-4, the conformations represented in the molecular dynamics studies are overwhelmingly represented within the top four clusters of the dockings from each state. Backbone RMSDs for all paired structures are <1.1Å (for a visual comparison see

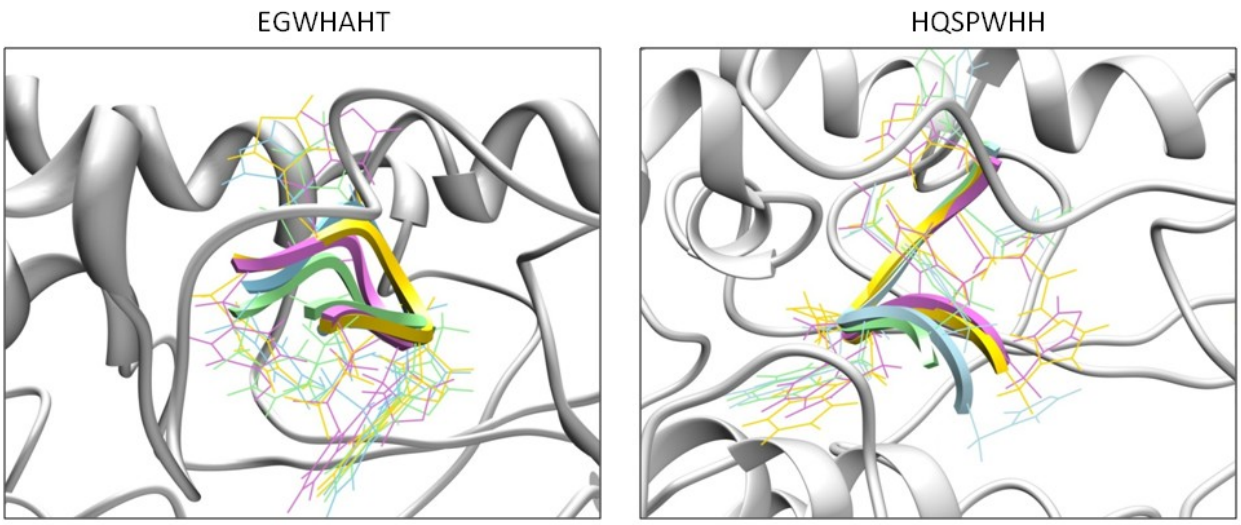


Figure **0-7**). The backbone structure of representative members of the top four clusters from the molecular dynamics is shown as a block ribbon, while the side chains are shown as wire.

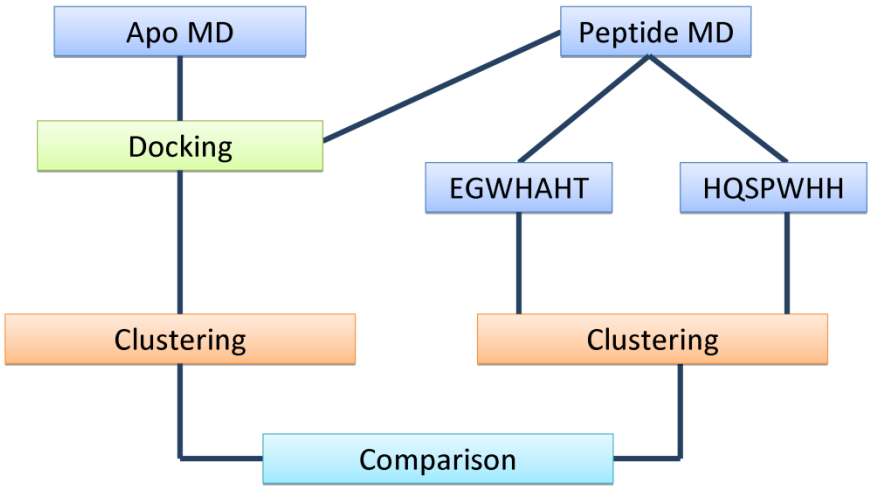


Figure 0-6 **Workflow of the clustering comparison.**

Following Apo MD, peptide docking was performed. The peptide bound structures were simulated, resulting structures were clustered, and comparison between the docking and the molecular dynamics was performed.

Table 0-4 Comparison between the MD and Docking clusters

Following superposition, RMSDs were calculated, and the cluster to which each structure belonged was identified. There is strong agreement between the molecular dynamics structures and the docking clusters.

EGWHAHT								
Crystal			**60 ns**			**80 ns**		
MD cluster	**Docking cluster**	**RMSD (Å)**	**MD cluster**	**Docking cluster**	**RMSD (Å)**	**MD cluster**	**Docking cluster**	**RMSD (Å)**
1	3	0.849	1	2	0.969	1	3	0.849
2	1	0.986	2	1	0.804	2	1	0.986
3	2	1.028	3	2	0.902	3	1	0.969
4	2	1.013	4	2	1.013	4	2	1.013
HQSPWHH								
Crystal			**60 ns**			**80 ns**		
MD cluster	**Docking cluster**	**RMSD (Å)**	**MD cluster**	**Docking cluster**	**RMSD (Å)**	**MD cluster**	**Docking cluster**	**RMSD (Å)**
1	4	0.919	1	1	0.941	1	3	0.818
2	2	1	2	3	0.958	2	1	1.05
3	2	1.042	3	1	0.76	3	1	1.099
4	2	0.93	4	1	0.754	4	2	0.959

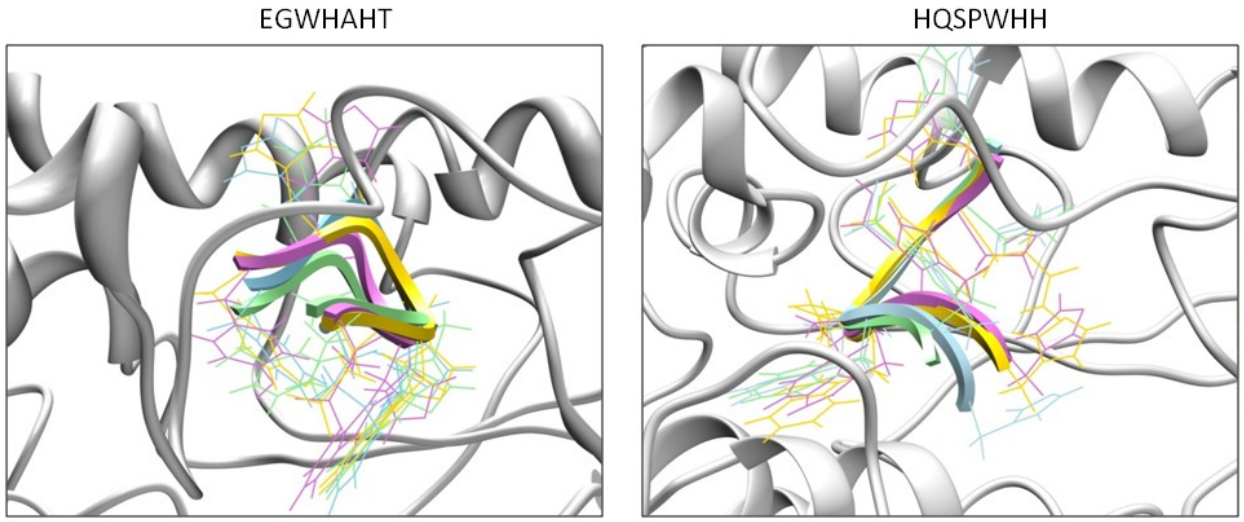


Figure 0-7 Visualization of the MD clustering results.

Representative members of the top four MD clusters from the peptide bound simulations. Clusters 1-4 are shown in blue, green, pink and gold respectively. Backbones are represented as block ribbon and sidechains as wire. In both simulations, good clustering was observed, and when compared to the docking clusters described in table S1 and S2, a high degree of similarity was present. We interpret this as good agreement between the docked conformations selected for study, and the conformational cohort described by the molecular dynamics

3.4.3 Solvent interaction analysis

An analysis of hydrogen bonding and salt bridges was performed to look for solvent interactions and other significant contributions to the stability and coordinated motions of the protein. All interactions present in more than 90% of the frames were subjected to further analysis and are listed in Table 1. While the overall number of H-bonds fluctuates from frame-to-frame, solvation of the active site behaves differently. Hydrogen bonds related to the regions described above have been tabulated separated. The "active site" for the purposes of this analysis was defined in the same way as it was for the docking. The Apo and P1 bound simulations show higher numbers of H-bonds overall, while the UPG and HQ bound simulations show fewer interactions. In all simulations, one water molecule remains stationary, interacting with residue E472 on the TcdB-RhoA recognition face. In the UPG and P1 bound simulations, no stationary waters are observed in the active site. The Apo simulation contains one active site water, and the P2 bound simulation contains two. Hydrogen bonding is observed between solvent water and residue D286 of the DXD motif in both cases. This indicates that P2 is not interacting with the active site in the same way that P1 is, and that P2 preserves the active site hydration observed in the Apo simulation. Salt bridge analysis echoes these results with a higher overall number of salt bridge interactions in the Apo and P2 simulations, and fewer in the UPG and P1 simulations. Again, this reiterates that P1 is mimicking the UPG bound behavior, while P2 is not.

Table 0-5 Interaction analysis of MD simulations.

Interactions	Apo	UPG	P1	P2
Total solvent H-bonds	81	77	87	74
Active site solvent H-bonds	1	0	0	2
Protein binding interface solvent H-bonds	1	1	1	1
Salt bridges	77	58	58	74

3.4.4 Principal component analysis

Principal component analysis was performed to determine the effects of peptide binding on protein structure and flexibility. Following simulation, the principal components of each trajectory were extracted and plotted along with the contribution of each eigenvalue to the total variance as shown in Figure 3. It is apparent that the binding of the three substrates (UPG, P1 and P2) each has an effect on the conformational space that TcdB explores. In columns 1 and 2, principal component structures are overlaid for each simulation. A widened ribbon in these plots indicates motion, whereas narrow ribbons indicate residues that remain relatively stationary. Column 3 contains the cross-plots of principal components 1 and 2 (PC1 and PC2), in essence, giving a two-dimensional representation of the conformational space that the protein structure is occupying. Since the simulations were projected onto the same core residues for the PC decomposition, all plots in column three are comparable. Column four breaks down the variance in the simulation into contribution by each individual eigenvalue, i.e. the point with the highest proportion of variance is principal component one. The distribution of points along the plotted line indicates the relative contribution to the overall motion from each component eigenvalue.

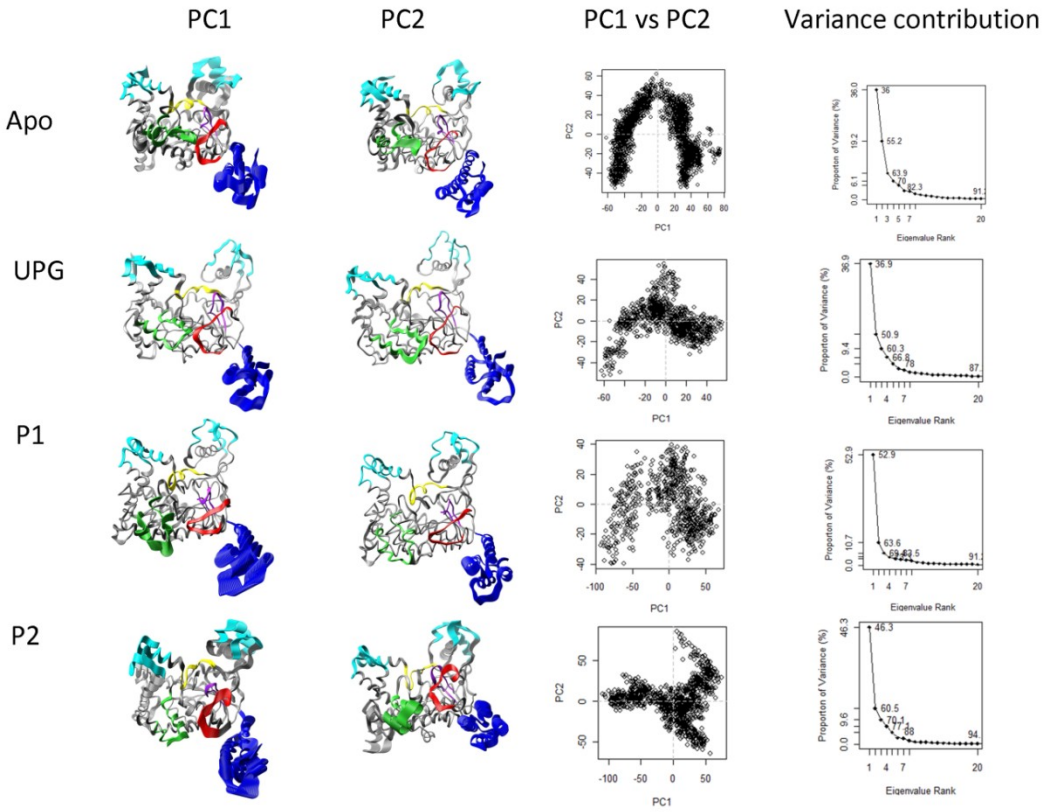


Figure 0-8 PCA analysis of MD simulations; Apo-TcdB, UPG, P1 and P2 bound

Simulations are organized by row, analyses by column. Principal component 1 of each simulation is shown in column 1, principal component 2 is shown in column 2. All structures are colored as in Figure 1 for comparison. Crossplots of the first two principal components are shown in column three. PC1 and PC2 are plotted on the X- and Y-axes respectively. Column four presents the contribution of all calculated principal components to the total variance as a percentage. Proportion of variance is plotted against eigenvalue rank to allow assessment of the relative weight of each component

The Apo simulation shows a high degree of flexibility in both mobile loops, as well as considerable "wagging" of the four-helix bundle at the N-terminus of the structure. The first principal component, PC1 largely describes this motion, while the second captures the side-to-side scissoring of the two promontories shown in cyan (described in Figure 1). The cross-plot of these two principal components shows an organized set of conformations connected by smooth transitions. The relative contribution of these principal components shows that 36% of the

variance in conformation is captured by the wagging motion of the four-helix bundle, and 19% by the scissoring motion. All other motions are captured in lower rank eigenvalues.

The UPG-bound simulation shows less flexibility than the Apo simulation, which is to be expected upon binding of a natural substrate. PC1 is again the wagging of the N-terminal four-helix bundle, while PC2 is a distributed motion, not specific to any single region. The cross-plot of principal components shows contraction of the conformational space, particularly with respect to the second principal component. The proportion of variance between the first and second eigenvalues is comparable to the Apo structure.

Peptide 1 binding appears to induce modifications in both the nature of the principal components and the distribution of variance. PC1 again is the wagging of the four-helix bundle, but the motion becomes exaggerated relative to both the Apo and UPG-bound simulations. PC2 is very similar to that of the UPG bound simulations, with very little motion in the active site apparent in either principal component. The cross-plot of principal components shows a pattern that appears to be somewhat intermediate between the cross-plots of the Apo and UPG-bound structures described above. We interpret this result as an indication that peptide 1 is inhibiting TcdB by mimicking UPG to a great extent. It has previously been shown by Abdeen et al. that Peptide 1 is competitive with UPG and can be displaced at high concentrations of UPG (407). Interestingly, the proportion of variance of the first principal component is considerably higher than for either the UPG or Apo structures, with 52.9% of the variance in the conformational space due to the wagging of the four helix bundle and moderate motion around the active site.

The simulation of peptide 2 bound to TcdB shows considerable alteration in both the principal components and distribution of variance. PC1 shows motion in the upper promontories, as well as the RhoA recognition site, something unseen in prior simulations. PC2 indicates scissoring of the promontories, albeit in a different direction than observed previously.

Additionally, major rearrangements of the RhoA recognition site are observed. The cross-plot of the first and second principal components bears little resemblance to any of the other simulations, and the contributions to the variance are moderately distributed. Abdeen at al. previously showed that peptide 2 is not competitive with UDP-glucose, thus using a distinctly different mechanism for inhibition. This evidence suggests that TcdB uses a conformational selection mechanism (449) and that deformation of the substrate binding site, rather than direct substrate competition is sufficient to achieve inhibition. This avenue for inhibitor may be effective since avoiding direct competition with a natural substrate is desirable to achieve maximum efficacy. The sum of these analyses leads us to believe that peptide 2 inhibits TcdB by perturbing the RhoA binding site, rather than by mimicking UDP-Glucose while peptide 1 represents a relatively classical competitive binding mode of inhibition.

Analysis of the simulations by PCA indicates that the UPG and peptide bound conformers have restricted flexibility relative to the apo conformation as expected. However the inhibitory peptides display differing behaviors with respect to their conformational restriction. As was previously shown by Abdeen *et al*, peptides show inhibitory potential by either interfering with RhoA or UPG binding (407). The spatial freedom exhibited by the apo toxin indicates that it is likely that TcdB is subject to a conformational selection and induction mechanism similar to that of the small G-proteins (419). Since TcdB must recognize RhoA and it is undergoing significant motions in response to its own conformational selection process, it makes logical sense that its binding partners, TcdB in this case, might also exhibit similar conformational selection behavior. The dramatic perturbation of the conformational space of TcdB upon contact with the inhibitory peptides, may illustrate a good way to identify proteins involved in conformational selection and study the way they interact with their substrates and/or targets.

3.4.5 Local RMSD analysis

To quantitate these results, total as well as local backbone RMSDs of the regions described in Figure 1 were calculated and are shown in. The RMSD overall for the Apo simulation is higher in all regions with the exception of the two protein-protein interface regions, which are dramatically perturbed in the P2 bound simulation. This result is in agreement with the PCA analysis where perturbation of these regions was a major component of the motion. The UPG bound simulation shows a low RMSD in all regions, again in agreement with the PCA analysis. The P1 bound simulation shows some increased movement in the protein-protein interface regions, while both the mobile loop and active site flap behave similarly to the UPG bound simulation. The combination of the qualitative PCA with the quantitative local RMSD breakdown shows good agreement.

RMSD in angstroms was calculated for the overall trajectory, as well as for each region described in Figure 1. The Apo simulation shows the highest overall RMSD as well as generally higher local RMSDs for all regions with the exception of the two protein-protein interfaces. The green and red binding sites are greatly perturbed by the binding of the P2 peptide with their RMSD reaching over 3Å. The UPG bound simulations shows the lowest overall RMSDs with low activity in the mobile loop and active site flap. P1 shows similar behavior in the active site, while still perturbing the protein-protein binding sites to some extent

Table 0-6 Overall and Local RMSDs of MD simulations

RMSD (Å)	Apo	UPG	P1	P2
Overall	2.475	1.294	1.798	2.101
Promontories	2.09	1.746	1.759	1.967

Mobile Loop	1.297	1.247	1.21	1.458
Active Site Flap	2.509	0.964	0.989	1.172
Binding Site Green	1.237	1.297	2.464	3.42
Binding Site Red	1.058	1.216	2.217	3.105
Four Helix Bundle	2.09	1.629	2.464	2.147

3.4.6 Generalized Masked Delaunay analysis

Generalized Masked Delaunay analysis was used to determine the effects of substrate binding on the relative rate of conformational activity. GMD assesses the "activity" of a simulation, by creating a masked Delaunay representation of the protein structure, and using it to determine how frequently side chains exhibit significant and persistent motion (16). As this is a "sliding window" analysis, the presented data has been truncated to remove artifacts that occur at the end of these plots. We examined the activity pattern across all four simulations, to determine what effect each substrate had on the rate and degree of activity of TcdB. Figure 0-9 shows the results of GMD analysis on the 4 simulations. Column 1 shows plots of activity vs. scaled frame, all simulations were scaled to 5% of the total frame count for the finished simulation. Columns 2 and 3 show the decomposition of the activity into contact forming interactions and contact breaking interactions.

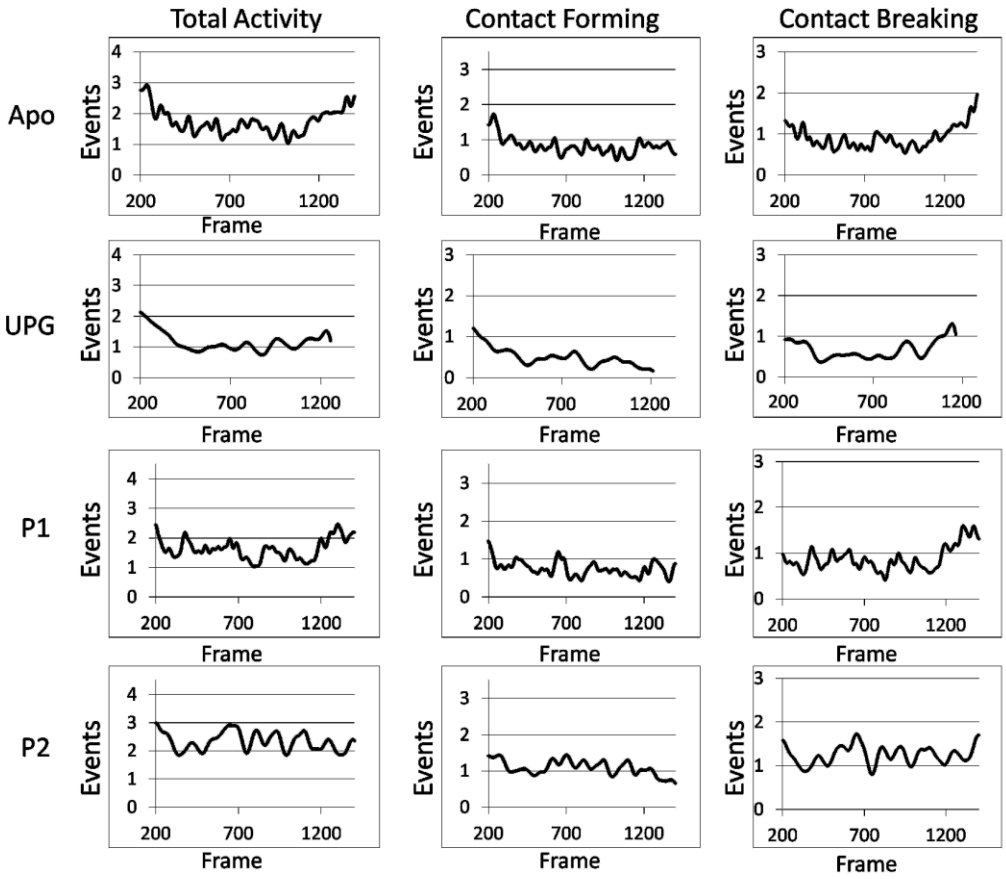


Figure 0-9 GMD analysis of MD simulations; Apo-TcdB, UPG, P1 and P2 bound.

As in the PCA analysis, simulations are organized by column, analyses are organized by row. Column 1 shows the total activity of each simulation as a function of frame. The count of significant persistent events are plotted on the Y- axis, against simulation progress on the X-axis. Columns 2 and 3 decompose the total activity into contact-forming events and contact-breaking events.

The Apo simulation shows a high level of flexibility throughout, along with a rapid event pattern. The decomposition of activity shows somewhat more contact breaking than forming, as well as a more steady level of contact breaking. The high number of activity spikes may be indicative of continuous conformational transitions, with low points in activity representing conformations amenable to the approach of a binding partner or substrate. Visual inspection of the trajectory shows a repetitive breathing motion that opens the active site somewhat, possibly accounting for the higher level of contact breaking.

The UPG bound simulation exhibits a dramatically different activity plot. Following an initial rearrangement, activity steadily declines, until reaching a relatively steady state roughly halfway through the trajectory. A few slow moderate rises in activity occur following this point, but the overall rate of activity remains moderate. The plot of contact forming events shows a steady decrease, suggesting that the bulk of the conformational change involves some degree of unfolding. Analysis of the conformational trajectory agrees with this interpretation. The active site flap moves away from the catalytic center, presenting the region near the manganese ion for catalysis, presumably to allow binding of a glucosylation partner such as RhoA. It has been previously shown that in the absence of UPG, the active site flap folds down, precluding protein-protein binding (449). The level of contact breaking throughout the UPG simulation shows an initial increase, likely associated with the initial rearrangements due to UPG binding.

Analysis of the P1 simulation shows the same initial rearrangements observed in both the UPG and Apo simulations, with an event pattern intermediate to both. A decrease in activity is observed, similar to the bound form with UPG, but remains at a higher level overall, roughly 1.5-1.75 events per frame. The number and frequency of activity spikes are also intermediate to the Apo and UPG-bound simulations. Interestingly the shape of the contact forming and contact breaking plots is similar to the Apo simulation, but with an increase in the number and frequency of activity spikes, similar to the Apo simulation. This seems to indicate that P1 is in some way acting as a UPG mimic; upon P1 binding, TcdB takes on activity characteristics of the UPG bound toxin.

The P2 simulation has an activity pattern disparate from all other simulations. The overall level of activity is higher, with smooth transitions between regions of high and low activity. This pattern is not seen in any of the other simulations. The level of contact forming is somewhat lower than the level of contact breaking overall. More rapid transitions in the level of contact making are apparent in the second half of the simulation, while no such pattern is

observed in the contact breaking activity. The distinct alteration in the event pattern may in some way be contributing to P2's ability to inhibit TcdB. While P2 is bound, it appears to disrupt the native pattern of conformational searching, but in an entirely different way than P1. No similarity is evident between the P2 bound and UPG bound simulations, suggesting that P2 is not acting as a UPG mimic.

3.5 Computational functionalization of peptide leads

Functionalization of HQSPWHH was considered necessary, while the peptides were inhibitory *in vitro* they did not provide cell protection in cell assays. The etiology of the toxin is such that we propose that the peptide inhibitors were becoming dislodged during the course of endosomal escape. To determine whether or not an irreversibly linked peptide could provide cell protection, initial studies with a photo-activatable crosslinker were performed, and cell protection was observed. However, as photo-activation of an inhibitor is at the very least inconvenient in a clinical setting, we chose to pursue other methods of irreversible binding. Inclusion of an epoxide in the structure was selected, as nucleophilic attack would result in a covalently linked inhibitor.The studies described in Section 3.4 indicated that of the two peptides chosen for study HQSWPHH was a better choice for functionalization. EGWHAHT mimics UDP-Glucose in our studies, and we considered the risk of off-target effects problematic, however initial functionalization studies were still performed. Alanine scanning was performed to locate sites where modification was possible, followed by epoxide scanning. As *in vitro* work included assays on TcdA, docking was carried out against both TcdA and TcdB structures. Once a lead molecule was selected, synthesis and testing were carried out.

3.5.4 Alanine and epoxide scanning

In silico scanning was performed to determine the optimal site for modification based on docking energies to TcdB (PDB: 2BVL) (387) and reiterated in light of recent high resolution crystal structure for *C. difficile* TcdA (PDB: 3SS1) (436). This optimization was done in several

steps. The first step involved incorporation of an alanine at each position to determine the relative contribution of the parent side chain at each site. It was determined that positions 1 and 5 of EGWHAHT and positions 1, 5 and 7 of HQSPWHH were not detrimental to either binding conformation or docking score. Modifications at positions 2 and 3 of EGWHAHT or 2 and 4 of HQSPWHH perturbed binding, dramatically altered docking score or both. All other positions had intermediate effects. These positions were then assessed for both the ability to sterically accomodate an epoxide moiety, and availability of nearby nucleophiles.

Table 0-7 Docking scores of parent and derivatized peptides and number of surrounding nucleophiles

The two positions chosen for functionalization are highlighted in gray. Selection was made based on docking score and number of nearby nucleophiles

	WT	H-1-X	Q-2-X	S-3-X	P-4-X	W-5-X	H-6-X	H-7-X
Alanine	-36.01	-38.84	-31.93	-29.86	-35.18	-34.74	-35.47	-39.12
Epoxide R	-35.58	-32.53	-15.74	-24.94	-40.76	-32.04	-37.50	-43.38
Epoxide S	-36.04	-29.04	-2.06	-35.34	-20.89	-52.85	-33.01	-24.26
Nucleophiles	-	2	1	2	2	4	2	6

WT: Parent peptide HQSPWHH

X-substituent, either alanine or R/S epoxides

Due to the nature of the docking algorithm, slight fluctuations in scores occur as a consequence of sub-angstrom variations in the docked conformation, predominantly through rotation around bonds in the flexible side chains. While not on an absolute energy scale, this procedure provides reliable relative binding affinities to assess positions where the epoxide might be accepted. A docking score of greater magnitude than the parent peptide was considered an advantageous modification, while a docking score lower than the parent peptide was considered disadvantageous.

Peptides were selected for derivatization based on two criteria: a) epoxy modified peptides with tighter binding affinity but relatively few reactive nucleophiles nearby or b) with moderate binding affinity but with larger number of surrounding nucleophiles. Two peptides

derivatized at different residues were subsequently selected for synthesis shown in Figure 0-10, both with docking scores more favorable than the parent peptide (parent docking score, -35.58): HQSPGepoxyHH (H-epoxy-5) (Figure 0-10 Epoxidated peptide structures and conformationsA and C) and HQSPWHGepoxy (H-epoxy-7) (Figure 0-10 B and D). One reason for selecting H-epoxy-5 was that it displayed the overall tightest docking score, while the rationale for selecting H-epoxy-7, despite a more modest docking score, was due to the higher number of potential nucleophiles in its vicinity that might facilitate rapid cross-linking. In all cases the structures interact both with the active site catalytic Manganese and residues on the highly conserved mobile loop within the active site critical in glucosylation (456, 457). The parent peptide HQSPWHH adopts a curled conformation (407) as seen in Figure 0-4. The N-terminal histidine contacts a active site region comprised of an Asn 270 -Asp 273 pair, while the C-terminal histidines interacts with the charged loop region comprised of residues 513-526 (456, 457). The epoxide of H-epoxy-5 has close contact with Lys142, Leu 265 and Asn 139. The epoxide of H-epoxy-7 is within reach of Lys 452, Asp 523, Asp 461 and Ser 518. Even though the structural (436) and functional similarity between the GTD of TcdB and TcdA makes TcdB a suitable substitute for the purpose of docking, we revalidated the binding modes and surrounding nucleophiles in terms of GTD of TcdA

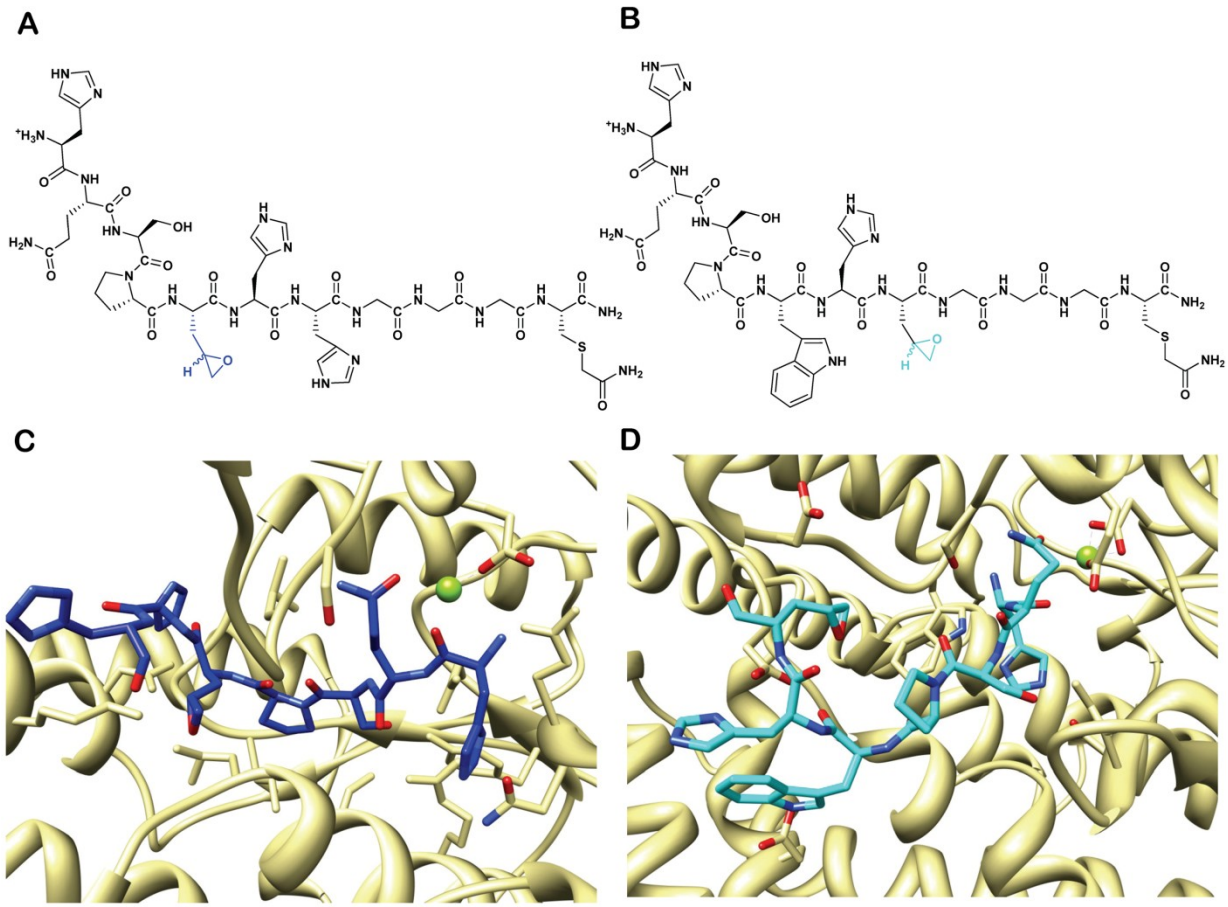


Figure 0-10 Epoxidated peptide structures and conformations.

Structures of selected epoxy-peptides and close-up view of ribbon structure of the TcdB active site, showing binding modes of peptides H-epoxy-5 and H-epoxy-7. Panels A and B show the structure of the epoxidated peptides. Panels C and D show the binding conformations of H-epoxy-5 and H-epoxy-7 respectively.The epoxide residue of H-epoxy-7 is in close proximity to more polar amino acids when compared to H-epoxy-5. The catalytic Manganese ion is indicated as a green sphere in both images

3.4.6 Comparison to TcdA

To verify that the active sites of TcdA and B were similar enough to be confident that our epoxidated peptide would react with both toxins, we compared the structures with respect to the bound peptide conformations. The docked structure of the TcdB-peptide complexes were structurally aligned with the TcdA crystal structure and spatial compatibility and active site similarity were assessed.

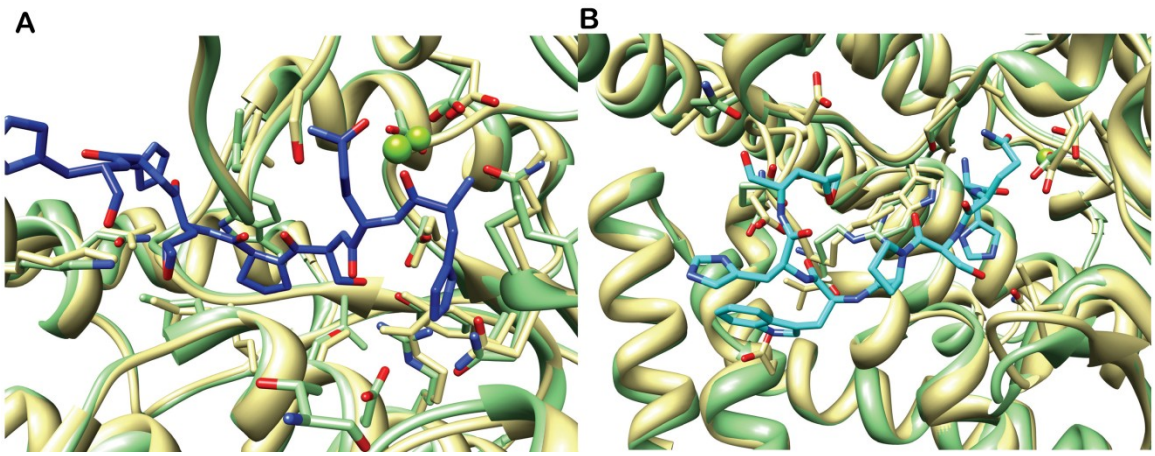


Figure 0-11 Overlay of TcdA on peptide bound TcdB structures.

Panel A shows H-epoxy-5 and panel B shows H-epoxy-7. TcdA is shown in green and TcdB in yellow. The aligned structures are similar in their overall active site organization, with a backbone RMSD of 0.85Å. No steric clashes were observed between TcdA and the epoxy-peptides.

No steric clashes were observed, so we continued to comparison of the sequence of TcdA and TcdB in the active site. o determine the extent of similarity between contacts in the bound conformations, the structural alignment was used to determine which residues are conserved in the putative binding orientation. All residues within 5Å of the peptides in the TcdB conformation were considered "contacts", the structurally cognate residues in TcdA were determined and compared for both the H-epoxy-5 and H-epoxy-7 structures.

Table 0-8 Comparison of residues in contact with H-epoxy-5.

Structurally cognate residues within 5Å of H-epoxy-5 in both TcdB and TcdA are compared. Residues highlighted in green are identical, those in blue are conservative substitutions, and those in red have non-conservative substitutions

H-epoxy-5			
TcdA		TcdB	
ASN	138	ASN	139
LYS	141	LYS	142
LEU	264	LEU	265
SER	268	SER	269
ARG	272	ARG	273
TYR	283	TYR	284
ASP	285	ASP	286
ILE	382	ASN	384
ASN	383	GLN	385
GLN	384	GLU	515
ASN	516	ALA	517
SER	517	SER	518
LEU	518	LEU	519

H-epoxy-7			
TcdA		TcdB	
ASP	288	ASP	287
MET	289	MET	288
ILE	382	VAL	381
ILE	383	ILE	382
ASN	384	ASN	383
GLN	385	GLN	384
MET	448	THR	447
GLU	449	LYS	448
LEU	450	ILE	449
PRO	460	GLU	460
GLU	515	GLU	514
MET	516	ASN	516
SER	518	SER	517
LEU	519	LEU	518
TRP	520	TRP	519
SER	521	SER	520
PHE	522	PHE	521
ASP	523	ASP	522
ASP	524	GLN	523

The conclusions from these comparisons were that the epoxidated peptides would likely have activity in both TcdA and TcdB. There were enough conserved residues in the proposed binding site, including conserved nucleophiles to move forward with synthesis.

3.6 Experimental validation of proposed inhibitors

3.6.1 Synthesis and *in cellulo* testing of epoxidated peptides

Dr. Sanofar Abdeen synthesized the epoxidated peptides by oxidizing allyl-glycine substituted peptides with mCPBA. Details of the synthesis and purification are described in Abdeen et al. (JBC submitted) As the parent peptides were capable of inhibiting TcdA and TcdB *in vitro,* but were unable to protect cells, the epoxidated peptides were tested *in cellulo*. Vero cells were challenged with 600μm TcdA and observed over a 48hr period. Cell viability was quantitated using an ATP-sensitive luminescent assay. All cell viability assays and imaging was performed by Dr. Stephanie Kern.

It was determined that the H-epoxy-5 was capable of providing 95% cell protection *in cellulo* and had no cellular toxicity. No cell protection was detected with H-epoxy-7. We propose that either the dissimilarity in the active site of TcdB and TcdA affected a critical contact in the H-epoxy-7 binding site, or that the overall binding geometry for this complex is more unfavorable than predicted.

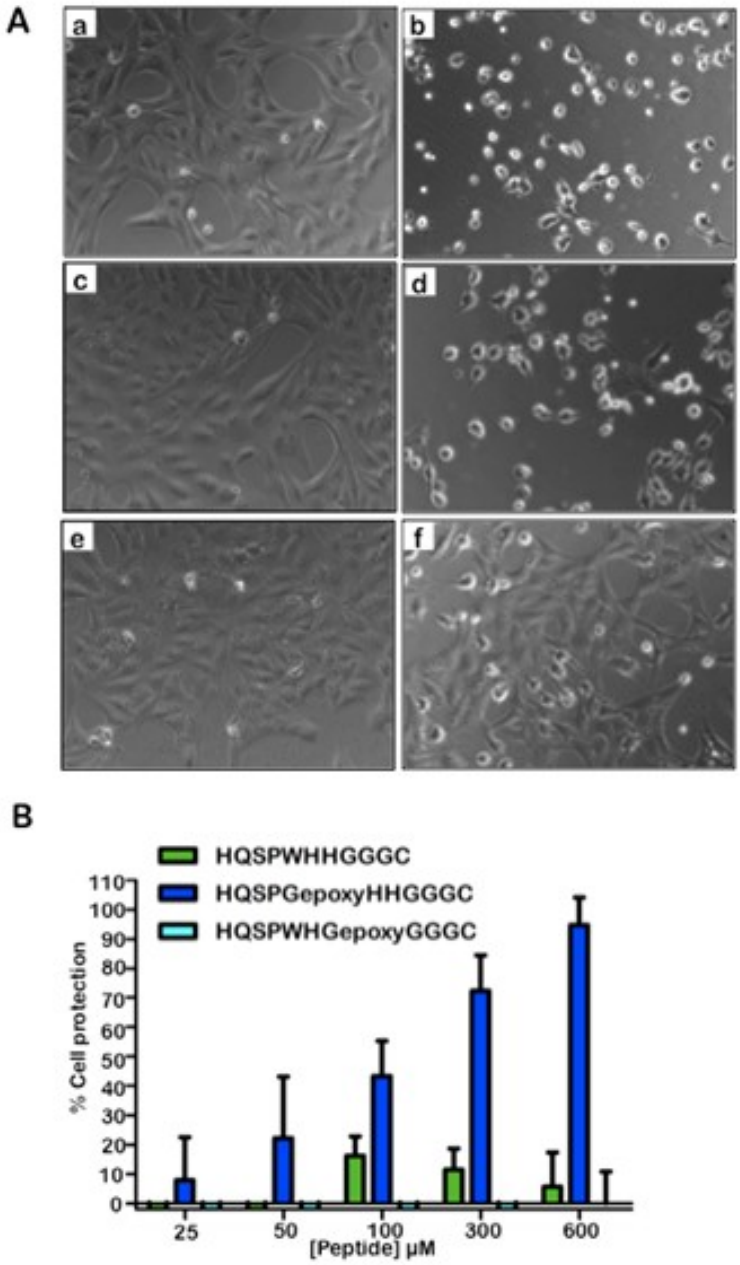


Figure 0-12 Cell protection and viability quantitation.

Panel A shows the morphological assessment of cells under various treatments. Sub-panel (a) shows healthy cells in PBS buffer, cells are elongated and adherent. Following treatment with TcdA, cells exhibit rounding and detach from the surface of the plate (b). Sub-panel (c) shows healthy cells in the presence of the parent peptide. No toxicity is observed. Upon challenge with TcdA, the parent peptide provides no protection and apoptosis is detected (d). Cells imaged in sub-panel (e) were treated with the H-epoxy-5, indicating that the epoxidated peptide is not inherently toxic. Lastly, treatment with both TcdA and H-epoxy-5 in (f) shows a dramatic reduction in observed apoptosis. Cell adherence and elongation are both preserved. Panel B describes the quantitation of these results. Cell protection as a percentage is plotted vs. increasing inhibitor concentration.

3.6.3 Validation of crosslinking site by mass spectrometry

To determine whether or not the epoxidated peptides were acting as designed, or if they were nonspecifically inhibiting TcdA by some other mechanism, mass spectrometry was used to determine the site of crosslinking. Recombinant TcdA was allowed to react with H-epoxy-5, then subjected to tryptic digestion and analyzed via FT-ICR mass spectrometry. Full details are available in Abdeen et al. (JBC in submission). Sequence coverage of roughly 70% was obtained, and the site of crosslinking was structurally mapped.

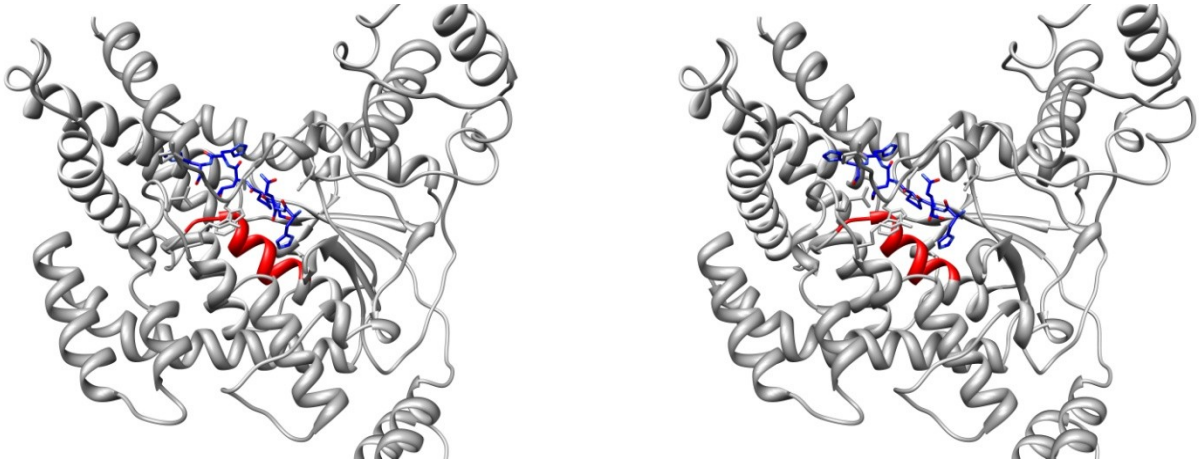


Figure 0-13 Stereoimage of TcdA/B crosslinking site.

The ribbon structure of TcdB in complex with H-epoxy-5 (in blue) is shown as a stereoimage. Shown in red is the TcdA sequence where crosslinking was detected by FT-ICR mass spectrometry. It is clear that the docking and epoxidation studies correctly predicted the bound conformation and crosslink region

From these studies it is apparent that the inhibitory activity was not due to random crosslinking, but that the predicted binding conformation was accurate. The epoxide is positioned directly facing the detected site of crosslinking.

3.7 Conclusions

In this study, we rationally designed an epoxide-containing peptide that acts as an irreversible inhibitor of the clostridial-glucosylating toxins sufficiently potent to protect cultured cells from intoxication. Previously described peptide inhibitors of TcdA and TcdB were found to employ differential mechanisms of inhibition *in vitro*. Upon docking, the peptides were found to

bind tightly to different regions of the active site. Long timescale simulations were applied to investigate possible mechanisms. Application of GMD analysis indicated that the peptides induced significantly different event patterns over the course of the simulations. Application of PCA analysis determined that flexibility in the active site is restricted when UPG and EGWHAHT are bound relative to the Apo structure. Binding of HQSPWHH, which proved to be the best inhibitor of glucosyltransferase activity, greatly deformed the protein-protein recognition face of TcdB. Evidence suggests that EGWHAHT acts as a UDP-Glucose mimic, while HQSPWHH interferes in the conformational selection mechanism. HQSPWHH was selected for further functionalization; computational alanine and epoxide scanning was used to determine which residues would perform best upon epoxidation. Following synthesis, H-epoxy-5 was able to provide 95% cell protection during *in cellulo* assays, and mass spectrometric analysis showed that the peptide covalently crosslinked exactly where the parent peptides were computationally predicted to bind.

3.8 Methods

All simulations were carried out using NAMD (428) with the CHARMM27 (439) force field. It is well known that this force field has a tendency to prefer helical structures (458, 459). Previous long timescale simulations did not show significant formation of helical structures in crystallographically unstructured regions (449), based on these results we decided to continue our studies with the CHARMM27 force field. The apo simulation was prepared by removal of the crystallographic UDP and Glucose as previously described. Bound conformations for UPG, HQSPWHH and EGWHAHT were generated using the LeadIT (338, 460-463) suite for docking. Docking parameters were tested using the crystallographic UDP conformation as proof of method, as published previously (407). All substrates were initially built and minimized at the AM1 level of theory using Spartan '03 (464) and docked into a sphere encompassing all

residues within 20 Å of the catalytic manganese. All crystallographic waters were retained and utilized as both fully rotatable and displaceable. Triangle matching was used for base fragment placement, and dockings were performed with two thousand solutions per each iteration and fragmentation. The standard scoring scale based on Böhm's scoring algorithm (465-467) was employed. Docking was carried out against dynamically relaxed structures following simulation of the apo toxin from a previous work (449) as well as the crystal structure. Following docking to the crystal, 60 ns and 80 ns structures, clustering was carried out.

Force field parameters for UDP glucose were created both de novo from single point calculations and by generalization. UDP-Glucose was built in Spartan and initially optimized at the AM1 level of theory. Restricted Hartree-Fock optimization at the 6-31G* level of theory was performed using Gaussian '03 (463). The optimized geometry was utilized for frequency calculation, NPA and ESP charge fitting. Paratool (431) was used to convert the Gaussian output into CHARMM format parameters for comparison with the manually determined parameters. Parameterization by generalization was performed using the parameters from UTP for the UDP after removal of the terminal phosphate. CHARMM parameters for glucose were readily available and parameters for the sugar-UDP linkage were obtained from those determined for phosphoserine. Comparison of parameters derived from these two methods indicated that they were identical within the decimal places utilized by the standard CHARMM force field.

MD simulations were run on the WSU rocks cluster. The canonical ensemble was maintained via periodic boundaries, with Langevin dynamics and thermostat (430). Simulation stability was verified by use of the trajectory analysis tools available with the VMD software (431). Stability was monitored by energy and RMSD. The systems were solvated with TIP3P water, neutralized with counter ions and subjected to 1000 steps of conjugate gradient minimization and temperature ramped to 300K. Frames from the trajectories were written every

1 ps. The solvation box includes a 15Å pad on each face of the box. Long-range electrostatic effects were taken into account using the smooth particle mesh Ewald method (441), and van der Waals interactions were calculated with a non-bonded cutoff of 8Å and a switching function between 7-8 Å. Results were analyzed by use of GMD graphs, via TimeScapes (442), and by PCA using the bio3d package for R. Hydrogen bond and salt-bridge analysis was performed using VMD. All hydrogen bonds and salt bridges occurring for more than 90% of the simulation time were noted. RMSDs were calculated using VMD.

Clustering was carried out on both MD and docked peptide conformations. Standard clustering in Chimera was performed (468). Cross comparison of the docked and MD conformations was carried out by superposing all conformers from the docking clusters onto representative structures from the MD clustering. RMSDs between best matching pairs were calculated, and the docking cluster to which they belonged was identified for rank comparison.

Chapter 4 Investigation of an allosteric circuit in the cysteine protease domain of *Clostridium difficile* Toxin B

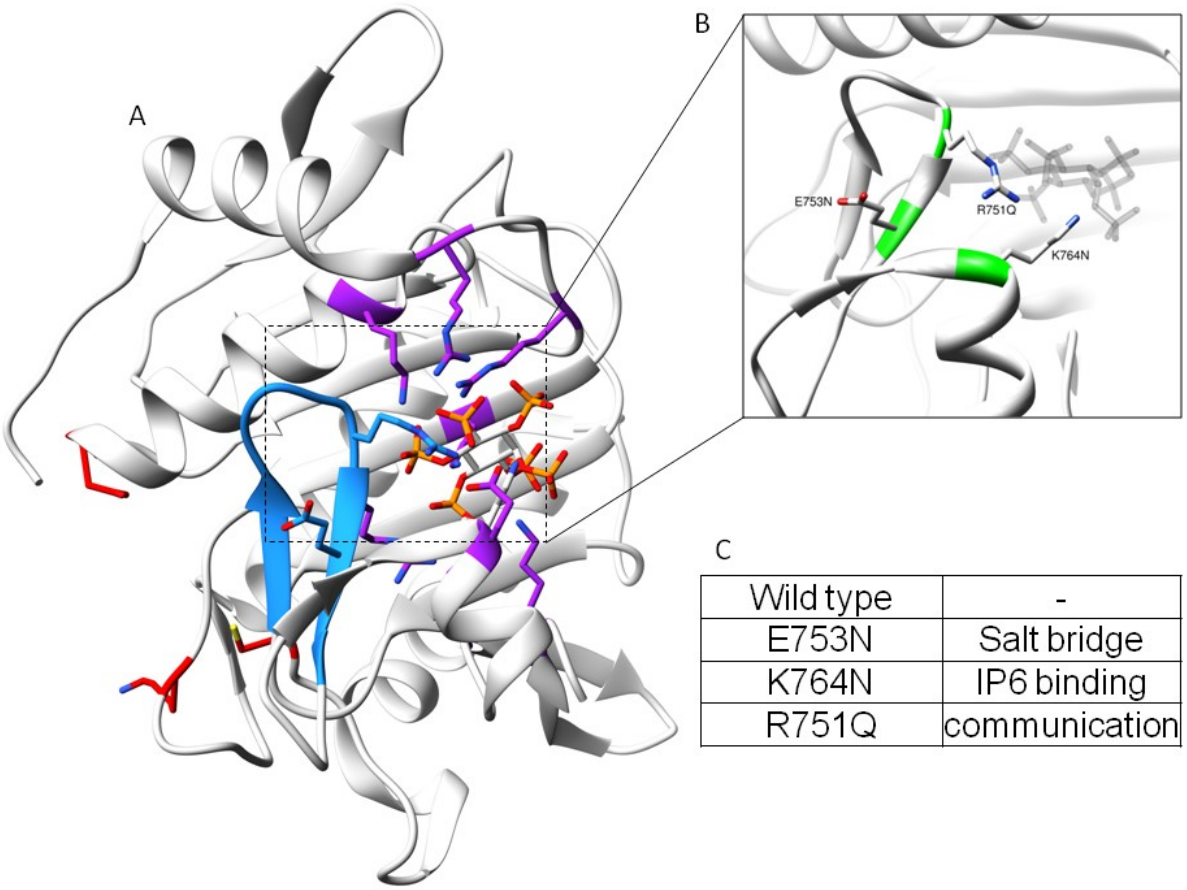
4.1 Background

The cysteine protease domain (CPD) of the *C. difficile* toxins A and B is responsible for cytosolic release of the glucosyltransferase domains, and was recently crystallized (PDBID:3PEE) (402) Once the toxin undergoes receptor mediated endocytosis, CPD and GT domains undergo translocation through a pore created by the translocation domain. Cleavage of the CPD releases the GT domain into the cytosol at this point. Cellular damage occurs when an endocytosed toxin transfers a glucose moiety from UDP-Glucose to any number of Rho-family GTPases. The CPD domain is a possible drug target since targeting virulence factors may be a way to abrogate the cellular damage that occurs during an active infection,. The CPD domain could be targeted at two points during cellular intoxication due to the etiology of the toxin,. Endosomal escape may be prevented if the CPD domain is bound irreversibly to a small molecule and tight folding prevents translocation,. The GT domain would potentially remain endosomal, no interaction with Rho-GTPases would be possible. If cleavage and release of the GT domain occurs before cell contact and endocytosis, there will be no GT domain translocated to damage the cells. Induction of CPD domain cleavage in the lumen of the intestine would preclude cell death.

4.2 Cysteine protease domain of TcdB

CPD cleavage is allosterically activated by the binding of inositol hexakisphosphate (IP6). Binding of IP6 triggers the organization of the CPD catalytic triad into an active conformation. Figure 0-1 shows the structure of the CPD from Clostridium difficile TcdB.

Residues comprising the catalytic triad are shown in red, residues experimentally determined to participate in IP6 binding are shown in purple, and the β-hairpin shown in blue has been determined experimentally to have an effect on rate of cleavage (402). We have introduced the mutations shown in Panel B in green, also experimentally explored (402), to determine their possible structural consequences both apo and IP6 bound. K764N is predicted to affect interactions with IP6,E764N potentially perturbs folding of the β-hairpin responsible for active site organization and R751Q is predicted to communicate IP6 binding allosterically to induce hairpin formation. The objective of studying these mutations is to determine which activities, IP6 binding, hairpin formation or communication, is most critical to folding and function. Further use of this work might include design of inhibitors based on IP6 with the purpose of manipulating one or more of these functions.



Wild type	-
E753N	Salt bridge
K764N	IP6 binding
R751Q	communication

Figure 0-1 Structure of the CPD domain and investigated mutations.

Structure of the CPD domain is shown in panel A. IP6 binding residues are shown in purple, The β-hairpin mutagenically shown to contain an allosteric communication circuit are shown in blue. Shown in red are the catalytic triad responsible for protease activity. Panel B shows the three mutations studied in this work, and Panel C indicates their putative roles. E753N participates in a salt bridge responsible for organizing the β-hairpin structure. K764N is directly involved in IP6 binding, and R751Q is a putative allosteric communicator.

4.3 Investigation of allosteric circuit through molecular dynamics

Previous work by Shen et al. determined the in vitro effects of introducing several mutations of the CPD domain. MD simulations of both the wild type and three mutant CPD structures both apo and IP6 bound were performed. Analysis of these simulations by PCA and GMD indicates differential behavior and may provide context for the development of targeted inhibitors or activators of function. Puri et al have designed cleavage inhibitors, however, these

molecules are active intracellularly (446). While some work has been done by Puri et al towards designing inhibitors of cleavage, however these molecules are active intracellularly. We hope that these studies may lead to the development of molecules that would induce CPD cleavage extracellularly, or prevent GT domain escape, as this would preclude issues of cellular transport. Cell protection by either of these avenues would allow for more efficient treatment of infection and alleviate many of the symptoms of CDAD.

A total of eight MD simulations were performed as listed in Table 0-1; Apo and IP6 bound structures of the wild type and each of the three mutants described above. Unbiased MD was performed as described in the Methods section, the Apo-WT simulation was run for 130ns, while all other trajectories were run for 25ns. The Apo-WT simulation was allowed to run for a longer time to provide better data density for analysis. Following preliminary analysis, the Apo-E753N simulation was extended to 90ns to determine if additional information on the mutant simulations was forthcoming with longer simulation times. Correlation analysis as well as PCA and GMD analyses were applied.

Table 0-1 Simulations performed on Apo and IP6 bound CPD domain.

A total of eight simulations were performed. The wild type and three mutant structures in both Apo and IP6 bound conformations were studied. The wild type was simulated longest to give more structural context for the molecular motions. E753N was extended following initial analyses.

Total simulation time		
Structure	Apo	IP6
Wild Type	130 ns	25ns
E753N	90 ns	25ns
R751Q	25ns	25ns
K764N	25ns	25ns

From visual inspection of the trajectories, it was clear that the apo structure exhibits a high degree of flexibility throughout the simulation. This is to be expected as IP6 binding is responsible *in vitro* for the organization of the active site. An overall loose fold was expected as Apo-CPD must be able to thread through a pore during endosomal escape. All IP6 bound

simulations are far less mobile relative to the Apo simulations and tight contacts with the IP6 phosphate groups are maintained throughout. All simulations were stable through their time-courses as was monitored by RMSD and total energy. Backbone RMSDs for all simulations were below 1.5Å.

4.3.1 Principal Component analysis of MD simulations

The PCA of both the IP6 bound and Apo simulations is very revealing. Figure 0-1 Panel A, shows the cross-plots of PC1 and PC2 for both the Apo and IP6 bound simulations. PC1 is on the x-axis and PC2 is on the y-axis. For each set (Apo vs IP6) the trajectories were projected onto the same core residues and may be directly compared. While Apo-IP6 cross-plot comparisons are not quantitatively reasonable, the breadth, localization, and transition patterns can be used to discern qualitative relationships. Panel B shows a broadened ribbon diagram for all first principal components, colored as in Figure 2. The longest timescale simulation Apo-WT shows a wide range of conformational freedom, which is to be expected. In both the Apo and IP6 bound sets of simulations, viewed horizontally across panels, the mutants all behave similarly to one another. One point of interest is the range of the axes on the IP6 bound plots relative to the Apo plots; there is considerable restriction in the conformational space explored in all IP6 containing simulations.

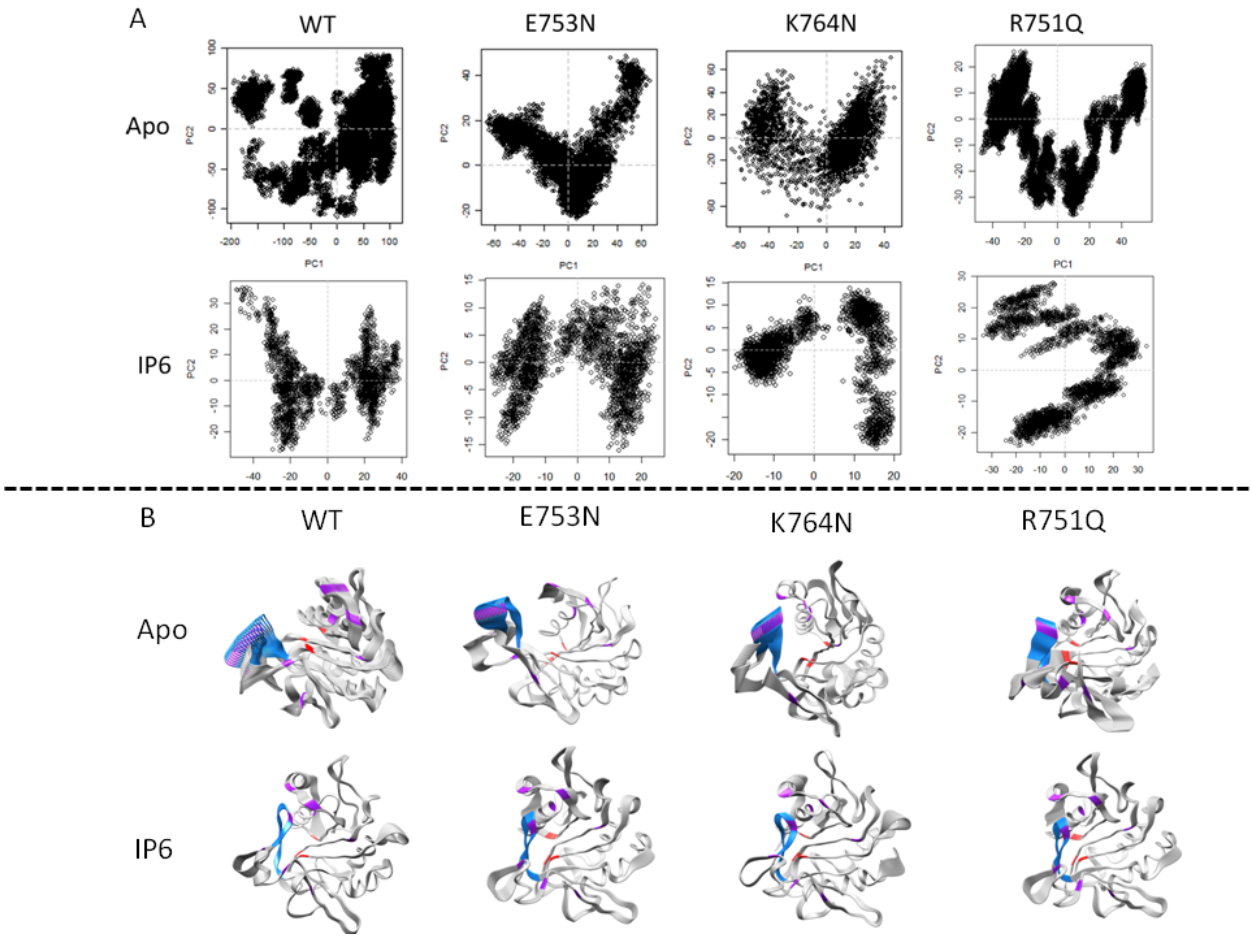


Figure 0-2 PCA analysis of Apo and IP6 bound CPD simulations.

Panel A: All apo simulations were projected onto the same invariant core for PC analysis, as were all IP6 simulations. Quantitative comparisons can be made left to right, and qualitative comparisons top to bottom. The greatest amount of flexibility was observed in the Apo-WT simulation. All Apo simulations seem to occupy the same general conformational space to varying degrees and in slightly more punctate or diffuse manners. The same is seen in the IP6 bound simulations. Again, the wild type has the highest degree of flexibility, and the mutant simulations are somewhat comparable. Panel B: Broadened ribbon diagrams of the first principal component projected onto the structures. All have been oriented similarly for comparison.

In the Apo simulations, while all mutants seem to exhibit similar behavior, Apo-K764N shows a bit more diffuse conformational exploration, while Apo-R751Q shows more clearly defined paths between conformational regions. Apo-E753N shows both consistent grouping and smooth transitioning.

In the IP6 bound mutant simulations, again all three behave similarly, though the IP6-E753N shows diffuse conformational space relative to IP6-K764N and IP6-R751Q. R751Q

shows transitioning between regions by clearly defined paths, showing some similarity to the Apo simulation.

Between the Apo and IP6 bound WT simulations, there is both restriction of conformational space, and alteration in the mode of transition. In all other simulations there is transitioning between conformational regions, while in the Apo-WT simulation there are a few isolated regions of localization. This may be due to the longer timescale of the simulation, possibly exploring several local minima that the other simulations do not investigate.

4.3.2 Interaction analysis of Apo and IP6 bound simulations

We performed hydrogen bond analysis, salt-bridge analysis, and assessed water-protein hydrogen bonds to quantitate the effects of IP6 binding on the simulations,. Tables 2-4 describe these metrics. With respect to intra-protein hydrogen bonds, the stabilizing effect of the IP6 dramatically increases organization.

Table 0-2 lists the total number of intra-protein hydrogen bonds for each simulation. In all cases the inositol bound simulations exhibit over twice the number of detected hydrogen bonds. This is physiologically sensible, as IP6 binding is predicted to stabilize the fold, theoretically allowing time for the optimization of the local hydrogen bonding network. This conclusion is supported by the data presented in Table 0-3 .

In all simulations, the apo structures exhibit greater hydrogen bonding to water throughout the simulation. As we observed considerable structural flexibility in the apo simulations, it would be logical that charged amino acids would be available to form more solvent-protein hydrogen bonds. Lastly, Table 0-4 presents the total number of salt bridge interactions detected in each simulation. This echoes the result of the intra-protein hydrogen bond analysis, where over twice as many self-interactions were detected. We again attribute this to the IP6 binding-induced stability allowing the protein structure to optimize it's self interactions.

Table 0-2 Intra-protein hydrogen bonds observed for both Apo and IP6 bound simulations.

In all IP6 bound simulations, we observe nearly double the total number of stable hydrogen bonds, we attribute this behavior to the additional backbone stability afforded by IP6 binding.

Intra-protein hydrogen bonds		
Structure	Apo	IP6
Wild Type	56	129
E753N	50	137
K764N	57	128
R751Q	57	131

Table 0-3 Solvent-protein hydrogen bonds observed in both Apo and IP6 bound simulations.

In agreement with the information in Table 4-2, without the stabilizing presence of IP6 the apo simulations fulfill their desired hydrogen bonding networks with solvent-protein hydrogen bonds. In the IP6 bound simulations, there are consistently fewer protein-solvent hydrogen bonding interactions.

Solvent-protein hydrogen bonds		
Structure	Apo	IP6
Wild Type	189	129
E753N	193	129
K764N	204	144
R751Q	166	143

Table 0-4 Intra-molecular salt bridges observed in both Apo and IP6 bound simulations.

Stable salt bridges were assessed for all trajectories. IP6 bound simulations show over twice as many salt bridged residues as the Apo simulations do in all cases. Again, we attribute this to the overall stability increase due to IP6 binding that allows to sidechains to find intra-protein interaction partners, rather than interacting with solvent during large scale conformational changes.

Salt bridges		
Structure	Apo	IP6
Wild Type	8	20
E753N	7	19
K764N	8	19
R751Q	7	17

4.3.3 Correlation analysis on Apo and IP6 bound simulations

Correlation analysis was applied to determine the effects of the mutations on the overall motions of the simulations. Difference plots are shown in Figure 0-3, where the correlation matrices of the mutant simulations have been subtracted from the correlation plot of the wild type. When plotted, this data indicates which residues show altered correlated motion relative to the wild type. Anti-correlation is shown as blue, correlated motions are shown as orange, and x and y axes are residue numbers. What is immediately apparent in these plots is that correlation is dramatically perturbed in the Apo-E753N simulation. Upon examination of the structure however, this is very logical. The mutation removes a salt bridge pair, replacing an Arg-Glu interaction with an Arg-Asn. This substitution causes the β-hairpin critical in forming the active site to partially unfold during the simulations, presumably damaging overall correlation as well as active site organization. The K764N position is involved in IP6 binding, and R751Q in communication of that binding, so it is not surprising that we see no effects on correlated motion in the Apo simulations.

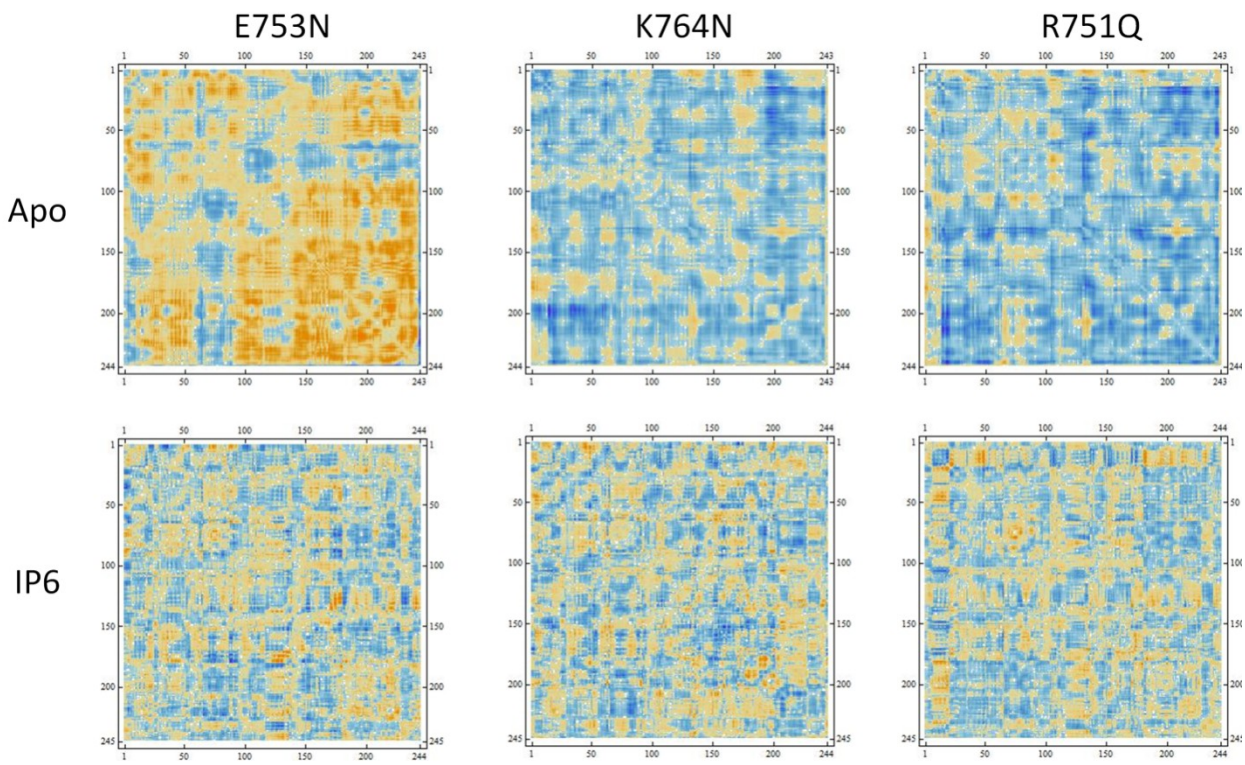


Figure 0-3 Difference correlation plots .

Correlation matrices of the mutant simulations have been subtracted from the wild type simulations to determine how the introduction of the mutation affects overall concerted motion. It is readily apparent that E753N dramatically perturbs how the CPD domain explores its conformational space. In all IP6 bound simulation, a low degree of motility is observed, and this low correlation overall is to be expected.

Once IP6 is bound we see little overall motion in all simulations, and consequently there can be a limited amount of correlated motion. As we see in the three mutant IP6 bound simulations, there is no discernible difference between the three other than a few slight point correlations. While IP6 binding precludes differentiating between these mutations by correlation, GMD was applied to determine the effects on the activity of the simulations.

4.3.4 Generalized Masked Delaunay analysis of simulations

Generalized masked Delaunay plots for all simulations are shown in Figure 0-4. Total activity is indicated in blue, contact forming is in green and contact breaking is in red. This analysis provides a number of events per frame, by monitoring the space explored by a residue during a window of the trajectory, and then determining whether or not motions that occur are significant or trivial. Please refer to the original manuscript for further information (442).

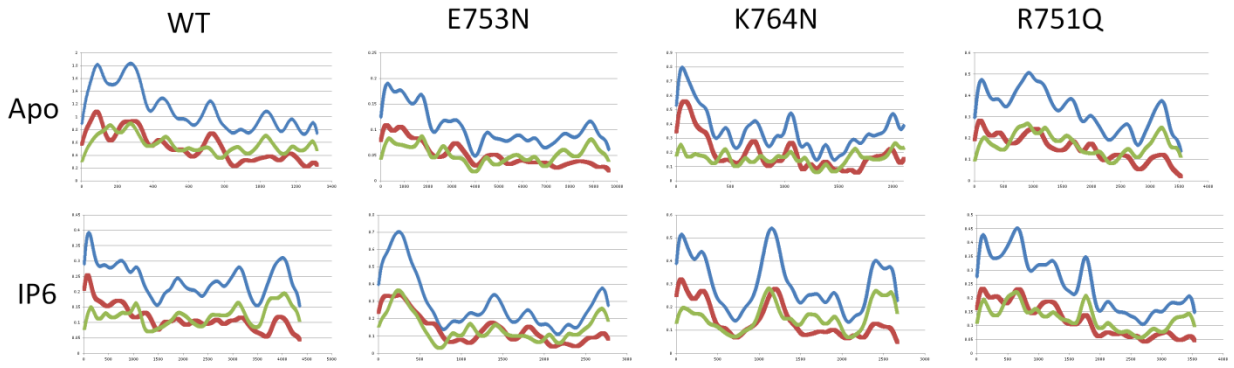


Figure 0-4 : GMD plots of all Apo and IP6 bound simulations.

Total activity is indicated in blue, contact breaking is in red, and contact forming is in green. This analysis provides events per frame and was used to monitor overall activity as well as investigate specific activity peaks, such as those seen in the IP6-K764N simulation.

The Apo-WT shows a high level of activity overall, two phases of bond breaking and forming are observed around 70ns and 115ns, these alternating contact forming and breaking periods indicate an open-close behavior when the trajectory is observed at these points and is likely a normal part of the conformational exploration of this protein. While the Apo-E753N behaves dissimilarly from the other simulations, we interpret this as associated with the breaking of the concerted motions of the β-flap that were previously described. The Apo-R751Q simulation shows slightly higher activity overall, between 0.2 and 0.5 events per frame, slightly higher than the Apo-E753N and K764N simulations.

The IP6-bound simulations all show a low level of activity relative to the Apo-WT simulation, which is partly due to simulation length, but is expected due to the low motion observed in the PCA analysis. The IP6-WT simulation shows the lowest level of activity overall, with some periodic fluctuations. The mutant IP6 bound simulations all share a common feature, a high activity peak somewhere between 7 and 20ns. As this peak occurs at different times in the simulations we believe it is not an artifact, but rather a consequence of forcing IP6 binding to these mutants. When the trajectory is observed, these time points correlate to rearrangement near the IP6 binding site.

The IP6-E753N simulation shows the lowest of these peaks, presumably because this mutation does not directly contact the IP6 molecule. The mutation, while removing an Arg-Glu interaction as discussed above, does not result in a repulsive pairing and the β-hairpin remains largely stable through the simulation. The IP6-K764N simulation shows the most pronounced peak, and when observed this also includes some rearrangement of the β-hairpin as well. The K764N mutation is directly interacting with the IP6 molecule, and therefore this would likely have the most dramatic effect on an IP6 bound conformation. The R751Q mutation does not directly interact with the IP6, but interacts with the K764 position to communicate IP6 binding. The GMD of this simulation shows a moderate peak, occurring later in the simulation than in either the E753N or K764N. Interestingly, the Apo-K764N simulation also shows this peak, and may be indicative of the K764N position having structural importance beyond IP6 binding.

4.3.5 Quantitation of active site organization

Organization of the active site was quantitated in two ways, by RMSF, and by calculating the area of the triangle defined by the alpha carbon of the Asp 44, Lys 112 and Cys 155, the residues of the catalytic triad.

Figure 0-5 plots RMSF of the three residues over the simulation time, normalized to wild type. The first four columns are the Apo simulations, while the last four are the IP6 bound

simulations. Interestingly, even in the highly flexible apo simulations, the active site remains fairly stable. What is more interesting is that on a per-residue basis, the mobility of the active site triad is not evenly distributed. In all simulations the Cys155 residue has the highest RMSF. This residue is positioned adjacent to the β-flap that gets organized upon IP6 binding, and we theorize that this residue may be an extension of the allosteric circuit, critical in organizing the active site.

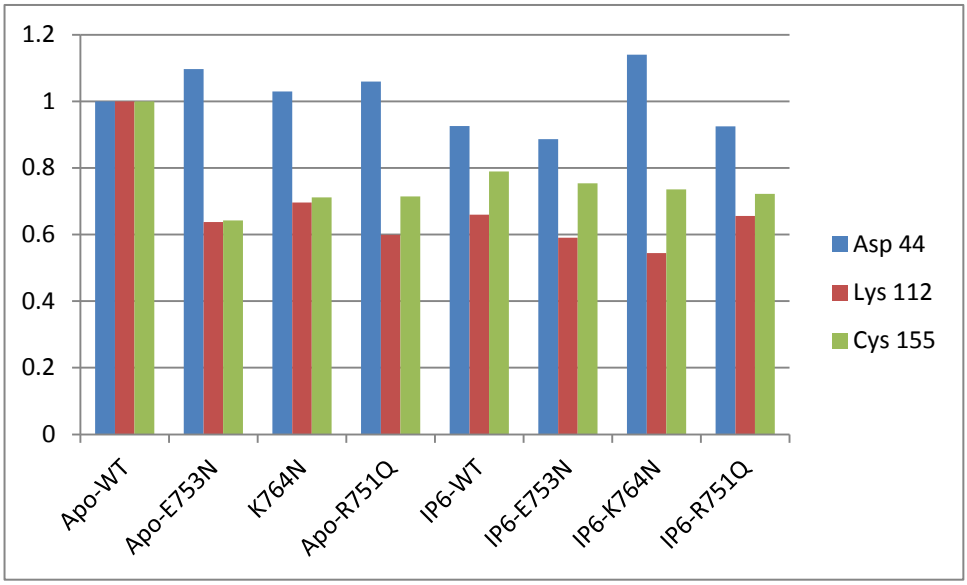


Figure 0-5 Rmsf vs. simulation for three catalytic residues normalized to wild type.

Cys 155 shows much higher flexibility than either Lys 112 or Asp 44, and as this residue is proximal to the β-flap critical for activity, this is to be expected. While overall the activities are fairly consistent, IP6-K764N shows a somewhat increased level of activity

To determine the extent of organization of the active site, a triangular distance metric was created, using the area as a measure of "openness" Measured at the C-alpha position, the average distance for each of the three active site triad residues was observed for all simulations. The area of the triangle defined by these distances was calculated by Heron's formula and is included in Table 0-5. A visualization of this metric is shown in Figure 0-6.

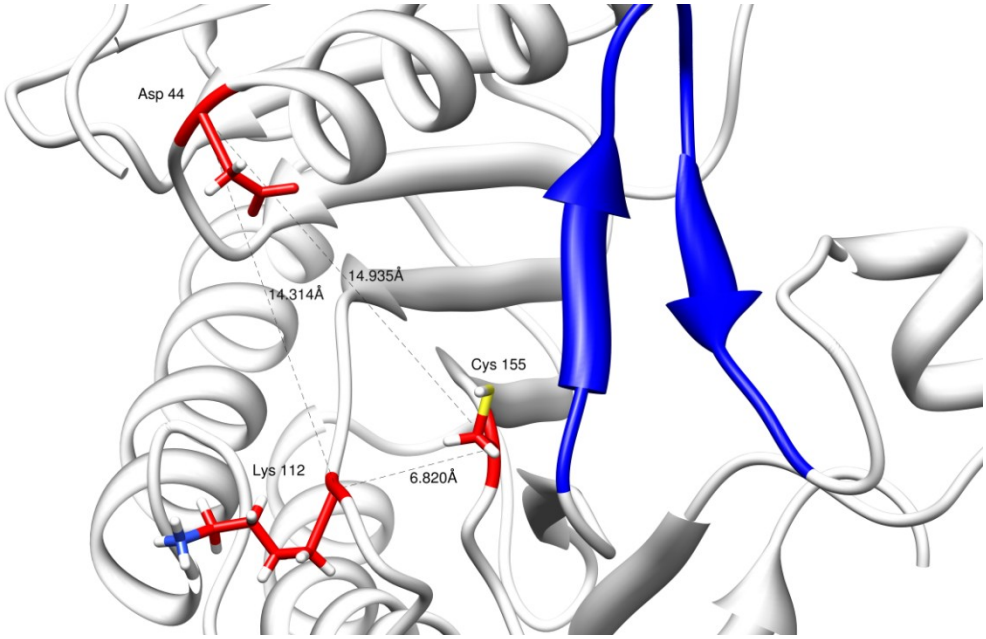


Figure 0-6 Visualization of the distance metric used to describe active site stability.

Distance between the alpha carbon of Asp 44, Cys 155 and Lys 112 were averaged through all trajectories. The triangle described by this region was used as a metric of active site stability.

Table 0-5 Distance between C-alpha of catalytic triad and area of triangle representing active site organization.

Apo				
Simulation	44 to 155(Å)	155 to 112(Å)	44 to 112(Å)	Area($Å^2$)
Wild Type	14.7	9.6	9.8	46.5
E753N	13.7	9	13.8	58.4
K764N	14.84	6.99	14.3	49.28
R751Q	16.15	11.32	10.76	60.75
IP6 bound				
Simulation	44 to 155(Å)	155 to 112(Å)	44 to 112(Å)	Area($Å^2$)
Wild Type	14.17	7.33	13.29	48.14
E753N	14.16	7.15	13.07	48.56
K764N	14.11	7.67	13.18	49.84
R751Q	15.21	6.17	14.1	43.47

The overall results are similar to that of the RMSF calculation, but there are some slight differences, in both the apo and IP6 bound simulations, the active site seems relatively stable. There are a few notable exceptions, which support the evidence for an allosteric pathway.

Residue R751Q has been previously described as being critical in the communication of IP6 binding to the active site. As previously shown in our RMSF calculations, we theorize that the active site residue with the largest amount of flexibility, and therefore the highest effect on active site organization is Cys 155. In the Apo-R751Q simulation, there is a dramatic increase in the Cys155 to Lys 112 distance. This causes an increase in the area of the catalytic triad from roughly 46-48 $Å^2$ to 60.75 $Å^2$, a phenomenon not observed in any of the other simulations. Again, in the IP6 bound simulations, we observe a relatively stable active site area, with the exception of IP6-R751Q. In this case the active site area is actually decreased relative to the other IP6 bound simulations. We propose that perturbing the allosteric circuit even in the absence of IP6 has effects on the organization of the active site. It is possible that an inhibitor that alters the distance between the catalytic triad may be effective in preventing protease activity, or an activator that stabilizes this section of the allosteric pathway might be able to induce cleavage extracellularly.

4.4 Conclusions

This work assesses the effects of three mutations of the CPD domain of TcdB on IP6 binding, overall conformational mobility, correlation and activity. Comparison of the effects of these mutations illustrates in greater depth the interactions required for the formation of the CPD active site. This information will be useful in the future design of inhibitors or activators of cleavage for this protein.

Previous experiments indicated that the selected mutations perturbed catalytic activity as well as IP6 binding. It was shown that positions E753, K764 and R751 were critical for activity in various capacities, and their roles were determined as β-hairpin folding, IP6-binding and

allosteric communication respectively. While some work has been done towards development of inhibitors of the pre-organized active site, these require cell penetration as their active site target does not exist until late in the process of cellular intoxication. Currently there are no inhibitors exploiting the allosteric site of this protease.

MD simulation, PCA, correlation and GMD analyses were used to determine the where, what and when of these structurally important mutations. PCA allowed us to determine the similarities and dissimilarities in the overall structure through the simulation timecourses. Correlation analyses were used to indicate which mutations altered the method of movement, and GMD was applied to determine how the introduced mutations affect event pattern.

MD simulations and PCA analysis indicate that in the IP6 unbound form the E753N mutation dramatically perturbs the conformational space of the protease, which may explain some of the *in vitro* findings with respect to IP6 binding and protease activity.

The correlation analysis dramatically illustrates that the E753N mutation in the apo conformation has strong effects on overall correlated motion. Strong correlation was observed in the Apo-WT simulation, and this is completely destroyed by the introduction of the E753N mutation. While later studies showed that the overall arrangement of the active site may not have been perturbed, our findings agree with the conclusions of Shen et al. that this mutation most dramatically interferes with protease activity. In the IP6 bound simulations, we observe a low level of activity and correlation as the simulations are quite stable. The difference correlation plots reflect this beautifully, as once IP6 is bound the protease should remain in a stable fold until proteolysis occurs.

GMD analysis was used to determine the effects of these mutations on overall activity and to discern critical events in simulation. While all three mutant IP6 bound simulations indicated an "event" during the course of the simulation, K764N was observed to have the greatest level of mid-simulation activity. This activity was determined to be a reorganization near the base of the β-flap. We propose that the alteration in IP6 interaction due to this mutation

partially destabilizes this region as the simulation progresses while the protein attempts to somewhat ameliorate the unfavorable contacts included by this mutation.

In observation of the RMSF and active site triad area, we determined that even through the highly flexible apo simulations, the active site remains fairly stable. On a per-residue basis, it was determined that Cys 155 appears to be an extension of the allosteric circuit defined by Shen et al.. As Cys 155 is proximal to the β-flap responsible for active site organization, this is logical. Both the K764N and R751Q mutations affect the behavior of this residue, albeit in different ways. During the IP6-K764N simulation, the RMSF of increases relative to wild type, and is in fact higher than all other simulations. The R751Q mutation affects the distances between the residues of the catalytic triad, directly affecting the shape of the active site pocket.

We consider the pre-organization of the active site to be in agreement with current *in vitro* work As there has been evidence for small molecule inhibition of the CPD domain. As some of these studies have been performed in the absence of IP6, we consider the active site stable enough to be considered a target for inhibition. However, as this work has been done at pH 7, it is possible that the range of pH found in the gut may affect organization or flexibility of the CPD domain, making active site targeting difficult. Additionally, as altered pH has been observed in individuals with severe bowel disturbances, pH sensitive studies would be advised in the future.

We suggest that small molecules targeting the IP6 binding region which stabilize the fold and allosteric circuit may potentially trigger extracellular cleavage. However, as we have determined that there is some extant pre-organization of the active site, targeting of the active site pocket is not an unreasonable goal, provided the inhibitors are not dislodged during endosomal escape. We propose that Cys 155 is the direct target of allosteric control in the CPD domain, and molecules perturbing the flexibility or position of Cys 155 would be capable of inhibiting protease activity. Furthermore, molecules capable of causing conformational alterations similar to those induced by the R751Q mutation may be able to prevent organization of the active site.

Investigation into allosteric activation as a potential mechanism to prevent cellular intoxication may prove fruitful.

It is clear that alteration of any step in the allosteric circuit has consequences. Introduction of the K764N mutation affects activity while IP6 is bound, introduction of the E753N mutation appears to perturb conformation, and introduction of the R751Q mutation alters organization of the active site. Small molecules exploiting these activities would potentially be powerful tools against *Clostridium difficile* infection.

4.5 Methods

The wild type crystal structure of CPD (PDBID:3PEE) was used as a scaffold for the creation of three mutants E753N, R751Q and K764N. Mutations were introduced using the structure editor in Chimera (469). Conformations were selected from the Dunbrack rotamer library (470), and taking into account the first two atoms of the wild type side chain conformation. All four structures were generated with and without the crystallographic inositol for a total of eight simulations. Structures were solvated and ionized using the MeadIonize plugin for highly charged systems using VMD (431). Simulations were carried out using NAMD (428) with the CHARMM force field (437, 438, 471).

Parameterization for IP6 was performed as previously described (449) using both ab initio and analogy methods. Parameters from both methods agree to within the decimal places used by the CHARMM force fields.

Correlation analysis was carried out using the ptraj module of AmberTools (472). All atom correlation of the protein structure was performed and results were plotted using Mathematica (473). Solvent, ions and IP6 were excluded. Difference plots were generated for ease of comparison. Unsubtracted plots are available in SI.

Generalized Masked Delaunay analysis was performed using TimeScapes (442). Simulations were strided to 1/10 of total simulation time, water, counterions and substrate was removed. The calculation was performed using a 5% frame window. Contact breaking, forming and total activity were plotted. Activity based trajectories were generated and used to examine the events detected in the IP6 bound simulations.

PCA was done using the bio3d package under R (474). Trajectories were reduced to C-alpha atoms only, each set of four trajectories was used to determine the spatially invariant residues. All trajectory frames were aligned onto the "core" residues and PCA was performed. Crossplots of PC1 vs. PC2, PC2 vs. PC3, and PC1 vs PC3 were generated as well as a plot of proportion of total variance captured vs. Eigenvalue rank. Broadened ribbon diagrams were generated using Chimera where the degree and direction of broadening indicates degree and direction of eigenvalue for each PC.

Hydrogen bond and salt bridge analyses were carried out using VMD (431). Active site distance metrics were calculated using python and the Chimera visualization suite (469). Area of active site was calculated using Heron's theorem. Rmsf of selected residues was performed using R (475).

Chapter 5 Comparative analysis of *Clostridium difficile* Toxins A and B

5.1 Background

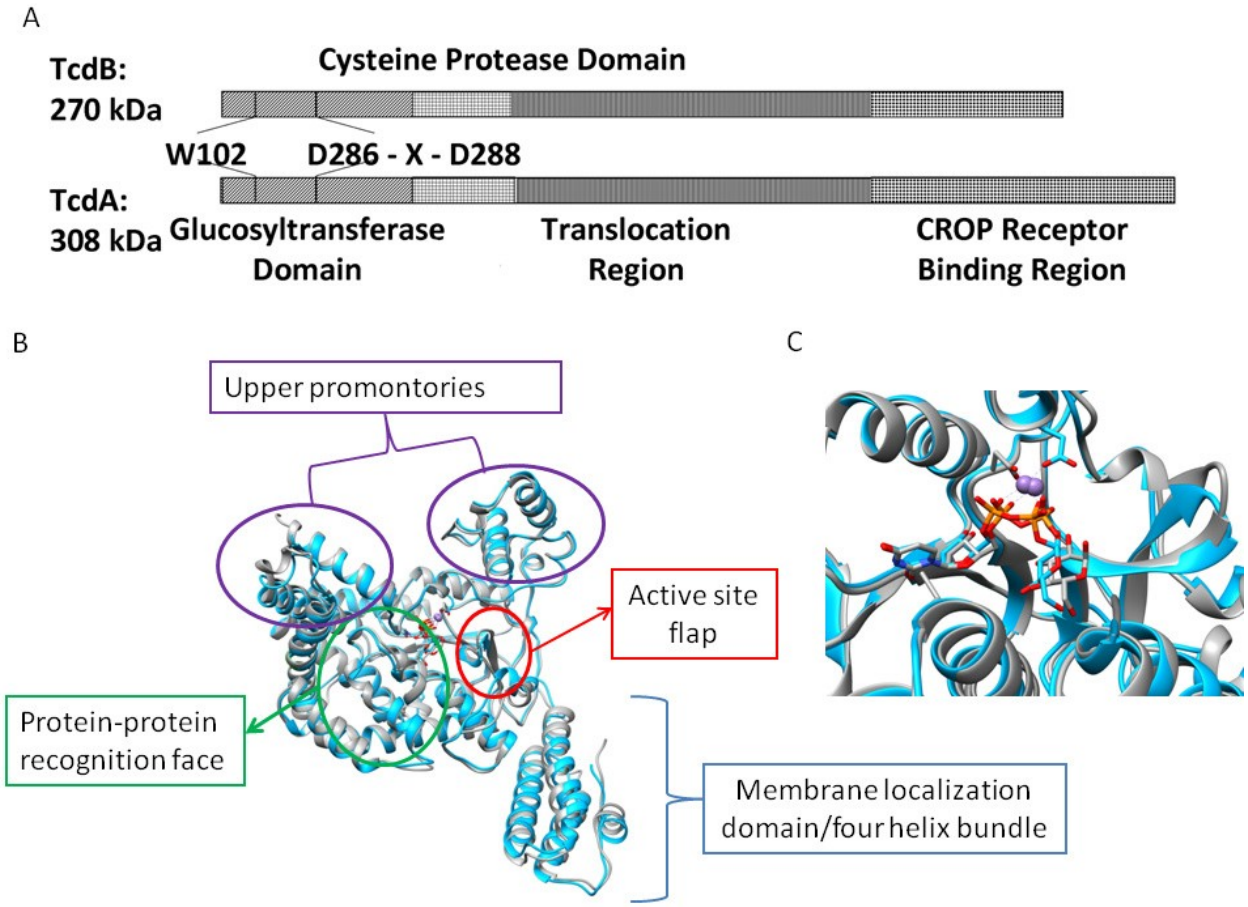
Clostridium difficile (C. diff) is a well known hospital-acquired infection that costs the American healthcare system millions of dollars and thousands of lives per year (374, 376). Infection occurs following treatment with broad-spectrum antibiotics during an inpatient stay.(370). While the broad-spectrum antibiotics perturb the natural colonic microflora, hospitalization provides exposure to C.diff bacteria and spores; creating the perfect opportunity for C. diff infection. Preventing patient-to-patient transmission would preclude the spread of C. diff infection, however complete control by this method is difficult in a real world setting. C. diff is a gram-positive, spore forming anaerobe, and the spores are notoriously resistant to antimicrobial agents, up to and including bleach (476). Once exposed to live cultures or spores, C. diff quickly takes over a compromised patients intestinal tract. The presence of the bacillus itself does not always result in disease and can be observed in a sub-clinical state in some individuals (477). However this is contingent on other gut bacteria keeping the local C. diff population in check. The more virulent strains involved in disease onset produce two large glucosylating toxins (GT's) which cause large scale tissue damage(370). Infection with C. diff causes a range of symptoms, ranging from diarrhea to pseudomembranous colitis, and toxic megacolon(370).

C. diff is not the only bacillus known to cause tissue damage by the action of glucosylating toxins. There are currently five known Large Clostridial Toxins (LCT's),

TcdA and B of C. diff, alpha toxin (Tcnα) from C. novyi, and hemorrhagic (TcsH) and lethal toxins (TcsL) from C. sordellii. In all cases these LCTs cause massive tissue damage, associated with the disregulation of cell signaling(478). The phenotypes associated with infection are typically edema, myonecrosis, and sepsis, usually in the form of gas gangrene, Toxic Shock Syndrome (TSS) or clostridium difficile associated diseases (CDAD). The glucosyltransfer step of these toxins places a glucose moiety from a nucleotide sugar onto a variety of Rho GTP-ases. This leads to disruption of cell signaling, followed by cell death. The propensity for massive tissue damage coupled with the fact that clostridial infections are difficult to treat with antibiotics has recently led to a novel approach. Treatment with anti-toxin therapies during an active infection may preclude the severe tissue damage that these toxins cause and improve clinical outcomes. Toxin binding polymers, monoclonal antibodies, and epoxide inhibitors are all currently being pursued for this purpose, with some success(383, 407, 446, 479-483).

To better design anti-toxin agents however, a thorough understanding of the causative agents is essential. An excellent comparison of the genetic structure and cellular targets of these toxins was performed by Busch and Aktories(478) in 2009, but a considerable amount of structural information has become available since those studies. Drug design based on structural and dynamic studies of the GT domain of TcdB have proved particularly productive (484). By incorporating information obtained through molecular dynamics studies, a potent and irreversible anti-toxin molecule was developed. It is with these studies in mind that we undertook a structural comparative analysis of the two glucosylating toxins of *C. diff,* TcdA and TcdB.

While the structure of TcdB was solved several years ago(408), TcdA has only recently been structurally characterized (436) provided a unique opportunity for study. While the toxins have a similar domain structure (Figure 1 Panel A), they have some common and some unique intracellular targets, and differing levels of toxicity. The overlay in Panel B is a superposition of the GT domains of TcdA and TcdB with relevant regions for analysis indicated. For the purposes of these studies, the toxins will always be presented in this orientation. The protein-protein recognition face, and active site flap are in the center, the four-helix bundle responsible for membrane localization is on the lower right, and the promontories unique to these toxins are shown at the top. Panel C shows a cutaway of the active site, with the glucose donor UDP-Glucose complexed to the catalytic manganese ion. In the TcdB crystal structure hydrolysis of this substrate occurred, producing a UDP and Glucose, it remains complete in the TcdA structure. While the simple structural overlay appears highly homologous, these proteins have dramatically different properties on a per-residue level. The initial TcdA structural analysis by Pruitt et al included some electrostatics images which seemed to disagree with my initial examination of the structure. Upon further examination, it became apparent that their "electrostatics" was simply coloring by residue type. This prompted a quick coloumbic surface calculation, which revealed a dramatically different surface for TcdA than what had been observed for TcdB. This fascinating result led to a deep comparative analysis of the two proteins investigating electrostatics, hydrophobicity, pH sensitivity and dynamic motion between TcdA and TcdB.



0-1 Comparison of TcdA and TcdB.

(A) Schematic diagram of the domain organization of TcdA/B. Both proteins have conserved glucosyltransferase, cysteine protease and translocation domains. The most significant difference is in the C-terminal repetitive oligopeptide receptor binding region. (B) Structural superposition of the toxins. (C) Structural overlay of the GT-domain active site. Both structures were crystallized with UDP-glucose in the active site, however hydrolysis occurred in the TcdB structure, resulting in separate UDP and Glucose moieties in the crystal.

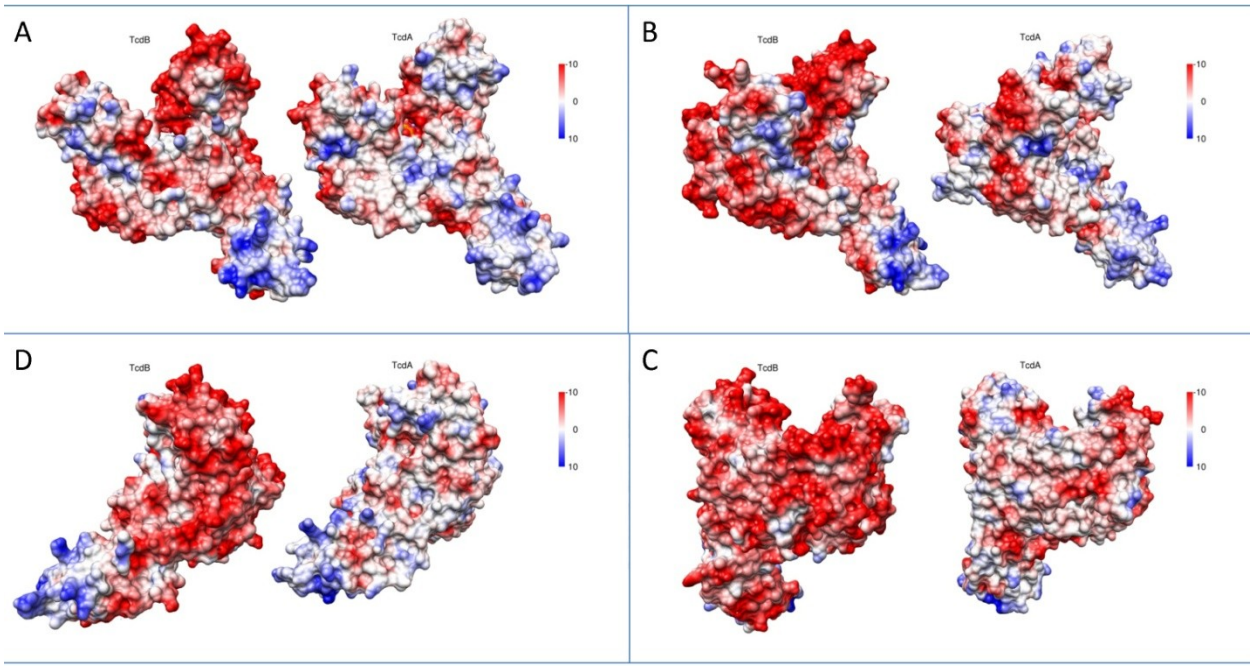
5.2 Comparison of surface properties

Full electrostatic surfaces were calculated for TcdA and TcdB using the Accelerated Poisson-Boltzmann Solver (APBS) (485). Hydrophobicity, sequence similarity, crystal structure RMSD, and charge differentials at pH 7 were all compared. Straightforward investigation of electrostatics as well as novel mapping techniques were used to determine regions on the structures where pH sensitivity may play a role in their

differential activity. Multi-Conformer Continuum Electrostatics (MCCE) calculations were carried out along with a computational pH titration to determine residues in both structures which show shifted pKa's.

5.2.1 Surface electrostatics, hydrophobicity and structural similarity at pH 7

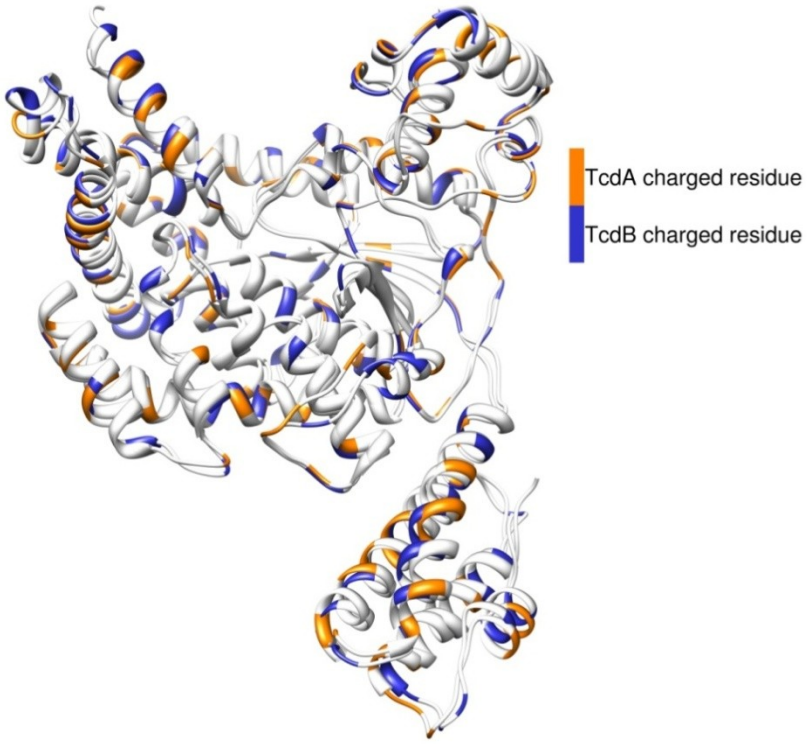
Figure 2 shows the full electrostatic APBS results of TcdA and TcdB at pH 7 projected onto an MSMS (Maximal Speed Molecular Surface) surface representation. Clockwise from top left, each progressive panel represents at 90° rotation from the previous panel. It is notable that on all faces, it appears that TcdB has significantly more charged residues on the surface compared to TcdA leading to more electrostatically charged surface on the former



0-2 APBS electrostatics at pH 7 projected onto the surfaces of TcdA and TcdB.

Panel A presents the active site and protein-binding face, with TcdA on the right and TcdB on the left. The four-helix membrane localization domain is positioned at the bottom right on this panel. TcdB shows considerably more electrostatic character than TcdA in all orientations, Panels B-D represent 90° rotations around the y-axis from each previous panel

Panel A orients the active site and protein-protein interface towards the viewer. It is notable that there is a prominent negatively charged patch directly on the active site interface. Also, between the two upper promontories there is a considerably greater amount of negative charge. The four-helix bundle in shows greater positive charge on the TcdB structure in all orientations. Panels C and D are striking in that there is a dramatic increase in negative character for TcdB relative to TcdA.

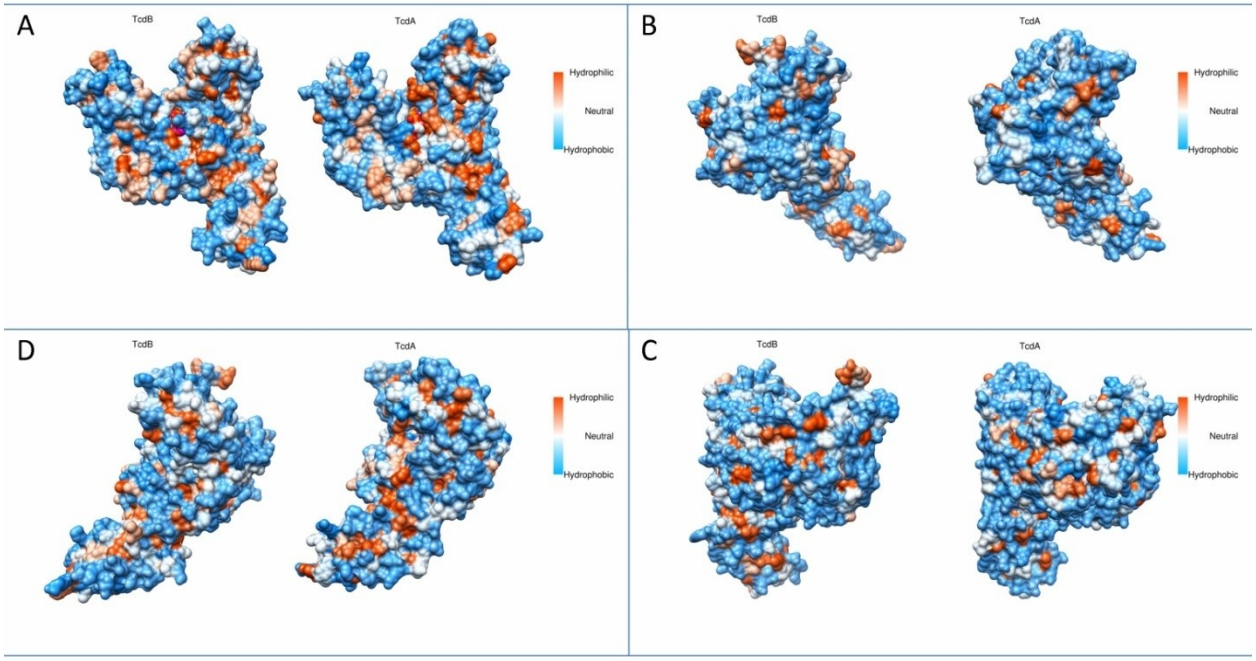


0-3 Charged residues on TcdA and TcdB

Charged residues of all types are colored on TcdA and TcdB. Those from TcdA are colored orange,a nd those from TcdB are colored blue. It is noticeable that there is a distinct lack of charged residues on the TcdA structure in and near the active site, while many of the surface charge has been maintained.

Structural superposition followed by sequence alignment indicates that overall, TcdA has far fewer charged amino acids than TcdB does. A full sequence alignment is available as Supplementary Figure S1. It was determined that TcdA has 49.54% identity with TcdB, increasing to 50.66% if gaps are omitted. Nearly all active site mutations (those occurring within 20Å of the crystallographic UDP-Glucose) replace charged amino acids with nonpolar residues as shown in Abdeen et al. in press. It is interesting to note that although TcdB has many more charged residues in its sequence, the tertiary structure for both proteins results in the arrangement of many charged residues on the surface for TcdA as well as TcdB. Additionally this results in a lack of charged residues in the active site of TcdA.

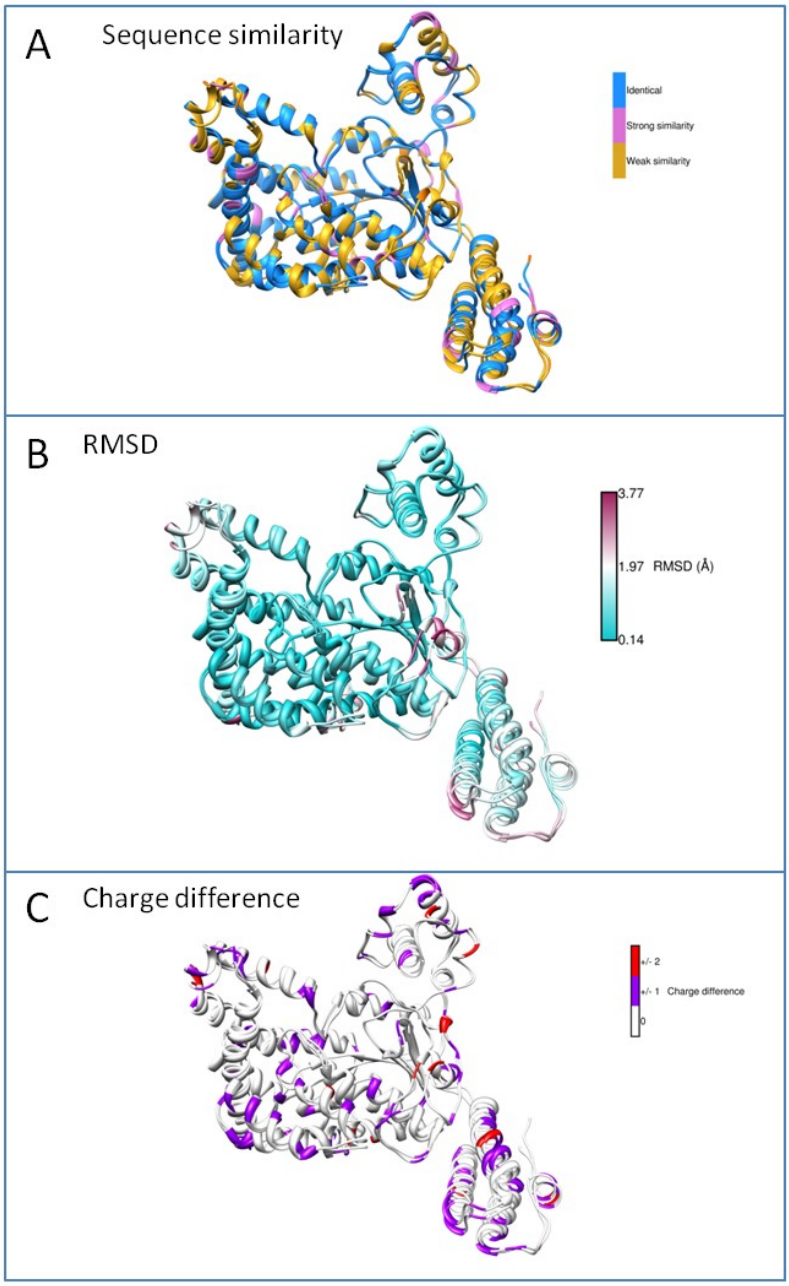
The overall charge distribution at pH 7 indicates that the two proteins, while structurally similar, have different surface properties. Figure 4 plots the Kyte-Doolittle hydrophobicity on an MSMS surface. Kyte-Doolittle hydrophobicity is based on a hydropathy scale derived from experimental observations, applied to a protein structure by using a moving-segment approach to determine average hydropathy within a segment. Panels A-D are each rotated 90° from the previous frame. Hydrophobic regions are shown in orange while hydrophilic regions are shown in blue. Neutral residues are shown in white. Panel A orients the toxins with both the protein recognition face and active site facing forward.



0-4 Kyte-Doolittle hydrophobicity surfaces of TcdA and TcdB

Toxins are in the same orientation as Figure 2 with the active site and protein-interface located in the center of the image. TcdA is on the right and TcdB is on the left. Panels B-D represent 90° rotations around the y-axis from each previous panel. There is a notable increase in hydrophilicity on the TcdA active site and protein interface regions, with an overall slight increase in hydrophilicity on the sides and rear of the protein.

More hydrophilic character is observed in the center of the TcdA structure, as seen in the center of the top left panel, and the amphipathicity of the membrane association bundle is greater than that of TcdB. The hydrophobic areas shown in Panel B are nearly identical, while an increase in hydrophilic residues in TcdB is observed in panel C. Panel D shows slightly greater hydrophilicity for TcdA on this face. The differential distribution of hydrophobic residues on various critical regions, notably the protein-protein interface and active site may be relevant for function. TcdA and TcdB have different preferential cellular targets(388), and this variability may correlate with the electronic or hydrophobic contacts on those proteins. TcdB targets Rac1, Rho(A/B/C), RhoG, TC10, and Cdc42 while TcdA targets Rac1, Rho(A/B/C) and Cdc42, so while there is some overlap in the targets, there are some unique targets as well.



0-5 Similarity metrics applied to TcdA and TcdB.

Panel A shows the sequence similarity projected onto the structure. Identical residues are colored blue, strongly similar residues are colored pink, and dissimilar residues are colored gold. Panel B describes the RMSD of the crystal structures for TcdA and TcdB. High RMSD is colored magenta, while low RMSD is colored cyan. Panel C projects the sequence-predicted charge differences. Residues with no change in charge are colored white, a change of +/- 1 is indicated by purple and a change of +/- 2 is indicated by red.

To better represent the differences in these two structures, we have done a direct comparison for several properties and overlaid them on the superimposed structures. We have depicted the properties of conservation, crystal structure RMSD and charge difference on a per-residue basis as shown in Figure 4 panels A-C.

Panel A indicates a measure of conservation between the two structures. Identical residues are colored blue, similar residues are colored pink, and dissimilar residues are colored gold. Similar and dissimilar residues are defined by the Al2CO method using the BLOSUM-62 matrix. The highest region of conservation is in the active site center, while high dissimilarity is seen near the protein-protein interface and four-helix bundle.

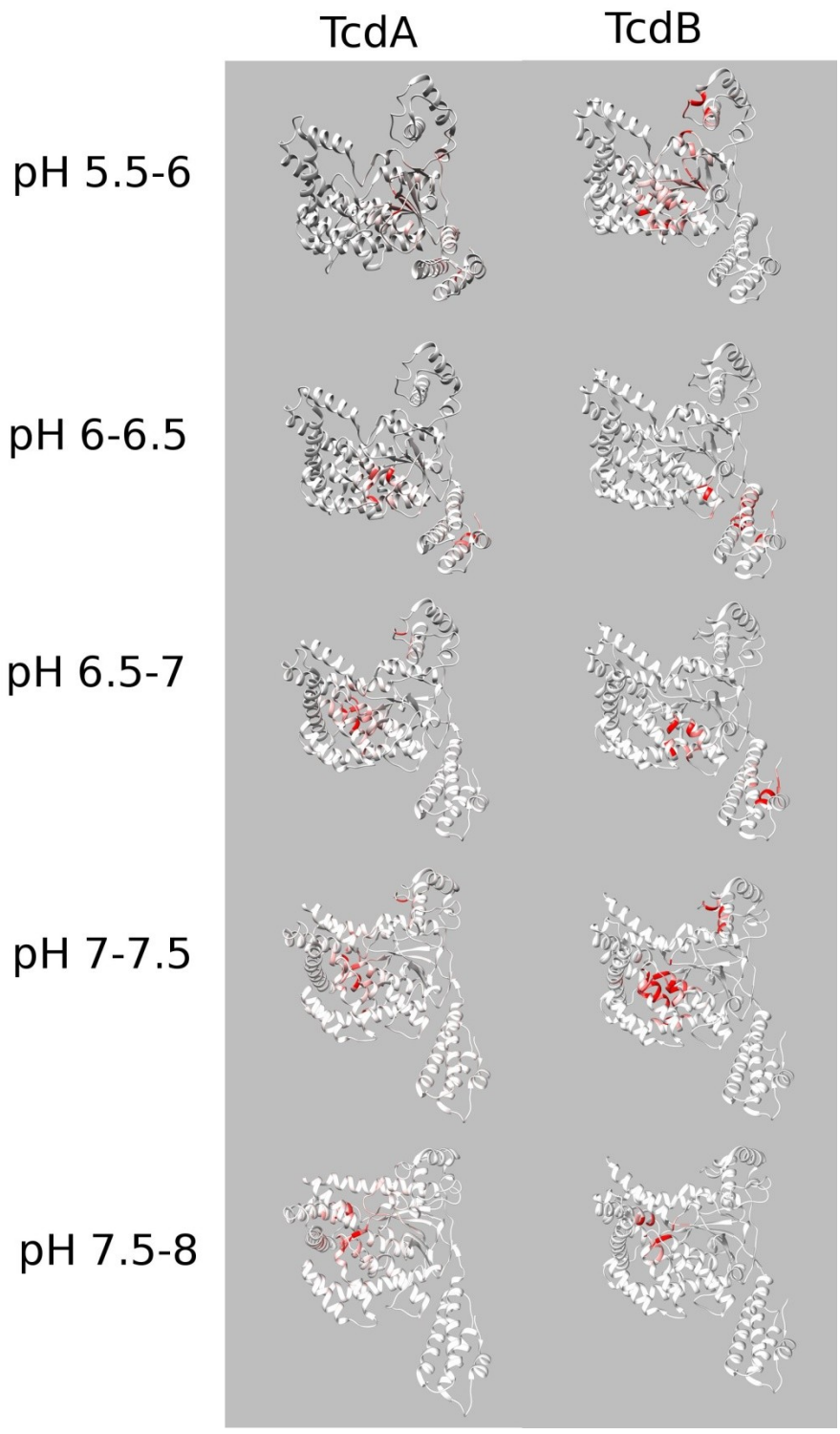
RMSD between the two crystal structures was measured and the ribbon structures are colored on a gradient from rose to cyan. Structurally, the two proteins are highly similar, with the major differences being near the active site flap. This flap is mobile during UDP-Glucose binding (Swett et al, Biophys J. in press) and presumably this difference is due to the fact that the UDP-Glucose observed in the TcdB structure underwent hydrolysis in the crystal and thus represents a EP complex rather than the ES state. Other regions with higher RMSDs are the edges of the promontories and the four helix bundle, presumably due to the inherent mobility in these regions or crystal contacts. Both the upper promontories and the four helix bundle of TcdB were observed to have high mobility in both normal mode and dynamics simulations(449).

Absolute value of charge difference is shown in Panel C. Residues with no charge difference are colored white, residues with a change of $\pm$ 1 are colored purple,

and ± 2 are colored red. Again, charge consistency is conserved in and near the active site, with high difference in charge on the upper promontories and on the four-helix bundle. There are some residues showing charge alteration near the active site flap, slightly more distant from the catalytic center.

5.2.2 Mapping of electrostatics for TcdA and TcdB

To better compare the internal electrostatics of TcdA and TcdB, a novel mapping method was applied to allow visualization of the sensitivity of electrostatics to pH changes. A pseudo-Voronoi approach was used to assign electrostatic potential as a property of each atom closest in space to the center of that Voronoi cell. The atom properties were summed per residue, allowing the accurate projection of electrostatic potential onto a ribbon structure as opposed to an isolevel volume projection or projection onto an MSMS surface. This was done from pH 5.5-8 in half pH unit intervals. Once assigned, subtracting the maps at a per-residue level allows us to point residues that are sensitive in a given pH range. Figure 5 shows the results of applying this mapping method to TcdA and TcdB. TcdA is shown on the top row and TcdB is shown on the bottom. Each map was subtracted from its nearest neighbor point, producing a spectrum representing pH sensitive residues at each point.



0-6 Mapping of pH sensitivity to the structures of TcdA and TcdB

The change in electrostatic potential across half pH unit ranges was mapped to the ribbon representations of the structures. Regions that show altered electrostatics in the given pH range are colored red, with color intensity indicating degree of change.

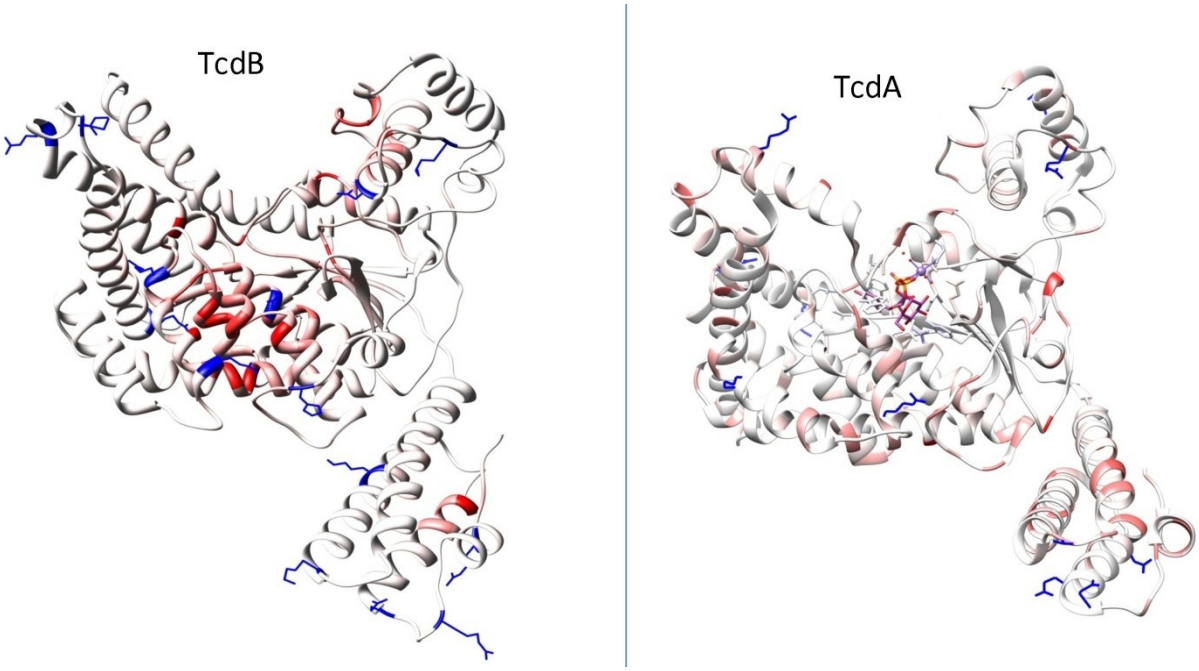
In both structures the region around the active site and protein-protein recognition face show the highest pH sensitivity, but interestingly, not in the same pH ranges. TcdA shows nearly no pH sensitivity between ph 5.5-6. In the pH ranges between 6-6.5, 6.5-7-7.5 and 7.5-8, we see pH response concentrated around the protein-protein interface and at higher pH's, near the active site. TcdB behaves radically differently, between pH 5.5-6 there is considerable response in the upper promontories and protein-protein interface. Between 6-6.5, this sensitivity almost completely disappears and a region near the c-terminus of the four-helix bundle becomes pH sensitive. Between pH 6.5-7, this sensitivity remains, and some pH response is seen proximal to the protein-protein recognition face. As the pH rises to 7-7.5 a number of residues in the active site and on the protein-protein interface are strongly pH responsive, and this sensitivity is concentrated within the active site at pH 7.5-8. We postulate that the differential pH sensitivity may be a factor in some of the functional variability between TcdA and TcdB. TcdB is far more lethal to cells and causes cell death far more quickly than TcdA. An increased rate of endosomal escape and refolding due to its higher pH sensitivity would be one plausible explanation. The pH sensitivity in the TcdB four-helix bundle that is not observed in TcdA may be of interest as well. If the unfolding and refolding of this helix is a bottleneck in the etiology of intoxication, the inclusion of pH sensitive residues might be helpful as TcdB transitions from the endosome to the cytosol.

5.2.3 Multi-conformer continuum electrostatics applied to pKa shift prediction

To further investigate the pH sensitivity of these two proteins, we applied multi-conformer continuum electrostatics (MCCE) to identify pK_a-shifted residues in both structures. In TcdB, 13 residues with pKa shifts of more than 2 pH units were found, and 11 were found in TcdA. Tables with the sequence numbers and quantitative pK_a shifts are in supplementary information Table 1. For our purposes we wanted to determine whether pKa shifted residues were in the regions we had determined as being sensitive to pH. Figure 6 shows side by side the comparison of TcdA and TcdB at pH 7, with pKa shifted residues colored in blue. There are a higher overall number of pKa shifted residues in TcdA and TcdB, but in both cases, the presence of pKa shifted residues corresponds with our predicted pH sensitive regions.

Table 0-1 MCCE determined pKa shifted residues in TcdA and TcdB

TcdB					TcdA				
Residue	Number	Calculated pKa	Standard pKa	Difference	Residue	Number	Calculated pKa	Standard pKa	Difference
ARG	194	0.098	12.48	12.382	ARG	215	0.334	12.48	12.146
ARG	6	1.284	12.48	11.196	LYS	346	0.459	10.53	10.071
ARG	158	5.063	12.48	7.417	ARG	302	4.071	12.48	8.409
LYS	303	4.045	10.53	6.485	ARG	405	4.318	12.48	8.162
ARG	165	7.379	12.48	5.101	ARG	224	4.563	12.48	7.917
ARG	16	7.898	12.48	4.582	ARG	67	8.161	12.48	4.319
HIS	492	2.581	6.1	3.519	LYS	428	6.267	10.53	4.263
ASP	270	0.592	3.86	3.268	ARG	535	9.07	12.48	3.41
ASP	310	1.345	3.86	2.515	GLU	22	1.721	4.07	2.349
ARG	445	10.101	12.48	2.379	GLU	20	1.874	4.07	2.196
LYS	64	8.317	10.53	2.213	ASP	31	1.731	3.86	2.129
ARG	455	10.374	12.48	2.106					
LYS	50	8.454	10.53	2.076					



0-7 Comparison of MCCE predicted pKa shifted residues and electrostatically mapped structures of TcdA and TcdB.

Regions with ph sensitivity are colored red, and residues with predicted pKa shifts are colored in blue

As both TcdA and TcdB transit from the lumen of the intestine, through endosomal acidification and into the cytosol, they are exposed to a range of pH's. We applied the Adaptive Poisson Boltzmann Solver to calculate the continuum electrostatics of both toxins across the pH range of 5.5-8. By applying a novel mapping method, we were able to highlight the regions that show pH sensitivity in half pH unit ranges. Multi-conformer continuum electrostatics calculations were performed to determine the positions of pK_a-shifted residues. These results were in agreement with the pH sensitivity mapping method. TcdB shows higher pH sensitivity than TcdA, likely due to the increase in total number of charged residues. Additionally, the regions that are pH sensitive differ in the two toxins, with a region near the active site broadly pH sensitive on TcdA, and a section in both the protein-protein binding face and four-helix bundle showing sensitivity on TcdB. This sensitivity in the four-helix bundle is

particularly interesting. If this pH sensitive region assists in the folding or unfolding of the membrane localization domain, this may enhance the ability of this toxin to exit the endosome or localize to the cell membrane. Further work needs to be done to valdiate the prediction that these specific residues affect the differential function of TcdA and TcdB during internalization and glucose acceptor binding/specificity.

5.3 Molecular dynamics applied to TcdA

All-atom molecular dynamics simulations of TcdA in the presence and absence of UDP-glucose were performed to determine whether there were any dynamic differences between TcdA and the dynamic analyses of TcdB present in the literature. PCA and GMD were applied to determine both the overall structural motions and event patterns in simulation. Solvent analysis, hydrogen bonding and salt bridge analyses were performed to determine whether or not the increased hydrophobicity of TcdA affects solvent dynamics relative to TcdB.

As computational studies on the dynamics of TcdB have already been discussed in chapter 2 and 3, we performed molecular dynamics simulations to assess the behavior of TcdA in comparison to TcdB. We simulated TcdA in the presence and absence of UDP-Glucose for 40 and 20ns respectively. We performed generalized masked Delaunay analysis (GMD), as well as PCA and hydrogen bond analysis. These experiments were undertaken to determine the extent of similarity in dynamic motion relative to the previously published data(449). It was previously shown that the Apo-TcdB structure had an increased range of motion, higher overall activity in GMD analysis, and increased active site solvation relative to the UDP-glucose-bound

conformer. The results of the TcdA simulations echo these results , and are discussed below.

5.3.1 Simulation of TcdA assessed with GMD, PCA and interaction analysis

Binding of UDP-glucose decreases the overall activity and flexibility of TcdA, in agreement with previous results for TcdB. Figure 7 shows the data from conformational analysis of this protein. Panels A and B are generalized masked Delaunay plots, which indicate the number of significant and persistent "events" per frame. This can be used as a rough gauge of the overall activity of a simulation. Panel A indicates that for the Apo simulation, the overall level of activity remains higher than the UDP-glucose-bound simulation shown in panel B. The PCA analysis shown in panels C and D largely agrees with these results. Crossplots of the eigenvalues for each simulation frame from principal component vectors 1, 2 and 3 are shown, along with a breakdown of the contribution to total variance expressed as a percentage of total. Overall Panel C indicates a lower amount of conformational space explored by the UDP-glucose bound simulation than that seen in panel D. This finding goes hand in hand with the data in panels A and B, we observe both lower overall activity, and a decreased amount of conformational space explored in the UDP-glucose-bound state. During visual inspection of the trajectories, we observed no large scale reorganization in either the Apo or bound form. RMSD's were calculated at 1.8Å and 1.6Å for the Apo and UDP-glucose-bound simulations respectively.

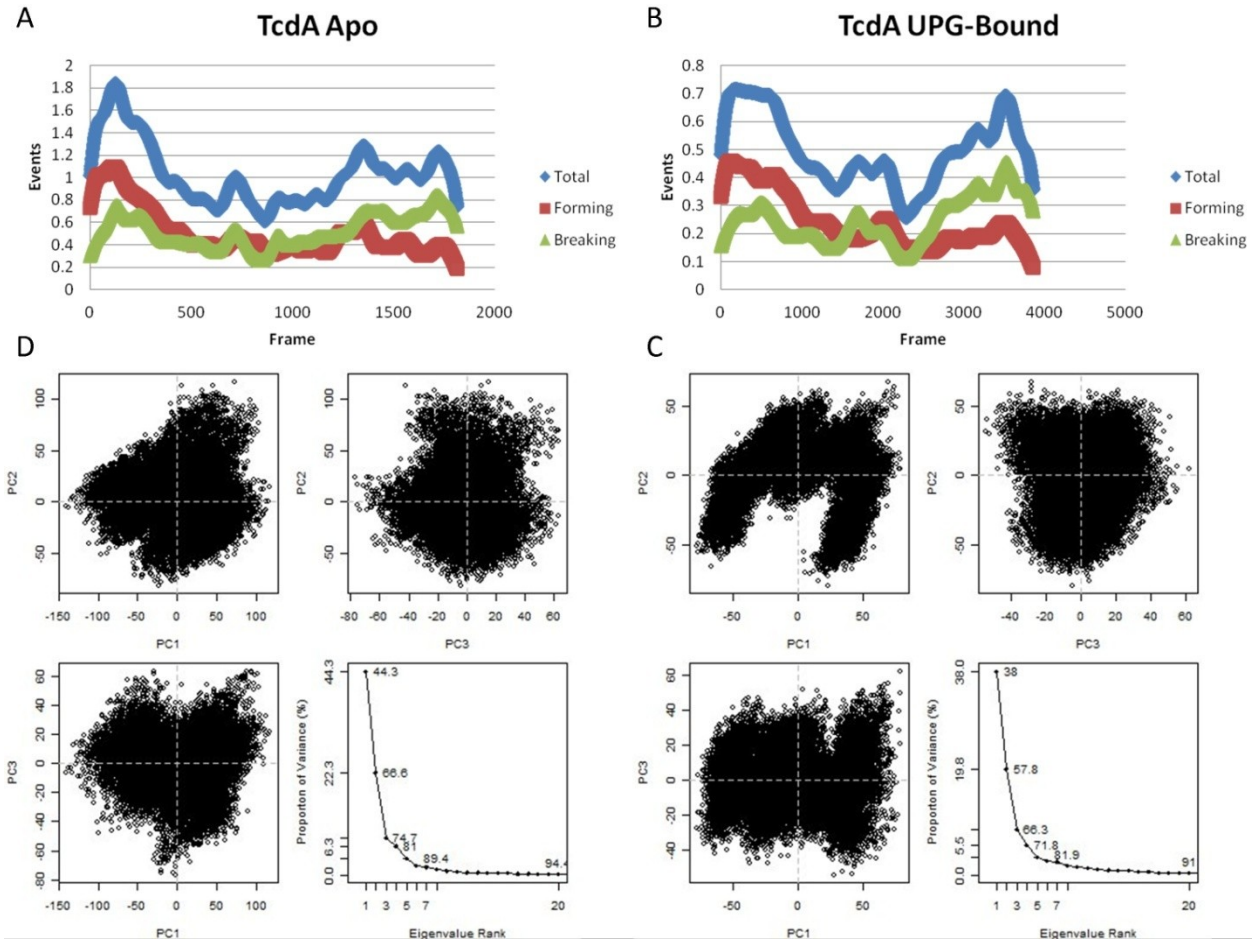


Table 0-2 GMD and PCA analysis of TcdA simulated both Apo and in the presence of UDP-glucose.

Panels A and B plot the number of significant events over the course of the trajectory. Panels C and D show the crossplots from the first three eigenvectors of the simulation, along with the proportion of variance that each contributes to the total.

Hydrogen bonding, solvent-interaction and salt bridge analyses suggest that the introduction of the UDP-Glucose increases the stability of the structure. Table 1 describes these interactions. Interaction type is broken down into intramolecular hydrogen bonds, solvent hydrogen bonds, and salt bridges. TcdA and TcdB simulations both in the presence and absence of UDP-glucose are presented for comparison. In the TcdA simulations, we see an increase in hydrogen bonds upon UDP-glucose-binding. This is largely due to the additional backbone stability contributed by ligand binding.

With UDP-glucose bound, the structure is more able to optimize intramolecular interactions. We also see a decrease in active site hydrogen bonds, as these are physically precluded by the presence of UDP-glucose. With respect to solvent interactions, we observe overall increased solvation in the active site and surface relative to TcdB, and again, several active site interactions are disturbed by the presence of UDP-glucose. During the TcdA simulations, there is an increase in salt-bridge formation upon UDP-glucose binding. This is in contrast to the TcdB simulations, however, several of the salt bridges disrupted in TcdB were in the active site, and are physically disrupted by the presence of UDP-glucose. Overall, we have determined that TcdA is stabilized by the presence of UDP-glucose, and we see a higher level of solvation, relative to TcdB. Differential behavior with respect to salt bridges is due to the location of interactions within the structures.

Table 0-3 Interaction analysis for the simulation of TcdA in the presence and absence of UDP-glucose.

Interaction type	TcdA		TcdB†	
	Apo	UPG bound	Apo	UPG bound
Intramolecular h-bond	28	39	-	-
Intra-active site h-bond	13	9	-	-
Intra-binding face h-bond	5	5	-	-
Solvent h-bond	68	62	81	77
solvent-active site h-bond	7	4	1	0
solvent-binding face h-bond	6	6	1	1
Salt-bridges	67	74	77	58

† Reproduced from Swett et al. (In press)

Molecular dynamics was applied to determine whether or not TcdA and TcdB behave similarly in solution. TcdA was simulated in the presence and absence of UDP-glucose, and analyzed with GMD, PCA and several interaction analyses. In comparison

to previous simulations of TcdB, the results are largely similar. We observe higher activity in the GMD analysis when TcdA is in the Apo form, as well as an increased range of conformational motion as shown by PCA. The interaction analysis indicates that TcdA has a higher level of stationary water molecules, some of which are displaceable upon UDP-glucose binding. As observed in the GMD and PCA, TcdA is more conformationally stable when bound to UDP-glucose and this results in a higher level of intramolecular interactions. We suggest that the additional backbone stability allows the sidechains to optimize their hydrogen bonding networks and salt bridges more readily than the Apo form. A decrease in the overall number of salt bridges observed on TcdB upon UDP-glucose binding was previously described, and is due to the interruption of active site salt-bridges by the substrate.

5.4 Conclusions

We have performed a full electrostatic and dynamic comparison of the clostridial toxins TcdA and TcdB. While the proteins have a highly similar overall fold, there are significant sequence changes that result in radically different electrostatic properties. These were compared across multiple pH's to determine the location, if any of pH sensitive residues. A novel mapping method was applied to allow clear visualization of these regions. Multi-conformer continuum electrostatics was used to probe pKa shifted residues and determine their locations relative to pH sensitive regions. Molecular dynamics simulations of TcdA in the presence and absence of TcdB were performed to determine whether or not there were significant differences in the mobility of TcdA, as compared to literature information on TcdB.

In conclusion, we have delineated the effects of the limited sequence similarity between TcdA and TcdB on the surface properties of these two toxins. We suggest that while the toxins have similar dynamic properties, the alterations in pH sensitivity and electrostatics may play a role in both their differential targets, and catalytic efficiency. A novel mapping method was developed and applied to indicate regions of pH sensitivity across pH ranges. This has discovered a pH sensitive region in the four helix bundle of the toxins which may aid in the translocation step, by assisting in the unfolding or refolding of the toxins following endosomal acidification and escape. As it has already been shown that the cytotoxicity of these proteins can be alleviated by perturbing the protein-protein binding face (ref), we suggest that the design of an inhibitor based on surface properties of the toxins may be useful as well. Also, as the dynamic comparison of TcdA and TcdB indicated very similar behavior in response to UDP-glucose, and it has been previously shown that inhibitors targeting TcdA also affect TcdB, we suggest that active site binders will likely have cross-toxin activity.

5.5 Methods

Structural superposition was performed using the Needleman-Wunsch algorithm in conjunction with the BLOSUM-62 matrix. A gap extension penalty of 1 was selected. Secondary structure score was included as 30% of the total superposition score. Secondary structure assignments were calculated using ksdssp(Kabsch and Sander Define Secondary Structure of Proteins) with an energy cutoff of -0.5, a minimum helix length of 3, and a minimum strand length of 3. Within the secondary structure score, structural gap-opening penalties were applied as follows. In predicted helices, an intra-

helix or intra-strand gap required a penalty of 18, and any other gap was a penalty of 6. After superposition, a structure-based sequence alignment was generated. The total alignment score was 1775.8. The percent identity was calculated based on simple alignment (49.54%), and with non-gap columns only (50.66%)

Conservation was calculated using AL2CO methods(486). The Independent counts method was used for frequency estimation, and conservation was calculated by the sum of pairs method. The averaging window was set to 1 and the gap fraction was set to 0.5. The Sum-of-pairs matrix again was set to BLOSUM-62, with no matrix transformations applied. .

Sequence alignment is colored by the ClustalX standard. Conservation is represented by bar height. Full height indicates identical residues, 2/3 indicate strong similarity, 1/3 height indicate weak similarity. as per AL2CO standards. Charge variation is relative to TcdA. Blue bars indicate more positive charge, red bars indicate more negative charge. Full height bars indicate a change of +2 charge units, and half height bars indicate a change of +1 charge units. RMSD is represented as a gray bar with height relative to increasing RMSD per residue. Hydrophobicity coloring was projected onto an MSMS surface(487), using the KD method(488).

APBS was used for the initial electrostatic calculations(489). The AMBER force field was used for the calculation, and output was returned using the APBS internal naming scheme. The hydrogen bonding network was optimized and PropKa(490) was used to assign protonation states between 5.5-8 in half pH unit intervals. An automatically configured sequential focusing multigrid calculation was performed(491,

492). For the TcdB structure, this used 193 grid points in all directions for grid-based discretization. The focusing calculation was carried out with the following coarse mesh domain lengths: x-direction 136.583, y-direction 143.144, z-direction 123.15. Fine mesh lengths were: x-direction 100.343, y-direction 104.215, z-direction 92.441. For the TcdA structure, this used 225 x-direction, 161 y-direction, and 193 z-direction points per processor for grid-based discretization. The focusing calculation was carried out with the following coarse mesh domain lengths: x-direction 153.095, y-direction 114.974, z-direction 126.133. Fine mesh lengths were: x-direction 110.056, y-direction 87.632, z-direction 94.196.

Both grids were centered at the center of mass of the protein structure. No mobile ion species were included in the calculation. The biomolecular dielectric constant was set at 2, and the dielectric of the solvent was set at 78.54. Point charges were mapped to the grid using cubic B-spline discretization(493), and 10 grid points per square angstroms were used in surface construction. Ion accessibility coefficients were defined based on a molecular surface definition, with an ion accessibility coefficient defined by an "inflated" van der Waals model and smoothed by a 9-point harmonic averaging(485). The temperature was set at 298.15K for the PBE calculation.

Mapping of the density to structure was performed using a pseudo-Voronoi approach. All electrostatic potential closest to a given atom was defined as a property of that atom, and atom "electrostatics" were summed per residue. This allows representation on a ribbon projection or surface. By subtracting the allocated density maps it is possible to determine which residues experience changes in the nearby

electrostatic field across a pH range. This was performed using the APBS generated electrostatic potential maps for both TcdA and TcdB across the range of 5.5-8, in 0.5 pH unit intervals.

MCCE calculations were run at the standard full level of analysis. pH titration was carried out between pH 0-14 at 1 pH unit intervals. Generation of rotamers was carried out by creating heavy atom rotamers followed by protonation at low energy positions. Using the Dunbrack rotamer library conformers all possible conformations for all sidechains were generated. The protein structure was stripped of residues and generated with all sidechains in “allowed conformations". Clash tolerance was set at 2Å and multiple protonation states were allowed for titratable residues. Structures were created with 5000 repacks per conformer were allowed with an occupancy cutoff of 0.01. Delphi was used to calculate the reaction field energy and pairwise electrostatics. Solvent dielectric constant was set at 80, and the 65 grids per calculation were used with a target of 2 grid points per angstrom. The probe radius used was 1.4Å, with an ion radius of 2Å and a final salt concentration of 0.15 mMol. Monte Carlo sampling was used to determine which conformers dominate at each pH. The number of sampling steps was set to2000 times the number of conformers generated. Energy was traced for 50,000 steps per round with a maximum of 1 million microstates allowed during the analytical solution step. A protein microstate is a combination of one conformer for each residue, any cofactors present, and water. pKa determination is done by determining which conformers dominate when we perturb all possible microstates. This was performed from pH 0-14, to obtain the pKa of each titrating group.

Both molecular dynamics simulations were done using the CHARMM force field with NAMD, on the WSU rocks cluster. The canonical ensemble was used with periodic boundary conditions, Langevin dynamics and thermostat. Simulations were monitored by GMD, and simulation stability was assessed using the trajectory analysis tools under VMD monitoring energy and RMSD. The systems were solvated with TIP3P water, neutralized with counterions and minimized for 1000 steps of conjugate gradient minimization. A smooth ramp to 300K was used to bring the simulation to production temperature. A 1 fs timestep was used in the calculation and frames were written every 1ps. Both simulations consist of approximately 57,000 atoms, with approximate box dimensions of 100Å x 70Å x 80Å. The solvation box includes a 10Å pad on each face of the box. Short range electrostatics were calculated with a non-bonded cutoff of 8Å, with switching between 7-8Å. Long range electrostatics were calculated with the smooth particle mesh Ewald method. Results were analyzed by GMD, PCA and assessed for hydrogen bonds, solvent interactions and salt bridges. Timescapes, bio3d for R and VMD were used for these calculations, respectively. For the hydrogen bond analyses, intramolecular hydrogen bonds were considered stable if they were present for 50% of the simulation or more, and solvent-protein hydrogen bonds were considered stable if they were present for 90% of the simulation or more. For the purpose of categorizing these interactions, "active site" residues are those within 10Å of the crystallographic UDP-Glucose, and "protein-protein interaction face" residues are those between 440 and 497 as this region encompasses all known residues critical for protein-protein association(411, 413, 415).

Chapter 6 CoGent-Seq: Connecting genes to sequence. Relating motif sequences to annotated gene structures.

6.1 Background

As increasing interest in protein binding motifs and regulatory elements develops, determining which genes contain these elements is required. Genome is a program designed to quickly search regions of a genome near either the 5' or 3' UTR for a pattern, and to return a user defined segment of genome for analysis. It can be applied to any genome, and any given set of annotations. It is capable of matching multiple patterns and returning a true/false match for each query. This program is user-friendly, requires minimal user interaction with input, and returns data ready for analysis via blast, structural characterization or statistical analysis. To date it has been applied to searches for Dcm methylation sites, and Hfq binding sites in 5' UTRs of the E.Coli genome. This chapter will include a brief description of the application of CoGent-Seq to the search for Hfq mediated regulatory sRNA's.

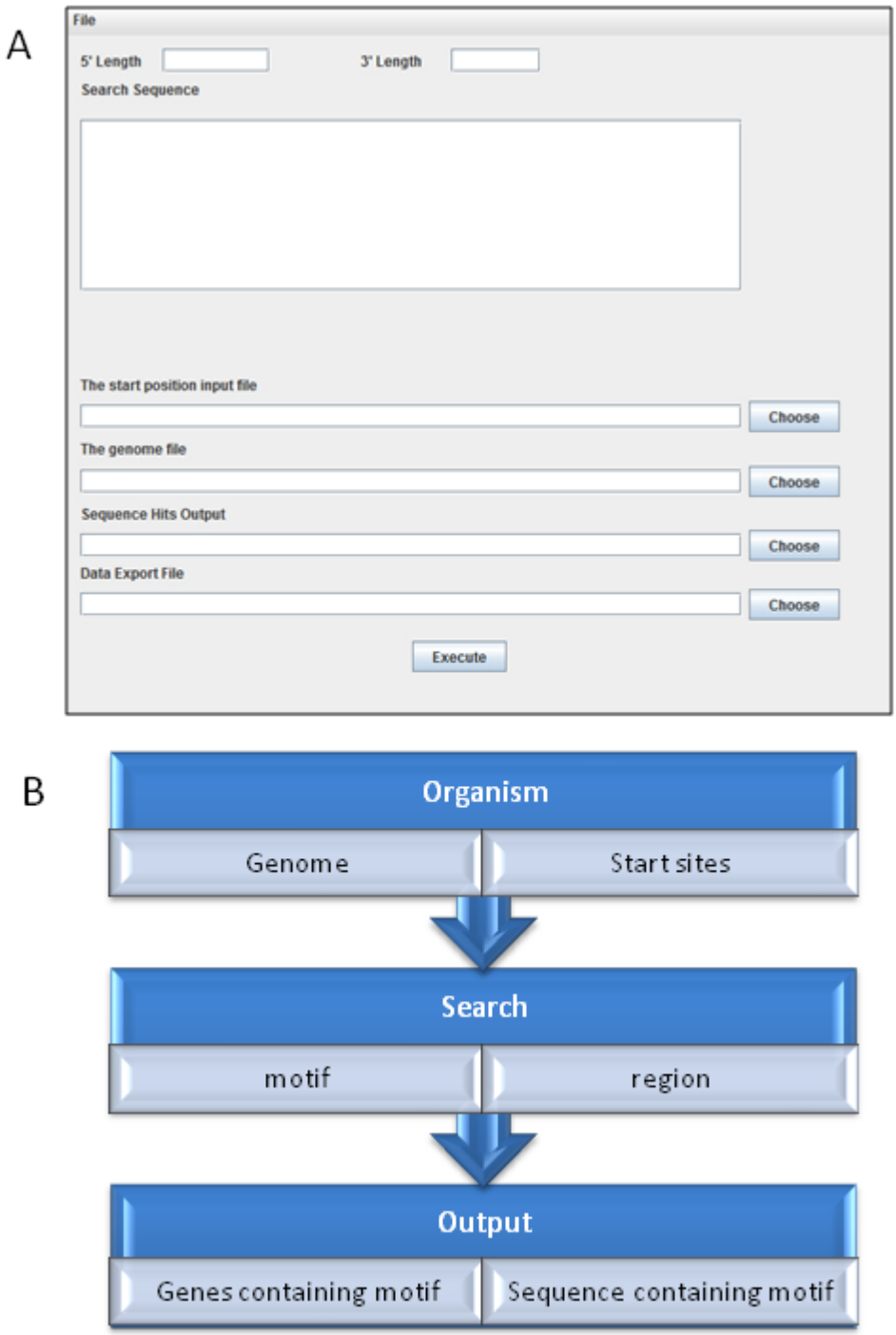
Detection of consensus sequences and recognition motifs can be performed by employing statistical methods to detect enrichment of various sequences in a genome. From the perspective of a molecular biologist however, detection of consensus sequences or motifs is largely informed from wet lab studies. Much of this research is focused on regulatory elements in either the 5' or 3' UTR, with novel sequences being reported frequently. Once found experimentally, exploration of these sequences within a genome may be attempted, but without simple, user-friendly tools, this may be a daunting prospect for a molecular biologist. It is tedious and time consuming to perform these searches by hand, and so here we present software capable of automating this task. Given a sequence such as a protein recognition motif, modification flag, or contextual region within a noncoding RNA, CoGent-Seq will return a list of

all genes containing that motif within a user specified region at either the 5' or 3' end of a gene, along with a segment of sequence to facilitate further analysis.

6.2 CoGent-Seq

A user need only supply the genome for query in the form of a FASTA formatted text file, the positions of interest for searching nearby as a .csv formatted list, the pattern or patterns to match (i.e. AATTCCGG) and the region the user would like excised for further study if desired. The entire program was written in Java, so as to facilitate cross-platform use. It can be used either command line or with a GUI, both of which are distributed as executable jar files. Results are given as a csv formatted table returning the start site, the motif or motifs searched with a true/false identifier, and the segment of sequence requested by the user. For negative sense genes, This is ideal for RNA studies where immediate secondary structure prediction of the region is required.

Figure 1A shows the graphical user interface, while Figure 1B shows a typical workflow for this software. Upon selecting the version of a genome appropriate to a researchers work, the number of nucleotides to the 5' and 3' regions are declared, as is the motif of interest. Any number of motifs may be searched simultaneously. The list of either 5' or 3' sites appropriate to the organism of study is required in .csv format, which is readily available from several databases. Output is returned in two .csv formatted files. The "*sequence hits"* file returns the list of start sites, and reports whether each motif has been found. The "*data export"* file returns the segment of sequence requested by the user for each start site region that contains the motif requested. These are semicolon separated for use with either BLAST searches or Mfold analysis.



6-1 Graphical user interface and workflow for CoGent-Seq

CoGent-Seq requires minimal user inputs. The graphical user interface in panel A, describes these fields. From top to bottom, length of search into 5' and 3' regions, sequence to match, a list of annotations for searching, a genome in fasta format, and then filenames for data export. A sample workflow would include selecting an organism of study, downloading the genome and annotations, applying the search by selecting a motif and local range for the search, and CoGent-Seq returns both a list of genes containing that motif, and the sequences queried.

6.3 Application to E. Coli Hfq-mediated sRNA regulation.

A first application of CoGent-Seq to a bacterial system involved the search for Hfq mediated sRNA-mRNA interactions in E. Coli. Hfq is known to assist in regulatory RNA function for several stress-response systems, and it has a known putative binding sequence. CoGent-Seq was applied to find all mRNAs in E. Coli containing the sequence AAYAA, to which Hfq binds. It was later determined that AAYAA is a subset of the Hfq binding sequence (ARN)x, where x=2 or more. The purpose of these studies was to attempt to find novel regulatory sRNA-mRNA pairs employing Hfq as a chaperone. Following the initial bioinformatics search, Martha Faner pursued the prospective mRNA's and performed both structural bioinformatic analysis and *in vitro* testing.

A list of all E.Coli gene start positions and sense were obtained from the EcoGene database and formatted as a .csv file(494). The genes were sorted by sense, and the start positions for both forward and reverse sense genes were output to separate files. A search was performed across the *E. coli* K-12 genome wherein the region from -200 to +60 was searched for the sequence AATAA or AACAA, setting zero as each gene start position iteratively. The 260 nucleotide range and start position were outputed into a .csv file by line for all lines containing either the AATAA or AACAA sequence. This process was repeated for all negative sense genes using the E.Coli K-12 genome complement strand sequence. Start position was matched back to gene name for systems analysis and the extracted 260 nucleotide region was submitted to mfold for structural analysis. Annotated transcription start sites for the biocyc database were used to discard any mRNA that contained and AAYAA in the region -200 to +60 but within the start site(495).

Martha Faner subjected the results of this CoGent-Seq search to structure approximation, assessment of sRNA potential binding partners, followed by *in vitro* testing. Structural assessment of known Hfq mediated mRNAs indicates that Hfq prefers a single stranded ARNx motif, flanked by structure. The mRNAs returned from this search were subjected to the folding predictor mfold, and mRNAs without ARNx sites meeting this structural requirement were discarded. This leaves a pool of roughly 20% of the annotated E.Coli mRNA's with potential Hfq mediated sRNA regulation. These were then subjected to sRNA matching using IntaRNA, resulting in a list of mRNAs with potential sRNA binding partners (M. Faner unpublished data).

Two mRNAs were selected for further study, *nhaA* and *mak*, both having logical connections to the type of stress-response typically mediated by Hfq. *nhaA* encodes a sodium antiporter used to maintain pH and sodium levels (496). *mak* endcoes a mano(fructo)kinase which participates in fructose metabolism (497). Their IntaRNA predicted partners were RhyeB for *mak* and RyfA for *nhaA.* Using electromobility gel shift assays it was determined that in the presence of Hfq, ternary complexes of the two RNA binding partners in complex with Hfq were stable. An *in vivo* GFP assay was applied to determine whether or not these interactions were regulatory in nature. It was determined that in the presence of RyfA, the expression of NhaA was downregulated. This constitutes the successful discovery of a novel sRNA-mRNA regulatory pair.

6.4 Conclusions

The discovery of numerous novel regulatory elements in recent years, along with the increasing interest in regulatory pathways is placing a demand on computational chemist to produce user friendly tools that can both inform and be informed by experimental data. CoGent-Seq is a simple, yet powerful tool that facilitates broad study of regulation at the 5' or 3' UTR, by individuals that may have varying levels of bioinformatic experience. Basic user input

requirements, along with a streamlined GUI and easy to parse output allow users with little to no interest in computational methods to easily perform genome wide analyses. Following the experimental detection of a regulatory or protein binding sequence, the results from a CoGent-Seq analysis will indicate which other systems may be involved through the same regulatory element. In application to *E. Coli* regulatory RNA systems chaperoned by Hfq, this led to the discovery and validation of one completely novel interaction. The purpose of CoGent-Seq was to provide a simple interface that provides initial motif searching for deeper bioinformatic or *in vitro* discovery. While this is only one of many possible applications of CoGent-Seq, it is an excellent illustration of the potential of this tool.

Chapter 7 Hypothesis Driven Single Nucleotide Polymorphism Search (HyDn-SNP-S)[3]

7.1 Background

In recent years, the amount of genomic data on disease phenotypes has increased exponentially. The decreasing cost of genotyping, along with the future promise of personalized medicine has resulted in a boom in individual genomic data (498-502). Most bioinformatic techniques determine clusters of mutations that may be followed and used as a diagnostic tool in various diseases (503-506). Traditional analysis of genome wide association studies (GWAS) focus on a single phenotype, and aim to find SNPs that show statistically significant association with the phenotype in any of the measured genes. In most cases these analysis do not have an *a priori* hypothesis of the locations of the SNPs. Therefore, very stringent statistical criteria are needed to obtain SNPs that are predictive, resulting in only a small number of SNPs being identified. Few studies have leveraged the vast information generated to identify new SNPs with clear functional impact on disease onset (507-510).

Moreover, tracking a mutation resulting from these SNPs through transcription and translation to their ultimate effects in a cell is largely left to the scientific community at large. In addition, correlating a mutation to a phenotype is a daunting task for researchers who typically work at a cellular level. Most biochemists or molecular biologists have a biosystem of interest, and broad sweeping GWAS studies are typically intractable for their purposes. It was our intent to create a tool that would allow a user to query genome-level data for their system of interest. There have been examples previously where this has been done on a single GWAS study, but not on combined data sets(511). To this end, we developed and implemented an algorithm with which a researcher can directly query one or several GWAS studies for their gene region of

[3] Sections of Chapter 7 have been previously published (DNA Repair *In Press*)

interest. We term this method Hypothesis Driven SNP Search (HyDn-SNP-S). The software returns all SNP mutations within their gene region, along with regarding the phenotype associated with the mutation. Further statistical methods can then be applied to the GWAS data. Additionally, by returning a focused set of mutations, tracking the consequences of the mutations through transcription, RNA processing and translation becomes trivial.

The workflow shown in Figure 1 outlines the process used in HyDn-SNP-S. Researchers can select one or many studies associated with the phenotype(s) of interest, apply HyDn-SNP-S to search for their gene of interest, and further analysis can be performed, as desired, on the output. Our program returns both intronic and exonic SNPs, allowing the investigation of impact on RNA, processing or protein sequence. The simplicity of the software implementation makes this method ideal for researchers uncomfortable with large-scale bioinformatic analyses, or those who lack the resources to perform such studies.

Figure 1

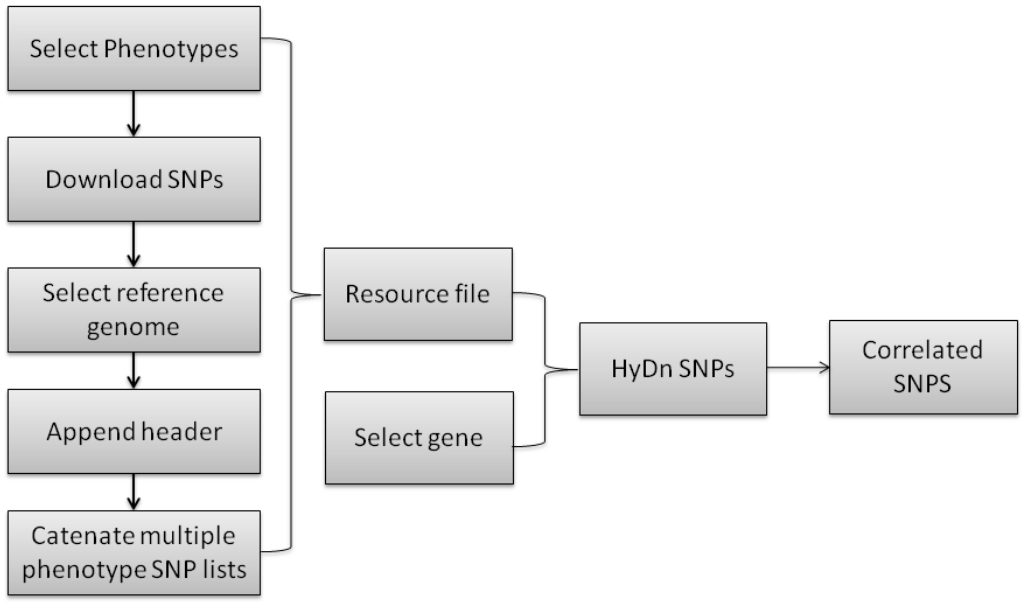


Figure 7-1 : Flowchart of the HyDn-SNPs method.

Upon development of a hypothesis, researchers select GWAS studies with relevant phenotypes, and obtain locations of the genes of interest. Following application of the algorithm, SNPs can be separated by intronic or exonic. Further analysis can be performed by *in vitro* validation or computational studies.

The database of Genotypes and Phenotypes (dbGaP, http://www.ncbi.nlm.nih.gov/gap) was developed to "archive and distribute the results of studies that have investigated the interaction of genotype and phenotype". Results are distributed in the form of either raw genotyping data tagged with individual specific data such as gender, race and onset of phenotype; or as catenated lists of SNPs. This repository provides an ideal source of GWAS data useful to researchers with a targeted interest. Users are able to freely download the sets of SNPs, in a standard format for use with our software. Any phenotype of interest that is represented in this database would be a possible point of study for a HyDn-SNP-S study.

In this chapter, we present the development of HyDn-SNP-S and its application to search for cancer related SNPs on DNA polymerases and Histone Deacetylases. These enzymes are involved in all processes related to DNA duplication. The efficiency and fidelity of the processes involved with duplication of DNA are critical since errors can lead to carcinogenesis. Numerous studies indicate that mutations in DNA polymerases affect characteristics ranging from fidelity, to nucleotide incorporation rate, to cell proliferation (450, 512-522). However, a direct link has not been established between these mutations and cancer onset.

Our results uncovered a large number of cancer related SNPs on DNA polymerases. Statistical analysis on selected studies reveals for the first time the possibility that DNA POLL could be a major contributor to cancer risk. Molecular dynamics simulations were performed on wild-type and a SNP mutant on POLL to further investigate the functional impact of the mutation.

7.2 Hy-Dn-SNP-S

The HyDn-SNP-S method returns results from whole genome genotyping studies rapidly, far faster than traditional bioinformatic methods. Pre-screening with HyDn-SNP-S

dramatically decreases the time required to perform statistical analysis on GWAS data, by excluding all mutations not relevant to a researcher's hypothesis. As proof of concept four genotyping studies have been statistically analyzed following application of the HyDn-SNP-S method. Additionally, one mutation, determined to be both statistically significant and of structural interest, was subjected to molecular dynamics studies and subsequent analysis.

7.2.1 Application to cancer phenotypes

Hy-Dn-SNP-S was applied to four cancer phenotype studies, melanoma, breast, lung and prostate cancer (523-528). A search for mutations in all polymerase genes was performed, resulting in a total of 708 cancer associated mutations. Of these mutations, 491 were intronic, and 217 were exonic. Additionally, four of the exonic mutations were found to be at splice sites. As per the workflow described above, all four searches were carried out simultaneously, and results were available within a few minutes. The four studies were subjected to traditional biostatistical analysis (529) following application of the HyDn-SNP-S. The focused nature of the search allows for relaxation of the more stringent mathematical methods, and facilitates more thorough analysis of the resulting mutations. Haplotype analysis on whole genome genotyping data is frequently not performed since the combinatorial nature of these studies across all mutations would be prohibitively expensive, computationally speaking. As the dataset used for analysis following the HyDn SNP-S method has significantly reduced complexity, these targeted studies can detect mutations of moderate significance that would be overlooked in traditional bioinformatic analyses and perform these searches more rapidly than is typically possible.

7.2.2 Statistical analysis

Logistic regression and haplotype analysis was performed on the resulting SNPs to determine their statistical significance. Results are presented in Table Table 7-1:

Table 7-1 DNA polymerase SNPs correlated to four cancer studies

Phenotype	Total SNPs	Significant
Prostate	69	11
Melanoma	215	26
Breast	100	22
Lung	51	20

This data is shown in The GWAS for prostate cancer yielded 69 SNPs in the genes of interest. Association of 11 SNPs with prostate cancer was statistically significant for at least one genetic model. The melanoma cancer case/control database examined yielded 215 SNPs in the genes of interest. Twenty-six of them were significantly associated with disease status for at least one genetic model after controlling for age and gender. The breast cancer study yielded 100 SNPs in the genes of interest. Twenty-two of them were statistically significantly associated with breast cancer status for at least one genetic model. The lung cancer case/control database examined yielded 51 SNPs in the genes of interest. Twenty of them were statistically significantly associated with lung cancer status for at least one genetic model. Table 7-2 reports the identities of the significant SNPs, their p-value and corresponding POL gene.

Analysis to determine the association between the derived haplotypes from each gene and disease status was performed. No haplotypes were predictive of disease status for the lung cancer study nor the melanoma study using any of the three genetic models. However, the haplotypes constructed from SNPs on POLL were borderline significant for the breast cancer study using a recessive (p-value = 0.048) or additive (p-value = 0.091) model formulation. This haplotype is constructed from two SNPs: rs3730477 (C>T; R438W) and rs3730463 (A>C: T221P). The odds ratios from individual significant and borderline significant contrasts within

each model type are reported below. In the case of the additive model, for each additional C-A haplotype observed, the odds of breast cancer are multiplied by 1.15, (p-value =0.029). Similarly, for each additional C-C haplotype observed, the odds of breast cancer are multiplied by 0.812 (p-value =0.062), i.e., a protective genotype. For the recessive model, having 0 or 1 copy of the C-A haplotype results in the odds of breast cancer being multiplied by 0.829 relative to having 2 copies of the C-A haplotype (p=0.026). Having 0 or 1 copy of the C-C haplotype results in the odds of breast cancer being multiplied by 3.01 relative to having two copies of the C-C haplotype (p=0.099). The haplotypes constructed from SNPs on the PolG genes were significant for prostate cancer. This haplotype is constructed from three SNPs: rs3087374, rs2351000 and rs2247233. The odds ratios from individual significant contrasts within each model type are reported below. For the recessive model, having 0 or 1 copy of the G-T-G haplotype results in the odds of prostate cancer being multiplied by 1.33 relative to having 2 copies of the G-T-G haplotype (p=0.005). Having 2 copies of the G-C-A haplotype results in 9.64 of the odds of prostate cancer as compared to having 0 or 1 copy of the G-C-A haplotype (p=0.008).

A literature search indicates that only one of these statistically evaluated mutations has been explored *in vitro* (530). Experimental analysis of the mutations we report here is outside the scope of this work. However, the mutations arising from these SNPs present interesting targets for further experimental studies.

7.2.3 Edge-Node Analysis

The data resulting from a HyDn-SNP-S search can be discussed at the molecular level, and in the context of predictive power, but due to the nature of these studies, it also allows a much broader basis of understanding. Relating many phenotypes to many polymerases generates a network of data best represented by an edge-node interactive diagram. Using the number of SNPs as a weighting property, it is possible to broadly examine the complete network

of phenotype-polymerase interactions. Figure 2 shows a flattened version of this data, limited to statistically explored phenotype data, which is available in interactive form online at http://chem.wayne.edu/cisnerosgroup/gexf-js2/index2.html

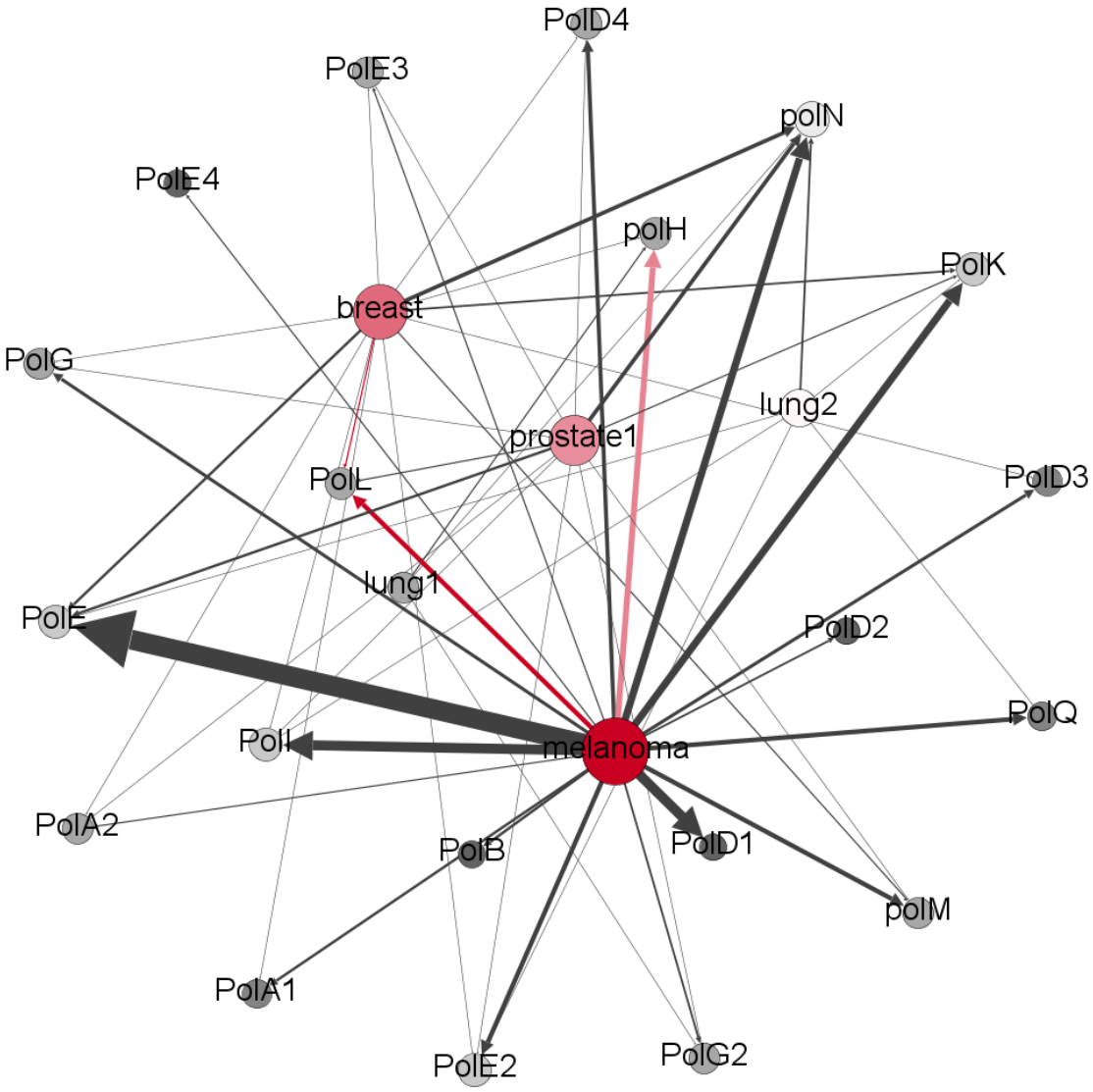


Figure 7-2 Edge-node network of the HyDn-SNPs results.

Phenotypes and polymerases are shown as nodes, edges are weighted by total number of SNPs connecting each phenotype to each polymerase. Statistically validated connections are shown in red.

The complete dataset is also available in this form. Both polymerases and phenotypic studies are represented as nodes, sized according to SNP density. By clicking on a node, all

connections to that node will be listed. Clicking on any of those connections will return all connections to that selected node.

The interactive map allows investigation of the network associations of various phenotypes and polymerases and the complete list of connections is available for download from the top of the page. This file also includes the translated mutations. Many diseases are not caused by a single point mutation, but rather by a collection of factors. As the formatting for the results of HyDn-SNP-S is well suited to network analysis, and additional data can be garnered as desired from the genotyping studies, this approach may have critical importance in searching for combinations of factors that may be predictive for disease. Due to the targeted nature of the search, there is a significant reduction in the analytic space and thus, more thorough analysis can be performed. The haplotype described above is one example; individually the two mutations would have been overlooked by traditional analysis, but in combination they are strongly predictive.

7.2.4 Molecular Dynamics investigations of DNA Polymerase λ mutant R438W

To further validate that hypothesis driven analysis of whole genome genotyping data is valuable to researchers, we sought to study a mutation with statistical significance that would have been overlooked by traditional methods. Of the two mutations that comprise the haplotype linking POLL to breast cancer, only the mutation R438W is in the polymerase domain. This position is not close to the active site, but it is within 14 Å of Loop 1, which has been shown to be critical for fidelity (531). The R438W SNP mutation has been previously shown to contribute to decreased fidelity in vitro, increased mutation frequency, and generation of chromosomal abnormalities (532). An 8-fold increase in inaccurate substitutions was observed in base substitution assays and karyotypic analysis of several cell lines carrying this mutation also reported a high level of spontaneous or IR-induced chromosomal aberrations. With ample

evidence to suggest a molecular basis for these results, we selected DNA polymerase lambda R438W for further study.

Four MD simulations were performed using crystal structures 1RZT and 2PFQ. These structures were selected as they represent the binary and ternary complexes of Pol Lambda, respectively. The change in Loop 1 conformation between the two structures is shown in Figure 7-3 panel A. Panel B shows a closer view of the loop conformations, indicating both binary and ternary conformations. Panels C and D illustrate the relative proximity to the R438W mutation. As the structure transits between the two loop conformations, the mutation ranges from roughly 12.6 Å to 14.3 Å away. Mutations in this loop have been shown to have no effect on catalytic rate while simultaneously increasing the number of misincorporations, thus Loop 1 is critical for polymerase fidelity (531). Following a 14 ns simulation, correlation analysis was carried out to determine whether the residues in Loop 1 were affected by the mutation.

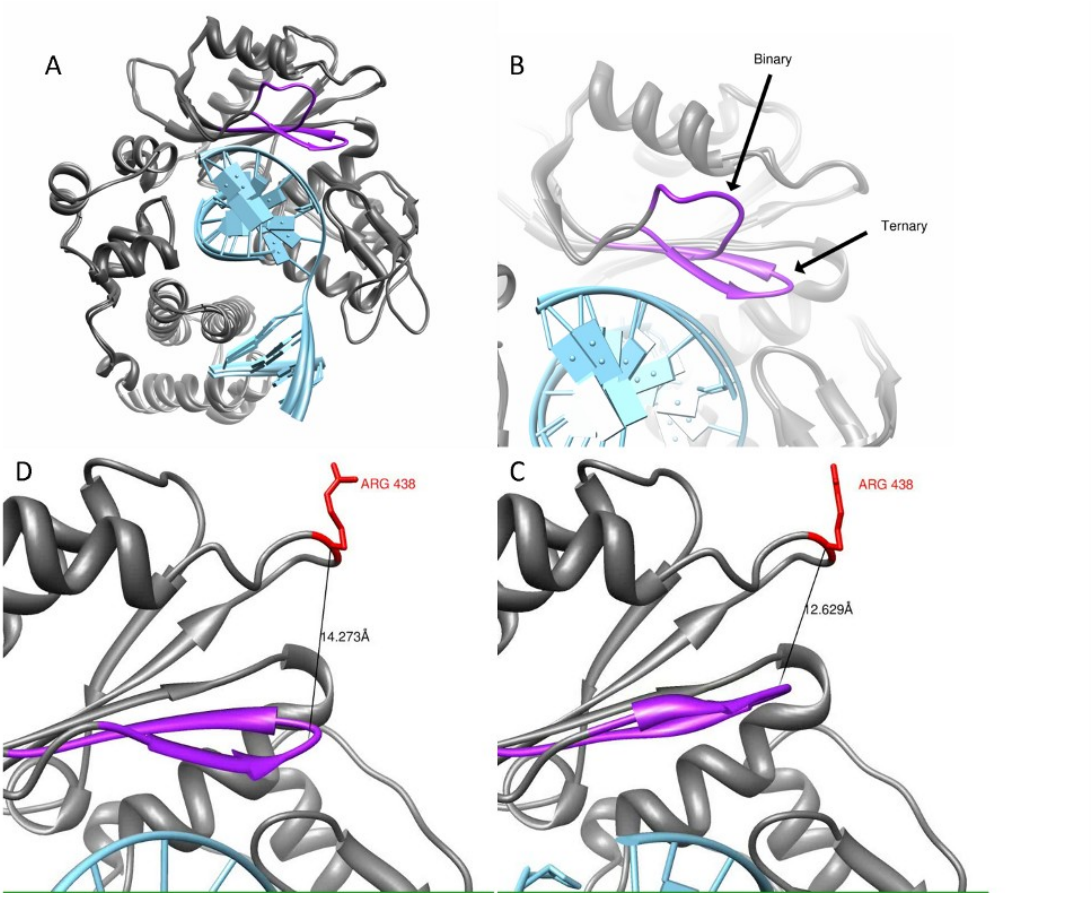


Figure 7-3 Structure and relevant regions of PolL

(A) Overlay of Pol□□in the binary and ternary conformations. DNA is shown in light blue, and the Loop 1 is shown in purple. (B) Differences in loop 1 orientation between the two conformations. Distance between position 438 and loop 1 following an interpolation between the two structures at its furthest (Panel D) and closest (Panel C) approaches.

7.2.5 Correlation and GMD analysis of PolL R438W

These simulations were analyzed by both GMD and correlation analysis. Details of GMD analysis were described in Chapter 2.7.4. Briefly, it calculates the number of significant and persistent motions experienced by a structure, and plots them against frame number. This gives an approximation of the overall level of activity and can be used for event detection. Correlation analysis can determine whether residues are moving in the same direction in space

(correlation), or opposite directions at the same time (anti-correlation) through the course of a trajectory. Specific residues or regions on a protein that show correlation frequently have structural implications for function(533). This is plotted on an orange to blue scale with color intensity depicting high correlation or anti-correlation. Residues whose motion is unrelated are presented in white.

As shown in Figure 7-4, the binary complex shows little change in correlation between the wild-type and mutant structures. Conversely, the ternary complex shows high correlation and anti-correlation in two regions. The highest points of correlation are between residues at position 438 and 569, as well as between 438 and 420. Residues showing the greatest change in correlation in the ternary complex were mapped to the structure and colored orange as shown in Figure 4C. It is notable that a majority of these residues are on Loop1. To further understand the impact of the SNP mutation on Loop 1, the correlation data between position 438 and all residues in Loop 1 was extracted and plotted. Figure 4D shows that although there is higher correlation in the ternary complex between the wild-type and mutant, both complexes show altered correlation between the wild-type and mutant. The sum of these analyses suggests that the introduction of the R438W mutation alters the overall correlation pattern in the ternary complex, but more importantly, directly affects the motions of Loop 1 in both complexes. Since loop 1 regulates fidelity, and transit between the two conformations shown in Figure 3 is required for catalysis, the SNP leading to the R438W mutation likely has direct effects on polymerase activity *in vitro* and *in vivo*.

In addition to the correlation analysis, generalized masked Delaunay (GMD) analysis was performed to determine the impact, if any, on the overall dynamics of the protein (during the simulation time). The results are shown in Figure 5, events are plotted on the Y-axis, and frame number is plotted on the X-axis. GMD defines events as persistent motions across the masked Delaunay reduced representation of the protein structure. Panels A and B show the wild type

activity for both the binary and ternary complexes, with average activity levels of roughly 0.2 events per frame. These patterns are typical of stable simulations, where no major rearrangement is occurring. The alteration in the correlation plots combined with the stability of the GMD indicates that the mutation induces only local alterations in activity.

The sum of the correlation and GMD analysis indicates that the R438W mutation appears to modify only the movement of Loop 1, while the overall dynamics are not significantly altered. This provides context for the experimental work by Terrados et al (532). Their experiments indicated that the R438W mutation increases the error rate of Pol lambda, but does not alter the overall rate of polymerization. Our results indicate that the R438W mutation alters only the behavior of Loop 1, while leaving the overall conformational motions of Pol lambda unperturbed. This would agree with the behavior observed experimentally. The R438W mutation alters the behavior of Loop 1, thus decreasing fidelity, while the overall behavior of the polymerase is unaffected, allowing it to maintain a normal rate of polymerization.

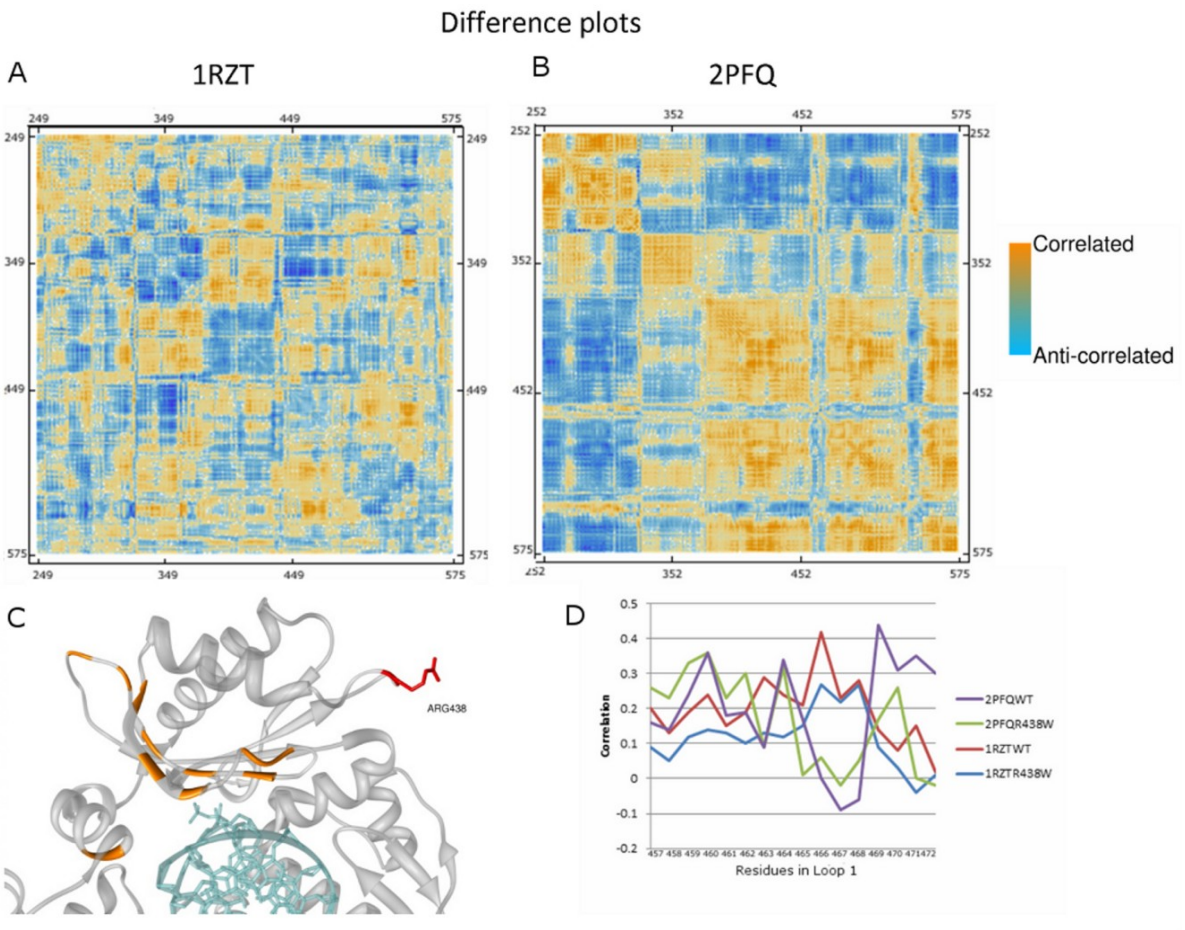


Figure 7-4 Correlation difference plots for the binary (A) and ternary (B) conformations relative to the wild type. Highly correlated residues are shown in orange (C) and total correlation for residues in Loop 1 is shown in (D)

Increases in correlation are shown in orange, while increases in anti-correlated motions are shown in blue. In both cases, alterations in the correlation plots are visible, more notably in the ternary complex. The highest values from the ternary complex correlation plots were mapped back to the residues affected, and are colored orange in Panel C. Notably many of these residues are on Loop 1. Panel D shows the individual correlation values for each of the residues in Loop 1. While the binary complex shows moderate alteration on several , the ternary complex shows considerable differences for several residues, particularly between residues 469 and 472.

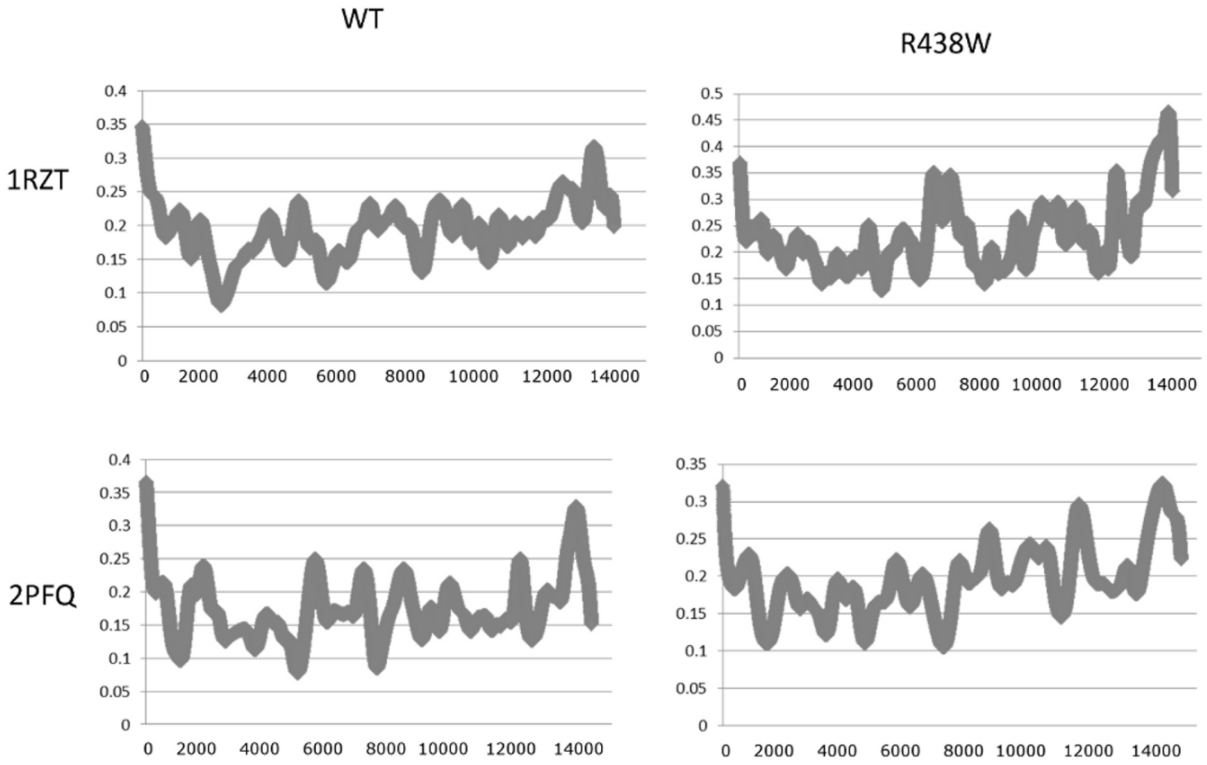


Figure 7-5 : GMD plots for the binary and ternary complex simulations in both wild type and mutant form.

No drastic differences are apparent between the four simulations indicating that all four are showing the same general level of physical activity. This result indicates that the overall motions of the polymerase are not perturbed by the presence of the mutations. In light of the data presented in Figure 4, this indicates that the significant alterations in conformational space are restricted to the Loop 1 region.

7.2.6 Associated Studies

During the development of this project, HyDn-SNP-S was also applied to Histone Deacetylases (HDACs). Further analysis of many of these projects is pending, but the work on HDACs has proved interesting. HDACs are responsible for the modification of histone tails during the process of chromatin repacking, and several have been known to be implicated cancer, including breast and colon (534-540). Following application of HyDn SNPs to HDACs 1,3, and 6, several mutations were both detected and pursued in various manners. Mutations found in HDAC 3 correlated to breast cancer were determined to be near the active site, and thus docking studies were performed to determine whether or not substrate binding was affected. Figure 7-6 shows the locations of the two mutations.

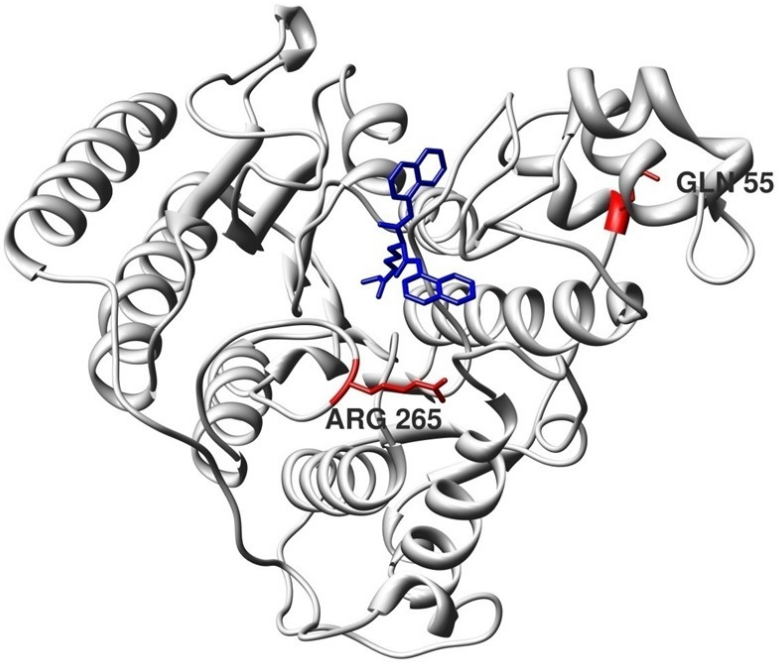


Figure 7-6 Two mutations on correlated to cancer mapped to HDAC3.

The Arg 265 mutation was determined to be in proximity to the typical SAHA binding site. Comparative dockings were performed to determine the possibility of altered substrate binding. Shown in blue is the inhibitor SK-691.

The R265W was determined to be adjacent to the natural substrate binding site, and a library of known HDAC3 inhibitors were docked to the structure (PDBID:4A69) to determine whether or not this mutation might have implications in the treatment of breast cancer. Following docking to both the wild type and mutant HDAC3, the docking scores were calculated. As shown in Table 7-2, several of the inhibitors show preference for either the mutant or wild type. Negative scores indicate preferred binding to the wild type, while positive scores indicate preferred binding to the mutant.

Table 7-2 Table of inhibitors showing preferential binding to wild type or mutant HDAC 3†

Inhibitor	WT	R265P	Δbinding
SK-691	-33.01	-36.86	3.85
SK-692	-31.36	-33.64	2.28
SK-683	-30.89	-30.73	-0.16
SK-658	-27.57	-30.8	3.23
TSA	-25.25	-22.96	-2.29
APHA-1	-22.47	-24.06	1.59
APHA-8	-20.63	-24.67	4.04
SAHA	-18.83	-21.55	2.72
CG1521	-16.18	-15.3	-0.88

† Inhibitor structures were retrieved from Wang et al (541). Docking scores of Wild type and the R265P mutants are shown, Δbinding is the difference in LeadIT docking scores.

These studies indicate that a breast cancer-associated SNP in HDAC3 may have implications for treatment based on differential substrate binding. Isoform specificity is a problem for drug design with respect to HDACs as their high degree of similarity leads to off-target inhibition. It may be possible in the future to determine if a specific drug may be preferable for treatment depending on the gene sequence of the individual patient.

7.3 Conclusions

We have developed a powerful method that allows researchers to interact with whole genome genotyping data in a focused, hypothesis driven way. By allowing researchers to find data on their own systems of interest, we will expedite the study of any mutations that may logically be connected to a phenotype. Also, the focused nature of these searches will allow more thorough statistical analysis, and appropriate recognition to combinations of factors that would be difficult to fully assess in an extremely broad GWAS analysis. By applying this methodology to our system of interest we were able make the first direct statistical link between DNA polymerases and cancer, define two haplotypes with strong predictive power, and trace a cancer-associated mutation to a structural effect in the translated protein and investigate its functional impact by computational simulations. Furthermore we were able to extend these studies to several other systems, one of which has shown promise in the development of isoform specific histone deacetylase inhibitors.

7.4 Methods

In this section we describe the algorithm to search for disease related SNPs based on a given hypothesis and its implementation in an easy to use software package. Subsequently we describe the statistical methods to determine the association of the SNPs with the phenotype. This is followed by a description of the graph analysis of the resulting data from HyDn-SNP-S for the present studies. Finally, the details of molecular dynamics (MD) simulations on DNA polymerase lambda structures are described.

HyDn-SNP-S: SNP collections deposited through studies on the database of genotypes and phenotypes are obtained. A header is appended to the data declaring the phenotype associated with each individual mutation. Mutations listed relative to the HuRef and Celera

genomes are removed, as we are working within the frame of reference of the GrCh37 human genome reference build. The SNP collections are then catenated into a searchable resource file. Following generation of this resource file, the program HyDn SNPs was used to search for mutations within the gene region of interest. Users enter the chromosome, and gene range for searching, and point the program to the resource file. Sample resource files are available with the HyDn SNPS download. Further information and instructions are available in the documentation for this program. Any SNPs found that match the chromosome and gene location range are deposited into a results file. This file lists all the SNP associated information, such as ss and rs number, allele, chromosome, chromosomal location, contig number, and contig location, and type of chip used in the original genotyping experiment. These can then be categorized by location; intronic, exonic, or at a splice site. For our purposes, exonic SNPs were then compared to reference SNPs to ascertain the extent of prior investigation, as well as relative allele frequency in the natural population. The consequence of any given SNP was determined either by use of the reference SNP database, or in the case of previously unreported SNPs, translated by use of a DNA codon table in conjunction with the gene sequence and protein sequence. HyDnSNPs is available for download at the link provided in the abstract.

Statistical Analysis: We utilized four publically available case/control genome wide association studies (GWAS) from dbGAP (access request # 1961) across multiple cancer types (including breast, melanoma. lung and prostate cancers)(14, 523, 524, 526-528) to determine if SNPs or haplotypes constructed from SNPs in our genes of interest are associated with a disease phenotype. Additionally, we determined if any synergistic results across multiple databases exist that may imply a common cancer genesis. Multiple genetic modes of inheritance were examined: additive, dominant, recessive and genotypic in a covariate-adjusted logistic regression analysis associating each SNP with the disease phenotype. The maximum

likelihood estimate of the posterior probabilities of haplotypes for each observation was produced using the EM algorithm. Score statistics for the association of the haplotypes with the cancer phenotype were constructed using these posterior probabilities. We use the R package “haplo.stats“ to implement these haplotype functions (529). Logistic regression is also used to estimate the association of a haplotype with the disease phenotype, given the genetic context. As we focus on the SNPs in only a few genes, we avoid issues with multiple testing, which are burdensome when trying to evaluate the association between genetic markers and a disease phenotype when measuring thousands or millions of genetic variants.

Graph analysis: For ease of visual analysis, the data resulting from the HyDn-SNP-S search has been transformed into edge-node format to allow visual interpretation of the networks of phenotypes and polymerases involved in tumorigenesis. Frequently more than one polymerase was found to have single point mutations within a cohort of cancer patients; network analysis allows for easy visual interpretation. Edge-node tables were csv formatted for use in Gephi (542), visualization was performed with a Fruchterman-Reingold (543) algorithm using an area of 15,000 and a gravity of 7.0. Nodes and edges were weighted by degree; for these analyses, weight was the number of mutations occurring between each phenotype and polymerase.

MD simulations: MD was performed on wild-type and the R438W mutant of POLλ in the binary and ternary conformations (PDBID: 1RZT, 2PFQ) using NAMD. The simulations were performed using a parallel build of NAMD(428) employing the CHARMM(429) force field on the XSEDE Teragrid. The structures were solvated, and appropriate counterions were added to reach 0.5mM NaCl. A timestep of one femtosecond was used, a Langevin thermostat was used to maintain temperature at 300K, and a Nose-Hoover Langevin combination method was used to control pressure. The systems were solvated with TIP3P water, neutralized with counter ions and subjected to 1000 steps of conjugate gradient minimization and temperature ramped to

300K. After equilibration, the systems were run for at least 14 ns of production time. Frames from the trajectories were written every 1 ps. The solvation boxes included a 15 Å pad on each face of the box. Long range electrostatics were calculated using particle mesh Ewald (441), and van der Waals were calculated with a nonbonded cutoff of 8 Å and a switching function between 7-8 Å.

Correlation analysis: Correlation analysis by residue was carried out for each system using the ptraj module of Amber11, across the entire simulation. An all residue correlation was performed and difference plots were calculated using Mathematica (544). Correlation between the mutated residue and the residues in Loop 1 were also calculated and plotted. Generalized Masked Delaunay analysis was carried out using the TimeScapes software from the D.E. Shaw group (442). Trajectories were prepared using VMD (431), and all solvent and nucleic acids were excluded from analysis. A sliding window of 5% of the total number of frames was used, and total events per frame were calculated and plotted against frame number.

HDAC 3 Dockings: The R265W mutation was introduced to HDAC3 (PDBID:4A69) via the structure editor under Chimera. Substitution was made using the Dunbrack Rotamer library, and 1000 steps of conjugate gradient minimization were applied to relieve unfavorable interactions. All HDAC inhibitors built and minimized at the AM1 level of theory using Spartan '03 (464) and docked into a sphere encompassing all residues within 20 Å of the center of the SAHA binding pocket. All crystallographic waters were retained and utilized as both fully rotatable and displaceable. Triangle matching was used for base fragment placement, and dockings were performed with two thousand solutions per each iteration and fragmentation. The standard scoring scale based on Böhm's scoring algorithm (465-467) was employed.

REFERENCES

1. Lyskov, S., and J. J. Gray. 2008. The RosettaDock server for local proteinprotein docking. Nucleic Acids Research 36:W233-W238.
2. NASA. 2009. Rocket Basics. http://www.nasa.gov/audience/foreducators/diypodcast/rocket-evolution-index-diy.html.
3. K., P. 2002. Beyond MP3s: IPod Holds Genome. wired.com. http://www.wired.com/medtech/health/news/2002/2011/56223.
4. Hol, W. G. J. 1986. Protein Crystallography and Computer-Graphics toward Rational Drug Design. Angew Chem Int Edit 25:767-778.
5. Berger, B., J. Peng, and M. Singh. 2013. Computational solutions for omics data. Nature reviews. Genetics 14:333-346.
6. Muhlberger, I., J. Wilflingseder, A. Bernthaler, R. Fechete, A. Lukas, and P. Perco. 2011. Computational analysis workflows for Omics data interpretation. Methods Mol Biol 719:379-397.
7. Friedman, L. S., E. A. Ostermeyer, C. I. Szabo, P. Dowd, E. D. Lynch, S. E. Rowell, and M. C. King. 1994. Confirmation of BRCA1 by analysis of germline mutations linked to breast and ovarian cancer in ten families. Nature genetics 8:399-404.
8. Rowell, S., B. Newman, J. Boyd, and M. C. King. 1994. Inherited predisposition to breast and ovarian cancer. American journal of human genetics 55:861-865.
9. King, M. C., and S. Rowell. 1994. Genetic analysis of breast and ovarian cancer in families. Cancer treatment and research 71:51-62.
10. Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang. 2012. Five years of GWAS discovery. American journal of human genetics 90:7-24.
11. Hunter, D. J., and P. Kraft. 2007. Drinking from the fire hose--statistical issues in genomewide association studies. The New England journal of medicine 357:436-439.

12. Vilhjalmsson, B. J., and M. Nordborg. 2013. The nature of confounding in genome-wide association studies. Nature reviews. Genetics 14:1-2.

13. Cambien, F. 2011. Heritability, weak effects, and rare variants in genomewide association studies. Clinical chemistry 57:1263-1266.

14. Dickson, S. P., K. Wang, I. Krantz, H. Hakonarson, and D. B. Goldstein. 2010. Rare variants create synthetic genome-wide associations. PLoS biology 8:e1000294.

15. Zhang, J., W. L. Rowe, A. G. Clark, and K. H. Buetow. 2003. Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. American journal of human genetics 73:1073-1081.

16. Huang, B. E., C. I. Amos, and D. Y. Lin. 2007. Detecting haplotype effects in genomewide association studies. Genetic epidemiology 31:803-812.

17. Lettre, G., and J. D. Rioux. 2008. Autoimmune diseases: insights from genome-wide association studies. Human molecular genetics 17:R116-121.

18. Angeleri, E., B. Apolloni, D. de Falco, and L. Grandi. 1999. DNA fragment assembly using neural prediction techniques. International journal of neural systems 9:523-544.

19. Xu, B., J. Gao, and C. Li. 2012. An efficient algorithm for DNA fragment assembly in MapReduce. Biochemical and biophysical research communications 426:395-398.

20. Pevzner, P. A., H. Tang, and M. S. Waterman. 2001. An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences of the United States of America 98:9748-9753.

21. Batzoglou, S., D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. P. Mesirov, and E. S. Lander. 2002. ARACHNE: a whole-genome shotgun assembler. Genome research 12:177-189.

22. Huson, D. H., K. Reinert, S. A. Kravitz, K. A. Remington, A. L. Delcher, I. M. Dew, M. Flanigan, A. L. Halpern, Z. Lai, C. M. Mobarry, G. G. Sutton, and E. W. Myers. 2001.

Design of a compartmentalized shotgun assembler for the human genome. Bioinformatics 17 Suppl 1:S132-139.

23. Sommer, D. D., A. L. Delcher, S. L. Salzberg, and M. Pop. 2007. Minimus: a fast, lightweight genome assembler. BMC bioinformatics 8:64.

24. Wu, X. L., Y. Heo, I. El Hajj, W. M. Hwu, D. Chen, and J. Ma. 2012. TIGER: tiled iterative genome assembler. BMC bioinformatics 13 Suppl 19:S18.

25. Brandon, M. C., D. C. Wallace, and P. Baldi. 2009. Data structures and compression algorithms for genomic sequence data. Bioinformatics 25:1731-1738.

26. Daniels, N. M., A. Gallant, J. Peng, L. J. Cowen, M. Baym, and B. Berger. 2013. Compressive genomics for protein databases. Bioinformatics 29:i283-i290.

27. Loh, P. R., M. Baym, and B. Berger. 2012. Compressive genomics. Nature biotechnology 30:627-630.

28. Irizarry, R. A., C. Wang, Y. Zhou, and T. P. Speed. 2009. Gene set enrichment analysis made simple. Statistical methods in medical research 18:565-575.

29. Murohashi, M., K. Hinohara, M. Kuroda, T. Isagawa, S. Tsuji, S. Kobayashi, K. Umezawa, A. Tojo, H. Aburatani, and N. Gotoh. 2010. Gene set enrichment analysis provides insight into novel signalling pathways in breast cancer stem cells. British journal of cancer 102:206-212.

30. Zhao, H., Q. Wang, C. Bai, K. He, and Y. Pan. 2009. A cross-study gene set enrichment analysis identifies critical pathways in endometriosis. Reproductive biology and endocrinology : RB&E 7:94.

31. Tilford, C. A., and N. O. Siemers. 2009. Gene set enrichment analysis. Methods Mol Biol 563:99-121.

32. Luo, W., M. S. Friedman, K. Shedden, K. D. Hankenson, and P. J. Woolf. 2009. GAGE: generally applicable gene set enrichment for pathway analysis. BMC bioinformatics 10:161.

33. Anders, S., and W. Huber. 2010. Differential expression analysis for sequence count data. Genome biology 11:R106.

34. Ng, S., E. A. Collisson, A. Sokolov, T. Goldstein, A. Gonzalez-Perez, N. Lopez-Bigas, C. Benz, D. Haussler, and J. M. Stuart. 2012. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. Bioinformatics 28:i640-i646.

35. Faulon, J. L., M. Misra, S. Martin, K. Sale, and R. Sapra. 2008. Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. Bioinformatics 24:225-233.

36. Pang, Y. P. 2007. In silico drug discovery: solving the "target-rich and lead-poor" imbalance using the genome-to-drug-lead paradigm. Clinical pharmacology and therapeutics 81:30-34.

37. Jiang, Z., and Y. Zhou. 2005. Using bioinformatics for drug target identification from the genome. American journal of pharmacogenomics : genomics-related research in drug development and clinical practice 5:387-396.

38. Hagiwara, M. 2005. Alternative splicing: a new drug target of the post-genome era. Biochimica et biophysica acta 1754:324-331.

39. Walke, D. W., C. Han, J. Shaw, E. Wann, B. Zambrowicz, and A. Sands. 2001. In vivo drug target discovery: identifying the best targets from the genome. Current opinion in biotechnology 12:626-631.

40. Turner, P. R., and W. A. Denny. 2000. The genome as a drug target: sequence specific minor groove binding ligands. Current drug targets 1:1-14.

41. Stiglic, G., M. Bajgot, and P. Kokol. 2010. Gene set enrichment meta-learning analysis: next- generation sequencing versus microarrays. BMC bioinformatics 11:176.

42. Haeberle, H., J. T. Dudley, J. T. Liu, A. J. Butte, and C. H. Contag. 2012. Identification of cell surface targets through meta-analysis of microarray data. Neoplasia 14:666-669.

43. Ning, Q. Y., J. Z. Wu, N. Zang, J. Liang, Y. L. Hu, and Z. N. Mo. 2011. Key pathways involved in prostate cancer based on gene set enrichment analysis and meta analysis. Genetics and molecular research : GMR 10:3856-3887.

44. Evangelou, E., and J. P. Ioannidis. 2013. Meta-analysis methods for genome-wide association studies and beyond. Nature reviews. Genetics 14:379-389.

45. Panagiotou, O. A., C. J. Willer, J. N. Hirschhorn, and J. P. Ioannidis. 2013. The Power of Meta-Analysis in Genome-Wide Association Studies. Annual review of genomics and human genetics.

46. Hesper B, H. P. 1970. Bioinformatica: een werkconcept. . Kameleon 1:28-29.

47. Hogeweg, P., and B. Hesper. 1984. The Alignment of Sets of Sequences and the Construction of Phyletic Trees - an Integrated Method. J Mol Evol 20:175-186.

48. Rizk, G., and D. Lavenier. 2010. GASSST: global alignment short sequence search tool. Bioinformatics 26:2534-2540.

49. Margelevicius, M., and C. Venclovas. 2005. PSI-BLAST-ISS: an intermediate sequence search tool for estimation of the position-specific alignment reliability. BMC bioinformatics 6.

50. Lipman, D. J., S. F. Altschul, and J. D. Kececioglu. 1989. A Tool for Multiple Sequence Alignment. Proceedings of the National Academy of Sciences of the United States of America 86:4412-4415.

51. Yu, Q., H. W. Huo, Y. P. Zhang, and H. Z. Guo. 2012. PairMotif: A New Pattern-Driven Algorithm for Planted (l, d) DNA Motif Search. Plos One 7.

52. Leibovich, L., and Z. Yakhini. 2012. Efficient motif search in ranked lists and applications to variable gap motifs. Nucleic Acids Res 40:5832-5847.

53. Stich, M., and S. C. Manrubia. 2011. Motif frequency and evolutionary search times in RNA populations. J Theor Biol 280:117-126.

54. Fu, W. J., P. Ray, and E. P. Xing. 2009. DISCOVER: a feature-based discriminative method for motif search in complex genomes. Bioinformatics 25:I321-I329.

55. Rajasekaran, S. 2009. Computational techniques for motif search. Front Biosci 14:5052-5065.

56. Dhillon, B. K., T. A. Chiu, M. R. Laird, M. G. Langille, and F. S. Brinkman. 2013. IslandViewer update: improved genomic island discovery and visualization. Nucleic Acids Res 41:W129-132.

57. Langille, M. G., and F. S. Brinkman. 2009. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics 25:664-665.

58. Chiapello, H., I. Bourgait, F. Sourivong, G. Heuclin, A. Gendrault-Jacquemard, M. A. Petit, and M. El Karoui. 2005. Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops. BMC bioinformatics 6:171.

59. Yoon, S. H., Y. K. Park, S. Lee, D. Choi, T. K. Oh, C. G. Hur, and J. F. Kim. 2007. Towards pathogenomics: a web-based resource for pathogenicity islands. Nucleic Acids Res 35:D395-400.

60. Waack, S., O. Keller, R. Asper, T. Brodag, C. Damm, W. F. Fricke, K. Surovcik, P. Meinicke, and R. Merkl. 2006. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC bioinformatics 7:142.

61. Hsiao, W. W., K. Ung, D. Aeschliman, J. Bryan, B. B. Finlay, and F. S. Brinkman. 2005. Evidence of a large novel gene pool associated with prokaryotic genomic islands. PLoS genetics 1:e62.

62. Tu, Q., and D. Ding. 2003. Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. FEMS microbiology letters 221:269-275.

63. Karlin, S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends in microbiology 9:335-343.

64. Pundhir, S., H. Vijayvargiya, and A. Kumar. 2008. PredictBias: a server for the identification of genomic and pathogenicity islands in prokaryotes. In silico biology 8:223-234.

65. Ou, H. Y., X. He, E. M. Harrison, B. R. Kulasekara, A. B. Thani, A. Kadioglu, S. Lory, J. C. Hinton, M. R. Barer, Z. Deng, and K. Rajakumar. 2007. MobilomeFINDER: web-based tools for in silico and experimental discovery of bacterial genomic islands. Nucleic Acids Res 35:W97-W104.

66. Perumal, D., C. S. Lim, K. R. Sakharkar, and M. K. Sakharkar. 2007. Differential genome analyses of metabolic enzymes in Pseudomonas aeruginosa for drug target identification. In silico biology 7:453-465.

67. Sun, H., H. F. Chen, and R. Chen. 2013. Genome comparisons as a tool for antimicrobial target discovery. Methods Mol Biol 993:31-38.

68. Zhao, J., Y. Li, C. Zhang, Z. Yao, L. Zhang, X. Bie, F. Lu, and Z. Lu. 2012. Genome shuffling of Bacillus amyloliquefaciens for improving antimicrobial lipopeptide production and an analysis of relative gene expression using FQ RT-PCR. Journal of industrial microbiology & biotechnology 39:889-896.

69. Yu, M., K. Tang, X. Shi, and X. H. Zhang. 2012. Genome sequence of Pseudoalteromonas flavipulchra JG1, a marine antagonistic bacterium with abundant antimicrobial metabolites. Journal of bacteriology 194:3735.

70. Schneider, J., O. Rupp, E. Trost, S. Jaenicke, V. Passoth, A. Goesmann, A. Tauch, and K. Brinkrolf. 2012. Genome sequence of Wickerhamomyces anomalus DSM 6766 reveals genetic basis of biotechnologically important antimicrobial activities. FEMS yeast research 12:382-386.

71. Rio-Alvarez, I., J. J. Rodriguez-Herva, R. Cuartas-Lanza, I. Toth, L. Pritchard, P. Rodriguez-Palenzuela, and E. Lopez-Solanilla. 2012. Genome-wide analysis of the response of Dickeya dadantii 3937 to plant antimicrobial peptides. Molecular plant-microbe interactions : MPMI 25:523-533.

72. Imperi, F., L. C. Antunes, J. Blom, L. Villa, M. Iacono, P. Visca, and A. Carattoli. 2011. The genomics of Acinetobacter baumannii: insights into genome plasticity, antimicrobial resistance and pathogenicity. IUBMB life 63:1068-1074.

73. Ding, R., Y. Li, C. Qian, and X. Wu. 2011. Draft genome sequence of Paenibacillus elgii B69, a strain with broad antimicrobial activity. Journal of bacteriology 193:4537.

74. Kuroda, M. 2006. [Whole genome sequence analysis reveals staphylococcal pathogenesis and antimicrobial resistance]. Nihon saikingaku zasshi. Japanese journal of bacteriology 61:235-241.

75. Hacker, J., L. Bender, M. Ott, J. Wingender, B. Lund, R. Marre, and W. Goebel. 1990. Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates. Microbial pathogenesis 8:213-225.

76. Hasan, M. S., Q. Liu, H. Wang, J. Fazekas, B. Chen, and D. Che. 2012. GIST: Genomic island suite of tools for predicting genomic islands in genomic sequences. Bioinformation 8:203-205.

77. Juhas, M., J. R. van der Meer, M. Gaillard, R. M. Harding, D. W. Hood, and D. W. Crook. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS microbiology reviews 33:376-393.

78. Zhu, L., Z. Yan, Z. Zhang, Q. Zhou, J. Zhou, E. K. Wakeland, X. Fang, Z. Xuan, D. Shen, and Q. Z. Li. 2013. Complete Genome Analysis of Three Acinetobacter baumannii Clinical Isolates in China for Insight into the Diversification of Drug Resistance Elements. Plos One 8:e66584.

79. Saha, S., and M. Lindeberg. 2013. Bound to succeed: Transcription factor binding site prediction and its contribution to understanding virulence and environmental adaptation in bacterial plant pathogens. Molecular plant-microbe interactions : MPMI.

80. Backofen, R., and W. R. Hess. 2010. Computational prediction of sRNAs and their targets in bacteria. RNA biology 7:33-42.

81. Gardy, J. L., C. Spencer, K. Wang, M. Ester, G. E. Tusnady, I. Simon, S. Hua, K. deFays, C. Lambert, K. Nakai, and F. S. Brinkman. 2003. PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Res 31:3613-3617.

82. Quistgaard, E. M., and S. S. Thirup. 2009. Sequence and structural analysis of the Asp-box motif and Asp-box beta-propellers; a widespread propeller-type characteristic of the Vps10 domain family and several glycoside hydrolase families. BMC structural biology 9:46.

83. Sharan, R., and E. W. Myers. 2005. A motif-based framework for recognizing sequence families. Bioinformatics 21 Suppl 1:i387-393.

84. Zhang, Y., J. Xuan, B. G. de los Reyes, R. Clarke, and H. W. Ressom. 2008. Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data. BMC bioinformatics 9:203.

85. Conforti, V. A., D. M. de Avila, N. S. Cummings, R. Zanella, K. J. Wells, H. Ulker, and J. J. Reeves. 2008. CpG motif-based adjuvant as a replacement for Freund's complete adjuvant in a recombinant LHRH vaccine. Vaccine 26:907-913.

86. Voss, B., G. Gierga, I. M. Axmann, and W. R. Hess. 2007. A motif-based search in bacterial genomes identifies the ortholog of the small RNA Yfr1 in all lineages of cyanobacteria. BMC genomics 8:375.

87. Tang, Y., L. Q. Zhang, and F. C. He. 2004. [A motif-based scanning approach for prediction of protein phosphorylation]. Sheng wu gong cheng xue bao = Chinese journal of biotechnology 20:623-626.

88. La, D., M. Silver, R. C. Edgar, and D. R. Livesay. 2003. Using motif-based methods in multiple genome analyses: a case study comparing orthologous mesophilic and thermophilic proteins. Biochemistry 42:8988-8998.

89. Ben-Hur, A., and D. Brutlag. 2003. Remote homology detection: a motif based approach. Bioinformatics 19 Suppl 1:i26-33.

90. Li, N., and M. Tompa. 2006. Analysis of computational approaches for motif discovery. Algorithms for molecular biology : AMB 1:8.

91. Das, M. K., and H. K. Dai. 2007. A survey of DNA motif finding algorithms. BMC bioinformatics 8 Suppl 7:S21.

92. Rajasekaran, S., S. Balla, C. H. Huang, V. Thapar, M. Gryk, M. Maciejewski, and M. Schiller. 2005. High-performance exact algorithms for motif search. Journal of clinical monitoring and computing 19:319-328.

93. Hu, J., B. Li, and D. Kihara. 2005. Limitations and potentials of current motif discovery algorithms. Nucleic Acids Res 33:4899-4913.

94. Csermely, P., T. Korcsmaros, H. J. Kiss, G. London, and R. Nussinov. 2013. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. Pharmacology & therapeutics 138:333-408.

95. dAmore, F., and R. Giaccio. 1995. Incremental hive graph. Lect Notes Comput Sc 1017:49-61.

96. Lin, C. C., and H. C. Yen. 2012. A new force-directed graph drawing method based on edge-edge repulsion. J Visual Lang Comput 23:29-42.

97. Fruchterman, T. M. J., and E. M. Reingold. 1991. Graph Drawing by Force-Directed Placement. Software Pract Exper 21:1129-1164.

98. McGuffin, M. J., and I. Jurisica. 2009. Interaction Techniques for Selecting and Manipulating Subgraphs in Network Visualizations. Ieee T Vis Comput Gr 15:937-944.

99. Kuzmin, S. L. 1991. The Problem of the Shortest-Path in an Oriented Graph with Arbitrary Arc Lengths. Sov J Comput Syst S+ 29:161-164.

100. Rao, A. S., and C. P. Rangan. 1989. Linear Algorithms for Parity Path and 2 Path Problems on Circular-Arc Graph. Lecture Notes in Computer Science 382:267-290.

101. Masuda, S., and K. Nakajima. 1988. An Optimal Algorithm for Finding a Maximum Independent Set of a Circular-Arc Graph. Siam J Comput 17:41-52.

102. Levi, G., and F. Luccio. 1973. Technique for Graph Embedding with Constraints on Node and Arc Correspondences. Inform Sciences 5:1-24.

103. Yilmaz, F., and D. Bozkurt. 2012. The Adjacency Matrix of One Type of Directed Graph and the Jacobsthal Numbers and Their Determinantal Representation. J Appl Math.

104. Qiu, H. J., and E. R. Hancock. 2005. A robust graph partition method from the path-weighted adjacency matrix. Graph-Based Representations in Pattern Recognition, Proceedings 3434:362-372.

105. Acciani, G., G. Fornarelli, and L. Liturri. 2004. Graph adjacency matrix associated with a data partition. Computational Science and Its Applications - Iccsa 2004, Pt 2 3044:979-987.

106. Hashimshony, R., E. Shaviv, and A. Wachman. 1980. Transforming an Adjacency Matrix into a Planar Graph. Build Environ 15:205-217.

107. Harary, F. 1962. Determinant of Adjacency Matrix of a Graph. Siam Rev 4:202-&.

108. Dumas, M., J. M. Robert, and M. J. McGuffin. 2012. AlertWheel: Radial Bipartite Graph Visualization Applied to Intrusion Detection System Alerts. Ieee Network 26:12-18.

109. Delbem, A. C. B., A. De Carvalho, and N. G. Bretas. 2004. Graph chain representation associated to an evolutionary algorithm for restoration of radial distribution systems. Eng Intell Syst Elec 12:3-12.

110. Carlisle, J. C., and A. A. El-Keib. 2000. A graph search algorithm for optimal placement of fixed and switched capacitors on radial distribution systems. Ieee T Power Deliver 15:423-428.

111. Boronin, A. B., and O. Y. Pershin. 1988. Dynamic Location-Problems with Radial Structure of the Components of the Connecting Graph. Cybernetics+ 24:195-203.

112. Krzywinski, M., I. Birol, S. J. Jones, and M. A. Marra. 2012. Hive plots--rational approach to visualizing networks. Briefings in bioinformatics 13:627-644.

113. Darzentas, N. 2010. Circoletto: visualizing sequence similarity with Circos. Bioinformatics 26:2620-2621.

114. Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. 2009. Circos: an information aesthetic for comparative genomics. Genome research 19:1639-1645.

115. Duan, Z., M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble. 2010. A three-dimensional model of the yeast genome. Nature 465:363-367.

116. Kairam, S., D. MacLean, M. Savva, and J. Heer. 2012. GraphPrism: compact visualization of network structure. In Proceedings of the International Working Conference on Advanced Visual Interfaces. ACM. 498-505.

117. Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. Nature 461:747-753.

118. Araujo, R. P., L. A. Liotta, and E. F. Petricoin. 2007. Proteins, drug targets and the mechanisms they control: the simple truth about complex networks. Nature reviews. Drug discovery 6:871-880.

119. Hambly, K., J. Danzer, S. Muskal, and D. A. Debe. 2006. Interrogating the druggable genome with structural informatics. Molecular diversity 10:273-281.

120. Russ, A. P., and S. Lampel. 2005. The druggable genome: an update. Drug discovery today 10:1607-1610.

121. Hopkins, A. L., and C. R. Groom. 2002. The druggable genome. Nature reviews. Drug discovery 1:727-730.

122. Plewczynski, D., and L. Rychlewski. 2009. Meta-basic estimates the size of druggable human genome. Journal of molecular modeling 15:695-699.

123. Imoto, S., Y. Tamada, H. Araki, K. Yasuda, C. G. Print, S. D. Charnock-Jones, D. Sanders, C. J. Savoie, K. Tashiro, S. Kuhara, and S. Miyano. 2006. Computational strategy for discovering druggable gene networks from genome-wide RNA expression profiles. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing:559-571.

124. Yu, H., P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein. 2007. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS computational biology 3:e59.

125. Barabasi, A. L., and Z. N. Oltvai. 2004. Network biology: understanding the cell's functional organization. Nature reviews. Genetics 5:101-113.

126. Agoston, V., P. Csermely, and S. Pongor. 2005. Multiple weak hits confuse complex systems: a transcriptional regulatory network as an example. Physical review. E, Statistical, nonlinear, and soft matter physics 71:051909.

127. Csermely, P., V. Agoston, and S. Pongor. 2005. The efficiency of multi-target drugs: the network approach might help drug design. Trends in pharmacological sciences 26:178-182.

128. Luni, C., J. E. Shoemaker, K. R. Sanft, L. R. Petzold, and F. J. Doyle, 3rd. 2010. Confidence from uncertainty--a multi-target drug screening method from robust control theory. BMC systems biology 4:161.

129. Nacher, J. C., and J. M. Schwartz. 2012. Modularity in protein complex and drug interactions reveals new polypharmacological properties. Plos One 7:e30028.

130. Yang, K., H. Bai, Q. Ouyang, L. Lai, and C. Tang. 2008. Finding multiple target optimal intervention in disease-related molecular network. Molecular systems biology 4:228.

131. Arooj, M., S. Sakkiah, G. Cao, and K. W. Lee. 2013. An innovative strategy for dual inhibitor design and its application in dual inhibition of human thymidylate synthase and dihydrofolate reductase enzymes. Plos One 8:e60470.

132. Xie, L., T. Evangelidis, and P. E. Bourne. 2011. Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. PLoS computational biology 7:e1002037.

133. Gattrell, W., C. Johnstone, S. Patel, C. S. Smith, A. Scheel, and M. Schindler. 2013. Designed multiple ligands in metabolic disease research: from concept to platform. Drug discovery today.

134. Morphy, R., and Z. Rankovic. 2005. Designed multiple ligands. An emerging drug discovery paradigm. Journal of medicinal chemistry 48:6523-6543.

135. Morphy, R., C. Kay, and Z. Rankovic. 2004. From magic bullets to designed multiple ligands. Drug discovery today 9:641-651.

136. O'Boyle, N. M., and M. J. Meegan. 2011. Designed multiple ligands for cancer therapy. Current medicinal chemistry 18:4722-4737.

137. Geldenhuys, W. J., M. B. Youdim, R. T. Carroll, and C. J. Van der Schyf. 2011. The emergence of designed multiple ligands for neurodegenerative disorders. Progress in neurobiology 94:347-359.

138. Zhan, P., and X. Liu. 2009. Designed multiple ligands: an emerging anti-HIV drug discovery paradigm. Current pharmaceutical design 15:1893-1917.

139. Alvarado, M., P. Goya, M. Macias-Gonzalez, F. J. Pavon, A. Serrano, N. Jagerovic, J. Elguero, A. Gutierrez-Rodriguez, S. Garcia-Granda, M. Suardiaz, and F. Rodriguez de Fonseca. 2008. Antiobesity designed multiple ligands: Synthesis of pyrazole fatty acid amides and evaluation as hypophagic agents. Bioorganic & medicinal chemistry 16:10098-10105.

140. Achenbach, J., P. Tiikkainen, L. Franke, and E. Proschak. 2011. Computational tools for polypharmacology and repurposing. Future medicinal chemistry 3:961-968.

141. Mount, D. W., and R. Pandey. 2005. Using bioinformatics and genome analysis for new therapeutic interventions. Molecular cancer therapeutics 4:1636-1643.

142. Kingsmore, S. F., I. E. Lindquist, J. Mudge, D. D. Gessler, and W. D. Beavis. 2008. Genome-wide association studies: progress and potential for drug discovery and development. Nature reviews. Drug discovery 7:221-230.

143. Sanseau, P., P. Agarwal, M. R. Barnes, T. Pastinen, J. B. Richards, L. R. Cardon, and V. Mooser. 2012. Use of genome-wide association studies for drug repositioning. Nature biotechnology 30:317-320.

144. Greulich, K. O. 2004. Single molecule techniques for biomedicine and pharmacology. Current pharmaceutical biotechnology 5:243-259.

145. Garcia-Saez, A. J., and P. Schwille. 2007. Single molecule techniques for the study of membrane proteins. Applied microbiology and biotechnology 76:257-266.

146. Skinner, G. M., and K. Visscher. 2004. Single-molecule techniques for drug discovery. Assay and drug development technologies 2:397-405.

147. Cornish, P. V., and T. Ha. 2007. A survey of single-molecule techniques in chemical biology. ACS chemical biology 2:53-61.

148. Zhu, R., A. Rupprecht, A. Ebner, T. Haselgrubler, H. J. Gruber, P. Hinterdorfer, and E. E. Pohl. 2013. Mapping the nucleotide binding site of uncoupling protein 1 using atomic force microscopy. Journal of the American Chemical Society 135:3640-3646.

149. Kao, F. S., W. Ger, Y. R. Pan, H. C. Yu, R. Q. Hsu, and H. M. Chen. 2012. Chip-based protein-protein interaction studied by atomic force microscopy. Biotechnology and bioengineering 109:2460-2467.

150. Bornschlogl, T., and M. Rief. 2011. Single-molecule protein unfolding and refolding using atomic force microscopy. Methods Mol Biol 783:233-250.

151. Lu, Z., G. Chen, and J. Wang. 2010. [Atomic force microscopy involved in protein study]. Sheng wu yi xue gong cheng xue za zhi = Journal of biomedical engineering = Shengwu yixue gongchengxue zazhi 27:692-695.

152. Zhang, L., and G. Ren. 2012. IPET and FETR: experimental approach for studying molecular structure dynamics by cryo-electron tomography of a single-molecule structure. Plos One 7:e30249.

153. Zhang, Z., S. Witham, and E. Alexov. 2011. On the role of electrostatics in protein-protein interactions. Physical biology 8:035001.

154. Kukic, P., and J. E. Nielsen. 2010. Electrostatics in proteins and protein-ligand complexes. Future medicinal chemistry 2:647-666.

155. Ohlendorf, D. H., and J. B. Matthew. 1985. Electrostatics and flexibility in protein-DNA interactions. Advances in biophysics 20:137-151.

156. Vizcarra, C. L., and S. L. Mayo. 2005. Electrostatics in computational protein design. Current opinion in chemical biology 9:622-626.

157. Sinha, N., and S. J. Smith-Gill. 2002. Electrostatics in protein binding and function. Current protein & peptide science 3:601-614.

158. Heifetz, A., E. Katchalski-Katzir, and M. Eisenstein. 2002. Electrostatics in protein-protein docking. Protein science : a publication of the Protein Society 11:571-587.

159. Guest, W. C., N. R. Cashman, and S. S. Plotkin. 2010. Electrostatics in the stability and misfolding of the prion protein: salt bridges, self energy, and solvation. Biochemistry and cell biology = Biochimie et biologie cellulaire 88:371-381.

160. Matyushov, D. V., and A. Y. Morozov. 2011. Electrostatics of the protein-water interface and the dynamical transition in proteins. Physical review. E, Statistical, nonlinear, and soft matter physics 84:011908.

161. Olsson, M. H. 2011. Protein electrostatics and pKa blind predictions; contribution from empirical predictions of internal ionizable residues. Proteins 79:3333-3345.

162. Neves-Petersen, M. T., and S. B. Petersen. 2003. Protein electrostatics: a review of the equations and methods used to model electrostatic equations in biomolecules--applications in biotechnology. Biotechnology annual review 9:315-395.

163. Holst, M., R. E. Kozack, F. Saied, and S. Subramaniam. 1994. Protein electrostatics: rapid multigrid-based Newton algorithm for solution of the full nonlinear Poisson-Boltzmann equation. Journal of biomolecular structure & dynamics 11:1437-1445.

164. Strickler, S. S., A. V. Gribenko, T. R. Keiffer, J. Tomlinson, T. Reihle, V. V. Loladze, and G. I. Makhatadze. 2006. Protein stability and surface electrostatics: a charged relationship. Biochemistry 45:2761-2766.

165. Liu, L. P., and C. M. Deber. 1999. Combining hydrophobicity and helicity: a novel approach to membrane protein structure prediction. Bioorganic & medicinal chemistry 7:1-7.

166. Eisenberg, D., W. Wilcox, and A. D. McLachlan. 1986. Hydrophobicity and amphiphilicity in protein structure. Journal of cellular biochemistry 31:11-17.

167. Sael, L., D. La, B. Li, R. Rustamov, and D. Kihara. 2008. Rapid comparison of properties on protein surface. Proteins 73:1-10.

168. Bloemendal, M., Y. Marcus, A. H. Sijpkes, and G. Somsen. 1989. Role of hydrophobicity in protein structure is overestimated. International journal of peptide and protein research 34:405-408.

169. Perot, S., O. Sperandio, M. A. Miteva, A. C. Camproux, and B. O. Villoutreix. 2010. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. Drug discovery today 15:656-667.

170. Le Guilloux, V., P. Schmidtke, and P. Tuffery. 2009. Fpocket: an open source platform for ligand pocket detection. BMC bioinformatics 10:168.

171. Chen, B. Y., and B. Honig. 2010. VASP: a volumetric analysis of surface properties yields insights into protein-ligand binding specificity. PLoS computational biology 6.

172. Jordan, R. A., Y. El-Manzalawy, D. Dobbs, and V. Honavar. 2012. Predicting protein-protein interface residues using local surface structural similarity. BMC bioinformatics 13:41.

173. Gu, S., P. Koehl, J. Hass, and N. Amenta. 2012. Surface-histogram: a new shape descriptor for protein-protein docking. Proteins 80:221-238.

174. Carl, N., M. Hodoscek, B. Vehar, J. Konc, B. R. Brooks, and D. Janezic. 2012. Correlating protein hot spot surface analysis using ProBiS with simulated free energies of protein-protein interfacial residues. Journal of chemical information and modeling 52:2541-2549.

175. Adikaram, P. R., and D. Beckett. 2012. Functional versatility of a single protein surface in two protein:protein interactions. Journal of molecular biology 419:223-233.

176. Konc, J., and D. Janezic. 2007. Protein-protein binding-sites prediction by protein surface structure conservation. Journal of chemical information and modeling 47:940-944.

177. Gruber, J., A. Zawaira, R. Saunders, C. P. Barrett, and M. E. Noble. 2007. Computational analyses of the surface properties of protein-protein interfaces. Acta crystallographica. Section D, Biological crystallography 63:50-57.

178. Tan, C., L. Yang, and R. Luo. 2006. How well does Poisson-Boltzmann implicit solvent agree with explicit solvent? A quantitative analysis. The journal of physical chemistry. B 110:18680-18687.

179. Godschalk, F., S. Genheden, P. Soderhjelm, and U. Ryde. 2013. Comparison of MM/GBSA calculations based on explicit and implicit solvent simulations. Physical chemistry chemical physics : PCCP 15:7731-7739.

180. Lee, M. S., and M. A. Olson. 2005. Evaluation of Poisson solvation models using a hybrid explicit/implicit solvent method. The journal of physical chemistry. B 109:5223-5236.

181. Xu, Z., and W. Cai. 2011. Fast Analytical Methods for Macroscopic Electrostatic Models in Biomolecular Simulations. SIAM review. Society for Industrial and Applied Mathematics 53:683-720.

182. Unni, S., Y. Huang, R. M. Hanson, M. Tobias, S. Krishnan, W. W. Li, J. E. Nielsen, and N. A. Baker. 2011. Web servers and services for electrostatics calculations with APBS and PDB2PQR. Journal of computational chemistry 32:1488-1491.

183. Feig, M., A. Onufriev, M. S. Lee, W. Im, D. A. Case, and C. L. Brooks, 3rd. 2004. Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. Journal of computational chemistry 25:265-284.

184. Rocchia, W., S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera, and B. Honig. 2002. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. Journal of computational chemistry 23:128-137.

185. Boschitsch, A. H., and M. O. Fenley. 2011. A Fast and Robust Poisson-Boltzmann Solver Based on Adaptive Cartesian Grids. Journal of chemical theory and computation 7:1524-1540.

186. Gabb, H. A., R. M. Jackson, and M. J. Sternberg. 1997. Modelling protein docking using shape complementarity, electrostatics and biochemical information. Journal of molecular biology 272:106-120.

187. Mandell, J. G., V. A. Roberts, M. E. Pique, V. Kotlovyi, J. C. Mitchell, E. Nelson, I. Tsigelny, and L. F. Ten Eyck. 2001. Protein docking using continuum electrostatics and geometric fit. Protein engineering 14:105-113.

188. Cheng, T. M., T. L. Blundell, and J. Fernandez-Recio. 2007. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins 68:503-515.

189. Jimenez-Garcia, B., C. Pons, and J. Fernandez-Recio. 2013. pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. Bioinformatics 29:1698-1699.

190. Gao, M., and J. Skolnick. 2012. The distribution of ligand-binding pockets around protein-protein interfaces suggests a general mechanism for pocket formation. Proceedings of the National Academy of Sciences of the United States of America 109:3784-3789.

191. de Vries, S. J., and A. M. Bonvin. 2006. Intramolecular surface contacts contain information about protein-protein interface regions. Bioinformatics 22:2094-2098.

192. Norel, R., S. L. Lin, H. J. Wolfson, and R. Nussinov. 1995. Molecular surface complementarity at protein-protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking. Journal of molecular biology 252:263-273.

193. Yeung, T., G. E. Gilbert, J. Shi, J. Silvius, A. Kapus, and S. Grinstein. 2008. Membrane phosphatidylserine regulates surface charge and protein localization. Science 319:210-213.

194. Goldenberg, N. M., and B. E. Steinberg. 2010. Surface charge: a key determinant of protein localization and function. Cancer research 70:1277-1280.

195. Moreira, I. S., P. A. Fernandes, and M. J. Ramos. 2006. Unraveling the importance of protein-protein interaction: application of a computational alanine-scanning mutagenesis to the study of the IgG1 streptococcal protein G (C2 fragment) complex. The journal of physical chemistry. B 110:10962-10969.

196. Hurley, J. K., M. Medina, C. Gomez-Moreno, and G. Tollin. 1994. Further characterization by site-directed mutagenesis of the protein-protein interface in the ferredoxin/ferredoxin:NADP+ reductase system from Anabaena: requirement of a negative charge at position 94 in ferredoxin for rapid electron transfer. Archives of biochemistry and biophysics 312:480-486.

197. Pokala, N., and T. M. Handel. 2004. Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. Protein science : a publication of the Protein Society 13:925-936.

198. Krebs, W. G., V. Alexandrov, C. A. Wilson, N. Echols, H. Yu, and M. Gerstein. 2002. Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. Proteins 48:682-695.

199. Nogales-Cadenas, R., S. Jonic, F. Tama, A. A. Arteni, D. Tabas-Madrid, M. Vazquez, A. Pascual-Montano, and C. O. Sorzano. 2013. 3DEM Loupe: analysis of macromolecular dynamics using structures from electron microscopy. Nucleic Acids Res 41:W363-367.

200. Tama, F., and Y. H. Sanejouand. 2001. Conformational change of proteins arising from normal mode calculations. Protein engineering 14:1-6.

201. Hinsen, K., A. Thomas, and M. J. Field. 1999. Analysis of domain motions in large proteins. Proteins 34:369-382.

202. Hinsen, K. 1998. Analysis of domain motions by approximate normal mode calculations. Proteins 33:417-429.

203. Delarue, M., and Y. H. Sanejouand. 2002. Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. Journal of molecular biology 320:1011-1024.

204. Wang, M., R. T. Borchardt, R. L. Schowen, and K. Kuczera. 2005. Domain motions and the open-to-closed conformational transition of an enzyme: a normal mode analysis of S-adenosyl-L-homocysteine hydrolase. Biochemistry 44:7228-7239.

205. Levitt, M., C. Sander, and P. S. Stern. 1985. Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. Journal of molecular biology 181:423-447.

206. Krol, M., I. Roterman, B. Piekarska, L. Konieczny, J. Rybarska, B. Stopa, and P. Spolnik. 2005. Analysis of correlated domain motions in IgG light chain reveals possible mechanisms of immunological signal transduction. Proteins 59:545-554.

207. Cui, H., C. Mim, F. X. Vazquez, E. Lyman, V. M. Unger, and G. A. Voth. 2013. Understanding the role of amphipathic helices in N-BAR domain driven membrane remodeling. Biophysical journal 104:404-411.

208. Cui, H., G. S. Ayton, and G. A. Voth. 2009. Membrane binding by the endophilin N-BAR domain. Biophysical journal 97:2746-2753.

209. Arkhipov, A., Y. Yin, and K. Schulten. 2009. Membrane-bending mechanism of amphiphysin N-BAR domains. Biophysical journal 97:2727-2735.

210. Ayton, G. S., E. Lyman, V. Krishna, R. D. Swenson, C. Mim, V. M. Unger, and G. A. Voth. 2009. New insights into BAR domain-induced membrane remodeling. Biophysical journal 97:1616-1625.

211. Ayton, G. S., P. D. Blood, and G. A. Voth. 2007. Membrane remodeling from N-BAR domain interactions: insights from multi-scale simulation. Biophysical journal 92:3595-3602.

212. Kulovesi, P., J. Telenius, A. Koivuniemi, G. Brezesinski, A. Rantamaki, T. Viitala, E. Puukilainen, M. Ritala, S. K. Wiedmer, I. Vattulainen, and J. M. Holopainen. 2010. Molecular organization of the tear fluid lipid layer. Biophysical journal 99:2559-2567.

213. Grant, L. M., and F. Tiberg. 2002. Normal and lateral forces between lipid covered solids in solution: correlation with layer packing and structure. Biophysical journal 82:1373-1385.

214. Hirn, R., B. Schuster, U. B. Sleytr, and T. M. Bayerl. 1999. The effect of S-layer protein adsorption and crystallization on the collective motion of a planar lipid bilayer studied by dynamic light scattering. Biophysical journal 77:2066-2074.

215. Fischer, T. M. 1992. Bending stiffness of lipid bilayers. I. Bilayer couple or single-layer bending? Biophysical journal 63:1328-1335.

216. Illya, G., and M. Deserno. 2008. Coarse-grained simulation studies of peptide-induced pore formation. Biophysical journal 95:4163-4173.

217. Takagi, F., and M. Kikuchi. 2007. Structural change and nucleotide dissociation of Myosin motor domain: dual go model simulation. Biophysical journal 93:3820-3827.

218. Spector, A. A., M. Ameen, and A. S. Popel. 2001. Simulation of motor-driven cochlear outer hair cell electromotility. Biophysical journal 81:11-24.

219. Shaw, D. E. 2009. Anton: A specialized machine for millisecond-scale molecular dynamics simulations of proteins. Abstr Pap Am Chem S 238.

220. Wang, Y., C. B. Harrison, K. Schulten, and J. A. McCammon. 2011. Implementation of Accelerated Molecular Dynamics in NAMD. Computational science & discovery 4.

221. Hamelberg, D., J. Mongan, and J. A. McCammon. 2004. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. The Journal of chemical physics 120:11919-11929.

222. Arrar, M., C. A. de Oliveira, M. Fajer, W. Sinko, and J. A. McCammon. 2013. w-REXAMD: A Hamiltonian Replica Exchange Approach to Improve Free Energy Calculations for Systems with Kinetically Trapped Conformations. Journal of chemical theory and computation 9:18-23.

223. Fajer, M., D. Hamelberg, and J. A. McCammon. 2008. Replica-Exchange Accelerated Molecular Dynamics (REXAMD) Applied to Thermodynamic Integration. Journal of chemical theory and computation 4:1565-1569.

224. Fajer, M., R. V. Swift, and J. A. McCammon. 2009. Using multistate free energy techniques to improve the efficiency of replica exchange accelerated molecular dynamics. Journal of computational chemistry 30:1719-1725.

225. Shim, Y., N. B. Callahan, and J. G. Amar. 2013. Localized saddle-point search and application to temperature-accelerated dynamics. The Journal of chemical physics 138:094101.

226. Abrams, C. F., and E. Vanden-Eijnden. 2012. On-the-fly free energy parameterization via temperature accelerated molecular dynamics. Chemical physics letters 547:114-119.

227. Shim, Y., and J. G. Amar. 2011. Adaptive temperature-accelerated dynamics. The Journal of chemical physics 134:054127.

228. Tsalikis, D. G., N. Lempesis, G. C. Boulougouris, and D. N. Theodorou. 2010. Temperature accelerated dynamics in glass-forming materials. The journal of physical chemistry. B 114:7844-7853.

229. Abrams, C. F., and E. Vanden-Eijnden. 2010. Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics. Proceedings of the National Academy of Sciences of the United States of America 107:4961-4966.

230. Vashisth, H., G. Skiniotis, and C. L. Brooks, 3rd. 2013. Enhanced sampling and overfitting analyses in structural refinement of nucleic acids into electron microscopy maps. The journal of physical chemistry. B 117:3738-3746.

231. Mitsutake, A., Y. Mori, and Y. Okamoto. 2013. Enhanced sampling algorithms. Methods Mol Biol 924:153-195.

232. Chen, C., Y. Huang, and Y. Xiao. 2013. Enhanced sampling of molecular dynamics simulation of peptides and proteins by double coupling to thermal bath. Journal of biomolecular structure & dynamics 31:206-214.

233. Kim, I., and T. W. Allen. 2012. Bennett's acceptance ratio and histogram analysis methods enhanced by umbrella sampling along a reaction coordinate in configurational space. The Journal of chemical physics 136:164103.

234. Higo, J., J. Ikebe, N. Kamiya, and H. Nakamura. 2012. Enhanced and effective conformational sampling of protein molecular systems for their free energy landscapes. Biophysical reviews 4:27-44.

235. Zhang, C., and J. Ma. 2010. Enhanced sampling and applications in protein folding in explicit solvent. The Journal of chemical physics 132:244101.

236. Wei, D., and F. Wang. 2010. Mimicking coarse-grained simulations without coarse-graining: enhanced sampling by damping short-range interactions. The Journal of chemical physics 133:084101.

237. Lin, I. C., and M. E. Tuckerman. 2010. Enhanced conformational sampling of peptides via reduced side-chain and solvent masses. The journal of physical chemistry. B 114:15935-15940.

238. Case, D. A., T. E. Cheatham, 3rd, T. Darden, H. Gohlke, R. Luo, K. M. Merz, Jr., A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. 2005. The Amber biomolecular simulation programs. Journal of computational chemistry 26:1668-1688.

239. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. Journal of molecular graphics 14:33-38, 27-38.

240. Ma, B., M. Shatsky, H. J. Wolfson, and R. Nussinov. 2002. Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. Protein science : a publication of the Protein Society 11:184-197.

241. Kumar, S., B. Ma, C. J. Tsai, H. Wolfson, and R. Nussinov. 1999. Folding funnels and conformational transitions via hinge-bending motions. Cell biochemistry and biophysics 31:141-164.

242. Ma, B., S. Kumar, C. J. Tsai, and R. Nussinov. 1999. Folding funnels and binding mechanisms. Protein engineering 12:713-720.

243. Tsai, C. J., S. Kumar, B. Ma, and R. Nussinov. 1999. Folding funnels, binding funnels, and protein function. Protein science : a publication of the Protein Society 8:1181-1190.

244. Chahine, J., H. Nymeyer, V. B. Leite, N. D. Socci, and J. N. Onuchic. 2002. Specific and nonspecific collapse in protein folding funnels. Physical review letters 88:168101.

245. Wong, S., and M. P. Jacobson. 2008. Conformational selection in silico: loop latching motions and ligand binding in enzymes. Proteins 71:153-164.

246. Hammes, G. G., Y. C. Chang, and T. G. Oas. 2009. Conformational selection or induced fit: a flux description of reaction mechanism. Proceedings of the National Academy of Sciences of the United States of America 106:13737-13741.

247. Changeux, J. P., and S. Edelstein. 2011. Conformational selection or induced fit? 50 years of debate resolved. F1000 biology reports 3:19.

248. Vogt, A. D., and E. Di Cera. 2012. Conformational selection or induced fit? A critical appraisal of the kinetic mechanism. Biochemistry 51:5894-5902.

249. Sullivan, S. M., and T. Holyoak. 2008. Enzymes with lid-gated active sites must operate by an induced fit mechanism instead of conformational selection. Proceedings of the National Academy of Sciences of the United States of America 105:13829-13834.

250. Kuzu, G., O. Keskin, A. Gursoy, and R. Nussinov. 2012. Expanding the conformational selection paradigm in protein-ligand docking. Methods Mol Biol 819:59-74.

251. Csermely, P., R. Palotai, and R. Nussinov. 2010. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends in biochemical sciences 35:539-546.

252. Schames, J. R., R. H. Henchman, J. S. Siegel, C. A. Sotriffer, H. Ni, and J. A. McCammon. 2004. Discovery of a novel binding trench in HIV integrase. Journal of medicinal chemistry 47:1879-1881.

253. Hazuda, D. J., N. J. Anthony, R. P. Gomez, S. M. Jolly, J. S. Wai, L. Zhuang, T. E. Fisher, M. Embrey, J. P. Guare, Jr., M. S. Egbertson, J. P. Vacca, J. R. Huff, P. J. Felock, M. V. Witmer, K. A. Stillmock, R. Danovich, J. Grobler, M. D. Miller, A. S. Espeseth, L. Jin, I. W. Chen, J. H. Lin, K. Kassahun, J. D. Ellis, B. K. Wong, W. Xu, P. G. Pearson, W. A. Schleif, R. Cortese, E. Emini, V. Summa, M. K. Holloway, and S. D. Young. 2004. A naphthyridine carboxamide provides evidence for discordant resistance between mechanistically identical inhibitors of HIV-1 integrase. Proceedings of the National Academy of Sciences of the United States of America 101:11233-11238.

254. Estiu, G., N. West, R. Mazitschek, E. Greenberg, J. E. Bradner, and O. Wiest. 2010. On the inhibition of histone deacetylase 8. Bioorganic & medicinal chemistry 18:4103-4110.

255. Weerasinghe, S. V., G. Estiu, O. Wiest, and M. K. Pflum. 2008. Residues in the 11 A channel of histone deacetylase 1 promote catalytic activity: implications for designing isoform-selective histone deacetylase inhibitors. Journal of medicinal chemistry 51:5542-5551.

256. Thangapandian, S., S. John, Y. Lee, S. Kim, and K. W. Lee. 2011. Dynamic structure-based pharmacophore model development: a new and effective addition in the histone deacetylase 8 (HDAC8) inhibitor discovery. International journal of molecular sciences 12:9440-9462.

257. Thangapandian, S., S. John, S. Sakkiah, and K. W. Lee. 2010. Docking-enabled pharmacophore model for histone deacetylase 8 inhibitors and its application in anti-cancer drug discovery. Journal of molecular graphics & modelling 29:382-395.

258. Thangapandian, S., S. John, S. Sakkiah, and K. W. Lee. 2010. Ligand and structure based pharmacophore modeling to facilitate novel histone deacetylase 8 inhibitor design. European journal of medicinal chemistry 45:4409-4417.

259. Thangapandian, S., S. John, Y. Lee, V. Arulalapperumal, and K. W. Lee. 2012. Molecular modeling study on tunnel behavior in different histone deacetylase isoforms. Plos One 7:e49327.

260. Gu, R. X., L. A. Liu, Y. H. Wang, Q. Xu, and D. Q. Wei. 2013. Structural comparison of the wild-type and drug-resistant mutants of the influenza a m2 proton channel by molecular dynamics simulations. The journal of physical chemistry. B 117:6042-6051.

261. Wang, J., C. Ma, G. Fiorin, V. Carnevale, T. Wang, F. Hu, R. A. Lamb, L. H. Pinto, M. Hong, M. L. Klein, and W. F. DeGrado. 2011. Molecular dynamics simulation directed rational design of inhibitors targeting drug-resistant mutants of influenza A virus M2. J Am Chem Soc 133:12834-12841.

262. Pieraccini, S., G. Saladino, G. Cappelletti, D. Cartelli, P. Francescato, G. Speranza, P. Manitto, and M. Sironi. 2009. In silico design of tubulin-targeted antimitotic peptides. Nature chemistry 1:642-648.

263. Read, M., R. J. Harrison, B. Romagnoli, F. A. Tanious, S. H. Gowan, A. P. Reszka, W. D. Wilson, L. R. Kelland, and S. Neidle. 2001. Structure-based design of selective and potent G quadruplex-mediated telomerase inhibitors. Proceedings of the National Academy of Sciences of the United States of America 98:4844-4849.

264. Huo, S., J. Wang, P. Cieplak, P. A. Kollman, and I. D. Kuntz. 2002. Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: insight into structure-based ligand design. Journal of medicinal chemistry 45:1412-1419.

265. Wang, D. F., P. Helquist, N. L. Wiech, and O. Wiest. 2005. Toward selective histone deacetylase inhibitor design: homology modeling, docking studies, and molecular dynamics simulations of human class I histone deacetylases. Journal of medicinal chemistry 48:6936-6947.

266. Martinez, L., P. Webb, I. Polikarpov, and M. S. Skaf. 2006. Molecular dynamics simulations of ligand dissociation from thyroid hormone receptors: evidence of the likeliest escape pathway and its implications for the design of novel ligands. Journal of medicinal chemistry 49:23-26.

267. Singh, N., M. A. Avery, and C. R. McCurdy. 2007. Toward Mycobacterium tuberculosis DXR inhibitor design: homology modeling and molecular dynamics simulations. Journal of computer-aided molecular design 21:511-522.

268. Braun, G. H., D. M. Jorge, H. P. Ramos, R. M. Alves, V. B. da Silva, S. Giuliatti, S. V. Sampaio, C. A. Taft, and C. H. Silva. 2008. Molecular dynamics, flexible docking, virtual screening, ADMET predictions, and molecular interaction field studies to design novel potential MAO-B inhibitors. Journal of biomolecular structure & dynamics 25:347-355.

269. Durdagi, S., T. Mavromoustakos, N. Chronakis, and M. G. Papadopoulos. 2008. Computational design of novel fullerene analogues as potential HIV-1 PR inhibitors: Analysis of the binding interactions between fullerene inhibitors and HIV-1 PR residues using 3D QSAR, molecular docking and molecular dynamics simulations. Bioorganic & medicinal chemistry 16:9957-9974.

270. Franca, T. C. C., M. D. M. Rocha, B. M. Reboredo, M. N. Renno, L. W. Tinoco, and J. D. Figueroa-Villar. 2008. Design of inhibitors for nucleoside hydrolase from Leishmania donovani using molecular dynamics studies. J Brazil Chem Soc 19:64-73.

271. Ko, S., W. Lee, S. Lee, and H. Park. 2008. Nanosecond molecular dynamics simulations of Cdc25B and its complex with a 1,4-naphthoquinone inhibitor: Implications for rational inhibitor design. Journal of molecular graphics & modelling 27:13-19.

272. Koyano, K., and T. Nakano. 2008. Interaction of HIV-1 aspartic protease with its inhibitor, by molecular dynamics and ab initio fragment molecular orbital method. Journal of synchrotron radiation 15:239-242.

273. Novak, W., H. Wang, and G. Krilov. 2009. Role of protein flexibility in the design of Bcl-X(L) targeting agents: insight from molecular dynamics. Journal of computer-aided molecular design 23:49-61.

274. Sheng, C., H. Ji, Z. Miao, X. Che, J. Yao, W. Wang, G. Dong, W. Guo, J. Lu, and W. Zhang. 2009. Homology modeling and molecular dynamics simulation of N-myristoyltransferase from protozoan parasites: active site characterization and insights into rational inhibitor design. Journal of computer-aided molecular design 23:375-389.

275. Tsai, C. W., N. Y. Hsu, C. H. Wang, C. Y. Lu, Y. Chang, H. H. Tsai, and R. C. Ruaan. 2009. Coupling molecular dynamics simulations with experiments for the rational design of indolicidin-analogous antimicrobial peptides. Journal of molecular biology 392:837-854.

276. Zhan, J. H., X. Zhao, X. R. Huang, and C. C. Sun. 2009. Molecular Dynamics and Free Energy Analyses of Erk2-Pyrazolylpyrrole Inhibitors Interactions: Insight into Structure-Based Ligand Design. J Theor Comput Chem 8:887-908.

277. da Silva, M. L., A. D. Goncalves, P. R. Batista, J. D. Figueroa-Villar, P. G. Pascutti, and T. C. C. Franca. 2010. Design, docking studies and molecular dynamics of new potential selective inhibitors of Plasmodium falciparum serine hydroxymethyltransferase. Mol Simulat 36:5-14.

278. Durdagi, S., M. G. Papadopoulos, P. G. Zoumpoulakis, C. Koukoulitsa, and T. Mavromoustakos. 2010. A computational study on cannabinoid receptors and potent bioactive cannabinoid ligands: homology modeling, docking, de novo drug design and molecular dynamics analysis. Molecular diversity 14:257-276.

279. Park, H., S. Ko, and Y. H. Jeon. 2010. Force field design and molecular dynamics simulations of factor-inhibiting HIF-1 and its complex with known inhibitors: Implications for rational inhibitor design. Journal of molecular graphics & modelling 29:221-228.

280. Sippel, M., and C. A. Sotriffer. 2010. Molecular Dynamics Simulations of the HIV-1 Integrase Dimerization Interface: Guidelines for the Design of a Novel Class of Integrase Inhibitors. Journal of chemical information and modeling 50:604-614.

281. Berhanu, W. M., and A. E. Masunov. 2011. Can molecular dynamics simulations assist in design of specific inhibitors and imaging agents of amyloid aggregation? Structure, stability and free energy predictions for amyloid oligomers of VQIVYK, MVGGVV and LYQLEN. Journal of molecular modeling 17:2423-2442.

282. Berhanu, W. M., and A. E. Masunov. 2011. Molecular dynamics simulations of the amylin oligomers for design of aggregation inhibitors. Abstr Pap Am Chem S 241.

283. Guimaraes, A. P., A. A. Oliveira, E. F. F. da Cunha, T. C. Ramalho, and T. C. C. Franca. 2011. Design of New Chemotherapeutics Against the Deadly Anthrax Disease. Docking and Molecular Dynamics Studies of Inhibitors Containing Pyrrolidine and Riboamidrazone Rings on Nucleoside Hydrolase from Bacillus anthracis. Journal of biomolecular structure & dynamics 28:455-469.

284. Perez-Castillo, Y., M. Froeyen, M. A. Cabrera-Perez, and A. Nowe. 2011. Molecular dynamics and docking simulations as a proof of high flexibility in E. coli FabH and its relevance for accurate inhibitor modeling. Journal of computer-aided molecular design 25:371-393.

285. Salo-Ahen, O. M., and R. C. Wade. 2011. The active-inactive transition of human thymidylate synthase: targeted molecular dynamics simulations. Proteins 79:2886-2899.

286. Schaefer, B., C. Kisker, and C. A. Sotriffer. 2011. Molecular dynamics of Mycobacterium tuberculosis KasA: implications for inhibitor and substrate binding and consequences for drug design. Journal of computer-aided molecular design 25:1053-1069.

287. Wang, J., C. L. Ma, G. Fiorin, V. Carnevale, T. Wang, F. H. Hu, R. A. Lamb, L. H. Pinto, M. Hong, M. L. Kein, and W. F. DeGrado. 2011. Molecular Dynamics Simulation Directed Rational Design of Inhibitors Targeting Drug-Resistant Mutants of Influenza A Virus M2. Journal of the American Chemical Society 133:12834-12841.

288. John, S., S. Thangapandian, and K. W. Lee. 2012. Potential Human Cholesterol Esterase Inhibitor Design: Benefits from the Molecular Dynamics Simulations and Pharmacophore Modeling Studies. Journal of biomolecular structure & dynamics 29:921-936.

289. Tamamis, P., A. Lopez de Victoria, R. D. Gorham, Jr., M. L. Bellows-Peterson, P. Pierou, C. A. Floudas, D. Morikis, and G. Archontis. 2012. Molecular dynamics in drug design: new generations of compstatin analogs. Chemical biology & drug design 79:703-718.

290. Thangapandian, S., S. John, M. Arooj, and K. W. Lee. 2012. Molecular Dynamics Simulation Study and Hybrid Pharmacophore Model Development in Human LTA4H Inhibitor Design. Plos One 7.

291. Hucke, O. T., R. Coulombe, M. Bertrand-Laperle, C. Brochu, J. Gillard, M. A. Joly, S. Landry, O. Lepage, M. Marquis, G. McKercher, M. Pesant, M. Poirier, P. L. Beaulieu, P. R. Bonneau, M. Llinas-Brunet, G. Kukolj, and T. A. Stammers. 2013. Molecular dynamics simulations and structure-based rational design lead to allosteric HCV NS5B polymerase thumb pocket 2 inhibitor with picomolar cellular replicon potency. Journal of medicinal chemistry.

292. Pradhan, D., V. Priyadarshini, M. Munikumar, S. Swargam, A. Umamaheswari, and A. Bitla. 2013. Para-(benzoyl)-phenylalanine as a potential inhibitor against LpxC of Leptospira spp.: homology modeling, docking, and molecular dynamics study. Journal of biomolecular structure & dynamics.

293. Sakkiah, S., M. Arooj, M. R. Kumar, S. H. Eom, and K. W. Lee. 2013. Identification of inhibitor binding site in human sirtuin 2 using molecular docking and dynamics simulations. Plos One 8:e51429.

294. Sun, S. X., X. B. Li, W. B. Liu, Y. Ma, R. L. Wang, X. C. Cheng, S. Q. Wang, and W. Liu. 2013. Design, Synthesis, Biological Activity and Molecular Dynamics Studies of Specific Protein Tyrosine Phosphatase 1B Inhibitors over SHP-2. International journal of molecular sciences 14:12661-12674.

295. Wang, Y. Y., L. Li, T. T. Chen, W. Y. Chen, and Y. C. Xu. 2013. Microsecond molecular dynamics simulation of Abeta and identification of a novel dual inhibitor of Abeta aggregation and BACE1 activity. Acta pharmacologica Sinica.

296. Zhang, Z., B. Wang, B. Wan, L. Yu, and Q. Huang. 2013. Molecular dynamics study of carbon nanotube as a potential dual-functional inhibitor of HIV-1 integrase. Biochemical and biophysical research communications.

297. Pradhan, D., V. Priyadarshini, M. Munikumar, S. Swargam, and A. Umamaheswari. 2013. 161 Discovery of potent KdsA inhibitors of Leptospira interrogans through homology modeling, docking, and molecular dynamics simulations. Journal of Biomolecular Structure and Dynamics 31:105-105.

298. Gacko, M., A. Minarowska, A. Karwowska, and Ł. Minarowski. 2008. Cathepsin D inhibitors. Folia Histochemica et Cytobiologica 45:291-290.

299. Lavecchia, A., C. Di Giovanni, and E. Novellino. 2009. CDC25A and B dual-specificity phosphatase inhibitors: potential agents for cancer therapy. Current medicinal chemistry 16:1831-1849.

300. Schaefer, B. 2013. Computergestützte Untersuchungen zur Inhibition und Dynamik der ß-Ketoacyl-ACP-Synthase I (KasA) aus Mycobacterium tuberculosis.

301. Zhao, X., Y. Jie, M. R. Rosenberg, J. Wan, S. Zeng, W. Cui, Y. Xiao, Z. Li, Z. Tu, and M. G. Casarotto. 2012. Design and synthesis of pinanamine derivatives as anti-influenza A M2 ion channel inhibitors. Antiviral research.

302. Topf, C. 2013. Design, Synthese und biologische Testung von KasA-Inhibitoren als potentielle Wirkstoffe gegen Mycobacterium tuberculosis.

303. Tintori, C., J. Demeulemeester, L. Franchi, S. Massa, Z. Debyser, F. Christ, and M. Botta. 2012. Discovery of small molecule HIV-1 integrase dimerization inhibitors. Bioorganic & medicinal chemistry letters 22:3109-3114.

304. Durdagi, S., C. T. Supuran, T. A. Strom, N. Doostdar, M. K. Kumar, A. R. Barron, T. Mavromoustakos, and M. G. Papadopoulos. 2009. In silico drug screening approach for the design of magic bullets: a successful example with anti-HIV fullerene derivatized amino acids. Journal of chemical information and modeling 49:1139-1143.

305. Caruana, M., T. Högen, J. Levin, A. Hillmer, A. Giese, and N. Vassallo. 2011. Inhibition and disaggregation of α-synuclein oligomers by natural polyphenolic compounds. FEBS letters 585:1113-1120.

306. Rennó, M. N., T. C. Costa França, C. B. Palatnik-de-Sousa, L. W. Tinoco, and J. D. Figueroa-Villar. 2012. Kinetics and docking studies of two potential new inhibitors of the nucleoside hydrolase from Leishmania donovani. European journal of medicinal chemistry.

307. Berhanu, W. M., and A. E. Masunov. 2010. Natural polyphenols as inhibitors of amyloid aggregation. Molecular dynamics study of GNNQQNY heptapeptide decamer. Biophysical chemistry 149:12-21.

308. Frearson, J. A., S. Brand, S. P. McElroy, L. A. Cleghorn, O. Smid, L. Stojanovski, H. P. Price, M. L. S. Guther, L. S. Torrie, and D. A. Robinson. 2010. N-myristoyltransferase inhibitors as new leads to treat sleeping sickness. Nature 464:728-732.

309. Bleicher, L., R. Aparicio, F. M. Nunes, L. Martinez, S. M. G. Dias, A. C. Figueira, M. A. Santos, W. H. Venturelli, R. da Silva, and P. M. Donate. 2008. Structural basis of GC-1 selectivity for thyroid hormone receptor isoforms. BMC structural biology 8:8.

310. Distinto, S., M. Yáñez, S. Alcaro, M. C. Cardia, M. Gaspari, M. L. Sanna, R. Meleddu, F. Ortuso, J. Kirchmair, and P. Markt. 2012. Synthesis and biological assessment of novel 2-thiazolylhydrazones and computational analysis of their recognition by monoamine oxidase B. European journal of medicinal chemistry 48:284-295.

311. Bianchi, M. T., and E. J. Botzolakis. 2010. Targeting ligand-gated ion channels in neurology and psychiatry: is pharmacological promiscuity an obstacle or an opportunity? BMC pharmacology 10:3.

312. Melamed, J. Y., M. S. Egbertson, S. Varga, J. P. Vacca, G. Moyer, L. Gabryelski, P. J. Felock, K. A. Stillmock, M. V. Witmer, W. Schleif, D. J. Hazuda, Y. Leonard, L. Jin, J. D. Ellis, and S. D. Young. 2008. Synthesis of 5-(1-H or 1-alkyl-5-oxopyrrolidin-3-yl)-8-hydroxy-[1,6]-naphthyridine-7-carboxamide inhibitors of HIV-1 integrase. Bioorg Med Chem Lett 18:5307-5310.

313. Harvey, M. J., and G. De Fabritiis. 2012. High-throughput molecular dynamics: the powerful new tool for drug discovery. Drug discovery today 17:1059-1062.

314. Buch, I., M. J. Harvey, T. Giorgino, D. P. Anderson, and G. De Fabritiis. 2010. High-throughput all-atom molecular dynamics simulations using distributed computing. Journal of chemical information and modeling 50:397-403.

315. Borhani, D. W., and D. E. Shaw. 2012. The future of molecular dynamics simulations in drug discovery. Journal of computer-aided molecular design 26:15-26.

316. Martin, E., P. Ertl, P. Hunt, J. Duca, and R. Lewis. 2012. Gazing into the crystal ball; the future of computer-aided drug design. Journal of computer-aided molecular design 26:77-79.

317. Kalyaanamoorthy, S., and Y. P. Chen. 2013. Modelling and enhanced molecular dynamics to steer structure-based drug discovery. Progress in biophysics and molecular biology.

318. Durrant, J. D., and J. A. McCammon. 2011. Molecular dynamics simulations and drug discovery. BMC biology 9:71.

319. Seddon, G., V. Lounnas, R. McGuire, T. van den Bergh, R. P. Bywater, L. Oliveira, and G. Vriend. 2012. Drug design for ever, from hype to hope. Journal of computer-aided molecular design 26:137-150.

320. Maggiora, G. M. 2006. On outliers and activity cliffs - Why QSAR often disappoints. Journal of chemical information and modeling 46:1535-1535.

321. Cramer, R. D., J. D. Bunce, D. E. Patterson, and I. E. Frank. 1988. Cross-Validation, Bootstrapping, and Partial Least-Squares Compared with Multiple-Regression in Conventional Qsar Studies. Quant Struct-Act Rel 7:18-25.

322. Ivanciuc, O. 2002. Design of topological indices. Part 28. Distance complement matrix and related structural descriptors for QSAR and QSPR models. Rev Roum Chim 47:577-594.

323. Wood, D. J., J. de Vlieg, M. Wagener, and T. Ritschel. 2012. Pharmacophore Fingerprint-Based Approach to Binding Site Subpocket Similarity and Its Application to Bioisostere Replacement. Journal of chemical information and modeling 52:2031-2043.

324. Murray, C. W., M. L. Verdonk, and D. C. Rees. 2012. Experiences in fragment-based drug discovery. Trends in pharmacological sciences 33:224-232.

325. Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of State Calculations by Fast Computing Machines. The Journal of chemical physics 21:1087-1092.

326. Goodsell, D. S., G. M. Morris, and A. J. Olson. 1996. Automated docking of flexible ligands: Applications of AutoDock. J Mol Recognit 9:1-5.

327. Trosset, J. Y., and H. A. Scheraga. 1999. PRODOCK: Software package for protein modeling and docking. Journal of computational chemistry 20:412-427.

328. Chen, H. M., B. F. Liu, H. L. Huang, S. F. Hwang, and S. Y. Ho. 2007. SODOCK: Swarm optimization for highly flexible protein-ligand docking. Journal of computational chemistry 28:612-623.

329. Namasivayam, V., E. Schild, R. Gunther, and A. G. Beck-Sickinger. 2008. Binding of peptides to GPCRs - A molecular docking approach with PSO@Autodock. J Pept Sci 14:126-126.

330. Namasivayam, V., and R. Gunther. 2007. PSO@AUTODOCK: A fast flexible molecular docking program based on swarm intelligence. Chemical biology & drug design 70:475-484.

331. Friesner, R. A., J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin. 2004. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. Journal of medicinal chemistry 47:1739-1749.

332. Halgren, T. A., R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, and J. L. Banks. 2004. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. Journal of medicinal chemistry 47:1750-1759.

333. Skillman, G., M. Stahl, and M. McGann. 2001. Docking and virtual screening with FRED. Abstr Pap Am Chem S 222:U384-U384.

334. Verdonk, M. L., J. C. Cole, M. J. Hartshorn, C. W. Murray, and R. D. Taylor. 2003. Improved protein-ligand docking using GOLD. Proteins-Structure Function and Genetics 52:609-623.

335. Khodade, P., R. Prabhu, N. Chandra, S. Raha, and R. Govindarajan. 2007. Parallel implementation of AutoDock. J Appl Crystallogr 40:598-599.

336. Boyd, S. 2007. FlexX suite. Chem World-Uk 4:72-72.

337. Cross, S. S. J. 2005. Improved FlexX docking using FlexS-determined base fragment placement. Journal of chemical information and modeling 45:993-1001.

338. Schellhammer, I., and M. Rarey. 2004. FlexX-Scan: Fast, structure-based virtual screening. Proteins 57:504-517.

339. Lemmen, C., S. A. Hindle, M. Gastreich, I. Dramburg, and H. Claussen. 2004. FlexX-docking: Past, present and planned technological advancements. Abstr Pap Am Chem S 228:U507-U507.

340. Kramer, B., M. Rarey, and T. Lengauer. 1999. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins-Structure Function and Genetics 37:228-241.

341. Sprous, D., R. Clark, D. Lowis, J. Leonard, and T. Heritage. 2001. Docking combinatorial libraries efficiently using FLEXX. Abstr Pap Am Chem S 222:U388-U388.

342. Yuriev, E., and P. A. Ramsland. 2013. Latest developments in molecular docking: 2010-2011 in review. J Mol Recognit 26:215-239.

343. Lauri, G., and P. A. Bartlett. 1994. Caveat - a Program to Facilitate the Design of Organic-Molecules. Journal of computer-aided molecular design 8:51-66.

344. Huggins, D. J., W. Sherman, and B. Tidor. 2012. Rational Approaches to Improving Selectivity in Drug Design. Journal of medicinal chemistry 55:1424-1444.

345. Cozzini, P., M. Fornabaio, A. Marabotti, D. J. Abraham, G. E. Kellogg, and A. Mozzarelli. 2004. Free energy of ligand binding to protein: Evaluation of the contribution of water molecules by computational methods. Current medicinal chemistry 11:3093-3118.

346. Schneider, N., G. Lange, S. Hindle, R. Klein, and M. Rarey. 2013. A consistent description of HYdrogen bond and DEhydration energies in protein-ligand complexes: methods behind the HYDE scoring function. Journal of computer-aided molecular design 27:15-29.

347. Schneider, N., S. Hindle, G. Lange, R. Klein, J. Albrecht, H. Briem, K. Beyer, H. Claussen, M. Gastreich, C. Lemmen, and M. Rarey. 2012. Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function. Journal of computer-aided molecular design 26:701-723.

348. Lemmen, C. 2011. HYDE: Scoring for lead optimization. Abstr Pap Am Chem S 241.

349. VanDrie, J. H. 2009. ReCore. Journal of the American Chemical Society 131:1617-1617.

350. Lemmen, C., C. Detering, M. Gastreich, and P. R. Oledzki. 2008. COMP 84-Recore: Instant 3-D scaffold hopping using replacement fragments. Abstr Pap Am Chem S 236.

351. Maass, P. C., T. Schulz-Gasch, M. Stahl, and M. Rarey. 2007. COMP 66-Recore: A fast and versatile method for scaffold hopping. Abstr Pap Am Chem S 234.

352. Maass, P., T. Schulz-Gasch, M. Stahl, and M. Rarey. 2007. Recore: A fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. Journal of chemical information and modeling 47:390-399.

353. Davis, B. C., and I. F. Thorpe. 2013. Thumb inhibitor binding eliminates functionally important dynamics in the hepatitis C virus RNA polymerase. Proteins 81:40-52.

354. Dang, Q., S. R. Kasibhatla, W. Xiao, Y. Liu, J. Dare, F. Taplin, K. R. Reddy, G. R. Scarlato, T. Gibson, P. D. van Poelje, S. C. Potter, and M. D. Erion. 2010. Fructose-1,6-bisphosphatase Inhibitors. 2. Design, Synthesis, and Structure-Activity Relationship of a Series of Phosphonic Acid Containing Benzimidazoles that Function as 5 '-Adenosinemonophosphate (AMP) Mimics. Journal of medicinal chemistry 53:441-451.

355. Dang, Q., B. S. Brown, Y. Liu, R. M. Rydzewski, E. D. Robinson, P. D. van Poelje, M. R. Reddy, and M. D. Erion. 2009. Fructose-1,6-bisphosphatase Inhibitors. 1. Purine Phosphonic Acids as Novel AMP Mimics. Journal of medicinal chemistry 52:2880-2898.

356. Mutyala, R., R. N. Reddy, M. Sumakanth, P. Reddanna, and M. R. Reddy. 2007. Calculation of relative binding affinities of fructose 1,6-bisphosphatase mutants with

adenosine monophosphate using free energy perturbation method. Journal of computational chemistry 28:932-937.

357. Erion, M. D., Q. Dang, M. R. Reddy, S. R. Kasibhatla, J. Huang, W. N. Lipscomb, and P. D. van Poelje. 2007. Structure-guided design of AMP mimics that inhibit fructose-1,6-bisphosphatase with high affinity and specificity. Journal of the American Chemical Society 129:15480-15490.

358. Dang, Q., S. R. Kasibhatla, K. R. Reddy, T. Jiang, M. R. Reddy, S. C. Potter, J. M. Fujitaki, P. D. van Poelje, J. Huang, W. N. Lipscomb, and M. D. Erion. 2007. Discovery of potent and specific fructose-1,6-bisphosphatase inhibitors and a series of orally-bioavailable phosphoramidase-sensitive prodrugs for the treatment of type 2 diabetes. Journal of the American Chemical Society 129:15491-15502.

359. Reddy, M. R., and M. D. Erion. 2005. Computer aided drug design strategies used in the discovery of fructose 1,6-bisphosphatase inhibitors. Current pharmaceutical design 11:283-294.

360. Wesolowski, S. S., and W. L. Jorgensen. 2002. Estimation of binding affinities for celecoxib analogues with COX-2 via Monte Carlo-Extended linear response. Bioorg Med Chem Lett 12:267-270.

361. Price, M. L. P., and W. L. Jorgensen. 2001. Rationale for the observed COX-2/COX-1 selectivity of celecoxib from Monte Carlo simulations. Bioorg Med Chem Lett 11:1541-1544.

362. Price, M. L. P., and W. L. Jorgensen. 2000. Analysis of binding affinities for celecoxib analogues with COX-1 and COX-2 from combined docking and Monte Carlo simulations and insight into the COX-2/COX-1 selectivity. Journal of the American Chemical Society 122:9455-9466.

363. Price, M. L. P., and W. L. Jorgensen. 2000. Origin of binding selectivity for celecoxib analogs with COX-1 and COX-2 from combined docking and Monte Carlo simulations. Abstr Pap Am Chem S 220:U286-U286.

364. Boger, D. L. 2009. Design, synthesis, and evaluation of fatty acid amide hydrolase inhibitors. Abstr Pap Am Chem S 237.

365. Boger, D. L., H. Miyauchi, W. Du, C. Hardouin, R. A. Fecik, H. Cheng, I. Hwang, M. P. Hedrick, D. Leung, O. Acevedo, C. R. W. Guimaraes, W. L. Jorgensen, and B. F. Cravatt. 2005. Discovery of a potent, selective, and efficacious class of reversible alpha-ketoheterocycle inhibitors of fatty acid amide hydrolase effective as analgesics. Journal of medicinal chemistry 48:1849-1856.

366. Guimaraes, C. R. W., D. L. Boger, and W. L. Jorgensen. 2005. Elucidation of fatty acid amide hydrolase inhibition by potent alpha-ketoheterocycle derivatives from Monte Carlo simulations. Journal of the American Chemical Society 127:17377-17384.

367. DeMartino, J. K., J. Garfunkle, D. G. Hochstatter, B. F. Cravatt, and D. L. Boger. 2008. Exploration of a fundamental substituent effect of alpha-ketoheterocycle enzyme inhibitors: Potent and selective inhibitors of fatty acid amide hydrolase. Bioorg Med Chem Lett 18:5842-5846.

368. Pearlman, D. A. 2005. Evaluating the molecular mechanics Poisson-Boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP kinase. Journal of medicinal chemistry 48:7796-7807.

369. Pepin, J., L. Valiquette, and B. Cossette. 2005. Mortality attributable to nosocomial Clostridium difficile-associated disease during an epidemic caused by a hypervirulent strain in Quebec. CMAJ 173:1037-1042.

370. Hookman, P., and J. S. Barkin. 2009. Clostridium difficile associated infection, diarrhea and colitis. World J Gastroenterol 15:1554-1580.

371. Bartlett JG, C. T., Gurwith M, Gorbach SL, Onderdonk AD. 1978. Antibiotic-accosicated, pseudomembranous colitis due to toxin-producing clostridia. The New England journal of medicine 298:531-534.

372. McFarland LV, M. M., Kwok RY, Stamm WE. 1989. Nosocomial acquisition of Clostridium difficile infection. The New England journal of medicine 320:204 210.

373. Dubberke, E. R., K. M. McMullen, J. L. Mayfield, K. A. Reske, P. Georgantopoulos, D. K. Warren, and V. J. Fraser. 2009. Hospital-associated Clostridium difficile infection: is it necessary to track community-onset disease? Infect Control Hosp Epidemiol 30:332-337.

374. Dubberke, E. R., and A. I. Wertheimer. 2009. Review of current literature on the economic burden of Clostridium difficile infection. Infect Control Hosp Epidemiol 30:57-66.

375. Kyne, L., M. B. Hamel, R. Polavaram, and C. P. Kelly. 2002. Health care costs and mortality associated with nosocomial diarrhea due to Clostridium difficile. Clin Infect Dis 34:346-353.

376. McGlone, S. M., R. R. Bailey, S. M. Zimmer, M. J. Popovich, Y. Tian, P. Ufberg, R. R. Muder, and B. Y. Lee. 2011. The economic burden of Clostridium difficile. Clin Microbiol Infect.

377. Huang, H., A. Weintraub, H. Fang, and C. E. Nord. 2009. Antimicrobial resistance in Clostridium difficile. Int J Antimicrob Agents 34:516-522.

378. Merrigan, M., A. Venugopal, M. Mallozzi, B. Roxas, V. K. Viswanathan, S. Johnson, D. N. Gerding, and G. Vedantam. 2010. Human hypervirulent Clostridium difficile strains exhibit increased sporulation as well as robust toxin production. Journal of bacteriology 192:4904-4911.

379. Cookson, B. 2007. Hypervirulent strains of Clostridium difficile. Postgrad Med J 83:291-295.

380. Gerding, D. N., S. Johnson, L. R. Peterson, M. E. Mulligan, and J. Silva, Jr. 1995. Clostridium difficile-associated diarrhea and colitis. Infect Control Hosp Epidemiol 16:459-477.

381. Vedantam, G., A. Clark, M. Chu, R. McQuade, M. Mallozzi, and V. K. Viswanathan. 2012. Clostridium difficile infection: toxins and non-toxin virulence factors, and their contributions to disease establishment and host response. Gut microbes 3:121-134.

382. Coia, J. E. 2009. What is the role of antimicrobial resistance in the new epidemic of Clostridium difficile? Int J Antimicrob Agents 33 Suppl 1:S9-12.

383. Gerding, D. N. 2004. Clindamycin, cephalosporins, fluoroquinolones, and Clostridium difficile-associated diarrhea: this is an antimicrobial resistance problem. Clin Infect Dis 38:646-648.

384. Wust, J., and U. Hardegger. 1988. Studies on the resistance of Clostridium difficile to antimicrobial agents. Zentralbl Bakteriol Mikrobiol Hyg A 267:383-394.

385. Gould, C. V., and L. C. McDonald. 2008. Bench-to-bedside review: Clostridium difficile colitis. Crit Care 12:203.

386. Voth, D. E., and J. D. Ballard. 2005. Clostridium difficile toxins: mechanism of action and role in disease. Clin Microbiol Rev 18:247-263.

387. Reinert, D. J., T. Jank, K. Aktories, and G. E. Schulz. 2005. Structural basis for the function of Clostridium difficile toxin B. J Mol Biol 351:973-981.

388. Just, I., J. Selzer, M. Wilm, C. von Eichel-Streiber, M. Mann, and K. Aktories. 1995. Glucosylation of Rho proteins by Clostridium difficile toxin B. Nature 375:500-503.

389. Just, I., and R. Gerhard. 2004. Large clostridial cytotoxins. Rev Physiol Biochem Pharmacol 152:23-47.

390. Rupnik, M., S. Pabst, C. von Eichel-Streiber, H. Urlaub, and H. D. Soling. 2005. Characterization of the cleavage site and function of resulting cleavage fragments after limited proteolysis of Clostridium difficile toxin B (TcdB) by host cells. Microbiology 151:199-208.

391. Giesemann, T., M. Egerer, T. Jank, and K. Aktories. 2008. Processing of Clostridium difficile toxins. J Med Microbiol 57:690-696.

392. Von Eichel-Streiber C, S. M., Thlestam M. 1996. Large clostridial cytotoxins a family of glycosyltransferases modifying small GTP-binding proteins. Trends in microbiology 4:375-382.

393. Jank, T., and K. Aktories. 2008. Structure and mode of action of clostridial glucosylating toxins: the ABCD model. Trends in microbiology 16:222-229.

394. Mathieu, R., J. Lim, P. Simpson, S. Prasannan, N. Fairweather, and S. Matthews. 2003. Resonance assignment and topology of a clostridial repetitive oligopeptide (CROP) region of toxin A from Clostridium difficile. J Biomol NMR 25:83-84.

395. Frisch, C., R. Gerhard, K. Aktories, F. Hofmann, and I. Just. 2003. The complete receptor-binding domain of Clostridium difficile toxin A is required for endocytosis. Biochemical and biophysical research communications 300:706-711.

396. Ho, J. G., A. Greco, M. Rupnik, and K. K. Ng. 2005. Crystal structure of receptor-binding C-terminal repeats from Clostridium difficile toxin A. Proceedings of the National Academy of Sciences of the United States of America 102:18373-18378.

397. Giesemann, T., T. Jank, R. Gerhard, E. Maier, I. Just, R. Benz, and K. Aktories. 2006. Cholesterol-dependent pore formation of Clostridium difficile toxin A. N-S Arch Pharmacol 372:125-125.

398. Pfeifer, G., J. Schirmer, J. Leemhuis, C. Busch, D. K. Meyer, K. Aktories, and H. Barth. 2003. Cellular uptake of Clostridium difficile toxin B. Translocation of the N-terminal

catalytic domain into the cytosol of eukaryotic cells. Journal of Biological Chemistry 278:44535-44541.

399. Kaiser, E., C. Kroll, K. Ernst, C. Schwan, M. Popoff, G. Fischer, J. Buchner, K. Aktories, and H. Barth. 2011. Membrane translocation of binary actin-ADP-ribosylating toxins from Clostridium difficile and Clostridium perfringens is facilitated by cyclophilin A and Hsp90. Infect Immun 79:3913-3921.

400. Genisyuerek, S., P. Papatheodorou, G. Guttenberg, R. Schubert, R. Benz, and K. Aktories. 2011. Structural determinants for membrane insertion, pore formation and translocation of Clostridium difficile toxin B. Mol Microbiol 79:1643-1654.

401. Egerer, M., T. Jank, T. Giesemann, and K. Aktories. 2008. Cysteine protease activity is responsible for autocatalytic cleavage of Clostridium difficile toxin A and B. N-S Arch Pharmacol 377:11-11.

402. Shen, A., P. J. Lupardus, M. M. Gersch, A. W. Puri, V. E. Albrow, K. C. Garcia, and M. Bogyo. 2011. Defining an allosteric circuit in the cysteine protease domain of Clostridium difficile toxins. Nat Struct Mol Biol 18:364-371.

403. Egerer, M., T. Giesemann, T. Jank, K. J. F. Satchell, and K. Aktories. 2007. Auto-catalytic cleavage of Clostridium difficile toxins a and B depends on cysteine protease activity. Journal of Biological Chemistry 282:25314-25321.

404. Ivarsson, M. E., J.-C. Leroux, and B. Castagner. 2012. Therapien gegen Bakterientoxine. Angewandte Chemie 124:4098-4121.

405. Dawson, L. F., E. Valiente, E. H. Donahue, G. Birchenough, and B. W. Wren. 2011. Hypervirulent Clostridium difficile PCR-ribotypes exhibit resistance to widely used disinfectants. Plos One 6:e25754.

406. Lanis, J. M., L. D. Hightower, A. Shen, and J. D. Ballard. 2012. TcdB from hypervirulent Clostridium difficile exhibits increased efficiency of autoprocessing. Mol Microbiol.

407. Abdeen, S. J., R. J. Swett, and A. L. Feig. 2010. Peptide inhibitors targeting Clostridium difficile toxins A and B. ACS chemical biology 5:1097-1103.

408. Reinert, D. J., T. Jank, K. Aktories, and G. E. Schulz. 2005. Structural basis for the function of Clostridium difficile toxin B. J Mol Biol 351:973-981.

409. Egerer, M., T. Jank, T. Giesemann, and K. Aktories. 2007. Involvement of cysteine residues in processing of Clostridium difficile toxins A and B. N-S Arch Pharmacol 375:93-93.

410. Müller, S., C. von Eichel-Streiber, and M. Moos. 1999. Impact of amino acids 22–27 of Rho-subfamily GTPases on glucosylation by the large clostridial cytotoxins TcsL-1522, TcdB-1470 and TcdB-8864. European Journal of Biochemistry 266:1073-1080.

411. Jank, T., T. Giesemann, and K. Aktories. 2007. Clostridium difficile glucosyltransferase toxin B-essential amino acids for substrate binding. Journal of Biological Chemistry 282:35222-35231.

412. Geissler, B., R. Tungekar, and K. J. F. Satchell. 2010. Identification of a conserved membrane localization domain within numerous large bacterial protein toxins. Proceedings of the National Academy of Sciences of the United States of America 107:5581-5586.

413. Jank, T., and K. Aktories. 2006. Chance of a single amino acid residue turns RhoD into a substrate of Clostridium difficile toxin B. N-S Arch Pharmacol 372:55-56.

414. Jank, T., U. Pack, T. Giesemann, G. Schmidt, and K. Aktories. 2006. Exchange of a single amino acid switches the substrate properties of RhoA and RhoD toward glucosylating and transglutaminating toxins. Journal of Biological Chemistry 281:19527-19535.

415. Jank, T., D. J. Reinert, T. Giesemann, G. E. Schulz, and K. Aktories. 2005. Change of the donor substrate specificity of Clostridium difficile toxin B by site-directed mutagenesis. Journal of Biological Chemistry 280:37833-37838.

416. Flores, S. C., K. S. Keating, J. Painter, F. Morcos, K. Nguyen, E. A. Merritt, L. A. Kuhn, and M. B. Gerstein. 2008. HingeMaster: Normal mode hinge prediction approach and integration of complementary predictors. Proteins 73:299-319.

417. Keating, K. S., S. C. Flores, M. B. Gerstein, and L. A. Kuhn. 2009. StoneHinge: Hinge prediction by network analysis of individual protein structures. Protein science : a publication of the Protein Society 18:359-371.

418. Suhre, K., and Y. H. Sanejouand. 2004. ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Research 32:W610-W614.

419. Grant, B. J., J. A. McCammon, and A. A. Gorfe. 2010. Conformational selection in G-proteins: lessons from Ras and Rho. Biophysical journal 99:L87-89.

420. Keating, K. S., S. C. Flores, M. B. Gerstein, and L. A. Kuhn. 2009. StoneHinge: hinge prediction by network analysis of individual protein structures. Protein Sci 18:359-371.

421. Krebs, W. G., V. Alexandrov, C. A. Wilson, N. Echols, H. Y. Yu, and M. Gerstein. 2002. Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic. Proteins-Structure Function and Genetics 48:682-695.

422. Von Eichel-Streiber C, S. M., Thlestam M. 1996. Large clostridial cytotoxins a family of glycosyltransferases modifying small GTP-binding proteins. Trends Microbiol 4:375-382.

423. D.W. Ritchie, D. K., and S. Vajda. 2008. Accelerating and Focusing Protein-Protein Docking Correlations Using Multi-Dimensional Rotational FFT Generating Functions. Bioinformatics in press.

424. Ritchie, D. W. K. D. a. V. S. 2005. High Order Analytic Translation Matrix Elements For Real Space Six-Dimensional Polar Fourier Correlations. J. Appl. Crystl 38:808-818.

425. Gray, J. J., S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker. 2003. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. Journal of molecular biology 331:281-299.

426. Offen, W., C. Martinez-Fleites, M. Yang, E. Kiat-Lim, B. G. Davis, C. A. Tarling, C. M. Ford, D. J. Bowles, and G. J. Davies. 2006. Structure of a flavonoid glucosyltransferase reveals the basis for plant natural product modification. The EMBO journal 25:1396-1405.

427. Jiang, J. Y., M. B. Lazarus, L. Pasquina, P. Sliz, and S. Walker. 2012. A neutral diphosphate mimic crosslinks the active site of human O-GlcNAc transferase. Nat Chem Biol 8:72-77.

428. Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. 2005. Scalable molecular dynamics with NAMD. Journal of computational chemistry 26:1781-1802.

429. MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. Journal of Physical Chemistry B 102:3586-3616.

430. Izaguirre, J. A., D. P. Catarello, J. M. Wozniak, and R. D. Skeel. 2001. Langevin stabilization of molecular dynamics. The Journal of chemical physics 114:2090-2098.

431. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: Visual molecular dynamics. J Mol Graphics 14:33-38.

432. Ma, J. 2005. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. Structure 13:373-380.

433. Debrunner, P. G., and H. Frauenfelder. 1982. Dynamics of Proteins. Annu Rev Phys Chem 33:283-299.

434. Petsko, G. A., and D. Ringe. 1984. Fluctuations in Protein-Structure from X-Ray-Diffraction. Annu Rev Biophys Bio 13:331-371.

435. Wei, Y., Y. Zhang, U. Derewenda, X. Liu, W. Minor, R. K. Nakamoto, A. V. Somlyo, A. P. Somlyo, and Z. S. Derewenda. 1997. Crystal structure of RhoA-GDP and its functional implications. Nature Structural Biology 4:699-703.

436. Pruitt, R. N., N. M. Chumbler, S. A. Rutherford, M. A. Farrow, D. B. Friedman, B. Spiller, and D. B. Lacy. 2012. Structural Determinants of Clostridium difficile Toxin A Glucosyltransferase Activity. Journal of Biological Chemistry 287:8013-8020.

437. Patel, S., and C. L. Brooks, 3rd. 2004. CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. Journal of computational chemistry 25:1-15.

438. Patel, S., A. D. Mackerell, Jr., and C. L. Brooks, 3rd. 2004. CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. Journal of computational chemistry 25:1504-1514.

439. Brooks, B. R., C. L. Brooks, 3rd, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. 2009. CHARMM: the biomolecular simulation program. Journal of computational chemistry 30:1545-1614.

440. Sapay, N., and D. P. Tieleman. 2010. Combination of the CHARMM27 force field with united-atom lipid force fields. Journal of computational chemistry.

441. Wang, H., F. Dommert, and C. Holm. 2010. Optimizing working parameters of the smooth particle mesh Ewald algorithm in terms of accuracy and efficiency. The Journal of chemical physics 133:034117.

442. Willy Wriggers, K. A. S., Yibing Shan, Stefano Piana, Paul Maragakis,, P. J. M. Kresten Lindorff-Larsen, Justin Gullingsrud, Charles A. Rendleman,, and R. O. D. Michael P. Eastwood, and David E. Shaw. (2009). "Automated Event Detection and Activity Monitoring in Long Time-Scale Molecular Dynamics". J. Chem. Theory Comput 5:2595–2605.

443. Wolfram Research, I. 2010. Mathematica Edition: Version 8.0. Wolfram Research, Inc. , Champaign, Illinois.

444. Sorg, J. A., and A. L. Sonenshein. 2010. Inhibiting the initiation of Clostridium difficile spore germination using analogs of chenodeoxycholic acid, a bile acid. Journal of bacteriology 192:4983-4990.

445. Britton, R. A., and V. B. Young. 2012. Interaction between the intestinal microbiota and host in Clostridium difficile colonization resistance. Trends in microbiology 20:313-319.

446. Puri, A. W., P. J. Lupardus, E. Deu, V. E. Albrow, K. C. Garcia, M. Bogyo, and A. Shen. 2010. Rational design of inhibitors and activity-based probes targeting Clostridium difficile virulence factor TcdB. Chemistry & biology 17:1201-1211.

447. Jank, T., M. O. Ziegler, G. E. Schulz, and K. Aktories. 2008. Inhibition of the glucosyltransferase activity of clostridial Rho/Ras-glucosylating toxins by castanospermine. FEBS Lett 582:2277-2282.

448. Kerzmann, A. 2009. Mechanistic Analysis of Clostridium difficile Toxin A. . In Chemistry. Indiana University, Bloomington, IN.

449. Swett, R., G. A. Cisneros, and A. L. Feig. 2012. Conformational analysis of Clostridium difficile toxin B and its implications for substrate recognition. Plos One 7:e41518.

450. Rogozin, I. B., Y. I. Pavlov, K. Bebenek, T. Matsuda, and T. A. Kunkel. 2001. Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. Nat Immunol 2:530-536.

451. Velec, H. F. G., H. Gohlke, and G. Klebe. 2005. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. Journal of medicinal chemistry 48:6296-6303.

452. Huang, S. Y., and X. Q. Zou. 2005. ITScore: A novel iterative knowledge-based scoring function to predict protein-ligand interactions. Biophysical journal 88:218a-218a.

453. Muegge, I. 2000. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. Perspect Drug Discov 20:99-114.

454. Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. Journal of computational chemistry 25:1605-1612.

455. Boehr, D. D., R. Nussinov, and P. E. Wright. 2009. The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol 5:789-796.

456. Ziegler, M. O., T. Jank, K. Aktories, and G. E. Schulz. 2008. Conformational changes and reaction of clostridial glycosylating toxins. J Mol Biol 377:1346-1356.

457. Belyi, Y., and K. Aktories. 2010. Bacterial toxin and effector glycosyltransferases. Biochimica et biophysica acta 1800:134-143.

458. Best, R. B., N. V. Buchete, and G. Hummer. 2008. Are current molecular dynamics force fields too helical? Biophysical journal 95:L07-09.

459. Cino, E. A., W. Y. Choy, and M. Karttunen. 2012. Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations. Journal of chemical theory and computation 8:2725-2740.

460. Rarey, M., B. Kramer, T. Lengauer, and G. Klebe. 1996. A fast flexible docking method using an incremental construction algorithm. Journal of molecular biology 261:470-489.

461. Miteva, M. A., W. H. Lee, M. O. Montes, and B. O. Villoutreix. 2005. Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. Journal of medicinal chemistry 48:6012-6022.

462. Degen, J., and M. Rarey. 2006. FlexNovo: structure-based searching in large fragment spaces. ChemMedChem 1:854-868.

463. M. J. Frisch, G. W. T., H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople,. 2004. Gaussian 03.

464. Wavefunction, I. V. K. A., Suite 370 Irvine, CA 92612, USA. Wavefunction, Inc.: 18041 Von Karman Avenue, Suite 370 Irvine, CA 92612, USA. Wavefunction, Inc.: 18041 Von Karman Avenue, Suite 370 Irvine, CA 92612, USA.

465. Huang, S. Y., and X. Q. Zou. 2010. Inclusion of Solvation and Entropy in the Knowledge-Based Scoring Function for Protein-Ligand Interactions. Journal of chemical information and modeling 50:262-273.

466. Huang, S. Y., and X. Q. Zou. 2006. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. Journal of computational chemistry 27:1866-1875.

467. Gohlke, H., M. Hendlich, and G. Klebe. 2000. Knowledge-based scoring function to predict protein-ligand interactions. Journal of molecular biology 295:337-356.

468. Kelley, L. A., S. P. Gardner, and M. J. Sutcliffe. 1996. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. Protein engineering 9:1063-1065.

469. Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. 2004. UCSF chimera - A visualization system for exploratory research and analysis. Journal of computational chemistry 25:1605-1612.

470. Dunbrack, R. L., Jr., and M. Karplus. 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. Journal of molecular biology 230:543-574.

471. Hatcher, E., O. Guvench, and A. D. Mackerell, Jr. 2009. CHARMM Additive All-Atom Force Field for Acyclic Polyalcohols, Acyclic Carbohydrates and Inositol. Journal of chemical theory and computation 5:1315-1327.

472. Case, D., T. Darden, T. Cheatham Iii, C. Simmerling, J. Wang, R. Duke, R. Luo, R. Walker, W. Zhang, and K. Merz. 2010. Amber Tools 1.5. AMBER 11, University of California, San Francisco, CA.

473. Wolfram, S. 1991. Mathematica: a system for doing mathematics by computer. Redwood City. CA: Addison–Wesley. 0 20:40-60.

474. Grant, B. J., A. P. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. Caves. 2006. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22:2695-2696.

475. Team, R. C. 2008. R: A language and environment for statistical computing. R Foundation Statistical Computing.

476. Perez, J., V. S. Springthorpe, and S. A. Sattar. 2005. Activity of selected oxidizing microbicides against the spores of Clostridium difficile: relevance to environmental control. American journal of infection control 33:320-325.

477. Ozaki, E., H. Kato, H. Kita, T. Karasawa, T. Maegawa, Y. Koino, K. Matsumoto, T. Takada, K. Nomoto, R. Tanaka, and S. Nakamura. 2004. Clostridium difficile colonization in healthy adults: transient colonization and correlation with enterococcal colonization. J Med Microbiol 53:167-172.

478. Busch, C., and K. Aktories. 2000. Microbial toxins and the glycosylation of rho family GTPases. Current opinion in structural biology 10:528-535.

479. Kurtz, C. B., E. P. Cannon, A. Brezzani, M. Pitruzzello, C. Dinardo, E. Rinard, D. W. Acheson, R. Fitzpatrick, P. Kelly, K. Shackett, A. T. Papoulis, P. J. Goddard, R. H. Barker, Jr., G. P. Palace, and J. D. Klinger. 2001. GT160-246, a toxin binding polymer for treatment of Clostridium difficile colitis. Antimicrobial agents and chemotherapy 45:2340-2347.

480. Deng, X. K., L. A. Nesbit, and K. J. Morrow, Jr. 2003. Recombinant single-chain variable fragment antibodies directed against Clostridium difficile toxin B produced by use of an optimized phage display system. Clinical and diagnostic laboratory immunology 10:587-595.

481. Ghose, C., A. Kalsy, A. Sheikh, J. Rollenhagen, M. John, J. Young, S. M. Rollins, F. Qadri, S. B. Calderwood, C. P. Kelly, and E. T. Ryan. 2007. Transcutaneous immunization with Clostridium difficile toxoid A induces systemic and mucosal immune responses and toxin A-neutralizing antibodies in mice. Infect Immun 75:2826-2832.

482. Safdar, A. 2010. Treatment with monoclonal antibodies against Clostridium difficile toxins. The New England journal of medicine 362:1444-1445; author reply 1445-1446.

483. Scheinfeld, N., and K. Biggers. 2008. Tolevamer, an orally administered, toxin-binding polymer for Clostridium difficile-associated diarrhea. Curr Opin Investig Drugs 9:913-924.

484. Abdeen, S., Feig, A.L., Swett, R. Kern, S. . 2011. Inhibition of C. difficile toxins. U.S. .

485. Bruccoleri, R. E., J. Novotny, M. E. Davis, and K. A. Sharp. 1997. Finite difference Poisson-Boltzmann electrostatic calculations: Increased accuracy achieved by harmonic dielectric smoothing and charge antialiasing. Journal of computational chemistry 18:268-276.

486. Pei, J. M., and N. V. Grishin. 2001. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17:700-712.

487. Sanner, M. F., A. J. Olson, and J. C. Spehner. 1996. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 38:305-320.

488. Kyte, J., and R. F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. Journal of molecular biology 157:105-132.

489. Baker, N. A., D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. 2001. Electrostatics of nanosystems: Application to microtubules and the ribosome. Proceedings of the National Academy of Sciences of the United States of America 98:10037-10041.

490. Rostkowski, M., M. H. M. Olsson, C. R. Sondergaard, and J. H. Jensen. 2011. Graphical analysis of pH-dependent properties of proteins predicted using PROPKA. BMC structural biology 11.

491. Oberoi, H., and N. M. Allewell. 1993. Multigrid Solution of the Nonlinear Poisson-Boltzmann Equation and Calculation of Titration Curves. Biophysical journal 65:48-55.

492. Holst, M., R. E. Kozack, F. Saied, and S. Subramaniam. 1994. Treatment of Electrostatic Effects in Proteins - Multigrid-Based Newton Iterative Method for Solution of the Full Nonlinear Poisson-Boltzmann Equation. Proteins 18:231-245.

493. Im, W., D. Beglov, and B. Roux. 1998. Continuum solvation model: computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation. Computer Physics Communications 111:59-75.

494. Rudd, K. E. 2000. EcoGene: a genome sequence database for Escherichia coli K-12. Nucleic Acids Res 28:60-64.

495. Keseler, I. M., J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muniz-Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A. G. Shearer, A. Mackie, I. Paulsen, R. P. Gunsalus, and P. D. Karp. EcoCyc: a comprehensive database of Escherichia coli biology. Nucleic Acids Res 39:D583-590.

496. Padan, E., E. Bibi, M. Ito, and T. A. Krulwich. 2005. Alkaline pH homeostasis in bacteria: new insights. Biochimica et biophysica acta 1717:67-88.

497. Sproul, A. A., L. T. Lambourne, D. J. Jean-Jacques, and H. L. Kornberg. 2001. Genetic control of manno(fructo)kinase activity in Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America 98:15257-15259.

498. 2010. The Human Genome Project: 10 years later. Lancet 375:2194.

499. Caskey, C. T. 1993. Presymptomatic diagnosis: a first step toward genetic health care. Science 262:48-49.

500. Caskey, C. T. 2010. Using genetic diagnosis to determine individual therapeutic utility. Annu Rev Med 61:1-15.

501. Peakall, D., and L. Shugart. 2002. The Human Genome Project (HGP). Ecotoxicology 11:7.

502. Rossiter, B. J., and C. T. Caskey. 1995. Presymptomatic testing for genetic diseases of later life. Pharmacoepidemiological considerations. Drugs Aging 7:117-130.

503. Margaritte, P., C. Bonaiti-Pellie, M. C. King, and F. Clerget-Darpoux. 1992. Linkage of familial breast cancer to chromosome 17q21 may not be restricted to early-onset disease. Am J Hum Genet 50:1231-1234.

504. King, M. C., J. H. Marks, and J. B. Mandell. 2003. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science 302:643-646.

505. King, M. C. 1991. Localization of the early-onset breast cancer gene. Hosp Pract (Off Ed) 26:121-126.

506. Austin, M. A., M. C. King, K. M. Vranizan, and R. M. Krauss. 1990. Atherogenic lipoprotein phenotype. A proposed genetic marker for coronary heart disease risk. Circulation 82:495-506.

507. Gudmundsson, J., P. Sulem, D. F. Gudbjartsson, T. Blondal, A. Gylfason, B. A. Agnarsson, K. R. Benediktsdottir, D. N. Magnusdottir, G. Orlygsdottir, M. Jakobsdottir, S. N. Stacey, A. Sigurdsson, T. Wahlfors, T. Tammela, J. P. Breyer, K. M. McReynolds, K. M. Bradley, B. Saez, J. Godino, S. Navarrete, F. Fuertes, L. Murillo, E. Polo, K. K. Aben, I. M. van Oort, B. K. Suarez, B. T. Helfand, D. Kan, C. Zanon, M. L. Frigge, K. Kristjansson, J. R. Gulcher, G. V. Einarsson, E. Jonsson, W. J. Catalona, J. I. Mayordomo, L. A. Kiemeney, J. R. Smith, J. Schleutker, R. B. Barkardottir, A. Kong, U. Thorsteinsdottir, T. Rafnar, and K. Stefansson. 2009. Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. Nat Genet 41:1122-1126.

508. Hunter, D. J., P. Kraft, K. B. Jacobs, D. G. Cox, M. Yeager, S. E. Hankinson, S. Wacholder, Z. Wang, R. Welch, A. Hutchinson, J. Wang, K. Yu, N. Chatterjee, N. Orr, W. C. Willett, G. A. Colditz, R. G. Ziegler, C. D. Berg, S. S. Buys, C. A. McCarty, H. S. Feigelson, E. E. Calle, M. J. Thun, R. B. Hayes, M. Tucker, D. S. Gerhard, J. F. Fraumeni, R. N. Hoover, G. Thomas, and S. J. Chanock. 2007. A genome-wide

association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39:870-874.

509. Hutter, C. M., A. M. Young, H. M. Ochs-Balcom, C. L. Carty, T. Wang, C. T. L. Chen, T. E. Rohan, C. Kooperberg, and U. Peters. 2011. Replication of Breast Cancer GWAS Susceptibility Loci in the Women's Health Initiative African American SHARe Study. Cancer Epidemiology Biomarkers & Prevention 20:1950-1959.

510. Manolio, T. A. 2010. Genomewide Association Studies and Assessment of the Risk of Disease. New England Journal of Medicine 363:166-176.

511. Manjarrez-Orduño, N., E. Marasco, S. A. Chung, M. S. Katz, J. F. Kiridly, K. R. Simpfendorfer, J. Freudenberg, D. H. Ballard, E. Nashi, and T. J. Hopkins. 2012. CSK regulatory polymorphism is associated with systemic lupus erythematosus and influences B-cell signaling and activation. Nature Genetics 44:1227-1230.

512. Copeland, W. C., N. K. Lam, and T. S. Wang. 1993. Fidelity studies of the human DNA polymerase alpha. The most conserved region among alpha-like DNA polymerases is responsible for metal-induced infidelity in DNA synthesis. J Biol Chem 268:11041-11049.

513. Copeland, W. C., and T. S. Wang. 1993. Mutational analysis of the human DNA polymerase alpha. The most conserved region in alpha-like DNA polymerases is involved in metal-specific catalysis. J Biol Chem 268:11028-11040.

514. Daee, D. L., T. M. Mertz, and P. V. Shcherbakova. 2010. A cancer-associated DNA polymerase Œ¥ variant modeled in yeast causes a catastrophic increase in genomic instability. Proceedings of the National Academy of Sciences 107:157-162.

515. Lang, T., M. Maitra, D. Starcevic, S.-X. Li, and J. B. Sweasy. 2004. A DNA polymerase Œ≤ mutant from colon cancer cells induces mutations. Proceedings of the National Academy of Sciences of the United States of America 101:6074-6079.

516. Longley, M. J., S. Clark, C. Yu Wai Man, G. Hudson, S. E. Durham, R. W. Taylor, S. Nightingale, D. M. Turnbull, W. C. Copeland, and P. F. Chinnery. 2006. Mutant POLG2

disrupts DNA polymerase gamma subunits and causes progressive external ophthalmoplegia. Am J Hum Genet 78:1026-1034.

517. Longley, M. J., P. A. Ropp, S. E. Lim, and W. C. Copeland. 1998. Characterization of the native and recombinant catalytic subunit of human DNA polymerase gamma: identification of residues critical for exonuclease activity and dideoxynucleotide sensitivity. Biochemistry 37:10529-10539.

518. Matsuda, T., K. Bebenek, C. Masutani, I. B. Rogozin, F. Hanaoka, and T. A. Kunkel. 2001. Error rate and specificity of human and murine DNA polymerase eta. J Mol Biol 312:335-346.

519. Ohashi, E., K. Bebenek, T. Matsuda, W. J. Feaver, V. L. Gerlach, E. C. Friedberg, H. Ohmori, and T. A. Kunkel. 2000. Fidelity and processivity of DNA synthesis by DNA polymerase kappa, the product of the human DINB1 gene. J Biol Chem 275:39678-39684.

520. Ohashi, E., T. Ogi, R. Kusumoto, S. Iwai, C. Masutani, F. Hanaoka, and H. Ohmori. 2000. Error-prone bypass of certain DNA lesions by the human DNA polymerase kappa. Genes Dev 14:1589-1594.

521. Sweasy, J. B., T. Lang, D. Starcevic, K.-W. Sun, C.-C. Lai, D. DiMaio, and S. Dalal. 2005. Expression of DNA polymerase Œ≤ cancer-associated variants in mouse cells results in cellular transformation. Proceedings of the National Academy of Sciences of the United States of America 102:14350-14355.

522. Wong, S. W., A. F. Wahl, P. M. Yuan, N. Arai, B. E. Pearson, K. Arai, D. Korn, M. W. Hunkapiller, and T. S. Wang. 1988. Human DNA polymerase alpha gene expression is cell proliferation dependent and its primary structure is similar to both prokaryotic and eukaryotic replicative DNA polymerases. EMBO J 7:37-47.

523. Amos, C. I., X. Wu, P. Broderick, I. P. Gorlov, J. Gu, T. Eisen, Q. Dong, Q. Zhang, X. Gu, J. Vijayakrishnan, K. Sullivan, A. Matakidou, Y. Wang, G. Mills, K. Doheny, Y.-Y.

Tsai, W. V. Chen, S. Shete, M. R. Spitz, and R. S. Houlston. 2008. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet 40:616-622.

524. Bishop, D. T., F. Demenais, M. M. Iles, M. Harland, J. C. Taylor, E. Corda, J. Randerson-Moor, J. F. Aitken, M.-F. Avril, E. Azizi, B. Bakker, G. Bianchi-Scarra, B. Bressac-de Paillerets, D. Calista, L. A. Cannon-Albright, T. Chin-A-Woeng, T. Debniak, G. Galore-Haskel, P. Ghiorzo, I. Gut, J. Hansson, M. Hocevar, V. Hoiom, J. L. Hopper, C. Ingvar, P. A. Kanetsky, R. F. Kefford, M. T. Landi, J. Lang, J. Lubinski, R. Mackie, J. Malvehy, G. J. Mann, N. G. Martin, G. W. Montgomery, F. A. van Nieuwpoort, S. Novakovic, H. Olsson, S. Puig, M. Weiss, W. van Workum, D. Zelenika, K. M. Brown, A. M. Goldstein, E. M. Gillanders, A. Boland, P. Galan, D. E. Elder, N. A. Gruis, N. K. Hayward, G. M. Lathrop, J. H. Barrett, and J. A. Newton Bishop. 2009. Genome-wide association study identifies three loci associated with melanoma risk. Nat Genet 41:920-925.

525. Wang, X., V. S. Pankratz, Z. Fredericksen, R. Tarrell, M. Karaus, L. McGuffog, P. D. P. Pharaoh, B. A. J. Ponder, A. M. Dunning, S. Peock, M. Cook, C. Oliver, D. Frost, EMBRACE, O. M. Sinilnikova, D. Stoppa-Lyonnet, S. Mazoyer, C. Houdayer, GEMO, F. B. L. Hogervorst, M. J. Hooning, M. J. Ligtenberg, HEBON, A. Spurdle, G. Chenevix-Trench, kConFab, R. K. Schmutzler, B. Wappenschmidt, C. Engel, A. Meindl, S. M. Domchek, K. L. Nathanson, T. R. Rebbeck, C. F. Singer, D. Gschwantler-Kaulich, C. Dressler, A. Fink, C. I. Szabo, M. Zikan, L. Foretova, K. Claes, G. Thomas, R. N. Hoover, D. J. Hunter, S. J. Chanock, D. F. Easton, A. C. Antoniou, and F. J. Couch. 2010. Common variants associated with breast cancer in genome-wide association studies are modifiers of breast cancer risk in BRCA1 and BRCA2 mutation carriers. Human molecular genetics 19:2886-2897.

526. Waters, K. M., L. Le Marchand, L. N. Kolonel, K. R. Monroe, D. O. Stram, B. E. Henderson, and C. A. Haiman. 2009. Generalizability of Associations from Prostate Cancer Genome-Wide Association Studies in Multiple Populations. Cancer Epidemiology Biomarkers & Prevention 18:1285-1289.

527. Weir, B. A., M. S. Woo, G. Getz, S. Perner, L. Ding, R. Beroukhim, W. M. Lin, M. A. Province, A. Kraja, L. A. Johnson, K. Shah, M. Sato, R. K. Thomas, J. A. Barletta, I. B. Borecki, S. Broderick, A. C. Chang, D. Y. Chiang, L. R. Chirieac, J. Cho, Y. Fujii, A. F. Gazdar, T. Giordano, H. Greulich, M. Hanna, B. E. Johnson, M. G. Kris, A. Lash, L. Lin, N. Lindeman, E. R. Mardis, J. D. McPherson, J. D. Minna, M. B. Morgan, M. Nadel, M. B. Orringer, J. R. Osborne, B. Ozenberger, A. H. Ramos, J. Robinson, J. A. Roth, V. Rusch, H. Sasaki, F. Shepherd, C. Sougnez, M. R. Spitz, M. S. Tsao, D. Twomey, R. G. Verhaak, G. M. Weinstock, D. A. Wheeler, W. Winckler, A. Yoshizawa, S. Yu, M. F. Zakowski, Q. Zhang, D. G. Beer, Wistuba, II, M. A. Watson, L. A. Garraway, M. Ladanyi, W. D. Travis, W. Pao, M. A. Rubin, S. B. Gabriel, R. A. Gibbs, H. E. Varmus, R. K. Wilson, E. S. Lander, and M. Meyerson. 2007. Characterizing the cancer genome in lung adenocarcinoma. Nature 450:893-898.

528. Yeager, M., N. Orr, R. B. Hayes, K. B. Jacobs, P. Kraft, S. Wacholder, M. J. Minichiello, P. Fearnhead, K. Yu, N. Chatterjee, Z. Wang, R. Welch, B. J. Staats, E. E. Calle, H. S. Feigelson, M. J. Thun, C. Rodriguez, D. Albanes, J. Virtamo, S. Weinstein, F. R. Schumacher, E. Giovannucci, W. C. Willett, G. Cancel-Tassin, O. Cussenot, A. Valeri, G. L. Andriole, E. P. Gelmann, M. Tucker, D. S. Gerhard, J. F. Fraumeni, R. Hoover, D. J. Hunter, S. J. Chanock, and G. Thomas. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 39:645-649.

529. Sinnwell, J. P., D. J. Schaid, and Z. Yu. 2005. haplo. stats: Statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous. R package version 1.

530. Woodbridge, P., C. Liang, R. L. Davis, H. Vandebona, and C. M. Sue. 2012. POLG mutations in Australian patients with mitochondrial disease. Intern Med J.

531. Bebenek, K., M. Garcia-Diaz, R. Z. Zhou, L. F. Povirk, and T. A. Kunkel. 2010. Loop 1 modulates the fidelity of DNA polymerase lambda. Nucleic Acids Res 38:5419-5431.

532. Terrados, G., J.-P. Capp, Y. Canitrot, M. Garc√≠a-D√≠az, K. Bebenek, T. Kirchhoff, A. Villanueva, F. o. Boudsocq, V. r. Bergoglio, C. Cazaux, T. A. Kunkel, J.-S. b. Hoffmann, and L. Blanco. 2009. Characterization of a Natural Mutator Variant of Human DNA Polymerase Œª which Promotes Chromosomal Instability by Compromising NHEJ. PLoS ONE 4:e7290.

533. Roe, D. R., and T. E. Cheatham Iii. 2013. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. Journal of chemical theory and computation.

534. Hagelkruys, A., A. Sawicka, M. Rennmayr, and C. Seiser. 2011. The Biology of HDAC in Cancer: The Nuclear and Epigenetic Components. Handb Exp Pharmacol:13-37.

535. Kawai, H., H. Li, S. Avraham, S. Jiang, and H. K. Avraham. 2003. Overexpression of Histone Deacetylase HDAC1 Modulates Breast Cancer Progression by Negative Regulation of Estrogen Receptor Alpha. International Journal of Cancer 107:353-358.

536. Noonan, E. J., R. F. Place, D. Pookot, S. Basak, J. M. Whitson, H. Hirata, C. Giardina, and R. Dahiya. 2009. miR-449a targets HDAC-1 and induces growth arrest in prostate cancer. Oncogene 28:1714-1724.

537. Ropero, S., M. F. Fraga, E. Ballestar, R. Hamelin, H. Yamamoto, M. Boix-Chornet, R. Caballero, M. Alaminos, F. Setien, M. F. Paz, M. Herranz, J. Palacios, D. Arango, T. F. Orntoft, L. A. Aaltonen, S. Schwartz, Jr., and M. Esteller. 2006. A truncating mutation of HDAC2 in human cancers confers resistance to histone deacetylase inhibition. Nature genetics 38:566-569.

538. Saji, S., M. Kawakami, S. Hayashi, N. Yoshida, M. Hirose, S. I. Horiguchi, A. Itoh, N. Funata, S. L. Schreiber, M. Yoshida, and M. Toi. 2005. Significance of HDAC6 regulation via estrogen signaling for cell motility and prognosis in estrogen receptor-positive breast cancer. Oncogene 24:4531-4539.

539. Weichert, W., A. Roske, V. Gekeler, T. Beckers, C. Stephan, K. Jung, F. R. Fritzsche, S. Niesporek, C. Denkert, M. Dietel, and G. Kristiansen. 2008. Histone deacetylases 1, 2 and 3 are highly expressed in prostate cancer and HDAC2 expression is associated with shorter PSA relapse time after radical prostatectomy. British journal of cancer 98:604-610.

540. Wynendaele, J., A. Bohnke, E. Leucci, S. J. Nielsen, I. Lambertz, S. Hammer, N. Sbrzesny, D. Kubitza, A. Wolf, E. Gradhand, K. Balschun, I. Braicu, J. Sehouli, S. Darb-Esfahani, C. Denkert, C. Thomssen, S. Hauptmann, A. Lund, J.-C. Marine, and F. Bartel. 2010. An Illegitimate microRNA Target Site within the 3,Ä≤ UTR of MDM4 Affects Ovarian Cancer Progression and Chemosensitivity. Cancer research 70:9641-9649.

541. Wang, D. F., O. Wiest, P. Helquist, H. Y. Lan-Hargest, and N. L. Wiech. 2004. On the function of the 14 A long internal cavity of histone deacetylase-like protein: implications for the design of histone deacetylase inhibitors. Journal of medicinal chemistry 47:3409-3417.

542. Bastian, M., S. Heymann, and M. Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks.

543. Fruchterman, T. M. J., and E. M. Reingold. 1991. Graph drawing by force-directed placement. Software: Practice and experience 21:1129-1164.

544. Wolfram, S., and A. Mathematica. 1991. System for Doing Mathematics by Computer. Addison_Wesley, ISBN 0:201-19334.

ABSTRACT

COMPUTATIONAL APPROACHES TO ANTI-TOXIN THERAPIES AND BIOMARKER IDENTIFICATION

by

REBECCA JANE SWETT

December 2013

Advisor: Dr. Andrew Lee Feig

Co-Advisor: Dr. Gerardo Andrés Cisneros

Major: Chemistry

Degree: Doctor of Philosophy

This work describes the fundamental study of two bacterial toxins with computational methods, the rational design of a potent inhibitor using molecular dynamics, as well as the development of two bioinformatic methods for mining genomic data.

Clostridium difficile is an opportunistic bacillus which produces two large glucosylating toxins. These toxins, TcdA and TcdB cause severe intestinal damage. As *Clostridium difficile* harbors considerable antibiotic resistance, one treatment strategy is to prevent the tissue damage that the toxins cause. The catalytic glucosyltransferase domain of TcdA and TcdB was studied using molecular dynamics in the presence of both a protein-protein binding partner and several substrates. These experiments were combined with lead optimization techniques to create a potent irreversible inhibitor which protects 95% of cells *in vitro.* Dynamics studies on a TcdB cysteine protease domain were

performed to an allosteric communication pathway. Comparative analysis of the static and dynamic properties of the TcdA and TcdB glucosyltransferase domains were carried out to determine the basis for the differential lethality of these toxins.

Large scale biological data is readily available in the post-genomic era, but it can be difficult to effectively use that data. Two bioinformatics methods were developed to process whole-genome data. Software was developed to return all genes containing a motif in single genome. This provides a list of genes which may be within the same regulatory network or targeted by a specific DNA binding factor. A second bioinformatic method was created to link the data from genome-wide association studies (GWAS) to specific genes. GWAS studies are frequently subjected to statistical analysis, but mutations are rarely investigated structurally. HyDn-SNP-S allows a researcher to find mutations in a gene that correlate to a GWAS studied phenotype. Across human DNA polymerases, this resulted in strongly predictive haplotypes for breast and prostate cancer. Molecular dynamics applied to DNA Polymerase Lambda suggested a structural explanation for the decrease in polymerase fidelity with that mutant. When applied to Histone Deacetylases, mutations were found that alter substrate binding, and post-translational modification.

AUTOBIOGRAPHICAL STATEMENT
Rebecca Jane Swett

Education

Ph.D., Computational Biochemistry, Wayne State University, Detroit, MI
- Advisors: Andrew L. Feig, G. Andrés Cisneros
- Research Focus: Computational investigation of *Clostridium difficile toxins. Molecular dynamics, small molecule docking and lead optimization applied to development of a potent irreversible inhibitor for TcdA and TcdB.* Development of bioinformatic software correlating phenotype to whole genome genotyping data.

Graduate Studies, Computational Chemistry, Northern Michigan University, Marquette, MI
- Advisors: Mark Paulsen, Gary Hiel
- Research Focus: Determining the mechanism of allosteric enhancement of epoxidation by P450 3A4 on sterol substrates. Natural light catalyzed synthesis of taxol precursors.

B.S., Chemistry, Northern Michigan University, Marquette, MI
- Advisor: Mark Paulsen
- Research Focus: Computational analysis of the interactions of P450 3A4 with natural and xenobiotic substrates, implications for protein-protein interactions.

Honors and Certifications
- Spooner Grant (NMU, 2005)
- Excellence in Education Grant (NMU 2007)
- Departmental Citation for Excellence in Teaching (2009)
- Esther and Stanley Kirschner Teaching Award (2010)
- Graduate School Citation for Excellence in Teaching (2011)
- Rumble Fellow (2011-2012)
- Rumble Fellow (2012-2013)
- Paul Barbara Fellowship (2013)
- Graduate Research Exhibition Winner (2013)

Affiliations
- American Chemical Society (2010)
- Sigma Xi (2007)
- Phi Lambda Upsilon (2009)

Publications and Patents

R. Swett, G. A. Cisneros, and A.L.. Feig. "Conformational Analysis of Clostridium difficile Toxin B and Its Implications for Substrate Recognition." *PloS one* 7.7 (2012): e41518.

R. Swett, G. A. Cisneros, and A.L.. Feig. "Disruption of intrinsic motions as a mechanism for enzyme inhibition" *Biophys J.* (July 2013)

R. Swett, A. Elias, J.Miller, G. Dyson, G. A. Cisneros. "Hypothesis Driven Single Nucleotide Polymorphism Search (HyDn-SNP-S)" *DNA Repair* (July 2013

S. Abdeen, **R. Swett**, S. Kern, D. Nei, M. T. Rodgers, A. L. Feig. "Rational design of an irreversible inhibitor of *Clostridium difficle* virulence factors TcdA and TcdB" *J. Biol Chem* (Submitted)

S. J. Abdeen**, R. Swett**, and A. L. Feig. "Peptide inhibitors targeting Clostridium difficile toxins A and B." *ACS Chem Biol* 5 (2010): 1097-1103.

Inhibition of *Clostridium difficile* Toxins, Provisional Patent 2066728.00067, December 201