Data-Driven Computational Approach to Study Bio-Molecular Interactions

Misagh Naderi
Louisiana State University and Agricultural and Mechanical College, misagh.naderi@gmail.com

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations

Part of the Biochemistry Commons, Structural Biology Commons, and the Systems Biology Commons

Recommended Citation
https://digitalcommons.lsu.edu/gradschool_dissertations/4545

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.
DATA-DRIVEN COMPUTATIONAL APPROACH TO STUDY BIO-MOLECULAR INTERACTIONS

A Dissertation

Submitted to the Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

The Department of Biological Sciences - Biochemistry

by

Misagh Naderi
B.S., Sharif University in Tehran, Iran, 2008
M.S., Louisiana State University in Baton Rouge, Louisiana, 2012
May 2018
ACKNOWLEDGEMENTS

I would like to thank my advisor and mentor Dr. Michal Brylinski for his unequivocal goodwill and his unwavering support. His guidance and encouragements are the most important factors in my success in the pursuit of the PhD degree.

I would also like to thank the members of my committee, Prof. Grover Waldrop, Dr. Aaron Smith, and Dr. Supratik Mukhopadhyay for sharing with me their wisdom and their wealth of knowledge; and, to Dr. Jeannette Cremer for her support as the dean’s representative on my committee. Without their help I would not be able to come this far.

Many thanks to my collaborators Dr. Chris Alvin, Tairan Liu, and Dr. Supratik Mukhopadhyay for their computer programing knowledge and support and the researchers at the Kousoulas lab: Dr. Paul Rider, Dr. Vladimir Chouljenko, Dr. Dmitry Chouljenko, Dr. Nithya Jambunathan, Dr. Ramesh Subramanian, S. Bergeron, Dr. Anu Charles, and Dr. Ahmad Saied; and special thanks to Prof. Gus Kousoulas for his guidance and support.

Special thanks go to my friends, colleagues, and teachers at the department of biological sciences and LSU graduate school: Anand Ramachandram, Dr. Rajiv Gandhi Govindaraj, Dr. Erin Schexnaydre, Dr. Nathan Gilbert, Dr. Wei Feinstein, Dr. Yun Ding, Dr. Surabhi Maheshvari, Lana Thaljeh, Ainsley Rothschild, Jeffery Lemoine, Omar Kana, Manali Singha, Abd-El-Monsif Ahmed Shawky, Dr. Muzammel Haque, Prof. Marcia Newcomer, Prof. Anne Grove, Dr. Michael Hellberg, Prof. Michelle Masse, Prof. Sean Lane.

Very special thanks to my family: Soosan Ansari, Mehdi Naderi, Mahsoo Naderi Dasoar, and my extended family Vishal Dasoar, Aryo Dasoar, Kaden Dasoar, Michael Benton, Vijay Dasoar, Mona Bhatia.
My sincere gratitude goes to my dearest Sogand Karbalaieali and dear friends Ava & Cordell Haymon, Dr. Azadeh Keyvani & Dr. Ashkan Balouchi, Melissa & Dave Dixon, Ms. Maureen Hewitt, Amanda & Dr. Ali Ghamsari, Arman Barekati, Donovan McClelland, Dr. Nathan Gilbert, Ege Adanir, Anika Campbell, Moghadam’s family, Rev. Pattie Snyder, Rev. Robin McCullough-Bade, for their friendship and support that kept me sane and helped me feel at home.

Much gratitude to my spiritual teachers and friends for their support and priceless teachings in characterbuilding: Master Stephen Co, Richard Gordon, Patricia Cole, Debra Tongue, Dr. Milad Ahmadi, Maryam Shad, Rita Riggio, Shruti Chowdhury, Masoumi Eshkevari, and to my teacher Master Choa Kok Sui.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS .............................................................................................................. ii

ABSTRACT .................................................................................................................................. vi

1. INTRODUCTION ...................................................................................................................... 1
   1.1 STATEMENT OF PROBLEM .............................................................................................. 1
   1.2 OVERVIEW OF THESIS ..................................................................................................... 9
   1.3 REFERENCES .................................................................................................................... 11

2. DRUG-BOUND PROTEIN MODELING .................................................................................. 13
   2.1 INTRODUCTION .................................................................................................................. 13
   2.2 METHODS .......................................................................................................................... 15
   2.3 RESULTS AND DISCUSSION ............................................................................................. 18
   2.4 CONCLUSIONS ................................................................................................................... 28
   2.5 REFERENCES ..................................................................................................................... 31

3. STRUCTURE-BASED DRUG REPURPOSING ....................................................................... 36
   3.1 INTRODUCTION .................................................................................................................. 36
   3.2 METHODS .......................................................................................................................... 40
   3.3 RESULTS AND DISCUSSION ............................................................................................. 43
   3.4 CONCLUSIONS ................................................................................................................... 64
   3.5 REFERENCES ..................................................................................................................... 65

4. TARGET-FOCUSED LIBRARIES FOR VIRTUAL SCREENING ........................................... 74
   4.1 INTRODUCTION .................................................................................................................. 74
   4.2 METHODS .......................................................................................................................... 78
   4.3 RESULTS ............................................................................................................................ 84
   4.4 DISCUSSION ....................................................................................................................... 98
   4.5 CONCLUSIONS ................................................................................................................... 102
   4.6 REFERENCES ..................................................................................................................... 102

5. DECOMPOSING MOLECULES FOR FRAGMENT-BASED DRUG DESIGN ........................... 108
   5.1 INTRODUCTION .................................................................................................................. 108
   5.2 METHODS .......................................................................................................................... 110
   5.3 RESULTS AND DISCUSSION ............................................................................................. 112
   5.4 CONCLUSIONS ................................................................................................................... 119
   5.5 REFERENCES ..................................................................................................................... 119

6. DOMAIN-BASED HSV-1 GLYCOPROTEIN-K PROTEIN MODEL ..................................... 123
   6.1 INTRODUCTION .................................................................................................................. 123
   6.2 MATERIALS AND METHODS ............................................................................................ 128
   6.3 RESULTS ............................................................................................................................. 133
   6.4 DISCUSSION ....................................................................................................................... 140
   6.5 REFERENCE ....................................................................................................................... 147
ABSTRACT

Proteins commonly convey their functions in coordination with other proteins, small molecules and/or other biological assemblies such as sugars, lipids, DNA or RNA. Understanding the nature of these interactions is therefore central to improving our knowledge of biological systems. This body of work is a consolidation of three different computational approaches to study bio-molecular interactions: large-scale protein-ligand modeling, fragment-based cheminformatics, and computational analysis of interactions in single proteins.

The first issue that is addressed in this study is the scarcity of atomic crystal structures of protein-drug complexes. In general, a drug molecule’s affinity for multiple protein targets may causes unsolicited side effects and therefore is considered a burden; nonetheless, drug promiscuity can be leveraged to repurpose known drugs to treat multiple diseases. We offer large-scale modeling of drug-protein complexes based on data curated in DrugBank, Binding Database and Protein Data Bank. eModelBDB, a library of 200k high-quality drug-protein models as well as a database of protein models bound to FDA-approved drugs explored for rare disease drug repurposing are presented.

Two fragment-based drug design software tools are built to generate targeted virtual screening libraries. These tools are based on minimalistic combinatorial chemistry implemented in a robust graph-based algorithm in C++ and python. eMolFrag deconstructs chemical compounds into small fragments, while eSynth can generate libraries of new compounds by conjoining those fragments. However, eSynth does not require any complicated chemistry rules and is independent of the structures of the protein targets. Computational performance and accuracy of these methods are discussed.
To analyze specific protein interactions, modeling of Herpes Simplex Virus type-1 (HSV-1) proteins, including a difficult case of a transmembrane glycoprotein, are tackled. These atomic models were used to complement and to guide experimental efforts answering important questions about virus protein functions as well as identifying new druggable targets. Identifying a domain in glycoproteinK in HSV-1 and its role in neuronal entry, recognizing a disulfide bond formation that is essential for viral reproduction \textit{in vitro}, characterizing a protein-protein interaction site in UL37, and discovery of a novel allosteric anti-HSV druggable site in the DNA-packaging motor are examined computationally and validated experimentally.
1. INTRODUCTION

1.1 Statement Of Problem

Systems level approaches investigating biomolecular interactions including protein/protein or protein/nucleic acids and protein-ligand interactions have contributed to basic biological research and drug discovery significantly. Computational reconstruction of protein assemblies from x-ray crystallography and NMR are currently the common practice in structure biology. Virtual screening amongst many other computational strategies is now an integral part of almost every drug discovery pipeline. Moreover, computer-aided systems biological approaches have invigorated interest into exploiting the natural promiscuity of drugs to repurpose known drugs, elucidate and develop drugs targeting complex pathways, and discover relationships between remotely related proteins. In this manuscript three diverse data-driven computational approaches to investigate protein-drug interactions, fragment-based drug design, and analysis of protein-structure functions are discussed.

Lack of experimental drug-bound complex structures

Over the past few decades, structural bioinformatics has become an increasingly important component of modern drug discovery. Despite significant advances in experimental methods such as X-ray crystallography to acquire protein structures, the limitations and expensive procedures make it unlikely to have experimental structures of all known protein sequences in the near future. For example, as of March 2018, the number of gene products in the Reference Sequence Database is $>1.02 \times 10^8$; in contrast, the number of experimentally determined structures deposited in the Protein Data Bank (PDB) is only 138,464. Considering a much slower pace of experimental structure determination compared to genome sequencing, this
disparity continues to grow. Thus, computational modeling can be utilized to generate high quality models to fill this gap and support large-scale rational drug design projects. It is suspected that these endeavors have far reaching consequences across a multitude of disciplines. Physiological responses and mechanisms of actions of therapies with multiple bioactive compounds lend themselves nicely to the philosophy of “one drug, multiple targets” (Reddy & Zhang, 2013) and “multiple drugs, similar targets”. Research on cancer (Peng et al., 2014) and neurodegenerative diseases (Geerts, Hofmann-Apitiu, Anastasio, & Brain Health Modeling, 2017) has much to gain from insights into the human interactome. As such, it is imperative to scrutinize the intricate networks of drug-protein interactions at a systems level.

Shortfall in Orphan drug discovery

Despite our advances in medicine and drug development, there are many diseases with inadequate or nonexistent treatment; some due to the complexity of the disease, but some only because they affect less than one person per 1,500. These diseases, affecting fewer than 200,000 people in the United States, are known as rare or orphan diseases. Rare diseases are sparse, but their patients are numerous; there are almost 7,000 rare diseases, affecting more than 25 million Americans and their families (NIH, 2013). The pharmaceutical industry does not conduct research on orphan diseases since it is unlikely to recover the costs of drug discovery and marketing from such. Thus, it is on academic researchers, those not limited by purely economic pressure, to address the issue and search for inexpensive solutions to help these patients. Accordingly, the FDA Office of Orphan Product Development has initiated the Orphan Drug Designation program to support rare disease drug discovery, including the repositioning of FDA approved drugs to treat such diseases (FDA). The significantly lower cost of drug repurposing in comparison to new discoveries makes this strategy very appealing and introduces a feasible
solution to treat almost 7,000 diseases. Yet, the extreme complexity of the biological systems imposes challenges in rational drug repurposing. For example, screening all pharmacological agents versus the many proteins involved in a disease pathway is simply impractical. Thus, computational approaches are required to find possible connections between drugs and new potential protein targets to narrow down the experimental space. Consequently, drug-repurposing experiments guided by computational research are more economical and have a significantly higher chance of success.

Low-yield virtual screening campaigns

Structure-based drug design is now an essential addition to drug discovery pipelines especially in the initial lead generation and optimization stages. This cost effective approach is many fold faster and cheaper than experimental drug screenings; Structure-based drug design also has the ability to virtually screen compounds that are not yet synthesized or are not available immediately for experimentation. The number of potential pharmacologically active molecules, according to the Lipinski rule of five, has been estimated to be as large as $10^{63}$. However, the libraries that are commonly being explored in virtual screening campaigns only cover a portion of possible chemical structures. For example, some of the largest libraries such as ZINC (Irwin, Sterling, Mysinger, Bolstad, & Coleman), Binding Database (Liu, Lin, Wen, Jorissen, & Gilson, 2007), and ChEMBL19 (Bento et al.) merely include 20M, 450K, and 1.5M compounds, correspondingly. Although exploring the entire chemical space might be an interesting option our current computational capabilities would not suffice such effort. The alternative is to build a smaller subset of the chemical space that is more relevant to the target protein. However, selecting random molecules or small representative portions of the chemical landscape will not lead to efficient sampling. To find an optimal algorithm to focus the search space to a small and
targeted subset of the chemical space a large-scale fragment-based drug design approach has been pursued in the current research. This is achieved by computationally constructing large quantities of compounds based on molecular fragments generated from prosthetic groups of the target protein/protein family.

Computational insights to inform experimental efforts

Understanding biomolecular interactions is essential to understanding biology. Insights generated from protein structures can help identifying the potential interaction sites, partners and subsequently guess the cellular localization and mechanism of action of a protein. The conventional approaches to explore a protein’s mechanisms of action include performing mutational analysis and protein-protein interaction studies through biochemical, in vitro, or in vivo experiments. These experiments have an inherent limitation of studying the protein indirectly in the context of interaction networks. Protein structures can be helpful to diagnose functional domains, possible conformational changes, binding sites, etc. in a protein. This strategy can guide experimental studies by suggesting critical mutations and specific amino acid cross linkers and small molecules that can disrupt or inhibit functions and interactions. However, our knowledge of protein structures is not distributed evenly between organisms. For example, in the field of virology, although protein structure analysis through x-ray crystallography, NMR, and cryoelectron microscopy has improved significantly in the past two decades many viral proteins’ structures are unknown. Currently, there are 138,464 experimentally determined structures deposited in the PDB as of March 2018; in comparison, only 9,362 structures belong to viruses. There is a large disparity among the viruses as well, where 18% of these protein structures belong to Human Immunodeficiency Virus (HIV) alone, while more than 400 genera of viruses exist. Specifically in herpesviridae with more than 100 species within the family, there are only
399 structures carrying 461 unique protein chains deposited into PDB where more than 20% of the structures were obtained with X-ray resolutions above 3 Å, and only 20% are below 2 Å. Furthermore, these structures represent no more than 10 different protein families. Considering that different herpes virus species genomes encode 100-200 gene products, our current data on herpes virus protein structures, <0.3% of PDB, is very limited. Viruses can cause life threatening and chronic diseases in humans and animals. World Health Organization (WHO) reported in January 2017 that 3.7 billion people under age 50 (67%) have HSV-1 infection globally. Understanding protein functions and exploring biomolecular interactions in between viral proteins themselves and with the host can not only lead to discovery of effective anti-viral treatments but also can further our understanding of the biological systems. In case herpes viruses proteins that are involved in virus entry into neurons or are involved in establishing neuronal latency can teach us about our neuronal systems, memory, and a plethora of neurodegenerative diseases. In this research, modeling of HSV-1 viral proteins were pursued to guide experimental efforts in understanding virus entry and to assist anti-HSV-1 drug design.

1.2 Project Objectives

In this manuscript the main goal is to generate protein models and bioactive compounds to support drug discovery efforts based on systems level computational methods in the context of biomolecular interactions. More specifically three main objectives have been purposed. A novel platform for rational drug repositioning

Among many approaches to study disease-related biological networks (Cho, Kim, & Przytycka, 2012), the physicochemical characterization of drug-binding pockets in macromolecular structures hold a significant promise to facilitate drug development efforts contributing to the understanding of protein molecular functions (Coleman & Sharp, 2010;
Duran-Frigola et al., 2017). Pocket matching algorithms assess whether pairs of binding sites are similar or not. Binding sites are considered similar if they function in the same way, and/or bind the same ligand. Under a classical, “one drug, one targets” view of pharmacology, there would be no similar drug-binding sites. Nonetheless, it became evident that the concept of one drug acting on a single receptor is not accurate (Medina-Franco, Giulianotti, Welmaker, & Houghten, 2013). Modern pharmacology now recognizes drugs as acting on biological systems, rather than only their intended targets. While it is possible for a drug to exclusively interact with its primary target, an estimated number of target proteins per drug is as high as 6.3 (Mestres, Gregori-Puigjane, Valverde, & Sole, 2008). Accordingly, the drug promiscuity is the principle phenomenon behind drug repurposing, side effects, and polypharmacology.

Following this framework in the current work, holo-protein models have been generated with the purpose to aid drug discovery especially for rare diseases. More specifically, a protein linked to an orphan disease carrying a binding pocket similar to a known binding site of an FDA approved drug will be suggested as a candidate for repositioning. This effort benefits not only the rare diseases but also the modern drug development as a whole. Initially two reliable, high-coverage libraries of computer generated structural models representing two large ligand-binding datasets the Binding Database and FDA approved drugs bound to their known target gene products were constructed. Although experimental crystal structures of these complexes are more desirable and will be added to the library, such structures are not available in most cases. Parallel to building a library of all available drug-protein complex models, another database including possible protein binding sites across the human proteome that are involved in rare diseases was constructed. Computationally predicted structures as well as available crystal structures contributed to building this database of protein binding sites. After constructing the necessary
libraries a sequence-order-independent searching algorithm, eMatchSite (Brylinski, 2017), was employed to perform the binding site comparisons. In other words, known drug-protein binding sites guided the search for similar binding pockets in the second library regardless of the overall structure of individual proteins. The theory behind this work is that in case a newly found binding pocket belongs to a protein which plays a role in a rare disease, that protein becomes a new alternative target for the pharmacological agent in the initial drug-protein complex.

Targeted virtual screening libraries:

In order to increase the yield in virtual screening (VS) campaigns, a novel method to build VS libraries tailored for any specific protein has been developed. These targeted VS libraries are enriched with chemical compounds that are predicted to have bioactivity against the target protein, inferred from other ligands of the same protein or similar homologous proteins. The VS library can similarly be targeted with respect to the binding pocket, as this method is independent of the protein structure and borrows molecular patterns from other ligands of interest. A fragment-based computational synthesis of chemical compounds, eSynth, is developed in C++, which is capable of assembling drug-like molecules from small molecular fragments generated from known bioactive molecules. The main goal was to generate larger and more complex molecules that would fit into Lipinksi rules for drug-likeliness. This task requires advanced knowledge in computer science to prevent combinatorial explosion by applying complex filtering algorithms to save memory space. This part of the project is achieved in collaboration with Dr. Chris Alvin and Dr. Supratik Mukhopadhyay from Computer Sciences.

These small molecular fragments not only are necessary for fragment-based drug design methods such as eSynth but also they can carry valuable information about their protein targets. Large-scale analysis of antibiotic, antifungal, or antiviral compounds and those compounds that
are active against different protein families can carry specific moieties that are pertinent to their bioactivity. In the current manuscript the second main goal in fragment-based drug is to decompose drug-like compounds to smaller units composed of a few atoms while recording the connectivity patterns of the fragments in the parent molecules. The goal was to produce a software to generate the fragments in a fast and efficient way. Adding parallel performance while using only the RDKit Python module to achieve this goal posed difficulties and required advanced programing skills. This goal was pursued in collaboration with Tairan Liu, Dr. Chris Alvin and Dr. Supratik Mukhopadhyay.

Virus protein modeling

To explore answer specific questions in herpes virology, the structure prediction for herpes virus glycoprotein K as well as HSV-1 protein UL37 were pursued. The goal was to attempt a difficult case of modeling in a membrane bound protein whose critical role in virus infection had been shown in vitro by this author. The goal included computational modeling of a detailed high quality protein structure of gK based on transmembrane protein templates. The objectives were to guide mutation studies to identify functional domains within gK protein, investigate the possibility of disulfide bond formations as well as other posttranslational modifications. This effort was to augment experiments that could further elucidate the role of the different elements predicted by computational modeling. In another effort, similar strategy was pursued to explore a region of interest within the UL37 protein, an interacting partner of gK, that also lacked the experimentally solved structure. The goal in this project was to guide mutation studies and to investigate the role of the critical amino acids in the context of biomolecular interactions and protein stability of UL37. To understand the roles of gK and UL37 from a
structure-based perspective and to identify active as well as allostERIC sites for drug design purposes was another goal in this study.

1.2 Overview of Thesis

The first chapter introduces the shortcomings in protein structure data generated by experimental methods, and the opportunities where computational efforts can augment the current knowledge of biomolecular interactions.

Chapter 2 describes a similarity-based docking method to model drug-bound protein structures as well as a new database, eModelBDB, compiled following this method. This chapter further includes quality assurance strategies, and case studies to evaluate the quality of the models in eModelBDB.

Chapter 3 presents a large-scale computational pipeline to explore possible drug repurposing candidates for rare diseases, eRepoORP (Brylinski, Naderi, Govindaraj, & Lemoine, 2017; Govindaraj, Naderi, Singha, Lemoine, & Brylinski, 2018). Moreover, similarity-based docking and threading strategies, similar to chapter 1, are used to construct drug-bound models for FDA-approved drugs and their known target proteins. These models are then compared with a sequence-order independent local binding-site alignment tool, eMatchSite, developed by the same group to identify similar binding sites. Quality control strategies cross validation and case studies are extensively discussed in this chapter.

Chapter 4 discusses a novel graph-based method, eSynth (Naderi, Alvin, Ding, Mukhopadhyay, & Brylinski, 2016), in computational combinatorial chemistry. This chapter investigates the possibility of constructing large yet targeted virtual libraries specific to a protein of interest based on previous knowledge of ligands of the same protein
family. Results including the computational performance, case studies, and cross validation experiments are shown in this chapter.

Chapter 5 introduces new software to generate molecular fragments from chemical compounds, eMolFrag (Liu, Naderi, Alvin, Mukhopadhyay, & Brylinski, 2017). The unique capability of eMolFrag to retain connectivity patterns of these fragments as well as its application in fragment-based drug discovery is laid out. The computational performances, case studies, and cross validation experiments to test the software are shown as well.

Chapters 6, 7, and 8 explore the capability of computational protein modeling to assist experimental efforts in the context of virology research. In chapters 6 and 7 computational and experimental results in investigating herpes virus glycoprotein-K (gK) (Jambunathan et al., 2015; Rider et al., 2017) are discussed. The first three-dimensional model for gK is proposed in this chapter. In chapter 8, herpes simplex virus type-1 UL-37 protein structure and its interaction with gK are studied by computational and experimental methods (Chouljenko et al., 2016). These chapters include highly collaborative works and the author of this manuscript has only performed the computational experiments.

Chapter 9 provides the summary of the work, conclusions from these projects to understand biomolecular interactions using data-driven approaches. Recommendations for future studies in computational drug discovery based on eModelBDB, eRepoORP, eSynth, and eMolFrag are also included. Moreover, the importance and future of computational modeling to understand and explain complex molecular interactions in biological systems have been discussed.
1.3 References


FDA. (2014). Developing Products for Rare Diseases & Conditions.


NIH. (2013). Rare Diseases Clinical Research Network.


2. DRUG-BOUND PROTEIN MODELING

2.1 Introduction

Structural bioinformatics is becoming an increasingly important component of modern drug discovery. Despite significant advances in experimental methods to acquire protein structures, such as X-ray crystallography and nuclear magnetic resonance, technical limitations and expensive procedures make it unlikely to have the experimental structures of all known protein sequences in the near future. For example, as of October 2017, the number of gene products in the Reference Sequence Database (O'Leary et al., 2016) is $9.5 \times 10^7$. In contrast, the number of experimentally determined protein structures in the Protein Data Bank (PDB) (Berman et al., 2000) is 130,750, which reduces to 46,893 structures after removing similar proteins at 95% sequence identity. Genome sequencing currently produces as many as 13 million protein sequences each year, whereas only an average number of 8,872 protein structures are solved experimentally at the same time. Since this disparity between the number of available sequences and structures will likely continue to grow, high-throughput computational modeling is expected to play a significant role in biomedical sciences by generating 3D models for those proteins whose structures will not be determined in the near future.

In addition to protein sequence and structure repositories, the Binding Database (BindingDB) provides comprehensive information on interactions between small, drug-like molecules and proteins considered to be drug targets collected from affinity measurements (Liu, Lin, Wen, Jorissen, & Gilson, 2007). The BindingDB can be used to identify protein targets for small molecules and bioactive compounds for new proteins, as well as to conduct virtual screening with ligand-based methods. As of October 2017, BindingDB contains 1,391,403
binding data, however, only 2,291 ligand-protein crystal structures with BindingDB affinity measurements are available in the PDB. To bridge this gap, we created eModel-BDB, a new database of 200,008 high-quality drug-protein complex models involving 108,363 unique drug-like compounds and 2,791 proteins from the BindingDB. This repository was constructed with a state-of-the-art protocol to generate protein models in their ligand-bound conformational state, employing meta-threading, pocket detection, and protein structure and ligand chemical alignment techniques. eModel-BDB roughly quadruples the current structural information on known drug-protein complexes.

To fully appreciate the immensity of the structural data included in eModel-BDB, we estimate the time required to solve an equal number of drug-protein assemblies. Figure 2.1 shows that at the current pace, 2,447 ligand-bound protein structures containing 607 non-redundant complexes are deposited to the PDB each month. Therefore, it would take about 329 months for 200,008 unique complex structures to be determined experimentally. In contrast to other databases comprising protein models in the unbound conformational state generated through traditional structure modeling (Castrignano, De Meo, Cozzetto, Talamo, & Tramontano, 2006; Sanchez et al., 2000), eModel-BDB includes annotated structure models of drug-protein complexes with known binding affinities. It provides high-quality data to support structure-based drug discovery as well as repurposing of known drugs based on binding pocket and ligand similarities. In addition, the information provided by eModel-BDB can be utilized to facilitate experimental structure determination by developing protocols to stabilize proteins with ligands. The protocol to construct eModel-BDB described in this communication is based entirely on open source software to ensure that any researcher is able to produce new holo-protein models as more data becomes available in the PDB and BindingDB.
Figure 2.1. Deposition rate of ligand-bound structures to the Protein Data Bank. At any given time, we counted the total number of protein chains binding small molecules (light gray squares and a dashed line) and the number of unique complex structures obtained by clustering individual chains at 80% sequence identity (dark gray circles and a solid line). $N_t$ and $N_u$ in the linear regression equations are the total and unique number of ligand-protein complexes, respectively, and $m$ stands for month.

2.2 Methods

Protein structure modeling

Drug-bound protein complexes in eModel-BDB are generated with a template-based approach. The first phase is to construct structure models for single protein chains 50-999 amino acids in length obtained from BindingDB with eThread (Brylinski & Lingam, 2012), which supports both close as well as remote homology modeling. eThread employs Modeller, a commonly used comparative modeling program (Sali & Blundell, 1993), to build apo-protein structures based on alignments produced by three fold recognition algorithms, HH-suite (Remmert, Biegert, Hauser, & Soding, 2011), SparksX (Yang, Faraggi, Zhao, & Zhou, 2011), and RaptorX (Ma, Wang,
Zhao, & Xu, 2013). Subsequently, side-chain positions and hydrogen-bonding networks in the initial models are improved with ModRefiner, a program to refine protein structures at the atomic-level with a composite physics- and knowledge-based force field (D. Xu & Zhang, 2011). The quality assessment of refined models is carried out with ModelEvaluator (Wang, Tegge, & Cheng, 2009) in terms of the estimated Global Distance Test score (GDT-score). Out of 5,501 BindingDB proteins, 4,906 were assigned an estimated GDT-score of ≥0.4 indicating good quality models (He, Alazmi, Zhang, & Xu, 2013; Mirabello, Adelfio, & Pollastri, 2014).

Ligand-binding site identification

Confident structure models with a GDT-score of ≥0.4 are further annotated with binding pockets and residues by eFindSite (Brylinski & Feinstein, 2013), which also computes a calibrated pocket confidence score. eFindSite detected 2,922 high-, 644 moderate-, and 776 low-confidence pockets in the eThread models of BindingDB targets. At this point, BindingDB drugs can be assigned to the predicted pockets with fingerprint-based virtual screening. Specifically, for a given drug-target pair in the BindingDB, we compute a rank of the drug against pockets detected by eFindSite, where the remaining BindingDB compounds are used as the background library. eFindSite conducts virtual screening with a representative set of molecular fingerprints and physicochemical properties calculated for ligands extracted from weakly homologous template structures (Feinstein & Brylinski, 2014). Top one, two and three pockets are considered for high-, moderate- and low-confidence targets, respectively. A drug matches the predicted pocket if it is ranked within the top 20% of the screening library. With this protocol, we matched 108,363 drugs to binding pockets identified in their target proteins.
Similarity-based ligand docking

In the next phase, drug molecules are positioned within the predicted pockets with a two-step, similarity-based docking protocol. This procedure exploits a significant structural conservation of ligand binding modes across remote homologs (Brylinski & Skolnick, 2009). First, globally similar ligand-bound templates from the PDB, identified by eFindSite to have a similar pocket as the BindingDB protein, are superimposed onto the apo-model. Proteins are aligned with Fr-TM-align (Pandit & Skolnick, 2008) employing the Template Modeling score (TM-score) (Zhang & Skolnick, 2004) to measure the global structure similarity. Subsequently, the BindingDB compound is aligned onto the template-bound ligand in order to place it in the predicted pocket of the apo-model. Here, we use chemical alignments constructed with kcombu (Kawabata, 2011), which also reports the chemical similarity between the BindingDB compound and the template-bound ligand measured by the Tanimoto coefficient (TC). Since a perfect case corresponds to both a TM-score and a TC of 1.0, we introduce a new metric, the Perfect Match Distance (PMD), combining protein structure and ligand chemical similarity values:

\[
PMD = \sqrt{(1 - \text{TM-score})^2 + (1 - \text{TC})^2}
\]

Eq. 1.

PMD is simply the Cartesian distance from the perfect match on the TM-score/TC plane. In order to generate only high-quality holo-models, those cases with a PMD of >0.6 are excluded from the modeling process. This PMD cutoff was chosen to ensure that both TM-score and TC for the selected templates are always above their individual significance threshold values of ≥0.4 (Kawabata, 2011; Zhang & Skolnick, 2004). Further, for those cases having multiple ligand-bound templates satisfying the PMD criterion of ≤0.6, a template with the shortest PMD is selected to build the holo-model of the BindingDB complex.

Complex structure refinement and assessment
In the final phase, protein models are rebuilt in the presence of the docked BindingDB compounds with Modeller. To make sure that the binding site is remodeled to accommodate the specific ligand, binding residues identified by eFindSite are removed from the initial model while enforcing the presence of secondary structure predicted by PSIPRED (Jones, 1999). The resulting models are further annotated with the ligand-protein interaction score according to the Distance-scaled Finite Ideal-gas REference (DFIRE) potential. The eModel-BDB database contains atomic-level structure models of 200,008 drug-protein interactions from BindingDB, comprising 2,791 non-redundant proteins and 108,363 drug-like compounds.

2.3 Results And Discussion

Data quality control

The quality control is pertinent to both protein structure modeling as well as binding site prediction. The quality of protein models is assessed with ModelEvaluator employing various structural features to compute the absolute quantitative score with a support vector regression. This approach assigns the GDT-score to a model by analyzing its secondary structure, relative solvent accessibility, contact map, and β-sheet structure. It has been demonstrated that GDT-scores estimated by ModelEvaluator for template-based models are highly correlated with the actual values with the Pearson correlation coefficient of 0.82 (Wang et al., 2009). The first violin in Figure 2.2 shows that eModel-BDB contains close as well as remote homology models with the median target-template sequence identity of 63%. The second violin indicates that the vast majority of these structures are accurate with the median estimated GDT-score for BindingDB proteins is 0.62. Further, as many as 78% of binding sites predicted by eFindSite to match BindingDB ligands have a high confidence of >0.8. We showed previously that confidence scores of >0.8 assigned by eFindSite correspond to the Mathews correlation coefficient (MCC).
of ≥0.6 for predicted binding residues (Brylinski & Feinstein, 2013). On that account, we expect that the majority of binding sites for BindingDB drugs are correctly annotated as well. Note that in contrast to other pocket predictors, eFindSite annotations and confidence assignments are, to some extent, independent on the accuracy of protein models.

Figure 2.2. Violin and box plots for model quality control. The distribution of the target-template sequence identity (SeqId) is compared to a Global Distance Test (GDT) score estimated for structure models.

The quality of complex models is controlled by imposing thresholds on the chemical similarity between BindingDB and PDB ligands as well as the global structure similarity between eThread models and ligand-bound templates from the PDB. Figure 2.3A shows the distribution of both parameters across eModel-BDB models. Encouragingly, the median TM-score and TC for ligand-bound templates used to build eModel-BDB are as high as 0.81 and 0.67, respectively. Previous studies show that the probability for a protein pair to belong to the
same fold is 98% when the TM-score is close to 0.8 (J. Xu & Zhang, 2010). In addition, it was demonstrated that the root-mean-square deviation (RMSD) over ligand non-hydrogen atoms for similarity-based docking conducted with the TC in the range of 0.6-0.8 is typically 2-3 Å (Brylinski, 2013). TM-score and TC values are combined into a single assessment score, the PMD, measuring the distance from the perfect match. Therefore, selecting template proteins with a lower TM-score to BindingDB targets requires their ligands to have a high TC and vice versa, selecting PDB ligands with a lower chemical similarity to BindingDB molecules requires a high global structure similarity between proteins. Figure 2.3B shows that eModel-BDB complex models have been constructed at the median PMD of 0.46.

Figure 2.3. Similarities between target and holo-template proteins. (A) The chemical similarity between BindingDB and PDB ligands measured with the Tanimoto coefficient (TC) is plotted against the global structure similarity of eThread models and ligand-bound templates from the PDB assessed with the TM-score. The 2D contour plot is generated by smoothing the data with the kernel density estimation technique. 1D histograms show the distribution of TC (top) and TM-score (right) values across eModel-BDB models. (B) Violin and box plot for the target-template Perfect Match Distance (PMD) combining TC and TM-score.
Data validation

In addition to the rigorous quality control maintained during dataset generation, eModel-BDB is validated retrospectively against experimental structures recently deposited to the PDB. The structure models of BindingDB interactions have been constructed with the PDB library as of January 31st 2017, therefore, we examined 7,012 experimental structures deposited to the PDB after February 2017 to validate eModel-BDB structures. The validation protocol is made more challenging by including only those models built from remote homology at a template-target sequence identity of <40%. In order to maximize the validation coverage, we use the recently determined structures of eModel-BDB targets and their homologs with at least 40% sequence identity. Recently solved experimental structures selected from the PDB validate 161 eThread models and 952 BindingDB reaction set IDs, comprising 39 target proteins, 52 pockets, and 881 compounds. The list of validation pairs is given in Supplementary File S2.

Protein structure modeling

The first violin in Figure 2.4 shows that the median TM-score of eModel-BDB vs. experimental structures is 0.85 with as many as 98.1% of the models having a TM-score of ≥0.4. Clearly, the majority of structures are modeled by eThread with a high accuracy. A representative example of the correctly predicted target structure is dihydrofolate reductase (DHFR) from *Streptococcus pyogenes* build on a crystal structure of DHFR from *Streptococcus pneumoniae* (PDB-ID: 3ix9, chain B, 36% sequence identity to the target) (Lee, Yennawar, Gam, & Benkovic, 2010). The eThread model, whose estimated GDT-score is 0.92, was then used to construct a structure model for the BindingDB reactant set ID 00267770 consisting of DHFR complexed with BDBM50329610. This model is validated against the crystal structure of DHFR-UCP1106 from *Staphylococcus aureus* (PDB-ID: 5isp, chain A, 43% sequence identity to
the target) released on 2017-06-28 (Reeve et al., 2016). Figure 2.5 shows the predicted weakly homologous model of DHFR-BDBM50329610 (purple) superposed on the experimental structure of DHFR-UCP1106 (gold). The eModel-BDB model is indeed highly accurate with a TM-score of 0.95 and a Ca-RMSD of 1.23 Å over 157 aligned residues. In addition, Figure 2.6A shows that the estimated GDT-score employed in this study as the confidence measure to control the quality of protein models correlates with the accuracy of final models evaluated with the TM-score. On that account, the estimated GDT-score provides a robust quality assessment measure, which can be used to control the quality of models in eModel-BDB.

Figure 2.4. Violin and box plots for the distribution of validation scores. The validation is conducted for remote homology protein models constructed by January 2017 against the experimental structures of either target proteins or their close homologs deposited to the PDB after February 2017. The global structure similarity is measured with the TM-score. The pocket distance is measured between the predicted pocket center and the geometric center of the ligand in the experimental structure superposed onto the eThread model.
Binding pocket prediction

The accuracy of pocket prediction is validated by superposing the experimental holo structure onto the eModel-BDB model and then calculating the distance between the geometric center of a bound ligand in the experimental complex and the pocket center predicted by eFindSite in the model. The second violin in Figure 2.4 shows that the median pocket distance is 5.5 Å with 59.6% of pockets predicted within 6 Å, therefore, most eFindSite annotations are highly accurate.

Figure 2.5. Representative example of a structure model constructed by eThread. The model of dihydrofolate reductase (DHFR, purple ribbons) complexed with BDBM50329610 is superposed onto the crystal structure of homologous DHFR from S. aureus (gold ribbons) complexed with UCP1106. Ligands bound to target proteins are shown as solid sticks (BDBM50329610 is purple and UCP1106 is gold) with non-carbon atoms colored by atom type (O – red, N – blue).
Figure 2.6. Analysis of the accuracy of structure modeling and ligand docking. The assessment is performed for remote homology complex models constructed by January 2017 against the experimental structures of either target proteins or their close homologs deposited to the PDB after February 2017. (A) Accuracy of global structure prediction evaluated by the TM-score with respect to the estimated GDT-score. (B) Accuracy of similarity-based docking with respect to the chemical similarity between BindingDB and PDB ligands measured by the Tanimoto coefficient (TC). The ligand RMSD is calculated over non-hydrogen atoms according to the chemical alignment reported by kcombu. Solid red lines show the average prediction accuracy for binned GDT-score values in A and the chemical similarity in B. Dotted black lines mark the median TM-score in A and RMSD in B across all benchmarking cases.

A binding site predicted in the model of vitamin D receptor (VDR) is a representative example of pocket prediction with eFindSite. This model was constructed by eThread based on human retinoic acid receptor RXR-alpha (PDB-ID: 4nqa, chain H, 38% sequence identity to the target) (Lou et al., 2014). Although the GDT-score estimated for the VDR model is 0.62 indicating a moderately accurate structure, the top ranked binding site annotated by eFindSite is assigned a high confidence of 94.2%. Figure 2.7 shows the VDR model (purple ribbons) superposed onto the crystal structure of vitamin D3 receptor A (gold ribbons) complexed with a synthetic analog of 1α,25-dihydroxyvitamin D3 (PDB-ID: 5nky, chain A, 66% sequence identity to the target) released on 2017-05-24 (Belorusova et al., 2017). Not only the VDR model aligns well to the experimental structure with a TM-score of 0.90 and a Cα-RMSD of 2.13 Å over 235
aligned residues, but also the predicted pocket center (purple sphere) is only 5.5 Å away from the geometric center of vitamin D analog (gold sphere).

Figure 2.7. Representative example of a binding site detected by eFindSite. The model of vitamin D receptor (VDR, purple ribbons) is superposed onto the crystal structure of homologous VDR from human (gold ribbons) complexed with a synthetic analog of vitamin D (gold and red sticks). Cα atoms of binding residues predicted in the VDR model by eFindSite are shown as small spheres. Large spheres connected by a dashed black line are placed at the location of the predicted pocket center (purple) and the geometric center of vitamin D analog (gold).

Ligand docking

Finally, we calculate the RMSD over non-hydrogen atoms between the BindingDB drug in the eModel-BDB structure and the bound ligand in the superposed experimental complex. The first violin in Figure 2.8 shows that the median ligand RMSD is 2.6 Å and it is ≤3 Å for 58.1% of BindingDB compounds. The model of the BindingDB reactant set ID 50103430 consisting of cytochrome P450 17A1 (CYP17A1) complexed with BDBM50061174 is selected to exemplify
the accuracy of complex structures in eModel-BDB. The model of CYP17A1 built on the crystal structure of human microsomal cytochrome P450 2A6 (PDB-ID: 1z11, chain A, 29% sequence identity to the target) (Yano, Hsu, Griffin, Stout, & Johnson, 2005) by eThread is assigned an estimated GDT-score of 0.69. Subsequently, the complex model of CYP17A1-BDBM50061174 was constructed by similarity-based docking employing the crystal structure of CYP17A1 bound to abiraterone, a steroidal prostate cancer drug (PDB-ID: 3ruk, chain D) (DeVore & Scott, 2012). The CYP17A1-abiraterone complex was selected as the best ligand-bound template based on the high TM-score of 0.84 and TC of 0.89, yielding the shortest PMD of 0.19. Figure 2.9 shows the validation of the modeled CYP17A1-BDBM50061174 by the experimental structure of CYP17A1-(R)-orteronel (PDB-ID: 5irq, chain B, 64% sequence identity to the target) released on 2017-03-15 (Petrunak, Rogers, Aube, & Scott, 2017). Kcombu reports a significant chemical alignment between steroidal BDBM50061174 and nonsteroidal (R)-orteronel with a TC of 0.54 (Figure 2.9A). Upon the superposition of CYP17A1 proteins, the RMSD between BDBM50061174 docked to the model and (R)-orteronel bound in the experimental structure calculated over the chemical alignment reported by kcombu is 2.95 Å (Figure 2.9B). These results verify that the computer-generated CYP17A1-BDBM50061174 model for the BindingDB reactant set ID 50103430 is correct.

Similarity-based docking procedure employed to construct ligand-bound structures in eModel-BDB superposes target ligands onto template molecules selected from the PDB according to the chemical alignment reported by kcombu. One may expect that superposing target compounds onto chemically similar template ligands yields more accurate binding poses than those generated from chemically less similar template molecules. Indeed, Figure 2.6B shows that the target-template chemical similarity measured with the TC correlates with the
resulting docking accuracy evaluated with the RMSD of ligand poses constructed based on target-template alignments. These results are in line with other studies reporting that the average RMSD values for similarity-based docking methods are generally below 2.0 Å when the target-template similarities are above 0.7 (Kawabata & Nakamura, 2014).

![Diagram](image.png)

Figure 2.8. Violin and box plots for the docking accuracy. The accuracy is assessed for remote homology complex models constructed by January 2017 against the experimental, ligand-bound structures of either target proteins or their close homologs deposited to the PDB after February 2017. The ligand RMSD is calculated over non-hydrogen atoms according to the chemical alignment reported by kcombu. The performance of similarity-based docking employed to construct eModel-BDB is compared to that of AutoDock Vina and rDock.

The performance of similarity-based docking employed to construct eModel-BDB is also compared to that of AutoDock Vina (Trott & Olson, 2010) and rDock (Ruiz-Carmona et al., 2014). In contrast to the median ligand RMSD of 2.6 Å for eModel-BDB complexes, the median RMSD values for BindingDB drugs docked to eFindSite pockets with AutoDock Vina and rDock are 6.7 Å and 7.2 Å, respectively (Figure 2.8). We note that similarity-based docking was
demonstrated to outperform traditional docking when the target-template similarity is larger than 0.4 (Kawabata & Nakamura, 2014), which was employed as the TC threshold to construct eModel-BDB complex models. Overall, the quality assessment as well as the independently obtained validation results demonstrate that the eModel-BDB database contains high-quality models closely resembling experimentally determined structures, not only at the global structure level, but also at the level of binding pockets and bound ligands.

Figure 2.9. Representative example of a complex structure constructed by similarity-based docking. (A) Chemical alignment between steroidal BDBM50061174 (left) and non-steroidal (R)-orteronel (right) reported by kcombu. 17 equivalent atom pairs constituting the maximum common substructure are numbered and outlined in purple in BDBM50061174 and in gold in (R)-orteronel. (B) The model of cytochrome P450 17A1 (CYP17A1, purple ribbons) is superposed onto the crystal structure of CYP17A1 from human (gold ribbons) complexed with (R)-orteronel (gold sticks). Cα atoms of binding residues identified in the CYP17A1 model by eFindSite are shown as purple spheres, whereas the target compound, BDBM50061174, docked into the predicted pocket is represented by purple sticks. Non-carbon atoms in BDBM50061174 and (R)-orteronel are colored by atom type (O – red, N – blue).

2.4 Conclusions

eModel-BDB is generated to support rational drug development projects. These data can directly aid structure-based drug discovery pipelines and protein function analysis by providing
atomic-level models of a large set of drug-protein interactions with known affinities. An important application of eModel-BDB is computational drug repositioning, i.e. finding new indications for existing drugs (Haupt & Schroeder, 2011). Although drug repurposing holds a significant promise to speed up drug development, particularly for diseases considered to be unprofitable, its major bottleneck is the scarce structural information on druggable pockets. On that account, a diverse dataset of drug-like small molecules bound to high-quality models with accurately annotated pockets provide an invaluable resource for drug repositioning employing sequence order-independent pocket matching algorithms (Brylinski, 2014, 2017; Konc & Janezic, 2010; Ren, Xie, Li, & Bourne, 2010). It is noteworthy that computational drug repurposing has recently revealed new opportunities to combat rare diseases (Brylinski, Naderi, Govindaraj, & Lemoine, 2018; Govindaraj, Naderi, Singha, Lemoine, & Brylinski, 2018).

Binding sites in eModel-BDB can also be matched to pockets predicted in potential drug targets in order to determine whether these proteins are druggable or not. If a new pocket aligns well with drug-bound pockets in eModel-BDB then it is likely going to be druggable. That being the case, our data can be utilized right at the outset of drug discovery, in the target identification phase. Finally, ligand binding can significantly help stabilize a protein, particularly from the point of view of the conformational stability (Deller, Kong, & Rupp, 2016). eModel-BDB can, therefore, inform crystallography efforts by suggesting possible compounds binding to certain protein targets at either the active or allosteric sites in order to increase the chances of successful crystallization.

Availability of supporting data and materials

Structure models in eModel-BDB are named according to the BindingDB reactant set IDs, which can be obtained by searching the BindingDB at https://www.bindingdb.org. This
procedure is illustrated in Figure 2.10. The BindingDB can be searched either by protein and compound names (Figure 2.10A) or by the target sequence (Figures 10B and 10C). Next, the complex of interest can be selected from the list of hits (Figure 2.10D) in order to download the corresponding SDfile of the complex (Figure 2.10E). The BindingDB reactant set ID, e.g. 00267770, is stored inside the SDfile (Figure 2.10F). The reactant set ID can then be used to find the detailed information on the BindingDB website, e.g. https://www.bindingdb.org/jsp/dbsearch/Summary_ki.jsp?reactant_set_id=00267770 (Figure 2.10G) as well as access the structure model in eModel-BDB, e.g. http://www.brylinski.org/pub/eModelBDB.php?reactant_set_id=00267770 (Figure 2.10H).
Figure 2.10. Procedure to obtain eModel-BDB complexes via the BindingDB. Target complex can be identified based on either the protein (blue arrows and boxes) or the ligand of interest (red arrows and boxes). Common actions that a user needs to perform are colored in green. (A) Specific ligands and proteins can directly be searched for on the BindingDB website. (B, C) Alternatively, target proteins can be found with the blast search. (D) A complex of interest can then be selected in order to (E) generate and download a SDfile file. (F) The BindingDB reactant set ID stored inside the SDfile is used to (G) view a web page containing detailed information about the target complex as well as (H) access the corresponding eModel-BDB structure model named according to the BindingDB reactant set ID.

2.5 References


32


3. STRUCTURE-BASED DRUG REPURPOSING

3.1 Introduction

Rare diseases are conditions afflicting a small subset of people in a population, where “small” is uniquely defined by each country. For example, the United States denotes disorders affecting fewer than 200,000 patients as rare diseases, also referred to as orphan diseases. Although each of approximately 7,000 orphan conditions has a tiny number of patients, they amount to 30 million patients in the U.S., 30 million in Europe, and around 350 million globally (Petersen et al., 2014). Because pharmaceutical companies seldom develop drugs for orphan diseases due to the lack of consumers, special attention needs to be placed on treating these conditions. After the success of the Orphan Drug Act signed into law in the U.S. by President Reagan in 1983, other governments adopted similar mechanisms to facilitate orphan drug development, mostly by granting market exclusivity and reducing research and development costs (Seoane-Vazquez, Rodriguez-Monguio, Szeinbach, & Visaria, 2008). These actions allow for not only sufficient financial incentives for pharmaceutical companies, but also manageable costs for non-profits. Fewer financial difficulties, various governmental inducements, increasing public awareness, together with advances in research techniques have stimulated a global interest in orphan drug development and rare disease research (Groft, 2013).

---

1 This chapter, previously published as:
Certainly, without the support of quality datasets and resources, the progress in orphan drug research might not be as consistent as it has been. For instance, Orphanet, the de facto rare disease reference source, contributes quality, robust data on rare diseases as well as reliable clinical practice guidelines (http://www.orpha.net). Most importantly, Orphanet enables researchers to share common language and information to undergo controlled scientific analysis and, ultimately, orphan drug discovery. Similar to Orphanet, the Genetic and Rare Diseases Information Center (GARD) at the National Institutes of Health provides comprehensive information regarding rare diseases and orphan drugs (Provost, 1968). Last but not least, the Developing Products for Rare Diseases & Conditions section of the U.S. Food and Drug Administration (FDA) website hosts freely accessible official legal documentation regarding orphan drug development and regulations (Reynolds & Garg, 1978). These rich resources on orphan diseases available to researchers worldwide facilitate the development of new treatments for rare conditions. For instance, a systems-level approach to find connections between existing drug products and orphan diseases, known as drug repositioning, holds a significant promise to greatly expand the repertoire of orphan drugs.

As an alternative strategy to drug discovery, compound repositioning finds new indications for existing drugs. This approach can dramatically improve the success rates by shortening the time of drug development to about 3-12 years at the reduced safety and pharmacokinetic uncertainty (Ashburn & Thor, 2004). Repurposing of already-approved drugs would most likely bypass initial clinical trials, especially if the corresponding dosage does not exceed the maximum approved by a regulatory agency. Although efficacy tests for the new treatment are still required, an existing drug is likely to have well characterized long-term toxicity and off-target effects. Further, the magnitude of side effects may be an important
determinant to repurpose a drug. For example, a drug with a high risk of significant side effects might not be appropriate when the primary goal is to maintain the quality of life of a patient, however, repurposing the same drug to treat a life-threatening disease may be acceptable.

Despite the fact that time- and cost-effective rational drug repositioning is expected to play a major role in the development of treatments for rare conditions (Sardana et al., 2011), it is not trivial and poses a number of onerous challenges. It is, therefore, not surprising that most of the repositioned drugs currently on the market are the result of serendipity. Perhaps the most recognizable example is sildenafil; originally intended to treat hypertension and angina pectoris in the 1980s, it was later repurposed to erectile dysfunction as well as pulmonary arterial hypertension (Boolell et al., 1996). Another instance is memantine (Witt, Macdonald, & Kirkpatrick, 2004), synthesized in the 1960s as a potential agent to treat diabetes, although it was found ineffective at lowering blood sugar. Its activity against the N-methyl-D-aspartate (NMDA) receptor was discovered in the 1980s and presently, memantine is used to treat Alzheimer's disease, vascular dementia and Parkinson's disease (Olivares et al., 2012). These examples show that even though drug repositioning is regarded as one of the most promising strategies for translational medicine, many new indications for existing drugs have been found serendipitously. Therefore, there is a clear need to establish rational, preferably computer-guided routines for drug repositioning.

This clear necessity for rational approaches to find alternative indications for existing therapeutics has stimulated the development of computational methods for drug repositioning (J. Li et al., 2016). Many currently available algorithms exploit the fact that proteins with similar pockets tend to have similar functions and recognize similar molecules (Ehrt, Brinkjost, & Koch, 2016). For instance, the sequence-order independent profile-profile alignment (SOIPPA)
program employs Delaunay tessellation of Cα atoms and geometric potentials to compare binding pockets (Xie & Bourne, 2008). Further, SiteAlign measures distances between druggable pockets with cavity fingerprints constructed by projecting eight topological and physicochemical properties onto a multidimensional, discretized space (Schalon, Surgand, Kellenberger, & Rognan, 2008). Both SOIPPA and SiteAlign have been used in drug repurposing, for example, SOIPPA helped reveal new targets for entacapone and tolcapone (Kinnings et al., 2009), whereas SiteAlign detected the cross-reaction of protein kinase inhibitors with a protein regulating neurotransmitter release in the synapse (Defranchi et al., 2010).

Notwithstanding the success of existing methods to recognize similar pockets, many of these algorithms perform well only against the experimental structures of proteins complexed with small molecules. Utilizing datasets of target structures with predicted binding sites poses a formidable challenge for pocket matching programs because of inevitable inaccuracies in the annotation of binding residues. To alleviate this issue, we recently developed eMatchSite, which offers a high tolerance to residue misannotations and, to some extent, structure imperfections in ligand-binding regions (Brylinski, 2014, 2017b). In this communication, we combine eMatchSite and structure-based virtual screening (VS) with AutoDock Vina (Trott & Olson, 2010) in order to enhance the accuracy of binding site matching.

In this chapter, eRepo-ORP, a new resource for orphan drug research. eRepo-ORP is a drug repositioning dataset that builds on the results of a large-scale pocket matching between target sites for known drugs and those binding pockets identified in proteins linked to rare diseases. Known drugs and their macromolecular targets are extracted from DrugBank, a unique bioinformatics and cheminformatics resource providing detailed chemical, pharmacological, and structural data on drug-target associations (Wishart et al., 2006), whereas proteins connected to
orphan diseases are obtained from Orphanet (http://www.orpha.net). Further, we designed a sophisticated protocol incorporating several state-of-the-art algorithms to find potential candidates for repositioning by modeling the high-quality structures of drug targets with eThread (Brylinski & Lingam, 2012), comprehensively annotating their binding sites with eFindSite (Brylinski & Feinstein, 2013; Feinstein & Brylinski, 2014), and effectively detecting similar drug-binding pockets with eMatchSite (Brylinski, 2014, 2017b). In general, this approach builds on ligand-binding homology, a technique previously employed in computer-aided drug development to detect binding sites (Brylinski & Skolnick, 2008) and to discover potential leads through virtual screening (Roy, Srinivasan, & Skolnick, 2015; Yang, Zhan, & Zhou, 2016). The results are discussed with respect to the structural data recently released in the Protein Data Bank (PDB) (Berman et al., 2002) for tyrosine-protein kinase HCK, Ras-associated autoimmune leukoproliferative disease as well as a possibility to repurpose a steroidal aromatase inhibitor to treat Niemann-Pick disease type C.

3.2 Methods

DrugBank dataset

FDA-approved drugs whose molecular weight is in the range of 150-550 Da and for which at least one target protein is known were selected from DrugBank (Wishart et al., 2006). Target structures composed of 50-999 amino acids were modeled with eThread, a template-based structure prediction algorithm (Brylinski & Lingam, 2012). eThread employs meta-threading with HH-suite (Remmert, Biegert, Hauser, & Soding, 2011), RaptorX (Ma, Wang, Zhao, & Xu, 2013), and SparksX (Yang, Faraggi, Zhao, & Zhou, 2011) to select structure templates in the non-redundant and representative subset of the PDB. Comparative structure modeling in eThread is carried out with Modeller (Webb & Sali, 2014) based on the top-ranked template and
incorporating secondary structure restraints from PSIPRED (Jones, 1999). Initial models assembled by Modeller were refined with ModRefiner (D. Xu & Zhang, 2011). Finally, each model was assigned an estimated GDT-score by ModelEvaluator (Cao & Cheng, 2016).

In the next step, drug-binding pockets were predicted by eFindSite (Brylinski & Feinstein, 2013) in confidently modeled target proteins whose estimated GDT-score is \( \geq 0.4 \). Pockets assigned by eFindSite a high and moderate confidence were then subjected to fingerprint-based virtual screening (Feinstein & Brylinski, 2014). Each target pocket was screened against a library containing drug molecules from DrugBank (Wishart et al., 2006) and a background collection of 244,659 non-redundant compounds selected from the ZINC database (Irwin & Shoichet, 2005). Only those drug-target pairs for which the drug molecule was ranked within the top 10% of the screening library were retained. Further, we devised a two-step alignment protocol to position drug compounds within the predicted binding pockets for each drug-target pair. First, holo-templates selected by eFindSite were structurally aligned onto the target protein with Fr-TM-align (Pandit & Skolnick, 2008) and then the drug molecule was superposed onto the template-bound ligand according to the chemical alignment constructed by kcombu (Kawabata, 2011).

Orphanet dataset

Genes associated with rare disorders were obtained from Orphanet (http://www.orpha.net) and the sequences of gene products were downloaded from UniProt (The UniProt, 2017). Subsequently, for those protein sequences composed of 50-999 amino acids, we employed a protocol described above for the DrugBank dataset to conduct comparative structure modeling with eThread (Brylinski & Lingam, 2012) followed by drug-binding pocket prediction by eFindSite (Brylinski & Feinstein, 2013). Finally, only protein structures with an estimated
GDT-score of ≥0.4 having binding sites predicted with a high and moderate confidence were retained.

Huang dataset

The Huang dataset was originally compiled to evaluate the performance of geometry-based methods to predict binding pockets (B. Huang & Schroeder, 2006) and then it was adopted to assess the accuracy of pocket comparison algorithms (Chikhi, Sael, & Kihara, 2010). From this dataset, we selected 107 proteins for which eFindSite correctly annotated binding sites within a distance of 8 Å from the geometric center of the bound ligand in the experimental complex structure. These target proteins bind the following ligands, adenosine, biotin, fructose-6-phosphate, α-L-fucose, β-D-galactose, guanine, α-D-mannose, O1-methyl-mannose, 4-phenyl-1H-imidazole, palmitic acid, retinol, and 2'-deoxyuridine 5'-monophosphate.

Pocket matching with eMatchSite

All-against-all matching of drug-binding pockets in DrugBank and Orphanet proteins was conducted with eMatchSite (Brylinski, 2014, 2017b). eMatchSite constructs sequence order-independent local alignments of pocket residues by solving the assignment problem with machine learning and the Hungarian algorithm (Kuhn, 1955). Subsequently, the local alignment is assigned a similarity score, called the eMS-score, calculated based on the overlap of various physicochemical features and evolutionary profiles. eMS-score ranges from 0 for completely dissimilar pockets to 1 for identical pockets, with an optimized threshold of 0.56 accurately distinguishing between pockets binding similar and dissimilar molecules (Brylinski, 2014). In addition to calculating the similarity score, eMatchSite superposes two pockets according to the constructed local alignments, so that a drug molecule bound to one pocket can be directly transferred to the other binding site. In this study, we use this feature of eMatchSite to transfer
drugs bound to DrugBank target to binding sites in Orphanet proteins. In the last step, the constructed complexes of drugs repositioned to Orphanet proteins are rebuilt with Modeller in order to refine drug-target interactions and eliminate steric clashes. The quality of the final complex models is assessed by a knowledge-based statistical energy function for protein-ligand complexes with the Distance-scaled Finite Ideal-gas REference (DFIRE) potential (C. Zhang, Liu, Zhu, & Zhou, 2005). Specific interactions between drugs and proteins, such as hydrogen bonds, hydrophobic and aromatic contacts, are identified by LPC (Sobolev, Sorokine, Prilusky, Abola, & Edelman, 1999), LigPlot+ (Laskowski & Swindells, 2011) and eAromatic (Brylinski, 2017a).

Virtual screening

A target binding site is subjected to VS with AutoDock Vina (Trott & Olson, 2010) against a non-redundant library of 1,515 FDA-approved drugs compiled previously (Govindaraj & Brylinski, submitted). MGL tools (Morris et al., 2009) and Open Babel (O'Boyle et al., 2011) were used to add polar hydrogens and partial charges, as well as to convert target proteins and library compounds to the PDBQT format. For each docking ligand, the optimal search space centered on the binding site annotated with eFindSite was calculated from its radius of gyration (Feinstein & Brylinski, 2015). Molecular docking was carried out with AutoDock Vina 1.1.2 and the default set of parameters.

3.3 Results And Discussion

Template-based drug repositioning

eRepo-ORP is constructed based on a large-scale drug repositioning conducted with accurate, template-based techniques according to a protocol presented in Figure 3.1. The first phase is to generate structural data for FDA-approved drugs and their molecular targets based on
information extracted from the DrugBank database (Figure 3.1A). Structure models of drug targets are constructed by eThread and annotated with drug-binding sites and residues by eFindSite (Figure 3.1B). Next, for each drug-target pair, we identify in the Protein Data Bank (PDB) (Berman et al., 2002) a globally similar template binding a ligand that is chemically similar to the DrugBank compound (Figure 3.1C). This holo-template is structurally superposed onto the DrugBank target (Figure 3.1D) and then the DrugBank compound is aligned onto the template-bound ligand (Figure 3.1E). This procedure produces 2,012 atomic models of drug-target complexes involving 348 unique proteins and 715 drugs (Figure 3.1F). The second phase is to model proteins associated with orphan diseases obtained from the Orphanet database (Figure 3.1G). Structure models of 922 Orphanet proteins with predicted drug-binding sites and residues (Figure 3.1H) are generated by a similar protocol to that used for DrugBank targets.

All-against-all pockets matching conducted with eMatchSite for DrugBank and Orphanet proteins produced 320,856 binding site alignments (Figure 3.1I), 5.6% of which yield a statistically significant eMS-score. It is noteworthy that the average TM-score between matched DrugBank and Orphanet targets is as low as 0.27 ±0.10 indicating that in the majority of cases, existing drugs are repositioned from proteins having globally unrelated structures. Based on 18,145 confident local alignments reported by eMatchSite, 31,142 unique putative complexes between DrugBank compounds and Orphanet proteins have been modeled. These complex models are subjected to all-atom refinement (Figure 3.1J). Refined structure models are included in the eRepo-ORP database. An analysis of the DrugBank→Orphanet repositioning data reveals that 381 existing drugs could be repurposed to target as many as 761 Orphanet proteins. These proteins link to 980 orphan diseases representing 32 classes including (ten the most common classes) 923 genetic, 428 neurological, 377 inborn errors of metabolism, 266 developmental
anomalies during embryogenesis, 170 eye, 117 skin, 102 bone, 93 neoplastic, 92 endocrine, and 85 hematological disorders.

Figure 3.1. Flowchart of the drug repositioning procedure employed to construct eRepo-ORP. This protocol utilizes data from three sources, DrugBank, Protein Data Bank (PDB), and Orphanet, shown in blue, red, and green, respectively. Databases are indicated by gray boxes. (A) For a given protein sequence from DrugBank, template-based structure modeling is conducted with eThread in order to construct (B) a 3D model subsequently annotated by eFindSite with drug-binding sites and residues represented by little circles. (C) A globally similar template binding a ligand that is chemically similar to the DrugBank compound is selected from the PDB. (D) The template carrying its ligand is structurally aligned onto the DrugBank apo-structure. (E) The DrugBank compound is then aligned onto the template-bound ligand generating (F) a 3D model of the drug-target complex. (G) For a given protein sequence from Orphanet, (H) a 3D model is constructed with eThread and annotated with eFindSite. (I) A local alignment is performed for a pair of binding sites in DrugBank and Orphanet models with eMatchSite. (J) The DrugBank compound is transferred to the Orphanet model when the similarity of binding pockets in DrugBank and Orphanet models is sufficiently high and the resulting complex is refined.

Quality of structural data generated for DrugBank and Orphanet

Structure models are generated for the DrugBank and Orphanet datasets with eThread, a meta-threading approach employing state-of-the-art fold recognition. Initial models constructed by Modeller from eThread alignments are refined with ModRefiner, which performs atomic-level energy minimization in a composite physics- and knowledge-based force field improving side-chain positions and hydrogen-bonding networks. An independent assessment of the quality
of protein models is carried out with ModelEvaluator utilizing the predicted secondary structure, relative solvent accessibility, residue contact map, and beta sheet structure. Statistics reported in Table 3.1 show that the template-based modeling protocol employed in this study produces highly confident structure models, whose mean estimated Global Distance Test (GDT)-score (Zemla, Venclovas, Moult, & Fidelis, 1999) values are 0.71 and 0.68 for DrugBank and Orphanet proteins, respectively. In addition, the mean confidence for the top-ranked binding sites predicted in these models by eFindSite is as high as 0.87 for DrugBank and 0.82 for Orphanet targets.

The structure models of DrugBank complexes are constructed by aligning the protein and the drug onto a holo-template selected from the PDB. Table 3.1 reports the mean Tanimoto coefficient (TC) (Tanimoto, 1958) between the DrugBank compound and the template-bound ligand of 0.49 and the mean Template Modeling (TM)-score (Y. Zhang & Skolnick, 2004) between receptor proteins of 0.65. Note that both TC and TM-score are even higher when only those cases producing statistically significant pocket alignments are considered. These numbers clearly indicate that globally similar templates binding chemically similar ligands are selected for the majority of drug-protein pairs from DrugBank to produce highly confident complex models. Table 3.1 also provides statistics for DrugBank→Orphanet pairs. Both TM-score and eMS-score values are very low for all data, basically showing that randomly selected pairs of proteins share neither global nor local structure similarity. However, considering the subset of 18,145 pairs producing statistically significant local alignments, the mean eMS-score is as high as 0.91, even though the mean TM-score is still only 0.27. These results demonstrate that the vast majority of similar binding sites included in eRepo-ORP are identified by eMatchSite in DrugBank and Orphanet proteins having unrelated global structures.
Table 3.1 Statistics on the confidence and quality of models included in eRepo-ORP. Columns labeled “All data” report mean values ± standard deviation and the number of cases for all DrugBank and Orphanet structures as well as PDB→DrugBank and DrugBank→Orphanet alignments, whereas “Pocket matches” correspond to those pairs of DrugBank and Orphanet proteins producing statistically significant pocket alignments.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Score(^e)</th>
<th>All data</th>
<th>Pocket matches</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean ± std.</td>
<td># of cases</td>
<td>mean ± std.</td>
</tr>
<tr>
<td>DrugBank(^a)</td>
<td>GDT-score</td>
<td>0.71 ±0.17</td>
<td>348</td>
</tr>
<tr>
<td></td>
<td>Pocket conf.</td>
<td>0.87 ±0.18</td>
<td></td>
</tr>
<tr>
<td>Orphanet(^b)</td>
<td>GDT-score</td>
<td>0.68 ±0.15</td>
<td>922</td>
</tr>
<tr>
<td></td>
<td>Pocket conf.</td>
<td>0.82 ±0.22</td>
<td></td>
</tr>
<tr>
<td>PDB→DrugBank(^c)</td>
<td>TM-score</td>
<td>0.65 ±0.14</td>
<td>2,012</td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>0.49 ±0.27</td>
<td></td>
</tr>
<tr>
<td>DrugBank→Orphanet(^d)</td>
<td>TM-score</td>
<td>0.26 ±0.08</td>
<td>320,714</td>
</tr>
<tr>
<td></td>
<td>eMS-score</td>
<td>0.13 ±0.20</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Structure models generated for DrugBank proteins by eThread with binding sites predicted by eFindSite. \(^b\) Structure models generated for Orphanet proteins by eThread with binding sites predicted by eFindSite. \(^c\) Alignment between DrugBank structures and holo-templates selected from the PDB to position drugs within binding sites. \(^d\) Pairs of DrugBank and Orphanet proteins. \(^e\) GDT-score is estimated by ModelEvaluator, pocket confidence is estimated by eFindSite for the top-ranked drug-binding sites, TM-score measures the global structure similarity, TC is the Tanimoto coefficient reported by kcombu, eMS-score is the pocket similarity reported by eMatchSite.

Repositioning multiple drugs through a single alignment

Drug repositioning conducted in this study includes two kinds of special cases. Figure 3.2 illustrates the first situation, in which complexes between multiple drugs (Figure 3.2A) and an Orphanet target (Figure 3.2B) are modeled based on a single pocket alignment. Employing this approach generates a series of structure models of drugs transferred from a DrugBank target to the binding site of an Orphanet protein (Figure 3.2C). For instance, catechol O-methyltransferase (COMT) produces a significant local alignment with guanine nucleotide-binding protein subunit
alpha-11 (GNA11), associated with a rare disease, autosomal dominant hypocalcemia (ADH) or hypoparathyroidism (Pollak et al., 1994) (ORPHA:428, GARD:2877). This condition is characterized by low levels of calcium in the blood and an imbalance of other molecules, such as phosphate and magnesium, leading to a variety of symptoms, although about half of affected individuals have no associated health problems (Kinoshita, Hori, Taguchi, Watanabe, & Fukumoto, 2014). ADH is primarily caused by mutations of a gene encoding the calcium-sensing receptor, however, activating mutations in GNA11 have also been reported (Nesbit et al., 2013; Roszko, Bi, & Mannstadt, 2016).

Figure 3.2. Multiple drugs repositioned through a single pocket alignment. Schematic of (A) a single DrugBank target binding two drugs (teal and yellow), (B) an Orphanet target (green), and (C) two modeled complexes of DrugBank drugs and an Orphanet protein (teal-green and yellow-green). (D) A real example of tolcapone (teal) and entacapone (yellow) repositioned to guanine nucleotide-binding protein subunit alpha-11 (green) based on its local alignment with catechol O-methyltransferase. Non-carbon ligand atoms in panel D are colored by atom type (blue – nitrogen, red – oxygen).

A binding site predicted in GNA11 by eFindSite aligns well to a pocket binding tolcapone and entacapone in COMT with an eMS-score of 0.97 and a Cα-RMSD of 4.5 Å calculated over 14 aligned binding residues. Based on this single alignment, tolcapone and entacapone, COMT inhibitors used as adjuncts to levodopa/carbidopa medication in the
treatment of Parkinson's disease (Guay, 1999; Najib, 2001), could be repositioned to GNA11. Figure 3.2D shows the putative binding poses of both compounds in the binding pocket of GNA11 modeled based on the local COMT→GNA11 alignment reported by eMatchSite. Interaction energies with GNA11 reported by DFIRE for tolcapone and entacapone are -355.7 and -311.7, respectively. For comparison, the interaction energies with COMT are -283.7 for tolcapone and -310.9 for entacapone. Overall, these results indicate that both molecules may favorably bind to GNA11 producing stable, low-energy assemblies.

Construction of multiple models of a single complex

The second special case is the modeling of a single complex based on multiple pocket alignments. More than one structure model of a drug repositioned to the Orphanet protein can be constructed if this drug has multiple targets in DrugBank producing significant pocket alignments with the Orphanet protein. This procedure is illustrated in Figure 3.4. Figure 3.4A shows three DrugBank targets binding the same compound, colored blue, orange and yellow. Assuming that pockets for this drug in all three proteins align to a binding site in an Orphanet target colored green (Figure 3.4B), three independent structure models can be constructed (Figure 3.4C). An example is ponatinib, a novel inhibitor of Bcr-Abl tyrosine kinase developed to treat chronic myeloid leukemia and Philadelphia chromosome-positive acute lymphoblastic leukemia (W. S. Huang et al., 2010). Ponatinib is a multi-targeted compound, which in addition to its primary target, Abelson tyrosine-protein kinase 1, binds to 14 other macromolecules according to DrugBank (Wishart et al., 2006). Binding sites of three of these proteins, Lck/Yes-related novel protein tyrosine kinase (LYN), lymphocyte cell-specific protein-tyrosine kinase (LSK), and proto-oncogene tyrosine-protein kinase Src (SRC), produce significant local alignments with a drug-binding pocket predicted in Ras-related protein Rab-23 (RAB23). The
corresponding eMS-score/Cα-RMSD values reported by eMatchSite for these alignments are 0.97/3.8 Å, 0.98/3.7 Å, and 0.98/3.8 Å, respectively. According to Orphanet, RAB23 is associated with Carpenter syndrome (Ben-Salem, Begum, Ali, & Al-Gazali, 2013; Haye et al., 2014) (ORPHA:65759, GARD:6003), a very rare disease with approximately 40 cases described in the literature (Robinson, James, Mubarak, Allen, & Jones, 1985). The repositioning of ponatinib to RAB23 can, therefore, be carried out through kinases LSK, LYN, and SRC, resulting in three independent models of a ponatinib-RAB23 complex structure. Figure 3.4D shows that the binding poses of ponatinib in the RAB23 pocket are very similar across these models. The heavy-atom RMSD between ponatinib molecules is 2.1 Å for LSK- and LYN-based models, 0.7 Å for LSK- and SRC-based models, and 2.3 Å for LYN- and SRC-based models, with similar drug-protein interactions present in all models. The interaction energy between ponatinib and RAB23 reported by DFIRE for LSK-, LYN-, and SRC-based models are -829.7, -727.7, and -723.6, respectively. These values are even lower than those calculated for the parent complexes of ponatinib and LSK (-587.9), LYN (-571.9), and SRC (-586.9) suggesting that ponatinib may form favorable interactions with the binding residues of RAB23.

Multiple structure models of the same complex of a drug repositioned to a Orphanet target can be used to estimate the confidence of the large-scale modeling reported in this study. Specifically, employing different DrugBank proteins to transfer the same drug to the Orphanet target should, in principle, produce similar complex models. To test this assumption, we selected 4,878 drugs repositioned to Orphan targets by matching binding sites of multiple DrugBank proteins. Figure 3.3 shows that up to 20 different models can be constructed for some drugs, however, two and three models are generated for the majority of cases (52.4% and 21.9%, respectively).
Figure 3.3. Histogram of the number of structure models generated for a subset of 4,878 drug-Orphanet complexes. Multiple structure models of the same complex are constructed using pocket alignments between the Orphanet target and different DrugBank proteins. Inset: Histogram of RMSD values calculated for different models of the same drug-target complex. RMSD is the root-mean-square deviation computed over ligand heavy atoms.

Next, we identified the most typical binding pose of each drug in the pocket of an Orphanet protein by calculating a ligand heavy-atom RMSD against all other models of the same drug-target complex. The distribution of these RMSD values across 4,878 DrugBank drugs repositioned to Orphanet targets is shown as inset in Figure 3.3. Encouragingly, the RMSD for most compounds is relatively low with a median value of 3.6 Å. One should keep in mind that these complex structures are constructed from the computer-generated models of target proteins with computationally predicted ligand-binding sites, and drug molecules are transferred according to fully sequence order-independent pocket alignments.
Figure 3.4. Multiple models of a single drug-target complex constructed based on multiple pocket alignments. Schematic of (A) three DrugBank targets binding the same drug (teal, orange, and yellow), (B) an Orphanet target (green), and (C) three poses of a DrugBank drug within the binding site of an Orphanet protein (teal/orange/yellow-green) modeled from different pocket alignments. (D) A real example of ponatinib repositioned to Ras-related protein Rab-23 (green) based on its local alignment with Lck/Yes-related novel protein tyrosine kinase (ponatinib is teal), lymphocyte cell-specific protein-tyrosine kinase (ponatinib is orange), and proto-oncogene tyrosine-protein kinase Src (ponatinib is yellow). Non-carbon ligand atoms in panel D are colored by atom type (blue – nitrogen, red – oxygen).

Binding affinity prediction for repositioned drugs

We also evaluate the binding affinity of drugs repositioned to Orphanet proteins in comparison with their complexes with primary targets from DrugBank. Figure 3.5 shows the relation between interaction energies estimated by DFIRE for DrugBank and Orphanet complexes. Because a single drug-target complex from DrugBank can be used to reposition the bound drug molecule to multiple Orphanet proteins, mean scores and the corresponding standard errors of the mean are plotted on the y-axis. Encouragingly, DFIRE energies calculated for DrugBank and Orphanet complexes involving the same drug are highly correlated with a Pearson correlation coefficient of 0.86.
Figure 3.5. Correlation between interaction energies calculated for DrugBank and Orphanet complex models. Each gray dot represents a drug-target pair from the DrugBank database, whose DFIRE score is displayed on the x-axis. Since a drug can be repositioned to multiple Orphanet proteins, the mean DFIRE score ± standard error is displayed on the y-axis. Linear regression is shown as a solid line.

This analysis indicates that the interaction strength of drug molecules repositioned to Orphanet proteins is generally comparable to that calculated for their complexes with primary targets. Therefore, those pairs of DrugBank and Orphanet proteins producing statistically significant pocket alignments also share similarities with respect to ligand binding as independently evaluated with knowledge-based statistical potentials.
Validation against recently determined X-ray structures

Repositioning prediction by eMatchSite is further validated against a complex structure released in the PDB several months after the modeling was completed. Three case studies are provided to show examples of cross validation against crystal structure models.

Ras-associated autoimmune leukoproliferative disease and vandetanib

Ras-associated autoimmune leukoproliferative disorder (RALD, ORPHA:268114) is a chronic, non-malignant condition characterized by monocytosis and often associated with leukocytosis, lymphoproliferation, and autoimmune phenomena (Calvo et al., 2015). RALD is linked to certain mutations in GTPase KRas (KRAS), which plays an important role in the regulation of cell proliferation promoting oncogenic events, thus it is considered a major target in anticancer drug discovery (Zimmermann et al., 2013). Specifically, amino acid substitutions in codons 12 and 13 of KRAS in RALD patients cause the constitutive binding of GTP and the activation of the KRAS protein inducing the Raf-MEK-ERK signaling pathway (Calvo et al., 2015). According to eRepo-ORP, KRAS produces a highly significant local alignment with protein-tyrosine kinase 6 (PTK6) implicated in the regulation of a variety of signaling pathways that control the differentiation and maintenance of normal epithelia, as well as tumor growth (S. H. Park, Lee, Kim, & Lee, 1997). PTK6 is a target for vandetanib, an oral kinase inhibitor of tumor angiogenesis and tumor cell proliferation approved by the FDA to treat non-resectable, locally advanced or metastatic medullary thyroid cancer in adult patients (Wedge et al., 2002).

Structure models of PTK6 (purple) and KRAS (gold) are shown in Figure 3.6. The model of PTK6 constructed with eThread from tyrosine-protein kinase HCK (PDB-ID: 1qcf, chain A, 42.6% sequence identity) (Schindler et al., 1999) is assigned a high estimated GDT-score of 0.74. Further, vandetanib (DrugBank-ID: DB05294) was transferred to PTK6 according to the
global structure alignment with cyclin-dependent kinase 6 bound to this inhibitor (PDB-ID: 2ivu, chain A, TM-score of 0.54) (Knowles et al., 2006).

Figure 3.6. Repositioning of vandetanib from protein-tyrosine kinase 6 (PTK6) to GTPase KRas (KRAS) according to eRepo-ORP. PTK6 and KRAS proteins are colored purple and gold, respectively, whereas ligands are colored by atom type (green/teal – carbon, blue – nitrogen, red – oxygen, yellow – sulfur, citron – chlorine, pink – fluorine, cyan – bromine). (A) Structure model of the complex between PTK6 (purple ribbons) and vandetanib (thick sticks) with predicted binding residues shown as spheres superposed onto the experimental structure of PTK6 (teal ribbons) bound to dasatinib (thin sticks). (B) Structure model of KRAS (gold ribbons) with predicted drug-binding residues shown as spheres superposed onto the experimental structure of KRAS (teal ribbons) bound to ADP (thin sticks). (C) Local superposition of PTK6 (purple ribbons) and KRAS (gold surface) according to the sequence order-independent pocket alignment by eMatchSite. Annotated binding residues in KRAS are solid, whereas the remaining surface is transparent. Vandetanib repositioned to KRAS is represented by thick sticks.

The final model of the vandetanib-PTK6 complex is shown in Figure 3.6A as solid ribbons and sticks. We selected this particular case because the vandetanib-PTK6 model was generated using the October 2016 version of the PDB and, in January 2017, a crystal structure of PTK6 kinase domain complexed with another inhibitor, dasatinib, was released (PDB-ID: 5h2u, chain A) (Thakur et al., 2017). This experimental structure superposed onto the vandetanib-PTK6 model is shown in Figure 3.6A as transparent ribbons and sticks. A TM-score between the PTK6 model and the experimental structure is as high as 0.92 with a Cα-RMSD of 2.3 Å. Further, the root-mean-square deviation (RMSD) calculated over dasatinib-binding residues is
only 0.7 Å demonstrating that not only the backbone, but also the binding pocket is modeled with a very high accuracy. Although vandetanib and dasatinib have a low chemical similarity with a TC of only 0.15, both inhibitors have a similar shape and the modeled binding pose of vandetanib resembles the experimental conformation of dasatinib. Moreover, the top-ranked binding site predicted with 99.7% confidence by eFindSite in the PTK6 model substantially overlaps with the dasatinib-binding pocket in the experimental complex structure. The Matthews correlation coefficient (MCC) (Matthews, 1975) between predicted and dasatinib-binding residues reported by Ligand-Protein Contacts (LPC) software is 0.62.

The model of KRAS was constructed from Ras-related protein Rap-1b (PDB-ID: 4m8n, chain G, 58.4% sequence identity) and assigned a high estimated GDT-score of 0.85. Although several inhibitors of KRAS are available, these compounds target the secondary binding site (Sun et al., 2014). In Figure 3.6B, a GDP-bound KRAS (transparent) is superposed onto the model structure (solid). This superposition yields a high TM-score of 0.93 and a low Cα-RMSD of 1.4 Å; furthermore, the RMSD calculated over GDP-binding residues is only 1.1 Å. The top-ranked drug-binding site comprising 27 residues, annotated by eFindSite with 95.7% confidence, has an MCC against GDP-binding residues of 0.61. Despite a very low global sequence identity of 12.9% and a structure similarity with a TM-score of 0.32 between PTK6 and KRAS, eMatchSite reports a significant local similarity of their binding sites with an eMS-score of 0.99. Figure 3.6C shows the conformation of vandetanib repositioned from PTK6 to KRAS according to the sequence order-independent pocket alignment by eMatchSite, which results in 4.3 Å Cα-RMSD over 25 aligned residues. Repositioned vandetanib fits well into a deep cavity in the KRAS structure forming hydrogen bonds with A18, N116 and K117, aromatic interactions with F28, and hydrophobic contacts with V8 and V9. The interaction energy between vandetanib and
KRAS calculated by DFIRE is -441.5, which is only slightly higher than -485.6 obtained for the vandetanib-PTK6 model. Altogether, these results suggest that the nucleotide-binding pocket of KRAS may be a suitable target for vandetanib. If so, we anticipate that the competitive binding of vandetanib to KRAS may subdue its gain-of-function caused by activating mutations, leading to the mitigation of RALD conditions.

Maturity-onset diabetes of the young and ibrutinib

Figure 3.7 shows ibrutinib (DrugBank-ID: DB09053), an anti-cancer drug primarily targeting B-cell malignancies (Gayko et al., 2015), predicted to bind to proto-oncogene, Src family tyrosine kinase Blk (UniProt-ID: P51451). According to Orphanet, Blk is linked to maturity-onset diabetes of the young (MODY, ORPHA:552, GARD:3697) (Reynolds & Garg, 1978) caused by mutations in at least 13 genes, 5 of which are placed within 100 kb corresponding to the Blk gene (Borowiec et al., 2009). Nonetheless, a reassessment study showed that Blk mutations, A71T in particular, unlikely cause highly penetrant MODY and may weakly influence type 2 diabetes risk in the context of obesity (Bonnefond et al., 2013). More recently, it was discovered that malignant T cells in the majority of patients with the cutaneous T-cell lymphoma (CTCL) display the ectopic expression of Blk (Imam, Shenoy, Flowers, Phillips, & Lechowicz, 2013). Since Blk functions as an oncogene promoting the proliferation of malignant T cells, it is a potential therapeutic target in CTCL (Petersen et al., 2014).

Although the full-length experimental structure of Blk is unavailable, a confident model of Blk, whose estimated Global Distance Test (GDT)-score (Zemla et al., 1999) is 0.72, was constructed by eThread based on proto-oncogene tyrosine-protein kinase Src (PDB-ID: 1y57, chain A, 64% sequence identity to Blk) (Cowan-Jacob et al., 2005). Further, the binding site annotated in the Blk model by eFindSite with a 99.2% confidence was matched to the ibrutinib-
binding pocket in tyrosine-protein kinase BTK with a high eMS-score of 0.99. In October 2017, tyrosine-protein kinase HCK co-crystallized with a 7-substituted pyrrolo-pyrimidine inhibitor, OOS (PDB-ID: 2bzg, chain A), sharing 69.7% sequence identity with Blk, was released in the PDB (Yuki et al., 2017). Figure 3.7A shows that ibrutinib and OOS have very similar chemical structures with a Tanimoto coefficient (Tanimoto, 1958) (TC) of 0.61 and 27 common atoms.

Figure 3.7. Example of a recently determined structure corroborating repositioning prediction by eMatchSite. Ibrutinib repositioned from tyrosine-protein kinase BTK to tyrosine-protein kinase Blk is compared to the X-ray structure of tyrosine-protein kinase HCK complexed with ligand OOS. (A) Chemical structures of the repositioned drug, ibrutinib, and the co-crystallized ligand, OOS. (B) The modeled structure of the ibrutinib-Blk complex, colored in purple, is globally superposed onto the experimental OOS-HCK structure, colored in gold. Proteins are shown as ribbons, ligands as sticks, and binding residues predicted by eFindSite in Orphanet models as spheres. Non-carbon atoms in ligands are colored by atom type (red – oxygen, blue – nitrogen, yellow – sulfur).

The global superposition of the modeled ibrutinib-Blk and experimental OOS-HCK structures is presented in Figure 3.7B. The Blk model (purple ribbons) has a globally correct
structure with a TM-score of 0.86 and a Cα-RMSD of 2.25 Å calculated against HCK (gold ribbons) over the kinase domain. Further, binding residues were accurately predicted by eFindSite in the Blk model (purple spheres) with a MCC of 0.60 against OOS-binding residues in the HCK complex structure. Encouragingly, the binding pose of ibrutinib repositioned to Blk based on the local BTK→Blk alignment closely resembles the conformation of OOS in HCK. The RMSD calculated over equivalent non-hydrogen atoms of these compounds is 2.57 Å and 1.43 Å upon the superposition of target proteins and ligands, respectively. Despite the fact that matching binding sites in a sequence-order independent manner is a challenging task, the modeled ibrutinib-Blk complex is noticeably similar to the experimental OOS-HCK structure recently released in the PDB.

Niemann-Pick disease, type C and exemestane

Niemann-Pick disease, type C (NPC, ORPHA:646) is a fatal hereditary disorder characterized by the accumulation of low-density, lipoprotein-derived cholesterol in lysosomes causing hepatosplenomegaly and severe progressive neurological dysfunction. Mutations in either of two lysosomal proteins, Niemann-Pick disease types C1 (NPC1) or C2 (NPC2), interrupt sterol transport from late endosomes and lysosomes to other cellular organelles resulting in cholesterol accumulation in lysosomes and the fatal NPC disease (Pentchev, 2004). As many as 22 mutations in NPC2 are associated with orphan NPC diseases, including adult, juvenile, late infantile, and severe early infantile neurologic onset. In particular, V30M, V39M, C47F, S67P, C93F, C99R, and P120S mutations in NPC2 have an effect on cholesterol binding (Chikh, Rodriguez, Vey, Vanier, & Millat, 2005; Klunemann et al., 2002; Millat et al., 2005; Millat et al., 2001; W. D. Park et al., 2003). Furthermore, mutations of M79, V81, and V83 block sterol transport making NPC2 a promising drug target to treat NPC diseases (Li, Saha, Li,
Blobel, & Pfeffer, 2016). Interestingly, eMatchSite detected a significant structure similarity between the cholesterol-binding pocket of NPC2 and the steroid-binding pocket of cytochrome P450 aromatase (CYP19A1), an enzyme involved in the biosynthesis of aromatic C18 estrogen from C19 androgen. CYP19A1 is a target for exemestane, an oral steroidal aromatase inhibitor approved by the FDA for the treatment of breast cancer in postmenopausal patients (Buzdar, Robertson, Eiermann, & Nabholtz, 2002).

The full-length model of CYP19A1 was generated by eThread from the crystal structure of an N-terminal-truncated recombinant human CYP19A1 (PDB-ID: 4kq8, chain A, 100.0% sequence identity with a coverage of 89.9%) (Lo et al., 2013). Subsequently, exemestane was placed in the steroid-binding pocket of CYP19A1 based on its global structure alignment with the X-ray structure of human placental CYP19A1 (PDB-ID: 3s79, chain A, TM-score of 0.89 and Cα-RMSD 0.55 Å) bound to androstenedione (Ghosh et al., 2012), another steroidal inhibitor with a TC to exemestane of 0.95. Although the experimental structure of CYP19A1 bound to exemestane is available (PDB-ID: 3s7s, chain A) (Ghosh et al., 2012), it is not included in the template library used to model DrugBank complexes. By reason of removing the redundancy in the library at 80% protein sequence identity (Brylinski & Lingam, 2012) and a TC of 0.9 for the ligand chemical similarity (Brylinski & Feinstein, 2013), androstenedione-bound CYP19A1 was identified as a cluster centroid to represent the entire group of similar complexes, including the exemestane-CYP19A1 structure.

We selected this case to demonstrate that a non-redundant library is adequate to build complex models fairly indistinguishable from experimental structures. The exemestane-CYP19A1 model constructed in this study is shown in Figure 3.8A as thick sticks colored by atom type representing exemestane and purple ribbons representing CYP19A1. Two other
structures are globally aligned onto the exemestane-CYP19A1 model, the androstenedione-CYP19A1 complex used as the template to position exemestane within the steroid-binding pocket and the experimentally determined exemestane-CYP19A1 complex; both structures are presented in Figure 3.8A as thin sticks colored by atom type and teal ribbons. Indeed, the Cα-RMSD as well as the RMSD calculated over binding residues between CYP19A1 model and experimental structure are below 1 Å. Further, RMSD calculated for exemestane upon the global structure superposition is as low as 0.06 Å demonstrating that the exemestane-CYP19A1 assembly is modeled with a very high accuracy. It is also noteworthy that eFindSite identified the binding site for exemestane with 96.2% confidence and the predicted binding residues, shown as purple spheres in Figure 3.8A, yield an MCC of 0.71 against exemestane-binding residues in the CYP19A1 model.

The full-length model of lysosomal protein NPC2 was constructed based on the crystal structure of the human NPC2 (PDB-ID: 5kwy, chain C, 100.0% sequence identity with a coverage of 87.4%) (X. Li et al., 2016). Figure 3.8B shows the global superposition of the NPC2 model represented by gold ribbons and two experimental NPC2 structures represented by teal ribbons, human (the template, 5kwyC) and bovine (PDB-ID: 2hka, chain A, 79% sequence identity to human NPC2) (S. Xu, Benoff, Liou, Lobel, & Stock, 2007), both complexed with cholesterol sulfate. These superpositions yield a Cα-RMSD of 0.92 Å against human and 1.06 Å against bovine structures. NPC2 has an Ig-like β-sandwich fold comprising seven β-strands forming a hydrophobic pocket that was suggested to become wider in order to accommodate cholesterol-like molecule (Friedland, Liou, Lobel, & Stock, 2003). This region was accurately identified by eFindSite with a high confidence of 95.2% as a highly hydrophobic binding site formed by 20 conserved residues. The prediction was made based on a non-redundant set of 21
holo-templates, including ganglioside GM2 activator (GM2A), lymphocyte antigen 96 (LY96), mite group 2 allergen Der f 2 (DERF2), and NPC2 itself. Selected template-bound ligands are shown in Figure 3.8B as a cluster of transparent, teal molecules upon the global alignment of template proteins onto the NPC2 model. In addition, eFindSite estimated that the average ± standard deviation molecular weight (MW), octanol-water partition coefficient (logP), and polar surface area (PSA) for molecules binding to this region on the NPC2 surface are 383 Da ± 225, 4.76 ± 1.97, and 90.2 Å² ± 77.9, respectively. The predicted physicochemical properties of putative binders of NPC2 are a good match for exemestane (and androstenedione), whose MW is 296 Da (286 Da), logP is 4.03 (4.09), and PSA is 34.1 Å² (34.1 Å²).

Although the global similarity between CYP19A1 and NPC2 is low as assessed by a TM-score of 0.14 and 5.2% sequence identity, eMatchSite predicted that their binding sites are in fact similar with a high eMS-score of 0.86. Figure 3.8C shows exemestane repositioned from CYP19A1 to the cholesterol-binding pocket of NPC2 based on the sequence order-independent local alignment reported by eMatchSite. Exemestane fits into a deep, non-polar cavity in the NPC2 structure forming a number of hydrophobic interactions with Y55, V57, V73, V74, F85, P88, Y109, N111, L113, V126, W128, and W141. Encouragingly, an interaction energy of -409.5 calculated with DFIRE for the exemestane-NPC2 complex is lower than a value of -381.4 for exemestane-CYP19A1 indicating that this drug may form favorable interactions with NPC2. Notably, exemestane adopts a conformation distinct from that of cholesterol sulfate in the crystal structure of NPC2. The latter is larger (MW of 466 Da) and has two moieties attached to the steroid scaffold, an aliphatic branched-chain interacting with the inner part of the NPC2 pocket and a polar sulfate group protruding from the pocket toward the cholesterol-transfer tunnel between NPC2 and the N-terminal domain of NPC1 (X. Li et al., 2016). In contrast, smaller
Exemestane may bind deeper in the NPC2 structure to inhibit conformational changes required for transporting cholesterol to NPC1.

Figure 3.8. Repositioning of exemestane from cytochrome P450 aromatase (CYP19A1) to Niemann-Pick disease type C2 protein (NPC2). CYP19A1 and NPC2 proteins are colored purple and gold, respectively, whereas ligands are colored by atom type (green/teal – carbon, red – oxygen, yellow – sulfur). (A) Global superposition of the modeled complex between CYP19A1 (purple ribbons) and exemestane (thick sticks), and two experimental structures of CYP19A1 (teal ribbons) bound to androstenedione and exemestane (thin sticks). Binding residues are shown as spheres. (B) Global superposition of the NPC2 model (gold ribbons) and two experimental structures of NPC2, human and bovine (teal ribbons), bound to cholesterol sulfate (thin sticks). Binding residues are shown as spheres. In addition, the steroid-binding pocket predicted by eFindSite is represented by a cluster of template-bound ligands (transparent sticks) extracted from the following template proteins superposed onto the NPC2 model (template-proteins are not shown): GM2A (PDB-IDs: 2ag2, 1tjj, 2agc), LY96 (PDB-IDs: 2e59, 2e56, 4g8a, 3fxi, 2x65, 3mu3, 3rg1, 5ijd, 3vq2, 3m7o), DERF2 (PDB-ID: 1xwv), and NPC2 (PDB-IDs: 5kwy, 2hka, 3web). (C) Cross-section of the internal cavity in the NPC2 structure exposing the repositioned exemestane (thick sticks). CYP19A1 (purple ribbons) and NPC2 (gold surface) are locally superposed according to the sequence order-independent pocket alignment by eMatchSite. Annotated binding residues in NPC2 are solid, whereas the remaining surface is transparent.

This conjecture is supported by several recent studies. For instance, U18666A, a cationic sterol similar to exemestane with a TC of 0.67, binds to NPC1, inhibiting cholesterol export (Lu et al., 2015). Further, FDA-approved ezetimibe was shown to target NPC1 decreasing the cholesterol level (Phan, Dayspring, & Toth, 2012). Another study independently suggests repurposing thiabendazole, a potent inhibitor of cytochrome P450 1A2 (CYP1A2), to NPC1 (Soufan et al., 2016). Note that CYP1A2 and CYP19A1 are members of the cytochrome P450
family (Bateman et al., 2000) (Pfam-ID: PF00067) and have highly similar structures with a TM-score of 0.87. Finally, NPC2 was demonstrated to bind a range of cholesterol-related molecules, leading to an alteration of its function in lysosomal cholesterol transport (Liou et al., 2006). On that account, we hypothesize that exemestane binding to NPC2 disrupts the dynamics of its hydrophobic cavity. This effect could be exploited as a viable strategy to impede sterol movement to NPC1 preventing the accumulation of cholesterol in lysosomes in NPC disease.

3.4 Conclusions

In this study, we employ a collection of state-of-the-art algorithms to match, at an unprecedented scale, binding sites for known drugs with those pockets identified in proteins associated with rare diseases. Based on these data, we created eRepo-ORP, a new resource for orphan drug research. eRepo-ORP comprises 31,142 putative complexes between DrugBank compounds and Orphanet proteins exposing vast opportunities to reposition existing drugs to rare diseases. It was also demonstrated that combining eMatchSite with structure-based virtual screening enhances the accuracy of the detection of similar binding pockets. This promising methodology was employed to match drug-binding pockets from DrugBank with those from Orphanet exposing a number of opportunities to combat orphan diseases with existing drugs. Examples of potential therapeutics for orphan diseases that can be identified with this novel method introduced as eRepo-ORP were illustrated. The examples included Ras-associated autoimmune leukoproliferative disease, Maturity-onset diabetes of the young, and Niemann-Pick disease type C. Freely available through the Open Science Framework at https://osf.io/qdjup/, eRepo-ORP provides a list of pairs of DrugBank and Orphanet proteins sorted by the matching score, structure models of DrugBank and Orphanet proteins with predicted drug-binding sites, sequence and secondary structure profiles, structure models of DrugBank complexes annotated
with energy scores, and complex models of DrugBank drugs repositioned to Orphanet proteins with the corresponding energy scores. We expect that eRepo-ORP will prove valuable to orphan disease research by providing a robust, rational drug-repositioning component.

3.5 References


Huang, W. S., Metcalf, C. A., Sundaramoorthi, R., Wang, Y., Zou, D., Thomas, R. M., . . . Shakespeare, W. C. (2010). Discovery of 3-[2-(imidazo[1,2-b]pyridazin-3-yl)ethynyl]-4-
methyl-N-{4-[(4-methylpiperazin-1-y l)methyl]-3-(trifluoromethyl)phenyl}benzamide (AP24534), a potent, orally active pan-inhibitor of breakpoint cluster region-abelson (BCR-ABL) kinase including the T315I gatekeeper mutant. *J Med Chem*, 53(12), 4701-4719. doi:10.1021/jm100395q


4. TARGET-FOCUSED LIBRARIES FOR VIRTUAL SCREENING

4.1 Introduction

Due to extreme costs of high-throughput screening, many drug discovery projects commonly employ inexpensive computations to support experimental efforts. In particular, virtual screening, a technique that shows great promise for lead discovery, has become an integral part of modern drug design pipelines (Jain, 2004; Leung & Ma, 2015). Here, the idea is to considerably reduce the number of candidate compounds that need to be tested experimentally against a protein target of interest. Due to advances in computer technology resulting in constantly increasing computational power, virtual libraries comprising many thousands of compounds can be rapidly evaluated in silico prior to experimental screens and at a fraction of the cost. Virtual screening approaches, historically divided into ligand- and structure-based algorithms, prioritize drug candidates by estimating the probability of binding to the target receptor (Lavecchia & Giovanni). Among many methods developed to date, docking-based techniques are valuable tools for lead identification (Gschwend, Good, & Kuntz). These algorithms rank compounds by modeling the binding pose of a query molecule in the binding pocket of the target protein, followed by the prediction of binding affinity from molecular interactions. There are many examples of successful applications of virtual screening to develop compounds with desired bioactivities (Cavasotto & W. Orry, 2007; Villoutreix, Eudes, & Miteva, 2009).

---

Computer-aided drug discovery traditionally utilizes large compound libraries. For example, the ZINC database is one of the most comprehensive repositories of commercially available compounds for virtual screening (Irwin & Shoichet, 2005). It currently features over 35 million compounds in ready-to-dock formats. These large generic collections of low molecular weight organic compounds provide a sufficient diversity to perform virtual screening against any molecular target, however, the vast majority of compounds will have a very low probability to exhibit the desired bioactivity for a specific target protein. Furthermore, considering the imperfections of compound ranking by virtual screening algorithms (Kitchen, Decornez, Furr, & Bajorath, 2004), the top-ranked subset of compound library may contain a very few active molecules. Consequently, the chances to identify novel, high-quality leads from large compound repositories are low. For instance, an internal analysis of the Abbott compound collection suggested that less than 4% of the compounds in their library have the potential to yield novel kinase hinge-binders (Akritopoulou-Zanze & Hajduk, 2009). To address these issues, there has been a significant effort to augment existing collections with those compounds having a higher potential to bind to a specific target of interest. On that account, the trend in library design has shifted to include a strong focus on the target class in addition to diversity and drug-likeness criteria (Lowrie, Delisle, Hobbs, & Diller, 2004).

A target-focused library is a screening collection of compounds specifically tailored to modulate the function of a particular target or a protein family (John Harris, D. Hill, W. Sheppard, J. Slater, & F.W. Stouten, 2011; Orry, Abagyan, & Cavasotto, 2006). There are a variety of approaches developed to date to design target-specific focused libraries against, e.g. protein kinases, ion channels, G-protein coupled receptors (GPCRs), nuclear receptors, and protein-protein interfaces. These libraries not only reduce waste by eliminating compounds that
are unlikely to bind to target proteins, but often lead to the increased potency and specificity of
binders, as demonstrated for c-Src kinase (Maly, Choong, & Ellman, 2000). Several approaches
employ molecular docking to determine target-specific thresholds that can be used as filters in
virtual screening. This strategy was experimentally validated on the kinase-targeted library of
1,440 compounds and 41 kinases from five different families, demonstrating a 6.7-fold higher
overall hit enrichment compared to a generic compound collection (Gozalbes et al., 2008).
Furthermore, a structure-based modeling was used to create a small, focused library against
Chlamydia pneumoniae, a common pathogen recently linked to atherosclerosis and the risk
of myocardial infarction (Alvesalo, Siiskonen, Vainio, Tammela, & Vuorela, 2006).
Encouragingly, the experimentally determined hit rate for the targeted library was 24.2%, which
is considerably higher than that expected for a generic library. Similar to structure-based
approaches, ligand-based techniques can also be used in the focused library design, as shown for
the GPCR family (Lowrie et al., 2004). Compared to large and diverse screening libraries, using
relatively small and targeted collections significantly improves the odds of finding potential drug
candidates, thus further reduces the costs of drug discovery.

Target-focused libraries are either designed or assembled upon some understanding of a
specific protein target or a protein family. These collections are often compiled from larger, more
diverse libraries using either molecular docking (structure-based approach) or ligand fingerprint
similarity (ligand-based approach). The former employs structural, sequence and mutagenesis
data, whereas the latter is based on the biomolecular properties derived from known ligands,
offering a useful way of “scaffold hopping” from one ligand class to another (Renner &
Schneider, 2006). Target-focused libraries are often constructed around a single scaffold with
one or more positions used to attach various chemical moieties or side chains. Although this
approach can result in millions of different compounds (Michal Brylinski & Waldrop, 2014), the chemical space remains largely unexplored, therefore truly novel compounds will not be discovered. On the other hand, combinatorial chemistry methods can produce a vast collection of diverse compounds, so vast that only a tiny fraction of them could be explored, even using supercomputers. One can hardly imagine screening the chemical universe containing from $10^{12}$ to $10^{180}$ drug-like compounds (Gorse, 2006). Therefore, techniques to design chemical libraries covering pharmacologically relevant regions are needed (Dobson, 2004). These methods hold a promise to advance our knowledge of biological processes leading to new strategies to treat diseases.

Compiling focused libraries by molecular synthesis is essentially a combinatorial problem that can be addressed using graph theory. These techniques have been already extensively used in computer science and artificial intelligence for the synthesis of plans (Traverso, Ghallab, & Nau, 2004), problems and solutions in geometry (Alvin, Gulwani, Majumdar, & Mukhopadhyay, 2016), hardware from specifications (Chu, 1987), and communication protocols (Afrati, Papadimitriou, & Papageorgiou, 1988). Graph-based approaches also have a wide range of applications in drug discovery including the analysis of chemical structures to better understand the common features of drug molecules (Bemis & Murcko, 1996), the design of novel bioactive compounds with the desired pharmacological profiles (Estrada, Peña, & García-Domenech, 1998), the structure-based modeling of protein flexibility upon ligand binding (Carlson, 2002), the investigation of systems-level drug-target interaction networks (Nikolsky, Nikolskaya, & Bugrim, 2005), and drug repositioning (Gramatica et al., 2014).
In this study, we propose a new method to computationally synthesize molecules for virtual libraries, called eSynth. In essence, an exhaustive graph-based search algorithm is used to reconnect chemical building blocks procured from bioactive compounds following realistic connectivity patterns. Rather than focusing on a certain scaffold, the moieties used for synthesis come from active ligands of a specific target protein. Thus, the resulting chemical space is highly diverse, yet targeted. Given a set of initial molecules, eSynth generates new compounds to populate the pharmacologically relevant space. In order to evaluate the performance of eSynth, we conducted a series of benchmarking calculations against the Directory of Useful Decoys, Enhanced (DUD-E) dataset. First, in a self-benchmarking test, we validate the correctness of the search algorithm with the objective to recover a molecule from its building blocks. Further, the capability to discover novel scaffolds is assessed in a cross-benchmarking test. Here, bioactive compounds for each DUD-E target are first clustered into chemically dissimilar groups. Subsequently, each group considered as the validation set is reconstructed using dissimilar molecules pooled from other clusters. This protocol mimics a real application, where one expects to discover novel compounds based on a small set of already developed bioactives. Equally important, eSynth allows adding active subunits to an existing compound in order to generate a large library of prototypes of the modified ligand. Such libraries can be examined by molecular docking to explore those modifications yielding the highest binding affinity to the protein target.

4.2 Methods

Molecular fragments

We developed a procedure for the automatic identification and extraction of molecular fragments from chemical compounds. An example decomposition procedure is shown in Figure 4.1. The extraction process utilizes the PDBQT file format containing a central rigid fragment,
labeled as ROOT, from which zero or more rotatable bonds protrude. The sets of atoms connected through rotatable bonds are organized as BRANCHes, and at the beginning and end of each BRANCH section, the serial numbers of the two atoms forming a rotatable bond are recorded. First, we identify all rigid moieties (Figure 4.1B), where a rigid fragment is defined as a set of at least 4 non-hydrogen atoms connected by non-rotatable bonds (Figure 4.1C).

![Figure 4.1. Decomposing organic compounds into molecular fragments.](image)

Figure 4.1. Decomposing organic compounds into molecular fragments. Assuming that organic compounds are composed of sets of rigid fragments connected by flexible linkers, a molecule can be decomposed into its building blocks tracking the atomic connectivity. (A) A stick representation of afatinib. Extracting rigid fragments: (B) All rigid fragments are shown as thick lines, (C) only those rigid fragments composed of four or more atoms are retained. Extracting flexible linkers: (D) Small fragments connected by rotatable bonds, (E) small linkers are merged to form longer fragments, a single atom can act as a linker as well. The following colors are used for atom types: carbon – green, nitrogen – blue, oxygen – red, fluorine – yellow, and chlorine – pink.

The remaining parts are extracted as flexible linkers (Figure 4.1D). If two linker fragments are attached to each other, these will be connected to form a longer linker (Figure 4.1E). Failing to construct longer linkers from shorter fragments would limit the library to contain only very short linkers. Furthermore, we track the connectivity between individual fragments, so that chemically feasible compounds can be synthesized using a graph-based
algorithm. Every fragment is stored in the Structure Data Format (SDF) containing the 3D coordinates of all atoms and the corresponding atomic types as well as the connectivity information. The following SYBYL chemical types (Clark, Cramer, & Van Opdenbosch, 1989) are used for ligand atoms: carbon (C.1, C.2, C.3, C.ar and C.cat), nitrogen (N.1, N.2, N.3, N.4, N.am, N.ar and N.pl3), oxygen (O.2, O.3 and O.co2), phosphorous (P.3), sulfur (S.2, S.3, S.O and S.O2), and halogens (Br, Cl, F, I).

Connectivity information

Figure 4.2 illustrates the graph representation of rigids and linkers. A rigid fragment carries connectivity information indicating those atoms from which a rigid fragment was originally branched and the corresponding atom types it was connected to (Figure 4.2A). On the other hand, linkers contain information only on the number of allowed contacts at every atom, which is sufficient to create bonds with rigid fragments (linkers cannot bind to each other). The number of connections in a linker cannot exceed the maximum number of covalent bonds. Thus, we saturate a linker with hydrogen atoms and report the maximum number of bonds allowed for each atom in the linker file, e.g. N.3 atom shown in Figure 4.2B can bind at most two atoms that belong to rigid fragments accepting N.3. Noticeably, long linkers with the extensive connectivity pose a risk of expanding the molecular search space to an unmanageable size. Therefore, unsaturated linkers can also be built to store only the number of original connections, regardless of the maximum capacity of their atoms to create covalent bonds. In contrast to saturated linkers, using unsaturated linkers with substantially less connectivity considerably restricts the search space.
Fragment consolidation and pruning

Redundancy is removed from molecular fragments extracted from multiple compounds by consolidating the connectivity information and deleting identical moieties. For instance, if different parent molecules have a similar fragment, it suffices to have only one file for this moiety. However, if this fragment is connected to different atom types in distinct parent molecules, all possibilities should be retained. Therefore, this information is deposited as one copy of the fragment representing all possible connections. This is shown in Figure 4.2A, where the aromatic ring is the rigid fragment (solid round boxes) that was connected to C.3 in one parent molecule and to C.ar in another (dashed round boxes); this information is consolidated to create only one rigid moiety with two possible connections.

*eSynth* considers molecular bonding over a given set of rigid and linker fragments restricted by the laws of chemistry. Molecular synthesis is a fixed-point approach to generate a complete set of molecules given a set of fragments. The component algorithms of *eSynth* are described in the following sections. A fragment-based approach to synthesis can result in an infinite molecular search space unless an upper bound for molecular size is specified. Even with reasonable upper bounds imposed on the molecule size, the synthesis process may result in $10^8$ molecules or more. It is therefore highly desirable to develop an efficient algorithm for molecular synthesis that is complete, i.e. all possible molecules that can be synthesized under chemical and physical constraints are guaranteed to be generated. In the following sections *eSynth*’s capability to construct molecules from their respective fragments as well as producing new molecules from fragments generated from different compounds are examined.

Implementation of *eSynth*
Figure 4.2. Graph representation of rigids and linkers. Sample molecular fragments: (A) A rigid fragment, pyridine, with six constituent atoms in the bold outline and two possible connections to C.3 and C.ar in the dashed outline, (B) a three-atom linking fragment containing C.3 carbon with up to 3 connections, C.3 carbon with up to 2 connections, and N.3 nitrogen with up to 2 connections. Examples of 2-molecules: (C) Two identical rigids connected to each other, (D, E) two possible ways of connecting rigid and linker fragments shown in A and B.

Exhaustive molecular synthesis with eSynth

The architecture of eSynth shown in Figure 4.3 reflects a simple input/output paradigm with a black-box synthesizer. The input to eSynth is a set of rigid and linker fragments in SDF format (Figure 4.3A). Each SDF file is parsed (Figure 4.3B) using some functionality of Open Babel into a graph-based representation of the corresponding rigids and linkers (Figure 4.3C). From the set of linkers and rigids, the Synthesizer (Figure 4.3D) implements Algorithm 2 to
construct new compounds (Figure 4.3F). Each synthesized molecule is output using Open Babel in the Writer component (Figure 4.3E).

![Diagram of the eSynth process]

Figure 4.3. Implementation of eSynth. Input rigid and linker fragments in SDF format (A) are parsed (B) into the graph-based representation (C). Synthesizer (D) is the main engine to generate new molecules, which are subsequently passed to the Writer (E) component and output in SDF format (F).

One significant challenge in developing eSynth was integrating Open Babel to handle molecules. It is clear from Open Babel documentation and our own experience that even a simple operation such as molecular addition is thread-unsafe. It is necessary, therefore, to treat Open Babel as a singleton resource; we use Open Babel in a limited capacity to handle input and output. We then employ a local representation of compounds and fragments (optimizing memory with bit fields) to ensure that the Synthesizer is independent of Open Babel. Open Babel creates an undesirable bottleneck in the multi-threaded implementation of eSynth and thus a serial execution is superior.

Validation datasets and procedures. In order to evaluate the performance of eSynth, a series of benchmarking calculations were conducted against the DUD-E dataset (Mysinger, Carchia, Irwin, & Shoichet, 2012). Here, we use 20,408 bioactive compounds for 101 receptor proteins representing many important drug targets. First, we validated the correctness of the search algorithm using a self-benchmarking test. Subsequently, we performed a cross-validation test to evaluate the capability of eSynth to generate bioactive molecules with novel chemical structures.

In the self-benchmarking test, each active compound in the DUD-E library was decomposed into fragments and the molecular synthesis was performed. Parent compounds are
compared to those constructed by eSynth using molecular fingerprint matching with the chemical similarity assessed by the Tanimoto coefficient (TC) (Bajusz, Rácz, & Héberger, 2015; Tanimoto, 1958). The cross-validation test mimics a real application, where novel compounds are expected to be discovered based on a small set of known bioactive molecules. Here, bioactive compounds for each DUD-E target were first clustered into a collection of chemically dissimilar groups using SUBSET (Voigt, Bienfait, Wang, & Nicklaus, 2001) and a TC similarity threshold of 0.7. Subsequently, each cluster was selected as a validation set and molecular fragments from the remaining clusters were used by eSynth to build new molecules. The performance of eSynth is evaluated using the fraction of successfully reconstructed validation compounds using fragments extracted from chemically different molecules. Due to the large size of compound datasets generated by eSynth, we first used Open Babel (O’Boyle et al., 2011) to filter out those molecules that are dissimilar to the validation compounds with TC <0.5. Next, 3D atomic coordinates were generated for the synthesized molecules using obgen from the Open Babel package. A build-up algorithm to find atomic correspondence between chemical structures that calculates 2D-TC based on the identified the maximum common substructure (kcombu) (Kawabata, 2011) was then applied to measure the topological similarity between the filtered subset of synthesized molecules and the validation compounds.

4.3 Results

Search algorithm and the computational efficiency

eSynth generates target-focused libraries directly form ligands known to bind to a particular target protein or a family of proteins. The synthesis protocol first decomposes bioactive compounds into the non-redundant sets of chemical building blocks and then exhaustively combines these fragments to generate new molecules. We define two types of
fragments, rigid moieties and flexible linkers; each unique fragment is accompanied by the connectivity information. In case of identical fragments extracted from different molecules, the connectivity information is consolidated to produce a complete list of possible connections for every atom in this fragment. As a result, the fragment library compiled for a given protein target is non-redundant and representative.

Each fragment is converted to a graph-based representation of a chemical entity, where nodes correspond to atoms and undirected edges represent chemical bonds. Nodes in the molecular graph are annotated with the connectivity information. For instance, a rigid fragment shown in Figure 4.2A can produce para-substitutions with two moieties attached through their C.3 and C.ar atoms at opposite positions on the heterocyclic aromatic ring, whereas Figure 4.2B shows a 3-atom linker that can form up to 7 bonds with rigid fragments (3 on top C.3 + 2 on middle C.3 + 2 on bottom N.3). From a set of unique fragments represented as graphs with annotated nodes, the synthesis algorithm constructs increasingly larger molecules, such as those presented in Figures 2C, 2D and 2E.

Our initial implementation of the molecular synthesis described in Algorithm 1 is non-optimal due to considerable limitations on the wall time and memory space. For example, implementing a strict, either serial or parallel level-based approach results in ever-increasing memory requirements. Experimentally, we encounter this sharp increase of the number of generated molecules around levels 9 through 11, which can be expected considering the mode in the distribution of molecular fragments in the DUD-E database presented in Figure 4.4. Due to the exponential growth and the width of the chemical space, a solution was required to overcome the strict level-based approach.
On that account, we optimized the molecular synthesis by (1) implementing the bounded, level-based algorithm described in Algorithm 2, (2) reducing redundancy with Bloom filters, (3) restricting the connectivity using unsaturated linkers to narrow the width of the search space in order to gain depth, and (4) eliminating compounds violating the Rule-of-Five. The bounded, level-based synthesis algorithm imposes an explicit capacity on each level, so that all levels can be explored; it also ensures the termination of the synthesis procedure. Bloom filters provide computationally efficient mechanisms to eliminate those molecules that have already been synthesized. Furthermore, long, saturated linkers pose a considerable risk of expanding the molecular search space to an unmanageable size, therefore, we introduced unsaturated linkers accepting only those connections that were originally present in their parent molecules. In contrast to saturated linkers that can form all chemically possible bonds with rigid fragments, unsaturated linkers significantly restrict the search space, dramatically improving the computational efficiency. Finally, using the Rule-of-Five ensures that the synthesized compounds have drug-like properties. However, in order to test the drug likeliness of a molecule prior to its synthesis, Lipinski’s descriptors need to be estimated from molecular fragments, which is discussed in the following section.
According to Lipinski’s Rule-of-Five, a drug candidate should have a molecular mass (MW) less than 500 Da, no more than 10 hydrogen bond acceptors (HBA), no more than 5 hydrogen bond donors (HBD), and the octanol-water partition coefficient (logP) no greater than 5 (Lipinski et al., 2001). It is important to note that nearly half of active compounds in the DUD-E dataset (9,332 out of 20,402) would not pass this filter with the default cutoffs. To mitigate this problem, we modified thresholds for MW and logP to ensure that the majority of actives comply with the drug-likeness criteria. Specifically, using MW ≤570, HBD ≤5, HBA ≤10 and logP ≤7.2
increases the number of compliant molecules to 18,104 (88.7% of DUD-E actives). We refer to these values as a modified Rule-of-Five.

Decomposing the library of 20,408 bioactive DUD-E compounds resulted in 67,801 linkers and 65,507 rigid fragments. Due to the number of possible combinations growing exponentially with the number of molecular fragments, a modified Rule-of-Five is applied to exclude those compounds that do not satisfy drug-likeness criteria. It is therefore critical to rapidly estimate these properties directly from molecular fragments used to construct chemical compounds in order to prevent the synthesis of non-compliant molecules. This approach restricts the molecular search only to those compounds having drug-like properties. The additive nature of Lipinski’s descriptors allows for MW, HBD, HBA and logP of the synthesized molecules to be estimated directly from values pre-calculated for the fragment library.

The correlation between exact and estimated values of Lipinski’s descriptors are shown in Figure 4.5. Figures 5A and 5B demonstrate a high degree of correlation for MW and HBA with the Pearson correlation coefficient (PCC) of 0.99 and 0.98, respectively. Figure 4.5C shows that the PCC between the exact and estimated logP is 0.75. The estimated values tend to be slightly lower than the exact logP, however, the distributions of the estimated and exact values are fairly similar (see the histogram in Figure 4.5C).
Figure 4.5 – Correlation between exact and estimated Lipinski descriptors. For a given molecule, the following Lipinski's descriptors are estimated from its fragments: (A) Molecular weight (MW), (B) the number of hydrogen bond acceptors (HBA), (C) the octanol-water partition coefficient (logP), and (D) the number of hydrogen bond donors (HBD). The distribution of the exact (gray) and estimated (black) properties are shown in histograms attached to correlation plots.

The PCC for HBD is 0.65 (Figure 4.5D) with an over-predicted number of hydrogen bond donors. This can be expected since fragments are saturated with hydrogen atoms at positions of covalent bonds in their parent molecules and some of these hydrogens are able to
form hydrogen bonds. In order to apply the modified Rule-of-Five, fragment-based physicochemical descriptors are linearly transformed using the regression analysis presented in Figure 4.5.

Next, we investigate whether Lipinski's descriptors estimated from molecular fragments can be used to effectively eliminate non-compliant molecules from the synthesis process. Figure 4.6 presents the receiver operating characteristic (ROC) analysis of a binary classifier based on the MW, HBA, logP and HBD estimates.

Figure 4.6 – Predicting drug-likeness for molecular synthesis. Receiver operating characteristic plot assessing the accuracy of the prediction of drug-likeness from molecular fragments. The following Lipinski's descriptors are considered: Molecular weight (MW), the number of hydrogen bond acceptors (HBA), the octanol-water partition coefficient (logP), and the number of hydrogen bond donors (HBD). TPR – true positive rate, FPR – false positive rate, the gray area corresponds to the accuracy of a random classifier.

Here, the accuracy is evaluated by the area under the ROC curve (AUC) calculated for each property. AUC ranges from 0.0 to 1.0, where 1.0 corresponds to the highest accuracy and
0.5 is the accuracy of a random classifier. Encouragingly, MW can be estimated with the highest AUC of 0.999, whereas AUC values for HBA and HBD are 0.967 and 0.948, respectively. Although the correlation between the estimated and exact logP is fairly high (Figure 4.5C), the AUC is only 0.717, thus it cannot be used as a reliable predictor of the drug-likeness. This poor discriminatory power is mainly due to the fact that logP values can be either positive or negative. For instance, during molecular synthesis, an intermediate non-polar molecule with logP >5 can be brought back to the logP below 5 by attaching a highly hydrophilic moiety. Therefore, we only use logP for the final filtering after molecules are synthesized.

![Figure 4.7 – Self-benchmarking example. An example of the successful reconstruction of a molecule from its fragments. (A) The parent molecule is first decomposed into two rigids, thiophene (C₄H₄S) and 2,5-dimethylfuran [(CH₃)₂C₄H₂O], and two linkers, sulfonamide (SO₂N) and carboxylic acid [C(O)OH]. Examples of constructed (B) 2-molecules, (C) 3-molecules, and (D) 4-molecules including the parent compound.](image)

Connecting these fragments through the locations marked by asterisks following the connectivity patterns of the parent molecule produces a series of 2-, 3- and 4-molecules shown in Figures 7B, 7C and 7D, respectively. The target molecule is correctly reconstructed at level 4 (Figure 4.7D).
In Figure 4.8, we assess the results obtained for the entire set of 20,408 active compounds from the DUD-E dataset using the highest TC between the synthesized and parent molecules. Using saturated linkers, 61.6% of actives are reconstructed at a TC of 1.0, whereas 83.1% have a TC of ≥0.8. Moreover, the fraction of actives generated by eSynth that match parent compounds increases to 70.9% when unsaturated linkers are used. Note that Open Babel calculates TC for a pair of ligands using their hashed fingerprints, therefore, a TC of 1.0 denotes identical fingerprints, but not necessarily identical chemical structures. The inset in Figure 4.8 shows the computational efficiency of eSynth. Here, over 60% of actives are reconstructed in less than a second using a single processor thread, whereas 90% compounds are generated within a minute. Note that the synthesis time is fairly similar when only successful cases at a TC of ≥0.8 are considered.

Figure 4.8 – Performance of eSynth in the self-benchmarking test. Cumulative fraction of compounds reconstructed with the Tanimoto coefficient (TC) shown on the x-axis (logarithmic scale). Saturated linkers (gray line) can form all possible connections with rigid fragments, whereas unsaturated linkers (black line) can only form as many connections as present in their parent molecules. The inset shows the cumulative fraction of all compounds (solid line) and successful cases with a TC of ≥0.8 (dashed line) computationally synthesized by eSynth in 1 hour.
Despite these encouraging results, eSynth fails to reconstruct certain molecules. To clarify why some compounds are not correctly generated, Figure 4.8 presents main scenarios leading to unsatisfactory results. The first example shown in Figure 4.9A is a bioactive compound made up of a single fragment that cannot be decomposed into smaller parts. Since the molecular synthesis is not executed, eSynth generates no output. Molecules shown in Figures 9B and 9C contain a long linker with a high degree of connectivity. In such cases, rigid fragments can potentially bind at multiple linker locations leading to a combinatorial explosion. In principle, the parent molecule will be reconstructed at some point, however, we limit the wall time for molecular synthesis to 1 hour by default. During that time, about 10% of actives will not be reconstructed as previously shown in Figure 4.8 (inset). Using unsaturated linkers, whose connectivity is limited to the original connections in the parent compounds, helps address this issue, nevertheless, those targets containing long and highly flexible linkers are still not generated in a reasonably short computing time. Finally, some compounds are actually correctly reconstructed, yet they are not recognized as similar to their parent molecules. This is a false negative in the assessment of chemical similarity performed by Open Babel; we find that the fingerprint-based matching by Open Babel occasionally fails to recognize the high chemical similarity. We investigate this issue further in the following section.

![Figure 4.9](image)

Figure 4.9 – Examples of molecules not reconstructed by eSynth. Unsuccessful cases in the self-benchmarking test: (A) A molecule composed of only one rigid fragment, (B, C) examples of molecules containing long linkers that exponentially increase the search space.
Cross-validation test

A cross-validation test was performed in order to evaluate the capability of eSynth to generate novel bioactive molecules. Here, we attempt to reconstruct molecules highly similar to target compounds using fragments extracted from chemically different molecules. A set of fragments obtained from clusters other than the target cluster may lack rigid fragment(s) necessary to rebuild some of the active compounds. Since the molecular synthesis algorithm builds on the provided set of fragments, reconstructing molecules without all necessary parts is impossible. Encouragingly, 76.1% of 23,964 active DUD-E compounds for 101 target proteins are, in principle, reconstructible. Moreover, we examined individual clusters of similar ligands and found out that out of 9,406 clusters, as many as 4,100 clusters (43.6%) contain at least one compound that is non-reconstructible because of missing rigid fragments. These numbers are likely underestimated, considering the fact that linker fragments can also be missing and the connectivity patterns may not allow for the correct reconstruction of the topology of target actives, leading to non-reconstructible cases. Interestingly, Figure 4.10A indicates that for the majority of DUD-E targets, non-reconstructible actives are typically distributed across clusters of similar molecules.

Figure 4.10B presents the results obtained for 9,406 chemically distinct groups of compounds compiled using active DUD-E ligands. Encouragingly, in 45.1% (99.3%) of the cases, the active ligand is reconstructed at a TC of ≥0.6 (≥0.5) using fragments extracted from different clusters associated with the same receptor protein. Here, we employ an ultra-fast implementation of hashed fingerprint-based chemical similarity using Open Babel to compute 1D-TC because of a large number of pairwise similarity calculations. Subsequently, a relatively small fraction of compound pairs, whose 1D-TC is ≥0.5 are subjected to more accurate
comparison using 2D-TC by kcombu. In contrast to fingerprint-based techniques, kcombu detects one-to-one chemical matching between two structures that can be used to assess the similarity of their biological activities. When the similarity is evaluated by kcombu, 34.9% (58.2%) are reconstructed at a 2D-TC of ≥0.6 (≥0.5). It has been shown that two ligands whose 2D-TC is ≥0.6 typically have similar binding modes with a root-mean-square deviation (RMSD) below 2.0 Å (M. Brylinski, 2013; Kawabata & Nakamura, 2014). Moreover, as depicted in the inset in Figure 4.10B, 2D-TC of 0.6 reported by kcombu roughly corresponds to a fingerprint-based 1D-TC of 0.7, which is a widely used threshold for similar bioactivity (Hendlich, Bergner, Gunther, & Klebe, 2003; Rodgers, Zhu, Fourches, Rusyn, & Tropsha, 2010; Vilar, Uriarte, Santana, Tatonetti, & Friedman, 2013).

Figure 4.10 – Performance of eSynth in the cross-validation test. (A) DUD-E targets depicted as gray triangles are positioned in the plot according to the fraction of reconstructible active compounds and the fraction of chemically similar clusters containing only reconstructible actives. Non-reconstructible actives are more uniformly distributed across clusters for those targets lying closer to the solid black diagonal line. (B) Cumulative fraction of compounds reconstructed with the Tanimoto coefficient (TC) shown on the x-axis. TC is calculated using Open Babel (dashed gray line) and kcombu (solid black line). The vertical dashed line delineates a TC threshold of 0.6. The inset shows a direct comparison between TC values computed by Open Babel (1D-TC) and kcombu (2D-TC) with a solid black regression line.
A direct comparison between 1D-TC from Open Babel and 2D-TC from kcombu is shown as the inset in Figure 4.10B. The correlation between 1D- and 2D-TC is 0.61 with a somewhat lower similarity assessed by 2D-TC compared to the fingerprint-based 1D-TC. This observation as well as missing fragments leading to non-reconstructible cases explain the lower fraction of compounds successfully reconstructed by eSynth when 2D-TC is used to measure the chemical similarity. Since 2D-TC is calculated only for those pairs having 1D-TC ≥0.5, the success rate of molecular synthesis is likely underestimated because there are numerous cases for which 2D-TC is actually higher than 1D-TC. This discrepancy between 1D- and 2D-TC also clarifies why a small fraction of target compounds are not recognized as correctly generated in the self-benchmarking test described in the previous section (false negatives). Nonetheless, the results from the cross-validation test performed against a large dataset of bioactive compounds from the DUD-E dataset clearly demonstrate that eSynth is capable of generating novel molecules with the desired bioactivities.

Assessment of the synthetic accessibility

Finally, we assess the synthetic accessibility of molecules generated by eSynth using SAscore (Ertl & Schuffenhauer, 2009). SAscore employs the synthetic knowledge extracted from already synthesized chemicals penalizing a high molecular complexity; its values range from 1 for easily synthesizable molecules to 10 for those compounds that are very difficult to make. The distribution of SAscore values calculated for molecules generated by eSynth in the cross-validation test is shown in Figure 4.11. The average SAscore across the DUD-E dataset varies from 2 to 6 (Figure 4.11A). In general, molecules with a high SAscore of >6 are difficult to synthesize (Ertl & Schuffenhauer, 2009), therefore, the majority of compounds constructed by eSynth can be considered as synthetically accessible. Specifically, the average SAscore for
compounds generated for 19.8%, 61.4%, and 87.1% of DUD-E targets is less than 3, 4, and 5, respectively. This analysis also demonstrates that the synthetic accessibility depends on a particular biological target and the associated set of bioactive compounds, which are used to extract molecular fragments for eSynth.

![Figure 4.11](image)

**Figure 4.11** – Synthetic accessibility of molecules generated by eSynth. (A) Average ± standard deviation synthetic accessibility score (SAscore) calculated for molecules constructed by eSynth for individual DUD-E targets (sorted on the x-axis). (B) Violin plots showing the distribution of SAscore values across several datasets: decoy and active DUD-E compounds, FDA approved drugs, molecules generated by eSynth for the DUD-E targets, and natural products (NP) from the NuBBE and UNPD databases. Red horizontal lines correspond to the median SAscore values.
We also compare the distribution of SAscore values for compounds generated by eSynth to those collected for several other datasets. Figure 4.11B shows that the median SAscore value is 2.75 for decoy and 2.87 for active compounds from the DUD-E dataset (catalogue molecules). Moreover, the median SAscore for FDA approved drugs obtained from DrugBank (Wishart et al., 2006) and compounds constructed by eSynth are 2.95 and 3.66, respectively. For comparison, another study reported that the majority of bioactive molecules collected from the Derwent World Drug Index and the MDL Drug Data Report databases have SAscore between 2.5 and 5 (Ertl & Schuffenhauer, 2009). In contrast, natural products are generally more difficult to synthesize than typical organic molecules. Encouragingly, the median SAscore for molecules constructed by eSynth is lower than those for natural products, which is 3.82 for the Nuclei of Bioassays, Biosynthesis and Ecophysiology (NuBBE) database of secondary metabolites and derivatives from the biodiversity of Brazil (Valli et al., 2013), and 4.30 for the Universal Natural Product Database (UNPD) (Gu et al., 2013). Compounds from the Dictionary of Natural Products were previously reported to have a broad distribution of SAscore values between 2 and 8. On that account, the synthetic accessibility of molecules generated by eSynth is fairly high. The resulting datasets can be further filtered using existing tools, such as SAscore, in order to exclude those compounds containing synthetically unfeasible elements, e.g. chiral centers, large rings and non-standard ring fusions.

4.4 Discussion

Exploring the chemical space to produce pharmacologically applicable compounds is a daunting task because of an enormous size of the search space and numerous biochemical criteria restricting compound generation, i.e. synthetic feasibility, drug-likeness, and the effective binding to the biological target. Using atom-based methods may create an enormous chemical
space that can easily surpass the available computing resources. For instance, the largest library generated by an atom-based approach is the GDB-17 dataset comprising 166 billion small molecules (Reymond, 2015). On that account, fragment-based methods can be used as an alternative. Here, reference molecules are used as a source of building blocks, which can be subsequently combined to produce new compounds that are to some extent related to the initial molecules (Kawai, Nagata, & Takahashi, 2014). Fragment-based algorithms typically employ certain rules for combining various moieties, e.g. linker-linker bonds are prohibited, while ring-linker-ring connections are allowed. In contrast to atom-based methods, fragment-based techniques have capabilities to explore much larger molecules.

To facilitate the construction of target-focused libraries for virtual screening, we developed eSynth, a new fragment-based approach to molecular synthesis that follows simple combinatorial chemistry steps using an optimized, graph-based algorithm. eSynth rapidly generates series of compounds with diverse chemical scaffolds complying with criteria for drug-likeness. Although, these molecules may have different physicochemical properties, the initial fragments are procured from biologically active and synthetically feasible compounds. Consequently, we demonstrated that the constructed libraries are enriched with pharmacologically relevant molecules synthesized under loose biochemical constraints.

Our effort simplifies the synthesis process by avoiding techniques such as click chemistry, e.g. AutoClickChem (Durrant & McCammon, 2012), and those relying on statistical restraints, e.g. Fragment Optimized Growth (FOG) (Kutchukian, Lou, & Shakhnovich, 2009). Moreover, in contrast to other methods designed for certain classes of compounds such as peptides generated from amino acid fragments, e.g. GrowMol (Bohacek, McMartin, Glunz, & Rich, 1999) and LUDI (Bohm, 1992), eSynth can construct any class of organic, drug-like
molecules. Several methods employ the binding site information in order to generate molecules with a binding affinity toward a given target protein, e.g. Multiple Copy Simultaneous Search (MCSS) (Miranker & Karplus, 1991), SPROUT: structure generating software using template (Gillet, Johnson, Mata, Sike, & Williams, 1993), and SMall Molecule Growth (SMoG) (Ishchenko & Shakhnovich, 2002). eSynth does not require protein structures, yet the cross-validation test clearly demonstrates that molecules highly similar to those compounds known to bind to the target protein are effectively generated.

Evolutionary algorithms that break fragments and make crossovers allow for an exhaustive exploration of the chemical space (Huang, Li, & Yang, 2010; Virshup, Contreras-Garcia, Wipf, Yang, & Beratan, 2013), however, using these techniques also requires applying chemical stability and synthetic feasibility rules, which, in turn, utilizes extra computational resources. For instance, the Algorithm for Chemical Space Exploration with Stochastic Search (ACSESS) was designed to construct representative universal libraries in an arbitrary chemical space (Virshup et al., 2013). This approach implements convergent evolutionary operations through bond and/or atom modifications on an initial library of molecules to acquire a maximally diverse subset of molecules. Although using evolutionary techniques does not guarantee a completeness of the space search, ACSESS systematically explores the small molecule universe, providing a near-infinite source of novel compounds. Differ from other techniques employing generic combinatorial algorithms, chemical rules and filters, eSynth was not designed to explore a broad chemical space; rather, it is purposely confined to a chemical sub-space around a particular drug target.

eSynth relies solely on fragments and their connectivity patterns extracted from parent molecules to generate a series of drug-like compounds. Thus, it is essential to use synthetically
feasible bioactive compounds as the source in order to generate molecules with similar chemical and pharmacological profiles. Importantly, eSynth is not restricted to a particular hypothesis, e.g. a pre-defined pharmacophore often used by synthesis algorithms. For example, a pharmacophore-based de novo design method of drug-like molecules (PhDD) ensures that molecules constructed from linker and rigid fragments fit a given pharmacophore model (Q. Huang et al., 2010). The search space in PhDD is not only confined to the fragment and linker libraries, but also it is limited to a user-defined template molecule in the form of a pharmacophore hypothesis. eSynth avoids such hypotheses in order to generate target-focused compound datasets, yet without any bias toward a specific scaffold.

Molecular synthesis methods often use knowledge-based rules to connect fragments. For example, combining the amine with the carbonyl to form the amide changes the preference of the nitrogen atom toward those moieties that might be more likely attached to an amide rather than an amine nitrogen (Kutchukian et al., 2009). On that account, FOG uses the statistical knowledge to create new branches and decide which branch to grow as an effective way to generate novel molecules. Similar to eSynth, FOG employs a construction algorithm using molecular fragments to generate synthetically tractable molecules, however, it grows molecules using a Markov Chain according to statistics on the frequency of specific connections in the database of chemicals. Moreover, the Topology Classifier algorithm is used to classify the constructed molecules as drugs or non-drugs. Given a set of fragments, the chemical search space in FOG may be somewhat limited to those molecules having similar characteristics as the training compounds. In contrast, eSynth creates new molecules by reusing fragments and following their connectivity patterns in the parent compounds. Therefore, it covers a larger chemical space and does not require constructing statistical databases of fragment connections.
4.5 Conclusions

eSynth is a new algorithm to generate large datasets of chemical compounds by connecting small molecular fragments. It first establishes the width of a search space with a diverse foundation of initial small molecules followed by the stochastic exploration of the depth of the chemical space by constructing multi-fragment molecules. This hybrid approach ensures a deeper exploration of the molecular space by synthesizing larger molecules while circumventing the necessity of a complete exploration through the synthesis of all possible molecules. eSynth can compile large libraries of drug-like molecules with the desired properties, which may be unfeasible using atom-based synthesis techniques. Moreover, the resulting libraries can be further filtered based on the geometry and energy of binding, and the biological activity toward specific targets. Finally, we demonstrated that eSynth has capabilities to generate novel, biologically active ligands for target proteins from chemically distinct parent molecules.

4.6 References


103


5. DECOMPOSING MOLECULES FOR FRAGMENT-BASED DRUG DESIGN

5.1 Introduction

Hit identification, lead generation and lead optimization are the key steps at the outset of a drug discovery process. Briefly, compounds showing promising activity identified by high-throughput screening as initial hits are filtered and modified to generate lead compounds, which satisfy basic drug-likeliness properties (C. A. Lipinski, Lombardo, Dominy, & Feeney, 2001). These lead compounds are further optimized to enhance the potency toward the target protein as well as to reduce their non-selectivity and toxicity (Palermo & De Vivo, 2014). Conventional hit identification is not only limited to already synthesized compounds often leading to low discovery rates, but it is also expensive and requires time consuming screening experiments (Leelananda & Lindert, 2016). Consequently, virtual screening that can rapidly evaluate millions of compounds has become an integral part of lead identification protocols (Mestres, 2002).

In order to enhance the chemical diversity of virtual screening libraries, large collections of drug-like compounds can be generated through combinatorial chemistry (Lipinski & Hopkins, 2004). Since constructing and screening the entire chemical space are not feasible even with the most advanced computers, building extensive yet targeted libraries is critical for the success of virtual screening. A number of fragment- and atom-based techniques have been developed to generate novel chemical compounds for virtual screening, including binding-site point connection

---

3 This chapter, previously published as Liu T, Naderi M, Alvin C, Mukhopadhyay S, Brylinski M. 2017. Break down in order to build up: Decomposing small molecules for fragment-based drug design with eMolFrag. J Chem Inf Model. 57(4):627-631. DOI: 10.1021/acs.jcim.6b00596, is reprinted here by permission of JCIM.
methods LUDI (Bohm, 1992), fragment connection methods LEA3D (Douguet, Munier-Lehmann, Labesse, & Pochet, 2005) LigBuilder (Yuan, Pei, & Lai, 2011), and eSynth (Naderi, Alvin, Ding, Mukhopadhyay, & Brylinski, 2016), sequential build-up algorithms LEGEND (Nishibata & Itai, 1991) and SPROUT (Gillet, Johnson, Mata, Sike, & Williams, 1993), and random connection techniques CoG (Brown, McKay, Gilardoni, & Gasteiger, 2004) and Flux (Fechner & Schneider, 2007). These *de novo* methods require an initial set of building blocks or molecular fragments, which ultimately control the properties of the resulting screening compounds and their affinity toward the target protein. Consequently, there is a great interest in efficient fragmentation techniques to generate sets of chemically feasible building blocks for the subsequent molecular synthesis. Retrosynthetic combinatorial analysis procedure RECAP (Lewell, Judd, Watson, & Hann, 1998) and breaking retrosynthetically interesting chemical substructures BRICS (Degen, Wegscheid-Gerlach, Zaliani, & Rarey, 2008) are examples of systematic fragmentation methods. In RECAP, compounds are dissected based on a set of 11 bond types, following simple rules such as leaving cyclic bonds and alkyl groups smaller than five carbons intact. These rules ensure that major structural features of organic compounds, such as ring motifs, are preserved. BRICS expands the bond type criteria used by RECAP from 11 to 16 taking into account the chemical environment of each bond type and the surrounding substructures. Additional filters are also applied in order to prevent generating small and unwanted fragments. Other methods extract and classify chemical scaffolds by pruning side chains and removing peripheral ring moieties (Schuffenhauer et al., 2007).

In general, the performance of fragment-based chemical synthesis tools such as eSynth (Naderi et al., 2016) CONFIRM (Thompson et al., 2008) and AutoGrow (Durrant, Amaro, & McCammon, 2009) could significantly be improved by employing building blocks annotated
with empirical connectivity patterns. Although this information could help explore pharmacologically relevant regions of the diverse chemical space (Naderi et al., 2016), many existing fragmentation tools, e.g. Fragmenter ("Fragmenter (ChemAxon)," http://www.chemaxon.com) and molBLOCKS (Ghersi & Singh, 2014), do not consider the chemical context of the fragments.

In other words, the connectivity information of a fragment is not stored while extracting building blocks. To address this issue, we developed eMolFrag, a new open-source molecular fragmentation software. eMolFrag decomposes either a single ligand or a library of compounds into two types of chemical building blocks, bricks and the connecting linkers. The resulting complete and non-redundant sets of building blocks are annotated with the comprehensive connectivity information in order to facilitate the construction of novel compounds with combinatorial synthesis software. eMolFrag has been parallelized to decrease the computing time required to analyze large collections of molecules.

5.2 Methods

eMolFrag employs a graph-based notation, where molecules are sets of nodes representing atoms connected by edges corresponding to chemical bonds. A fragment is a substructure, which has either all or only some atoms and bonds of a given molecule; fragments are categorized as either bricks or linkers. Given a collection of molecules, the complete set of unique fragments is constructed in two steps shown in Figure 5.1. The first step, labeled as Part I, involves creating an initial set of fragments, whereas the second step, labeled as Part II, guarantees the uniqueness of the resulting set of fragments.
Figure 5.1. Flowchart of eMolFrag. Part I: Input molecules are fragmented with the BRICS algorithm to generate a complete set of building blocks. Part II: Fragment redundancy is removed according to pairwise chemical alignments with the kcombu program. At the end, non-redundant sets of bricks and linkers are reported along with the consolidated connectivity information as well as lists of similar fragments that were removed.

Part I: Fragmentation

In eMolFrag, a set of molecules are first decomposed into constituent fragments with the BRICS algorithm (Degen et al., 2008), implemented in RDKit ("RDKit: Open-source cheminformatics," http://www.rdkit.org). Chemical compounds are broken down into larger moieties called bricks connected by linkers based on 16 chemical environments defined by the BRICS model(Degen et al., 2008); a pseudo-code for the fragmentation process is given in Supporting Information (Algorithm S1). A brick fragment is a molecular construct having at least four non-hydrogen atoms. Subsequently, bricks are removed from a molecule and the remaining fragments are classified as linkers (see Algorithm S2 in Supporting Information). Broken bonds are replaced by dummy atoms, which are placeholders for those atoms removed from a particular bond. The complete information, including the type of atoms involved in those bonds that were broken, is stored for each brick in order to provide empirical connectivity patterns. Linkers have different auxiliary connectivity information, i.e. these fragments are annotated only with the maximum number of bonds at various positions. Examples of bricks and linkers are provided in Supporting Information (Examples S1 and S2, respectively). We found that this approach allows to efficiently construct series of new molecules, whose chemistry is similar to that of parent compounds.

Part II: Mitigation of fragment redundancy
Since one of the objectives of an effective fragmentation procedure is to employ the resulting fragments in a synthesis procedure, the cardinality of the final set of fragments is critical. On that account, eMolFrag attempts to minimize the size of sets of bricks and linkers by removing redundancy with a partitioning and sieve-based removal scheme presented in Supporting Information (Algorithm S3). Two fragments are equivalent if the Tanimoto coefficient (TC) calculated for topologically constrained maximum common substructures by the kcombu program (Kawabata, 2011) is equal to 1.0. Information on equivalent atoms provided by kcombu as well as their connectivity information is then used to consolidate identical fragments into a single, unique construct.

5.3 Results and Discussion

Benchmarks against the DUD-E database

We validate the eMolFrag algorithm by conducting a self-reconstruction test as described previously (Naderi et al., 2016). Briefly, given an input molecule \( m \), a set of fragments extracted from \( m \) by eMolFrag are passed to a fragment-based construction procedure with eSynth employing its chemical rules (Naderi et al., 2016). A molecule with the highest chemical similarity to \( m \) measured by the TC calculated for Daylight fingerprints ("Daylight Chemical Information Systems Inc.," http://www.daylight.com/) is selected from a series of compounds constructed by eSynth. Here, we employ a fingerprint-based assessment of chemical similarity with OpenBabel (O'Boyle et al., 2011) because this technique is computationally much faster than kcombu. A TC of \( \geq 0.8 \) indicates that a molecule highly similar to \( m \) was generated, whereas a TC of 1.0 indicates that compound \( m \) has been reconstructed. As a testing set, we use 20,408 active compounds for 102 protein targets from the Directory of Useful Decoys, Enhanced (DUD-
E) database (Mysinger, Carchia, Irwin, & Shoichet, 2012) covering a diverse chemical space of pharmacologically relevant molecules.

The performance of eMolFrag is compared to molBLOCKS (Ghersi & Singh, 2014), another fragmentation software employing the RECAP algorithm (Lewell et al., 1998). Figure 5.2 shows a two-way box plot of the number of atoms per fragment and the number of fragments per molecule for these two programs. Fragments generated by molBLOCKS typically contain 6-10 (the default protocol) and 6-11 (an extensive mode) atoms, whereas most fragments extracted by eMolFrag consist of 2-7 atoms. The median numbers of fragments per molecule are 3, 8, and 6 for molBLOCKS (default), molBLOCKS (extensive), and eMolFrag, respectively. Finally, molecular synthesis with eSynth (Naderi et al., 2016) was conducted employing fragments generated by eMolFrag for active compounds in the DUD-E database. Encouragingly, 82.8% of active compounds were reconstructed with a TC of 1.0 and 92.2% with a TC of ≥0.8. An inspection of the failed cases revealed that the major reason for not generating a relatively small fraction of testing compounds is the fact that the synthesis software does not allow to directly connect two bricks. Overall, the self-reconstruction benchmarking results demonstrate that eMolFrag properly extracts molecular fragments providing sufficient connectivity information to rebuild the majority of parent molecules.
Figure 5.2. Two-way box plot of the number of fragments per molecule against the number of atoms per fragment. Bioactive compounds from the Directory of Useful Decoys, Enhanced database were fragmented with eMolFrag (green) and molBLOCKS (default protocol in red, extensive mode in grey). Computational performance

Decomposing large compound libraries can be time-consuming depending on the number of input molecules, therefore, we parallelized the eMolFrag code. The serial and parallel performance of eMolFrag is assessed by fragmenting subsets of DUD-E actives with sizes varying from 100 to 12,800 molecules. All tests were performed on a machine equipped with two 2.6 GHz 8-core Sandy Bridge Xeon 64-bit processors, 32GB 1666MHz RAM and 500GB HD, running Red Hat Enterprise Linux 6. Figure 5.3 shows that the wall time for eMolFrag scales linearly with the number of input molecules. The average processing speed of the serial code ranges from 8.7 molecules per second for the smallest dataset to 4.8 molecules per second for the largest dataset (see Table S1 in Supporting Information).
The actual decomposition speed (Part I in Figure 5.1) is faster for larger sets because the I/O overhead is reduced by efficient data caching. However, removing redundancy (Part II in Figure 5.1) from large datasets requires significantly longer computing times compared to small datasets, which in turn causes the overall speed to decrease with the increasing number of input molecules. Without removing redundancy, the average processing speed of serial eMolFrag is 9.8 molecules per second for the smallest dataset and 23.2 molecules per second for the largest dataset. For comparison, a serial version of molBLOCKS, which does not remove redundancy, is capable of processing 6.6 and 12.5 molecules per second for the smallest and the largest datasets, respectively. Thus, eMolFrag is 1.2-1.9× faster than molBLOCKS. Further, algorithms implemented in eMolFrag are polynomial in complexity; the best-fit curves in Figure 5.3 are 

\[ y = 0.022 \times x^{1.238} \quad (R^2 = 0.9989) \] 

for serial and 

\[ y = 0.013 \times x^{1.201} \quad (R^2 = 0.9998) \] 

for parallel execution. This near-linear scaling gives an empirical evidence of the efficient implementation of eMolFrag.

The impact of the number of computing cores on parallel processing is assessed by comparing the performance of parallel eMolFrag to the theoretical speedup estimated with Amdahl’s law.(Amdahl, 1967) The inset in Figure 5.3 shows that executing eMolFrag in parallel for a fixed input dataset of 3,200 molecules and the number of computing cores varying from 1 to 16 roughly corresponds to a hypothetical code consisting of 47-60% parallel calculations. Note that eMolFrag does not conform exactly to Amdahl’s law because the workload related to removing redundancy (Part II in Figure 5.1) is unevenly distributed across computing cores. Although the total execution time of eMolFrag diverges from Amdahl’s law, the parallel processing is faster than the serial execution. The average processing speed for the parallel code running on 16 computing cores ranges from 24 molecules per second for the smallest dataset to
11.8 molecules per second for the largest dataset (see Table S1 in Supporting Information). This shorter processing time for parallel eMolFrag becomes particularly beneficial for larger datasets. For instance, decomposing 20,408 active compounds from the DUD-E dataset for the self-benchmarking test takes 1 hour and 18 minutes on a single core compared to only half an hour on 16 computing cores.

![Figure 5.3](image.png)

Figure 5.3. Serial and parallel performance of eMolFrag. The main graph shows the wall time for the complete fragmentation procedure plotted against the number of input molecules. A serial code is compared to the parallel processing on 16 computing cores. Parallel scaling for a fixed size input dataset of 3,200 molecules is presented as the inset. Upper and lower bounds for the ideal speedup calculated according to Amdahl’s law are shown as dark and light gray lines, respectively.

Application to antagonists of the adenosine receptor

To illustrate the application of eMolFrag in *de novo* drug discovery, we show that bioactive compounds can successfully be constructed from molecular fragments extracted from chemically dissimilar binders of the same target protein. Here, we selected the human A2a adenosine receptor (AA2AR), a member of the G protein-coupled receptor (GPCR) superfamily
containing targets for about 27% of all FDA-approved drugs (Overington, Al-Lazikani, & Hopkins, 2006). Figure 5.4 presents individual steps of the cross-validation procedure, in which CHEMBL144979, a known bioactive ligand for AA2AR (Baraldi et al., 2000) is the target molecule. Four other AA2AR antagonists, called donors, are shown in Figure 5.4A. Since the chemical similarity of donors to the target measured by the TC reported by kcombu is lower than 0.5, CHEMBL144979 can be considered novel with respect to the donor molecules.

Unique sets of 10 bricks and 7 linkers extracted by eMolFrag from 4 donors are shown in Figures 4B and 4C, respectively. For instance, the triazolo-quinazoline fragment highlighted in pink carrying the chlorine moiety was obtained from CHEMBL95229. This compound is a member of a series of pyrazolo-triazolo-pyrimidines with subnanomolar affinity against ARs created via N5-phenylcarbamoyl substitutions (Kim et al., 1998). Bricks contain information on atom types that can be attached at various positions (small boxes in Figure 5.4B), whereas linkers are annotated with the maximum number of allowed bonds (small circles in Figure 5.4C). The sets of bricks and linkers are complete and non-redundant, i.e. each unique fragment carries the connectivity information extracted from multiple donor compounds. For example, the connectivity information for a benzene ring, which is present in all donors, is consolidated by eMolFrag into a single fragment shown in cyan in Figure 5.4B.
Figure 5.4. Example of the successful construction of a bioactive of the adenosine receptor by eMolFrag and eSynth. (A) Donor molecules with the chemical similarity to CHEMBL144979 measured by the Tanimoto coefficient (TC). (B) Bricks annotated with the list of atom types that can be attached at various positions. (C) Linkers annotated with the number of the maximum allowed connections. (D) Examples of new molecules synthesized using bricks and linkers. The first molecule shown in a box is a known bioactive of the adenosine receptor. Highlighted in different colors are essential building blocks to generate CHEMBL144979 that are extracted from donor molecules by eMolFrag and used in molecular synthesis by eSynth. Further, the connectivity information inferred from donors that is required to correctly assemble CHEMBL144979 is highlighted in bold in B and C.

Subsequently, molecular fragments extracted by eMolFrag were passed to eSynth (Naderi et al., 2016) in order to generate a series of compounds. A serial version of eSynth produced 4,492,609 virtual compounds in 12 hours. Encouragingly, the first compound in Figure 5.4D (shown in a box) is CHEMBL144979, therefore, the target molecule has been successfully constructed. Further, the set of virtual molecules comprises 845 compounds, whose TC to
CHEMBL144979 is ≥0.7 and as many as 239,656 molecules with a TC of ≥0.5. Three randomly selected virtual molecules are presented in Figure 5.4D to demonstrate the chemical diversity of compounds generated by eSynth. It is important to note that these retrospective cross-validation benchmarks are designed to mimic real applications by attempting to construct target molecules using building blocks extracted from chemically dissimilar compounds. This case study demonstrates that high-quality fragment sets generated by eMolFrag can be used in fragment-based drug discovery to create targeted screening libraries likely containing novel bioactives.

5.4 Conclusions

eMolFrag is a fast and robust tool to extract molecular fragments, classified as bricks and linkers, from small molecule datasets. Subsequently, these fragments can be used to construct targeted libraries for virtual screening. A unique feature of eMolFrag is that it stores the connectivity information for the extracted building blocks to help generate new series of chemically feasible compounds. Although eMolFrag was optimized to work with eSynth, a recently developed molecular synthesis algorithm, it can also be integrated into other cheminformatics toolkits utilizing chemical fragments. eMolFrag is freely available as a stand-alone software and a webserver at www.brylinski.org/emolfrag and https://github.com/liutairan/eMolFrag.

5.5 References


Daylight Chemical Information Systems Inc. (http://www.daylight.com/).


6. DOMAINS-BASED HSV-1 GLYCOPROTEIN-K PROTEIN MODEL

6.1 Introduction

Herpes simplex virus 1 (HSV-1) encodes at least 26 tegument proteins and 11 virally encoded glycoproteins, as well as several nonglycosylated membrane-associated proteins. Viral glycoproteins gD, gB, gH, and gL serve critical roles in virion entry (Cai, Gu, & Person, 1988; Desai, Schaffer, & Minson, 1988; Herold, WuDunn, Soltys, & Spear, 1991; Hutchinson, Browne, et al., 1992; Spear, 2004). Virion entry is initiated by the binding of glycoproteins gB and gC to glycosaminoglycan (GAG) moieties of cell surface proteoglycans (Spear, 2004). This initial attachment causes the interaction of gD with one or more of its specific receptors, including the herpesvirus entry mediator (HVEM) (HveA), nectin-1 (HVEC), and 3-O-sulfated HS. In addition, gB binds to PILR-α, NMHC-IIA, and myelin-associated glycoprotein (MAG) receptors (Chowdhury, Naderi, Chouljenko, Walker, & Kousoulas, 2012). HSV-1 enters into neurons strictly via a pH-independent fusion of the viral envelope with neuronal plasma membranes (David et al., 2012; Nicola, Hou, Major, & Straus, 2005; Qie, Marcellino, & Herold, 1999), while it can enter a wide range of nonneuronal cells via either pH-independent or pH-dependent endocytosis (Milne, Nicola, Whitbeck, Eisenberg, & Cohen, 2005). Fusion of the viral envelope with cellular, including neuronal, membranes causes deposition of the viral capsid into the cytoplasm, which is subsequently transported to the cell nucleus. Virus entry into all cells involves the coordinated functions of the glycoproteins gD, gB, gH, gL, and gC. Initial binding

---

of gD to the nectin-1 receptor is thought to alter interactions of the gH/gL complex with gB, triggering gB-mediated fusion of the viral envelope with plasma membranes (Connolly, Jackson, Jardetzky, & Longnecker, 2011).

The UL20 and UL53 (gK) genes are highly conserved in all alphaherpesviruses and encode proteins of 222 and 338 amino acids, respectively, each with four membrane-spanning domains (Foster, Alvarez, & Kousoulas, 2003; MacLean, Efstathiou, Elliott, Jamieson, & McGeoch, 1991; Melancon, Foster, & Kousoulas, 2004; Ramaswamy & Holland, 1992). HSV-1 gK is posttranslationally modified by N-linked carbohydrate addition at the amino terminus of gK, while the UL20 protein (UL20p) is not glycosylated (Hutchinson, Goldsmith, et al., 1992; Ramaswamy & Holland, 1992). HSV-1 gK and UL20 functionally and physically interact, and these interactions are necessary for their coordinate intracellular transport, cell surface expression, and functions in virus-induced cell fusion, virus entry, virion envelopment, and egress from infected cells (Dietz et al., 2000; Foster, Chouljenko, & Kousoulas, 2008; Foster & Kousoulas, 1999; Foster, Melancon, Baines, & Kousoulas, 2004; Foster, Melancon, & Kousoulas, 2001; Foster, Melancon, Olivier, & Kousoulas, 2004; Fuchs, Klupp, Granzow, & Mettenleiter, 1997; Jambunathan et al., 2011; Melancon, Fulmer, & Kousoulas, 2007). The gK/UL20 protein complex interacts with gB and gH and is required for gB-mediated cell fusion (Chouljenko, Iyer, Chowdhury, Chouljenko, & Kousoulas, 2009; Chouljenko, Iyer, Chowdhury, Kim, & Kousoulas, 2010). HSV-1 gK is a structural component of virions and functions in virion entry (Foster, Rybachuk, & Kousoulas, 2001; Jambunathan et al., 2011). Deletion of amino acids 31 to 68 within the amino terminus of gK inhibits virus-induced cell-to-cell fusion and virus entry without drastically inhibiting virion envelopment and egress. Moreover, deletion of gK amino acids 31 to 68 inhibited virus-induced cell fusion caused by syncytial mutations in gK and
entry into PILR-α-expressing Chinese hamster ovary cells (Chouljenko et al., 2009; Chowdhury, Chouljenko, Naderi, & Kousoulas, 2013). We have shown that gK is essential for neuronal infection and virulence (David, Baghian, Foster, Chouljenko, & Kousoulas, 2008). Specifically, we have reported that gK-null virus was unable to infect axonal termini and egress from neuronal cell bodies. In addition, we have recently shown that the HSV-1(McKrae) gKΔ31–68 virus, specifying gK with a deletion of amino acids 31 to 68, was unable to efficiently infect mouse trigeminal ganglia after ocular infection of scarified mouse eyes (Saied, Chouljenko, Subramanian, & Kousoulas, 2014). These results indicate that the amino terminus of gK plays a pivotal role in corneal infection and neuroinvasiveness.

HSV-1 as well as other viruses utilizes the intracellular microtubular network, which is utilized to move intracellular cargo in a retrograde manner toward the microtubule-organizing center (MTOC) and nucleus, as well as in an anterograde manner toward the cell periphery during maturation and cellular egress (Dohner, Nagel, & Sodeik, 2005; Greber & Way, 2006; Hammonds, Denyer, Jackson, & Irving, 1996; Luxton et al., 2005; Sodeik, Ebersold, & Helenius, 1997) (Figure 6.1). Cellular cargo is transported in a retrograde manner toward the MTOC and the cell periphery in conjunction with the dynein and kinesin motors, respectively. Intracellular cargo can simultaneously bind to both dynein and kinesin and move bidirectionally along microtubules. This dynein/kinesin competition for intracellular cargo transport is highly coordinated within cells to maintain subcellular organization (Kramer & Enquist, 2013). In a similar fashion, HSV-1, vaccinia virus, and adenovirus utilize the dynein-dynactin motor complex for intracellular transport (Diefenbach, Miranda-Saksena, Douglas, & Cunningham, 2008; Dohner et al., 2002; Johnson & Baines, 2011; Leopold et al., 2000; Mettenleiter, Klupp, & Granzow, 2009; Sanderson, Hollinshead, & Smith, 2000; Suomalainen et al., 1999).
substantial evidence suggests that a number of other viruses, including African swine fever virus (ASFV) (Alonso et al., 2001), canine parvovirus (CPV) (Suikkanen et al., 2003), influenza virus X-31 (Lakadamyali, Rust, Babcock, & Zhuang, 2003), human foamy virus (HFV) (Petit et al., 2003), Mason-Pfizer monkey virus (M-PMV) (Sfakianos, LaCasse, & Hunter, 2003), and rabies virus (RV) (Raux, Flamand, & Blondel, 2000), utilize the microtubular network for their intracellular transport.

Figure 6.1. Schematic representation of intracellular transport of virion capsids via the cellular microtubular network. The virus enters the cell via fusion or endocytosis and is transported by dynein motors toward the nucleus (retrograde transport), presumably by the interaction of dynein with one or more inner tegument proteins. Kinesin transports the cargo toward the cell membrane (anterograde transport).

After fusion of the HSV-1 envelope with the host plasma membrane, the tegumented capsids containing the viral genome are released into the cytosol. The majority of the outer
tegument proteins remain at the plasma membrane along with viral glycoproteins, while inner tegument proteins, such as the UL36, UL37, and US3 proteins remain attached to virion capsids (Granzow, Klupp, & Mettenleiter, 2005; Maurer, Sodeik, & Grunewald, 2008). HSV-1 tegument proteins UL36 and UL37 are strong candidates for binding to the dynein motor (Antinone et al., 2006; Radtke, Dohner, & Sodeik, 2006). The HSV-1 UL37 is a 120-kDa phosphorylated tegument protein expressed in both mature virions and light particles (Coller, Lee, Ueda, & Smith, 2007; Pasdeloup, Blondel, Isidro, & Rixon, 2009; Roberts et al., 2009). UL36 (also called VP1/2) is the largest capsid-bound tegument protein encoded by the Herpesviridae, containing 3,164 amino acids (Gibson & Roizman, 1972). The UL36 and UL37 proteins remain with capsids undergoing retrograde transport in neuronal and epithelial cells, while the majority of the other tegument proteins are absent from capsids during transport (Antinone & Smith, 2010; Copeland, Newcomb, & Brown, 2009). Also, a lack of either UL36 or UL37 causes failure of retrograde transport of capsids to the nucleus (Radtke et al., 2010). Recent evidence indicates that pseudorabies virus (PRV) UL36 binds to dynein/dynactin, facilitating microtubule transport, neuroinvasion, and pathogenesis (Zaichick et al., 2013).

Resolving the structures of membrane proteins is challenging mainly due to difficulties in overexpression, purification, and reassembly of membrane proteins into membrane-mimetic systems for structural analysis. Consequently, <2% of structures available in the Protein Data Bank (PDB) are membrane proteins. Computational methods, such as homology-based modeling, are instrumental in shedding light on the secondary and tertiary structures of membrane proteins (Koehler Leman, Ulmschneider, & Gray, 2015). Homology modeling is a well-established technique to study protein functions and mechanisms (Nayeem, Sitkoff, & Krystek, 2006; Ramachandran S, 2012). For example, it can be used to investigate the pandemic
potential of mutant influenza viruses and rational antiviral drug and vaccine design (Rajapaksha & Petrovsky, 2014), find important amino acid residues through computer-guided mutations (Morita et al., 2011), and elucidate protein-protein interactions that are important for host-pathogen interactions (Ader et al., 2012).

Here, we generated a new model for the predicted three-dimensional (3D) structure of gK by assembling individual domains modeled separately into a full-chain model. This gK model reveals highly conserved domains among alphaherpesviruses, including a β-sheet structure that spans the gKΔ31–68 deletion. We show that the HSV-1 (McKrae) gKΔ31–68 virus is unable to enter into neuronal axons in cell culture, in agreement with in vivo ocular infections of mice, which indicated that the virus was not transported into trigeminal ganglionic neurons (Saied et al., 2014). Based on the conservation of gK predicted domains among alphaherpesviruses, we suggest that gK may function to facilitate entry of other alphaherpesviruses into neurons.

6.2 Materials And Methods

Cell lines and viruses

African green monkey kidney (Vero) cells were obtained from the American Type Culture Collection (Manassas, VA) and were maintained in Dulbecco's modified Eagle's medium (Gibco-BRL, Grand Island, NY) supplemented with 10% fetal calf serum and antibiotics. HSV-1(F), VC1, wild-type (WT) McKrae, McKrae gKΔ31–68, and McKrae D4V5 viruses were used in this study. VC1 was engineered to express gK with the V5 epitope inserted in frame immediately after amino acid 68 of gK and a 3×FLAG epitope inserted in frame at the amino terminus of UL20, as described earlier (Jambunathan et al., 2011). HSV-1(McKrae ΔgK31–68) was engineered to express gK lacking 38 amino acids immediately after the gK signal sequence (Saied et al., 2014), and gKΔ31–68-D4V5 (D4V5) was constructed by insertion of the V5
epitope tag at the carboxyl terminus of gK of HSV-1(McKrae gKΔ31–68) using double-red mutagenesis implemented on the viral genome cloned as an bacterial artificial chromosome, as described previously (Tischer, von Einem, Kaufer, & Osterrieder, 2006).

Virus entry proximity ligation assay (VEPLA)

Proximity ligation assay (PLA) was performed as we have described earlier (Jambunathan et al., 2014). Briefly, Vero cells were grown on 8-well chamber slides (Nunc Lab-Tek II chamber slide system) and infected with strain F virus at a multiplicity of infection (MOI) of 10. The virus was allowed to attach to confluent monolayer of Vero cells at 4°C for 1 h and shifted to 37°C to allow virus entry, and the chamber slides were removed at time zero, 30 min, 2 h, 3 h, 6 h, 9 h, 12 h, and 16 h and fixed with ice-cold methanol for 10 min at −20°C. Mouse antidynein antibody against intermediate-chain I (Abcam) and rabbit anti-UL37 antibody (a gift from Frank J. Jenkins, University of Pittsburgh Cancer Institute) were used for dynein/UL37 detection. Mouse anti-gD antibody (Virusys, Inc.) and rabbit anti-nectin-1 antibody (Santa Cruz) were used as positive control for gD/nectin-1 interaction in Vero cells. Mouse antidynein antibody and rabbit anti-gM antibody (a gift from Joel Baines, Louisiana State University) were used as a negative control for detecting dynein/gM interaction. Mouse monoclonal antibody against α-tubulin (fluorescein isothiocyanate FITC) (Abcam), was used for microtubule detection. Duolink in situ PLA probes, (anti-rabbit plus and anti-mouse minus) were added to the samples (1:5 dilution) and incubated at 37°C for 1 h, followed by ligation at 37°C for 30 min after washing with Duolink in situ wash buffer A. Amplification solution (40 to 50 µl) was added, and slides were incubated for 1.5 h at 37°C. Texas Red-labeled oligonucleotide detection probes (Olink Bioscience) were used. The slides were subsequently washed with Duolink in situ wash buffer B twice for 10 min each and once with 0.1× Duolink buffer B, followed by mounting with
mounting medium containing DAPI (4′,6′-diamidino-2-phenylindole) (Duolink II), and stored at −20°C with protection from light until confocal images were taken. The images were taken using a 60× objective on an Olympus Fluoview FV10i confocal laser scanning microscope.

Quantification of PLA signal

The quantification of the fluorescent signals was done using the SlideBook5 digital microscopy imaging software (Intelligent Imaging Innovation, Denver, CO) (kind help was provided by Masami Yoshimura, Department of Comparative Biomedical Sciences, Louisiana State University). The fluorescent images were imported as TIF files into the SlideBook software. In order to quantify the PLA signals, segment mask was done for both the PLA signals and DAPI-stained nucleus. Both the size (number of pixels) and average intensity of PLA signals were measured. The sum intensity (average intensity × number of pixels) of the PLA signals was divided by the area (in pixels) of the nucleus. This value was calculated for WT McKrae gD/nectin, McKrae gKΔ31–68 gD/nectin-1, WT McKrae UL37/dynein, and gKΔ31–68 UL37/dynein. Unpaired Student t test results showed that the two-tailed P value for WT McKrae and McKrae gKΔ31–68 gD/nectin-1 is 0.0738 (P > 0.05), which is not statistically significant, and the two-tailed P value for WT McKrae UL37/dynein and gKΔ31–68 UL37/dynein is 0.0372 (P < 0.05), which is considered to be statistically significant. The efficiency of virus entry for the data shown in Figure 6.7 was calculated as E=number fluorescent spots×intensity (UL37/dynein PLA)/number of spots × intensity (gD/nectin-1 PLA); E (gKΔ31–68)/E (McKrae) = 0.58.

Infection of DRG neurons in cell culture

Embryonic day 18 (E18) Sprague-Dawley rat dorsal root ganglia (DRGs) in specialized medium including nerve growth factor (NGF) were obtained from BrainBits Inc. and seeded on poly-D-lysine-coated 8-well culture slides (catalog number 354632; Becton Dickinson, Inc.) per
the manufacturer's instructions. Neuronal cell cultures obtained from DRGs were highly enriched in neurons. The DRG is surrounded by a connective tissue capsule and is histologically composed of neuronal cell bodies, which are surrounded by supportive cells (satellite cells). Satellite cells provide electrical insulation for the pseudounipolar neurons in the ganglia. Neuronal extensions of these DRG cultures are practically devoid of DRG-associated fibroblast or glial cells. Consequently, the presence of fibroblast and epithelial cells surrounding neuronal extensions was very sparse, constituting less than one percent of the neuronal axonal projects in any microscope field examined. Cultures contained glial cells, while they were largely devoid of fibroblast and epithelial cells (not shown). Maintenance tissue culture medium consisted of neural basal medium with B-27 supplement at the manufacturer's recommended concentration (Invitrogen, Grand Island, NY). Medium was supplemented with 50 ng/ml neural growth factor 2.5s (Invitrogen), 2% normal rat serum (Invitrogen), 1% GlutaMAX (Invitrogen), and 0.2% Primocin (InvivoGen, San Diego, CA). The ganglia were monitored regularly for axonal growth and cultured with neural basal medium supplemented with neuronal growth factors. A healthy extension of axons was observed at 7 days postseeding, and the ganglia were ready to be infected. The medium was then removed, and 200,000 PFU of either HSV-1(McKrae) or gKΔ31–68 mutant virus was added to the ganglia. The virus was removed after 1 h, and the slides were fixed with ice-cold methanol for 10 min at −20°C. PLA assay was performed on these slides as described above.

Immunogold labeling for TEM

Purified virions immobilized on 400-mesh Butvar/carbon-coated nickel grids (Electron Microscopy Sciences, Inc., Hatfield, PA) were used to detect the presence of gK and gD on the virions using immunogold labeling, as we have described previously (Jambunathan et al., 2011).
Briefly, the grids were incubated with 5 µl of mouse anti-V5 antibody (Invitrogen) and anti-gD antibody (Virusys) at a dilution of 1:10,000 in 1% bovine serum albumin (BSA) in Tris-buffered saline (TBS) for 30 min to detect gK and gD, respectively, followed by 30 min of incubation with 5 µl of goat anti-mouse IgG (whole molecule)–10-nm colloidal gold (Sigma-Aldrich) at a 1:80 dilution in 1% BSA in TBS. A 2% solution of sodium phosphotungstate (pH 6.8) was added as a final step for contrast purposes. Grids were desiccated and visualized by transmission electron microscopy (TEM). The number of virions scanned for each panel ranged from 40 to 65. Not all virions were positive for immune gold particles. On average, 10% of virions did not show the presence of any gold particles.

Virus purification

The virus was purified as we have described earlier (Jambunathan et al., 2011). Briefly, supernatants and cells from 10 T-150 flasks of Vero cells infected with YE102-VC1(gK-EK-V5;UL20-3×FLAG), gKΔ31–68, and gKΔ31–68-D4V5 viruses were collected at 36 h postinfection (hpi) and purified by 50 to 20% discontinuous iodixanol gradients twice, followed by a 20% iodixanol cushion. The resulting pellet was resuspended in 250 µl of NP-40 lysis buffer (Invitrogen) and used for immunoblot assay.

Immunoprecipitation and immunoblot assays

The HSV-1(F) virus-infected lysate was immunoprecipitated with protein G magnetic Dynabeads bound to dynein antibody according to the manufacturer's instructions (Invitrogen). The protein was eluted from the magnetic beads in 40 µl of elution buffer and used for immunoblot assays. Immunoblot assays were carried out using anti-gB (Virusys), anti-UL37, goat anti-mouse–horseradish peroxidase (HRP) (Abcam), and goat anti-rabbit–HRP (Abcam).

Protein modeling of gK
The gK sequence was divided into domains, and each part was modeled separately. Atomic structures were built using Modeler (Fiser & Sali, 2003) from template-target alignments calculated by HHpred for transmembrane domains and by eThread (Brylinski & Lingam, 2012) for the N terminus and domain II. Subsequently, the individual components were assembled into a full-chain model using Chimera (Pettersen et al., 2004) according to the current understanding of the orientation of the domains with respect to each other and the lipid bilayer. Loops and gaps in the alignments were constructed by Modeler. Finally, the model of gK was embedded in a POPC membrane (140 Å by 50 Å) composed of 170 lipid molecules using VMD (Humphrey, Dalke, & Schulten, 1996) and PyMOL (PyMOL Molecular Graphics System, version 1.2r3pre; Schrödinger, LLC); the entire system comprises 18,887 nonhydrogen atoms.

6.3 Results

Molecular modeling of the effect of the gKΔ31–68 mutation on the predicted structure of gK.

We have reported previously that the gKΔ31–68 domain deletion did not adversely affect infectious virus production, although it inhibited the ability of syncytial mutations in gB to cause extensive virus-induced cell fusion. In contrast, the smaller amino acid deletion gKΔ31–48 and the larger gKΔ31–117 mutation inhibited infectious virus production drastically (Chouljenko et al., 2009). To better understand the structural aspects of gK, we constructed the 3D structure of gK by modeling individual domains separately, followed by assembly of a full-chain model according to the experimentally derived topography of gK in membranes (Foster et al., 2003), as described in Materials and Methods. The model revealed that the gKΔ31–68 deletion spanned a well-defined predicted β-sheet structure within the amino terminus of gK that when deleted did not appreciably change the overall structure of the remaining amino terminus of gK. Also, the
model predicted a prominent 41-amino-acid alpha-helical domain spanning gK domain II located intracellularly (Figure 6.2).

Figure 6.2. (A) Predicted structures of gK and gKΔ31–68. Transmembrane and intracellular domains are shown in cyan and green, respectively. (B) The 3D structure predicts a prominent β-sheet structure after the signal sequence. (C) Structure of the amino terminus of gKΔ31–68 (gK without 38 amino acids (Suikkanen et al.)).

Glycoprotein gKΔ31–68 is incorporated into virions

We have shown that gK is expressed on virion particles (Foster, Rybachuk, et al., 2001) and that insertion of the V5 epitope tag at the carboxyl terminus of gK, as well as within selected internal sites of gK, did not adversely affect overall viral replication and infectious virus production (Foster et al., 2003). To detect the presence of gKΔ31–68 within virions, we generated the HSV-1(McKrae) gKΔ31–68-D4V5 (D4V5) mutant virus expressing a V5 epitope
tag inserted at the carboxyl terminus of gK using double-red mutagenesis, as we have described previously (Saied et al., 2014) (see Materials and Methods). The presence of gK on virion particles was initially assessed using colloidal gold immunoelectron microscopy (Figure 6.3A). An average of 50 virion particles for each virus were inspected for the presence of gold particles. V5-tagged gK was readily detected on most virion envelopes (>45 out of 50 visualized virions contained gold particles) for recombinant viruses VC1 (containing a V5 epitope tag within the amino terminus of gK) and D4V5, while none of the gKΔ31–68 virions contained gold particles. Similarly, the anti-gD monoclonal antibody readily detected gD on all virions (Figure 6.3A). The presence of gK on virions was confirmed in immunoblots of extracts from purified VC1 and D4V5 virion extracts probed with anti-V5 antibody. VC1 gK protein species migrated with apparent molecular masses ranging from 38 to 55 kDa, while D4V5 gK protein species migrated with apparent molecular masses ranging from 45 to 55 kDa. There were no gK protein species detected in HSV-1(McKrae) gKΔ31–68 virion extracts (Figure 6.3B). The V5-tagged gK was also detected on the surface of infected cells via fluorescence-activated cell sorting (FACS) (not shown).

The UL37 protein interacts with cytoplasmic dynein

UL36 and UL37 are associated with capsids undergoing retrograde transport in neuronal and epithelial cells. To test whether the UL37 protein interacts with dynein, cell extracts from infected cells were immunoprecipitated with antidynein antibody, and the samples were immunoblotted and probed with both anti-UL37 and anti-gB antibodies. Dynein immunoprecipitates contained the UL37 protein, detected as a protein species migrating with molecular mass of 120 kDa (Figure 6.4), band marked with an asterisk), while dynein did not
coprecipitate gB, which was used as a negative control. Additional negative controls included lysates from uninfected Vero cells (Figure 6.4).

![Image](image-url)

**Figure 6.3.** Detection of gKΔ31–68 on virion particles. (A) VC1(F), McKrae gKΔ31–68, and McKrae D4V5 extracts from purified virions were reacted with anti-V5 antibody to detect V5-tagged gKs and visualized by immunogold transmission electron microscopy. Approximately 50 virion images were scanned for the presence of immunogold particles on virion envelopes. Detection of gD was used for positive-control purposes, while the parent gKΔ31–68 without a V5 tag on its gK was used as negative control. (B) Extracts from purified virions were electrophoretically separated and immunoblotted with anti-V5 antibody to detect V5-tagged gKs. The presence of gK species was detected in both VC1 (tagged with V5 within the amino terminus of gK) and D4V5 (tagged with V5 at its carboxyl terminus), while no gK was detected in the untagged parental virus gKΔ31–68.

**Detection of UL37 interaction with cytoplasmic dynein and development of the VEPLA**

Cell surface-bound virions were detected by monitoring interactions of virion gD with its cognate receptor nectin-1 using PLA. Similarly, PLA was utilized to detect UL37-dynein interactions to monitor viral capsids that have entered the cytoplasm and bound dynein as the prerequisite step for microtubular loading and retrograde transport (Figure 6.5). Bright red fluorescent signals were obtained when PLA detected colocalization of gD with nectin-1 immediately after adsorption of the virus for 1 h at 4°C (zero time point) on Vero cells. In
contrast, PLA with antibodies against dynein and UL37 did not detect colocalization of UL37 and dynein at the zero time point (Figure 6.6). UL37-dynein colocalization was readily visible as early as 30 min after virus entry, visualized as numerous bright red fluorescent spots in the cytoplasm of infected cells. Glycoprotein M (gM) did not colocalize with dynein at any time point tested (negative control). UL37-dynein colocalization was detected at all times postinfection tested except the zero time point. The distribution and density of fluorescent spots revealing colocalization of UL37 with dynein changed at 6 hpi, assuming a perinuclear distribution at 9 hpi. The overall number and intensity of fluorescent spots dissipated after 12 hpi, and they were undetected at 16 hpi (Figure 6.6).

Figure 6.4. Dynein interacts with UL37. Lysates from cells infected with HSV-1(F) were subjected to immunoprecipitation as described in Materials and Methods. Negative controls included detection of gB (*) in dynein immunoprecipitates and uninfected cell extracts.
We have shown previously that the gKΔ31–68 virus enters Vero cells with slower entry kinetics than the wild-type virus (Chowdhury et al., 2013; Jambunathan et al., 2011). Virus entry proximity ligation assay (VEPLA) performed at 1 hpi revealed that both the wild-type virus and the gKΔ31–68 virus attached equally well to Vero cell surfaces. In agreement with previous findings, VEPLA revealed a marked reduction in gKΔ31–68 virion entry in comparison to that of the wild-type virus, as evidenced by the significant reduction in the number and relative intensity of fluorescent spots produced by colocalization of UL37 with dynein. (Figure 6.7).
Figure 6.6. Kinetics of virus entry into Vero cells. Confluent monolayers of Vero cells seeded in microscopy chamber slides were infected at an MOI of 10 with wild-type virus HSV-1(F) and tested by VEPLA for cell surface-bound virions (anti-gD/anti-nectin-1 antibodies) and cytoplasmic capsids at different times postinfection at 37°C (anti-UL37/antidynein). Anti-gM and antidynein antibodies were utilized as negative controls. The microtubules were visualized using mouse monoclonal antibody against α-tubulin (FITC), shown in green. DAPI (blue) was used for visualization of the nucleus, and the PLA signals were seen as red spots.
The amino terminus of HSV-1 gK is required for the entry of the virus into the DRG axons

Recently, we showed that gK, specifically the amino-terminal 38 amino acids of gK, which are deleted in the gKΔ31–68 mutant virus, is required for the replication and efficient spread of the virus to the trigeminal ganglion after ocular infection of mice (David et al., 2008; David et al., 2012). In addition, lack of gK caused an inability of the virus to infect axonal termini separated from neuronal somata in specialized microfluidic devices harboring purified ganglionic neurons (David et al., 2012). To directly test the ability of the gKΔ31–68 virus to enter axonal termini of ganglionic rat neurons in cell culture, purified rat dorsal root ganglia (DRGs) seeded on fluorescence microscopy slides were infected with either the HSV-1(McKrae) virus originally isolated from patients suffering from keratitis (Wang, Davido, & Morrison, 2013) or the gKΔ31–68 virus, and VEPLA was performed at 1 h postinfection (hpi). Both viruses attached to neuronal surfaces equally well, as evidenced by the number of fluorescent spots produced by colocalization of viral gD with the nectin-1 receptor. In contrast, there were no fluorescent spots detected on neuronal axons infected with the gKΔ31–68 virus, while numerous spots were detected on neuronal axons infected with the McKrae wild-type strain. Both viruses appeared to infect equally well glia cells surrounding neuronal axons, as well as neuronal somata (Figure 6.8).

6.4 Discussion

Viral glycoprotein K (gK) and its interacting partner UL20 are highly conserved among all neurotropic alphaherpesviruses, while beta- and gammaherpesviruses do not specify gK or UL20 orthologues. Therefore, our working hypothesis is that gK and UL20 function to facilitate successful infection of neurons. Here, we show for the first time that a deletion of a predicted β-sheet structure within the amino terminus of HSV-1 gK, conserved among alphaherpesviruses,
prevents the virus from entering into ganglionic axons, in agreement with previous findings that this gK deletion caused the inability of mutant viruses to infect ganglionic neurons and establish latency after ocular infection of mice (Saied et al., 2014).

Figure 6.7. McKrae gKΔ31–68 virions enter more slowly than parental McKrae virions in Vero cells. VEPLA was utilized to detect wild-type McKrae and gKΔ31–68 entry into Vero cells. Confluent Vero cell monolayers were infected with either McKrae or gKΔ31–68 viruses at an MOI of 10, and capsids that entered the cytoplasm and virions bound to cell surfaces were detected at 1 h postinfection with anti-UL37/antidynein and anti-gD/anti-nectin-1 antibodies, respectively. The PLA signals were quantified using SlideBook 5 digital imaging software.
Figure 6.8. The McKrae gKΔ31–68 mutant virus is unable to enter DRG axons. VEPLA was utilized to assess McKrae and gKΔ31–68 virus entry into DRG neurons in cell culture. Embryonic day 18 (E8) rat DRGs were collected and seeded on polylysine-coated microscopy slides. DRGs were infected with WT HSV-1 strain McKrae or gKΔ31–68 at 6 days postseeding. The top and bottom left panels show interaction between gD and nectin-1 on DRG axons infected with WT McKrae and gKΔ31–68 mutant viruses, respectively, as seen by the presence of red spots along the axons (arrows). The top right panel shows the interaction between UL37 and dynein in DRG axons infected with WT McKrae virus. There was no interaction detected between UL37 and dynein in DRG axons infected with gKΔ31–68 (bottom right panel). Neurofilament marker (green) and DAPI (blue) were used to identify axons and the nuclei of glial cells, respectively.

We generated a new model for the predicted three-dimensional structure of gK by assembling individual domains modeled separately into a full-chain model. This methodology has been successfully applied to derive protein models of other viral membrane proteins. For example, modeling of the structure of the H7N9 hemagglutinin (HA) helped predict the potential
of the Chinese A/Hangzhou/1/2013 strain by predicting that the H7H9 strain could bind to human sialic acid receptors. The accuracy of the results of this study was confirmed after the crystal structure of the H7N9 HA protein was resolved later, leading to an almost identical structure with a root mean square deviation (RMSD) of only 0.721 Å over 305 atoms, while \textit{in vitro} results showed binding capabilities similar to the predictions (Rajapaksha & Petrovsky, 2014). Importantly, computer-generated protein models can be used not only to predict the overall function of a protein but also to elucidate the molecular mechanism of such function. For example, an X-ray structure of a four-helix bundle in the H protein in parainfluenza virus type 5 (PIV5) was used to model the H protein in the stalk region of the canine distemper virus to investigate the conformational changes in viral fusion (F) protein required for cell entry. Combining \textit{in vitro} experiments and computational structure modeling ultimately led to the discovery of a new model for activation of the fusion machinery utilized by morbilliviruses (Ader et al., 2012).

To overcome the difficulty in predicting the three-dimensional structures of multiple membrane-spanning proteins such as gK and UL20, we predicted the structures of each gK domain (I to IV) separately and assembled them into a full-chain structure. This gK model produced a number of important structural features, including a prominent β-sheet structure spanning the gKΔ31–68 deletion in the amino terminus of gK and a 41-amino-acid α-helical structure spanning gK domain II located intracellularly (Figure 6.2). These predicted domains are conserved among alphaherpesviruses such as HSV-2 and others (Figure 6.2). Syncytial mutations within the carboxyl terminus of gB do not cause fusion in the presence of the gKΔ31–68 deletion (Chouljenko et al., 2009).
Figure 6.9. Prediction of the three-dimensional structures of the amino termini of gK specified by HSV-1, HSV-2, VZV, and monkey B virus. The distal side of the amino terminus of gK is colored yellow and is retained for comparison purposes as background on all other frames. (A) HSV-1 gK. (B) HSV-2 gK. (C) VZV gK. (D) monkey B virus gK.

In addition, we have shown that the amino-terminal 82-amino-acid domain of gK when expressed separately binds gB and complements the inability of the gKΔ31–68 mutation to support gB-mediated membrane fusion. These results suggest that gB-mediated membrane fusion is regulated by interactions between the amino termini of gB and gK (Chouljenko et al., 2009; Chouljenko et al., 2010).

We have shown previously that gK is incorporated into virions and functions in virus entry (Chowdhury et al., 2013; Jambunathan et al., 2011). Furthermore, the gKΔ31–68 deletion
did not negatively affect infectious virus production (Chouljenko et al., 2009). Similarly, the
D4V5 virus replicated efficiently in Vero cells, achieving titers at 24 hpi that were similar to
those of the parental gKΔ31–68 virus (not shown). The gKΔ31–68 deletion includes two N-
glycosylation sites located within the amino terminus of gK. The D4V5 gKΔ31–68 gK migrated
with a molecular mass of 45 to 55 kDa despite lacking both N-glycosylation sites. It has been
shown that gK migrates anomalously (Foster et al., 2008) in SDS-PAGE due to its high
hydrophobicity and ability to multimerize (Foster et al., 2003; Hutchinson, Goldsmith, et al.,
1992). Characteristically, boiling of gK SDS-PAGE samples prevents gK from entering into
gels. Heating up to 45°C for 15 min is typically utilized to allow gK to enter into separating gels.
Therefore, it is likely that the observed gKΔ31–68 molecular mass range of 45 to 55 kDa is due
to the overall gKΔ31–68 structure, which causes anomalous migration in SDS-PAGE.

Biochemical analysis and live imaging have strongly suggested that the viral VP1/2
(UL36) and UL37 proteins interact with dynein (Radtke et al., 2010). In agreement with these
findings, we detected interactions of UL37 with dynein via PLA. PLA is a relatively simple tool
to detect potential protein-protein interactions. This technique has been utilized for detecting
infectious agents and proteins with a sensitivity rated higher than even that of PCR (Gustafsdottir
et al., 2006). Interestingly, PLA detected UL37-dynein interactions throughout the course of
infection of Vero cells. It is possible that UL37 may bind dynein to prevent dynein interference
and ensure efficient utilization of kinesin-mediated anterograde transport during virion
cytoplasmic envelopment and egress.

Most virus entry assays depend on the expression of viral proteins after deposition of
viral DNA into the nuclei of infected cells through detection of either viral antigens or marker
genes, such as green fluorescence protein expressed from the viral genome, or the induction of
marker genes such as β-galactosidase under HSV-1 promoter control responding to expression of HSV-1 immediate early proteins (Chowdhury et al., 2013; Reinhard, Le, Ohlin, Hengel, & Trilling, 2011; Simpson et al., 2005). However, these assays measure the overall successful rate of infection and not necessarily the ability of virions to enter into the cytoplasm of infected cells, since entering capsids may not be efficiently transported to the nucleus. Detection of capsids in the cytoplasm of infected cells can be achieved by high-power confocal microscopy and transmission electron microscopy, with great difficulty in obtaining semiquantitative results. We describe here the development of the virus entry proximity ligation assay (VEPLA) as a general method to efficiently visualize and quantify the relative efficiency of virus entry into the cytoplasm of infected cells in a time-dependent manner by monitoring both enveloped virions attached to cell surfaces and capsids in the cytoplasm that interact with dynein prior to loading onto the microtubular network for retrograde transport.

VEPLA was utilized to show that the gKΔ31–68 mutation prevented virions from entering into neuronal axons, in agreement with our published results that this mutation caused the inability to infect ganglionic neurons after ocular infection of mice (Saied et al., 2014). It is likely that inhibition of virus entry reflects the inability of gB to cause fusion of viral envelopes with axonal membranes in the presence of the gKΔ31–68 mutation. This is supported by our previous results that the amino terminus of gK spanning the gKΔ31–68 mutation interacts with the amino terminus of gB (Chouljenko et al., 2009). Alternatively, the amino terminus of gK may enable binding of virions to gK-specific receptors that are required for virus entry. HSV-1 and HSV-2, as well as all alphaherpesviruses, infect neuronal endings embedded into the epidermis and mucosal surfaces, including the highly innervated corneal epithelium. Prediction of the three-dimensional structures of gK specified by alphaherpesviruses shows remarkable
primary and predicted tertiary structures (unpublished data). Of particular interest to this study is conservation of predicted structural domains within the amino termini of gK specified by HSV-1, HSV-2, varicella-zoster virus (VZV), and monkey B virus (Figure 6.9). The β-sheet structure contained within the gKΔ31–68 deletion is conserved among these viruses as well as other alphaherpesviruses (not shown), suggesting that they may play a conserved role in infection of neuronal axons.

Understanding of the role of gK in neuronal entry will enable the production of new ways to ameliorate these infections, as well as assist in the development of safe live-attenuated vaccines and viral vectors to combat HSV and other infectious diseases. In support of this prediction, we have recently shown that the HSV-1 (VC2) vaccine strain containing the gKΔ31–68 mutation protected mice against lethal intravaginal challenge with either virulent HSV-1 or HSV-2 in mice (Stanfield et al., 2014).

6.5 Reference


Chowdhury, S., Naderi, M., Chouljenko, V. N., Walker, J. D., & Kousoulas, K. G. (2012). Amino acid differences in glycoproteins B (gB), C (gC), H (gH) and L (gL) are associated with enhanced herpes simplex virus type-1 (McKrae) entry via the paired immunoglobulin-like type-2 receptor alpha. *Virol J, 9*, 112. doi:10.1186/1743-422X-9-112


Foster, T. P., Melancon, J. M., & Kousoulas, K. G. (2001). An alpha-helical domain within the carboxyl terminus of herpes simplex virus type 1 (HSV-1) glycoprotein B (gB) is associated with cell fusion and resistance to heparin inhibition of cell fusion. *Virology, 287*(1), 18-29. doi:10.1006/viro.2001.1004


7.1 Introduction

Membrane fusion is an essential step for entry of enveloped viruses into host cells (Harrison, 2015; Smith & Helenius, 2004; White & Whittaker, 2016). The process of viral fusion involves the conversion of the viral fusion protein from a pre-fusion to a post-fusion conformation (Harrison, 2015). This conformational change ultimately results in fusion of the virus envelope with host membranes and deposition of the virion capsid into the cytoplasm of infected cells. Proper control of virus-induced membrane fusion is essential for efficient virus replication and spread. Mechanisms of membrane fusion regulation include receptor binding, proteolytic processing and pH-dependence (White & Whittaker, 2016). In the simplest system, as with vesicular stomatitis virus (VSV), the fusion protein is the only protein on the viral envelope and mediates both receptor binding and subsequent fusion (Kim et al., 2017). Other viruses, such as herpesviruses, require protein complexes made up of multiple viral proteins to mediate fusion (Eisenberg et al., 2012).

Herpesvirus fusion is a complex, highly coordinated process. All members of the herpesvirus family require a heterodimer of glycoprotein H and glycoprotein L (gH/gL), as well as the only fusogenic glycoprotein B (gB) to mediate membrane fusion (Connolly, Jackson, Jardetzky, & Longnecker, 2011; Eisenberg et al., 2012; Heldwein & Krummenacher, 2008; 5 This chapter, previously published as Rider P, Naderi M, Bergeron S, Chouljenko VN, Brylinski M, Kousoulas KG. 2017. “Cysteines and N-glycosylation sites conserved among all alphaherpesviruses regulate membrane fusion in herpes simplex virus 1 infection.” J Virol. 91(21), DOI: 10.1128/JVI.00873-17, is reprinted here based on ASM journals authors rights. Copyright © 2017, American Society for Microbiology.
Roizman, Knipe, & Whitley, 2006). Additionally, some species of herpesvirus such as herpes simplex virus -1 (HSV-1) and -2 (HSV-2) require the receptor binding glycoprotein D (gD). To facilitate membrane fusion in transient systems, it is sufficient to coexpress gD, gH/gL, and gB (Turner, Bruun, Minson, & Browne, 1998). However, the roles played by individual fusion complex members in mediating membrane fusion are poorly understood. Current evidence suggests that gD binds a receptor and transfers, presumably via a conformational change, an activation signal through gH/gL to gB, which then undergoes a conformational change from the pre-fusion to post-fusion state (Atanasiu et al., 2007).

Regulation of herpesvirus fusion is incompletely understood and findings from transient systems may be misleading. It is clear that in the context of viral infection there are more viral proteins than the minimal fusion complex involved in the fusion process. This is most apparent in HSV-1 mutants that are found to exhibit dysregulated fusion resulting in the formation of syncytia. Mutations causing significant syncytial phenotypes have been found predominantly in HSV-1 genes gB, gK, and UL20 (Baines, Ward, Campadelli-Fiume, & Roizman, 1991; Bond & Person, 1984; Dolter, Ramaswamy, & Holland, 1994; Haffey & Spear, 1980; Pogue-Geile & Spear, 1987). gB is the conserved and only fusogen of all herpesvirus, whereas gK and UL20 are conserved only in neurotropic alphaherpesvirus and are not part of the minimal fusion complex (Lamers et al., 2015). The identification of syncytial virus strains which possess mutations in either gK or UL20 suggests a role for these proteins in the process of alphaherpesvirus fusion. Indeed, early studies demonstrated that transient expression of gK and UL20 with the minimal fusion complex resulted in a decrease in fusion mediated transiently, while syncytial mutations in gK did not increase cell-to-cell fusion (Avitabile, Lombardi, & Campadelli-Fiume, 2003).
gK is a multiple transmembrane domain-containing glycoprotein that is found in the envelope of viral particles (Jambunathan et al., 2015; Neubauer & Osterrieder, 2004). We and others have shown that gK and UL20 form a complex in multiple alphaherpesvirus species (Dietz et al., 2000; Foster, Chouljenko, & Kousoulas, 2008; Guggemoos, Just, & Neubauer, 2006; Haque, Stanfield, & Kousoulas, 2016). Further, we have demonstrated that the gK/UL20 complex interacts with fusion complex members gB and gH/gL (Chouljenko, Iyer, Chowdhury, Kim, & Kousoulas, 2010). Specifically, we showed a direct interaction between the gK amino terminus and gB (Chouljenko, Iyer, Chowdhury, Chouljenko, & Kousoulas, 2009; Chouljenko et al., 2010). Importantly we have shown a role for the amino terminus of gK in HSV-1 neurotropism as an HSV-1 with a deletion in amino acids 31-68 of the amino terminus of gK is unable to enter neurons via the axonal termini both in vitro and in vivo (Jambunathan et al., 2015; Saied, Chouljenko, Subramanian, & Kousoulas, 2014; Stanfield et al., 2014). A virus with this mutation in gK is currently in development as VC2, an HSV-1 vaccine that has been shown to be protective against both HSV-1 and HSV-2 infection in multiple animal models (submitted and (Stanfield et al., 2014)). VC2 has also been demonstrated as highly immunogenic in monkey and mouse models (Liu et al., 2017; Stanfield, Pahar, Chouljenko, Veazey, & Kousoulas, 2017).

To determine the contribution of the extracellular N-linked glycosylation sites located at residues 48 and 59 to the function of HSV-1 gK we constructed three mutant viruses in which each asparagine residue was mutated to alanine independently as well as a double N-linked glycosylation site mutant. We found that virus lacking the N58 or lacking both sites were severely defective for regulation of viral fusion. We extended our analysis to include the amino terminal cysteines of gK. Our results suggest that glycosylation of HSV-1 gK may facilitate
proper gK intramolecular disulfide bond formation and that this disulfide bond formation may be
critical to gK structure and function in regulation of HSV-1 fusion.

7.2 Materials And Methods

Cells

African green monkey kidney cells (Vero) cells were obtained from ATCC (Rockville, MD). The Vero derived gK complementing cell line VK302 was originally obtained from David Johnson (Oregon Health Sciences University, Portland, OR). Both Vero and VK302 cells were
maintained in DMEM supplemented with 10% heat inactivated FBS and antibiotics.

Immunohistochemistry

Vero or VK302 cells were indicated with either VC1 or mutant virus for 48 hours and
fixed with 3.5% Formalin overnight at room temperature. After washing cells were incubated
with anti-HSV-1 antibody (Agilent, Santa Clara, CA) for one hour. Subsequently, Polyclonal
Goat Anti-Rabbit Immunoglobulins conjugated to HRP (Agilent, Santa Clara, CA) were added
for 30 minutes at room temperature. For detection NovaRED peroxidase HRP substrate (Vector,
Burlingame, CA) was applied until appropriate level of staining was obtained. Images were
obtained using an Olympus IX2 inverted microscope and cellSens software (Olympus, Waltham,
MA).

Virus and BAC mutagenesis

Bacterial artificial chromosome (BAC) plasmid pYEbac102 carrying the HSV-1(F)
genome (a gift from Y. Kawaguchi, University of Tokyo, Japan) was used to construct YE102-
VC1 (VC1) as previously described (Jambunathan et al., 2011). VC1 was used to construct all
recombinant BACs described in this study. High-efficiency markerless DNA manipulation of
VC1 was achieved using two step red-mediated recombination (Tischer, von Einem, Kaufer, &
Osterrieder, 2006). Recombinant HSV-1 was recovered after transfecting BACs into Vero cells using Lipofectamine™ according to manufacturer’s protocol. DNA was extracted from viral stocks and gK was sequenced to ensure the presence of the desired mutation.

Immunoprecipitation and western blot assays

Uninfected and infected cells were lysed using NP-40. Lysate was immunoprecipitated with protein G magnetic Dynabeads (Thermo Fisher) bound to mouse V5 antibody (46-1157, Invitrogen) according to the manufacturer's instructions. The protein was eluted from the magnetic beads in 40 µl of elution buffer and used for immunoblot assays. Immunoblot assays were carried out using anti-VP5 (ab6508, Abcam), Rabbit anti-beta tubulin (ab179513, Abcam), goat anti-mouse HRP (Abcam), and goat anti-rabbit HRP (Abcam).

Molecular Visualization

Structure of HSV-1 gK was predicted using assembly of separate domain models (Jambunathan et al., 2015). Molecular visualization was performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, to generate a hybrid surface-cartoon representation of the protein. The N-terminal and the third domain, an extra-cellular loop, are shown in orange ribbon diagrams with a transparent surface and the rest of the protein, including the trans-membrane domain, is shown in blue.

7.3 Results

Analysis of conserved structural elements in the amino terminus of alphaherpesvirus gKs. HSV-1 gK is a four transmembrane domain containing glycoprotein with three extracellular and two intracellular domains (Figure 7.1A). Comparative analysis of the primary structures for 17 alphaherpesvirus gKs (Table 7.1) reveals two conserved N-linked
glycosylation sites in all species, except for the Fruit Bat gK, which possesses only one N-linked glycosylation site (Figure 7.1B). Interestingly, the distance between the N-linked glycosylation sites appears to be nearly conserved among herpesvirus species, which infect similar hosts. This is most evident among the primate and non-human primate alphaherpesviruses in which the N-linked glycosylation sites are all separated by 11 amino acids with the most striking comparison that of between simplex and varicella alphaherpesvirus species. Furthermore, the avian alphaherpesviruses N-linked glycosylation sites are separated by 8 amino acids with the exception of gallid herpesvirus.

In our previously predicted three-dimensional structure of gK (Jambunathan et al., 2015), the N-linked glycosylation sites bracket the second predicted beta-pleated sheet (Figure 7.1C). This structural position of the N-linked glycosylation sites bracketing a beta strand is conserved in several alphaherpesvirus gKs for which we have predicted structures (Figure 7.2).

Construction and characterization of HSV-1 N-linked glycosylation mutants

To determine the role of gK N-linked glycosylation on infectious virus production and virus-induced membrane fusion, we generated three recombinant viruses: two in which either the first asparagine (N48) or the second asparagine (N58) of the predicted N-linked glycosylation sites was replaced with alanine and one recombinant virus having both asparagines replaced with alanines (Figure 7.1). Recombinant viruses were generated using the VC1-F-BAC. VC1-F-BAC is a modified HSV-1(F) strain cloned as a BAC that possesses a V5 epitope-tagged gK and a Flag-tagged UL20 that has been described previously (Jambunathan et al., 2011). BAC-derived virus was reconstituted on gK-complementing cell line VK302. Virus was stocked, titered and used to infect either Vero cells or gK-complementing cell line VK302. Forty-eight
hours post infection cells were fixed and immunohistochemistry (IHC) was performed to identify
HSV positive plaques (Figure 7.3). Readily apparent was the extensive syncytial formation in
Vero cells infected with the mutant virus N58 in which the second N-linked glycosylation (N58-
to-A59 change) was mutagenized and the virus N48N58 having both N-linked glycosylation sites
replaced with alanines. In contrast, plaques from virus in which the first N-linked glycosylation
(N48) was mutagenized were indistinguishable from wild type plaques (Figure 7.3).

Table 7.1. Accession numbers for sequences used in this study.

<table>
<thead>
<tr>
<th>gK Sequence</th>
<th>Accession Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatid herpesvirus 1</td>
<td>YP_003084366.1</td>
</tr>
<tr>
<td>Bovine herpesvirus_1</td>
<td>NP_045305.1</td>
</tr>
<tr>
<td>Cercopithecine herpesvirus_2</td>
<td>YP_164497.1</td>
</tr>
<tr>
<td>Chimpanzee alpha-1 herpesvirus</td>
<td>YP_009011041.1</td>
</tr>
<tr>
<td>Equid herpesvirus_1</td>
<td>YP_053051.1</td>
</tr>
<tr>
<td>Falconid herpesvirus_1</td>
<td>YP_009046554.1</td>
</tr>
<tr>
<td>Felid herpesvirus_1</td>
<td>YP_003331525.1</td>
</tr>
<tr>
<td>Fruit bat alphaherpesvirus_1</td>
<td>YP_009042116.1</td>
</tr>
<tr>
<td>Gallid herpesvirus_1</td>
<td>YP_182337.1</td>
</tr>
<tr>
<td>Human herpesvirus_2</td>
<td>YP_009137206.1</td>
</tr>
<tr>
<td>Human herpesvirus_3</td>
<td>NP_040128.1</td>
</tr>
<tr>
<td>Macacine herpesvirus_1</td>
<td>AIA09548.1</td>
</tr>
<tr>
<td>Meleagrid herpesvirus_1</td>
<td>NP_073348.1</td>
</tr>
<tr>
<td>Papiine herpesvirus_2</td>
<td>YP_443901.1</td>
</tr>
<tr>
<td>Psittacid herpesvirus_1</td>
<td>NP_944381.1</td>
</tr>
<tr>
<td>Saimiriine herpesvirus_1</td>
<td>YP_003933787.1</td>
</tr>
<tr>
<td>Suid herpesvirus_1</td>
<td>YP_068321.1</td>
</tr>
</tbody>
</table>
Figure 7.1. Conservation of alphaherpesvirus gK N-linked glycosylation sites. A) Experimentally determined topology of HSV-1 gK illustrating location of N-linked glycosylation sites (red) and conserved cysteine residues. B) Analysis of N-linked glycosylation site spacing across 17 distinct alphaherpesvirus gKs. C) Ribbon representation of the predicted 3D model of HSV-1 gK. Surface representations of lipid bilayer head groups are shown with disks. N-linked glycosylation sites are indicated by red spheres.
Figure 7.2. Structural predictions of selected simplex and varicella gKs. Cartoon representations of HSV-1 (P68333), HSV-2 (G9I276), and PRV (Q85230) gK protein models are shown in A-C, correspondingly. Fully conserved residues are shown in blue. Conserved Cysteine residues are shown in cyan. Predicted N-glycosylation sites are marked with red arrows.

Figure 7.3. Plaque phenotypes of HSV-1 gK N-linked glycosylation mutants. Vero cells infected with indicated virus for 48 hours were fixed and immunohistochemistry performed. A) Parental virus (VC1). N-linked glycosylation mutants B) N48A, C) N58A, and D) double mutant.
To determine the effect of glycosylation mutagenesis on the gK protein we performed a western blot analysis on protein extracted from cells infected with the parental VC1 virus, or the glycosylation mutant viruses. Vero cells were infected at an M.O.I of 5 for 20 hours, lysed and western blot analysis was performed. Infection of Vero cells with the parental virus yielded a single band of approximately 37 kDa, as previously described (Foster, Rybachuk, & Kousoulas, 2001; Hutchinson et al., 1992) (Figure 7.4A). However, gK from cells infected with either N48A or N58A appeared as two bands, migrating with apparent molecular masses of 33 and 30 kDa. N58A gK was produced consistently at lower levels than parental or N48A gKs.

![Figure 7.4](image-url)  
**Figure 7.4.** Mutation of N-linked glycosylation sites leads to increased mobility of HSV-1 gK protein. Uninfected vero cells (Un) are compared to vero cells infected with the wild type HSV-1 (F), parental strain (VC1) with V5-tagged gK, VC1 derived single N-linked glycosylation mutants (N48A or N58A), and double mutant (Dbl). A) Cells were infected at a multiplicity of infection of 5 and lysed 20 hours post infection. Lysates were separated via SDS-PAGE and analyzed by western blotting. VP5 (UL19) is the major capsid protein of HSV-1. VP5 and Tubulin serve as loading controls. B) Lysates were immunoprecipitated (IP) with V5 antibody and precipitates were separated via SDS-PAGE and analyzed by western blotting.

Finally, the double glycosylation mutant was not detectable via western blotting of protein lysates. After immunoprecipitation using the V5 tag, all gK were detected, including the
gK specified by the double glycosylation mutant, which was now visible as a faint protein species migrating with an apparent molecular mass of 29 kDa (Figure 7.4B).

To determine the effect of gK N-linked glycosylation on virus production, the replication kinetics of mutant viruses were examined. Vero cells were infected at an M.O.I of 0.1 and supernatants were separated from monolayers at 0, 4, 12, 24 and 36 hours post infection. Samples were frozen and titered via plaque assay. All glycosylation site mutant viruses reached similar titers as that of the parental VC1 virus in either the supernatant or monolayers (Figure 7.5). However, the N58 virus and double glycosylation mutant virus displayed approximately 1-log lower titers in both supernatant and cell pellet fractions at 24 hours suggesting a possible delay in envelopment of this virus.

![Figure 7.5](image)

Figure 7.5. Mutation of N-linked glycosylation sites does not affect peak viral titers. Vero cells were infected with the parental strain (VC1), VC1-derived single N-linked glycosylation mutants (N48A or N58A), or double mutant (N48A_N58A). (A) Supernatants and (B) cell pellets were separated at specified hours post infection (hpi) and plaque assays were performed to determine viral titers (plaque forming units, pfu/ml).

It is known that N-linked glycosylation plays a major role in the proper folding and processing of proteins. Specifically, N-linked glycosylation has been demonstrated to facilitate proper disulfide bond formation. In nearly all alphaherpesvirus gKs there are four extracellular cysteines: 3 in domain 1 and 1 in domain 3 (Figure 7.6A). Alignment of gK primary structures similar to the one shown in Figure 7.1B demonstrates the conservation of spacing between
cysteine residues and N-linked glycosylation sites across all human and non-human primate simplex viruses (Figure 7.6B). Remarkably, the spacing between the second and third cysteine residues of domain one is 33 amino acids in all but one alphaherpesvirus species that possess three or more cysteine residues in domain one. Further, in our predicted structure for gK all four extracellular cysteines are predicted to be near enough to one another to potentially form intramolecular disulfide bonds (Figure 7.6C). The predicted proximity of these cysteine residues suggest the existence of disulfide bonds required for the three-dimensional structure of the amino terminus of gK and its function in the regulation of fusion and may be responsible for the syncytial phenotype of the glycosylation mutant virus.

Figure 7.6. Conservation of cysteine residues in the amino terminus of alphaherpesvirus gKs. A) Experimentally determined topology of HSV-1 gK illustrating location of N-glycosylation sites and cysteine residues in HSV-1 gK. B) Analysis of cysteine residue spacing for the amino terminus of 17 alphaherpesvirus gKs. C) Surface representation of the predicted 3D model of HSV-1 gK with transparent amino terminus (orange) featuring close-up view of the cysteine and asparagine residues in the extracellular regions of HSV-1 gK. Colors in B and C match those in panel A.
To address the role of the extracellular cysteine residues of HSV-1 gK, we constructed four mutant viruses, each with a deletion of a single cysteine denoted as dC37, dC82, dC114, and dC243. Virus were recovered on the complementing cell line VK302 and their plaque morphologies were characterized on both Vero and VK302 cells. The dC37 and dC114 viruses formed very small plaques on Vero cells similar to those of the gK-null virus plaques, while the viruses with mutations in either C82 or C243 exhibited a strong syncytial phenotype (Figure 7.7). Importantly, both dC37 and dC114 were efficiently complemented on VK302 cells. gK syn virus have been reported to be complemented (reverted) to wild type plaque phenotype on VK302 cells (Hutchinson et al., 1993). However, both dC82 and d243 retained their syncytial phenotype even in VK302 suggesting that the cysteine mutations are dominant negative mutations.

![Figure 7.7. Plaque phenotypes for HSV-1 gK cysteine mutants. A) Immunohistochemistry of vero cells and gK complementing cells (VK302) infected for 48 hours with parental virus (VC1), or cysteine deletion mutant viruses. B. Topology of HSV-1 gK illustrating putative disulfide bonding between conserved cysteine residues indicated by solid lines.](image-url)
To determine the effect of the cysteine mutations on the level of gK production, western blot analysis was performed as described in Materials and Methods. Loss of any of the cysteine residues did not alter the apparent molecular mass of gK in comparison to the parental VC1 virus gK (Figure 7.8). Overall protein levels of gK were found to be reduced by the loss of individual cysteines. Specifically, loss of C37 and C114 most severely affected gK. While deletion of C82 or C243 did not appreciably affect gK.

![Figure 7.8](image)

**Figure 7.8.** Mutation of HSV-1 gK cysteine residues has limited effect on gK protein levels in infected cells. Uninfected vero cells (Un) are compared to vero cells infected with the parental strain (VC1) with V5-tagged gK or VC1-derived cysteine mutants. Cells infected with the indicated viruses at a multiplicity of infection of 5 were lysed 20 hours post infection. Lysates were separated via SDS-PAGE and analyzed by western blotting. VP5(UL19) is the major capsid protein of HSV-1 and serves as a loading control.

To determine the effect of cysteine mutation on viral growth we conducted growth analysis of all mutants in both Vero and the gK complementing cell line VK302 (Figure 7.9C-D). Supernatants and cell pellets were assayed separately. dC37 and dC114 exhibited the greatest defects in growth. dC37 and dC114 virus titers in the supernatant 36 hpi were four logs lower than the parental VC1 virus titers. dC82 and dC243 viruses had higher peak titers in the supernatant, but the differences between these viruses and dC37 and dC114 were not significant. A similar pattern was seen in cell pellet titers with dC37 and dC114 achieving approximately 3-logs lower peak titers than those of the VC1 parental virus. dC82 and dC243 viruses obtained from cell pellets were significantly lower, but reached higher peak titers than the dC37 and
dC114 viruses. Importantly, all viruses tested were complemented for growth on VK302 cells compared to in Vero cells (Figure 7.9A-B), and titers were indistinguishable from those of the parental VC1 virus at all time points assayed, both in supernatants and cell pellets.

![Graph A](image1)

![Graph B](image2)

![Graph C](image3)

![Graph D](image4)

Figure 7.9. Cysteine mutants are defective for growth in both cell pellet and supernatant fractions. Vero cells (A and B) and gK complementing (VK302) cells (C and D) were infected with parental virus (VC1) and indicated cysteine mutant viruses. Supernatants (A and C) and cell pellets (B and D) were separated at varying times post infection and plaque assays (plaque forming units, pfu/ml) were performed to determine viral titers.

7.4 Discussion

Wild-type HSV-1 strains cause a limited amount of cell-to-cell fusion, while mutations in three genes coding for gB, gK and UL20 cause extensive syncytia formation defining these genes as key players in virus-induced membrane fusion phenomena. All three proteins are membrane proteins found in mature virion particles. gB is the sole herpesviral fusogen, UL20 is a non-glycosylated four transmembrane domain envelope protein, which forms a complex with
gK in all alphaherpesvirus species tested and interacts with carboxyl terminus of gB. gK is the only HSV-1 protein in which most syncytial mutations map to the extracellular portion of the protein and specifically its amino terminus, which is also known to interact with the amino terminus of gB. In this regard, the gK/UL20 protein complex appears to regulate gB’s fusogenicity through direct binding to gB at both its extracellular and intracellular domains. Herein, we have explored the role of gK’s amino terminal N glycosylation and the conserved cysteine residues in infectious virus production and membrane fusion by generating single and double mutant viruses. The results indicated that at least one of the two N glycosylation site and two of the four conserved cysteine residues play crucial roles in the function of gK in infectious virus production and membrane fusion.

Glycoproteins embedded in viral envelopes serve important roles for entry, assembly, immune evasion, signaling and pathogenesis (Vigerust & Shepherd, 2007). Specifically, N-linked glycosylation of the influenza virus hemagglutinin and neuraminidase has been shown to influence receptor binding and virulence (Alymova et al., 2016; Li, Schulman, Itamura, & Palese, 1993; Liao et al., 2010; Reading et al., 2009; Zhao et al., 2017). A major focus of work in glycobiology is the role of N-linked glycosylation in protein processing (Breitling & Aebi, 2013; Ellgaard, McCaul, Chatsisvili, & Braakman, 2016; Moremen, Tiemeyer, & Nairn, 2012; Xu & Ng, 2015). N-linked glycosylation takes place in the endoplasmic reticulum and further modifications to glycans occur in the Golgi network. However, initial glycosylation in the endoplasmic reticulum occurs cotranslationally, which is thought to be critical for proper folding of nascent proteins. Seminal work from the Helenius laboratory demonstrated a critical role for N-linked glycosylation in viral glycoprotein folding (Hammond, Braakman, & Helenius, 1994), and it has since been shown that the positioning of glycosylation sites are important for optimal
protein expression (Daniels, Kurowski, Johnson, & Hebert, 2003). N-linked glycans are critical for directing proper folding of viral and cellular proteins and proper folding is important for correct disulfide bond formation (Ellgaard et al., 2016; Ishmael, Ishmael, Jones, & Bond, 2006; Qin, Wang, & Thirumalai, 2015; Zhu et al., 2015). Not surprisingly, due to the importance of glycosylation to protein processing mutations in glycosylation have profound effects on virion assembly. N-linked glycosylation sites have been shown to be important for proper localization and subsequent incorporation of Ebola virus envelope glycoprotein (GP) into pseudovirions and Lassa virus glycoprotein GP-C (Eichler, Lenz, Garten, & Strecker, 2006; Wang et al., 2017). N-linked glycosylation of West Nile and Zika virus envelope proteins influenced assembly and infectivity in a cell type specific manner (Hanna et al., 2005; Mossenta, Marchese, Poggianella, Slon Campos, & Burrone, 2017; Moudy, Zhang, Shi, & Kramer, 2009; Shirato et al., 2004). Mutations in VSV G protein that affect N-linked glycosylation sites were reported to lead to improper disulfide bonding affecting transport and maturation of the viral fusogen (Machamer & Rose, 1988).

The role of N-glycosylation on infectious virus produced and virus-induced cell fusion

A striking feature of N-glycosylation sites of gK is the largely conserved spacing of these sites among most alphaherpesviruses suggesting a conserved role in the overall structure and function of gK. N glycosylation may contribute to preservation of the overall structure of the amino terminus of gK, facilitate interactions of gK with the amino terminus of gB, or potentially with other viral and cellular proteins. Mutations in gK leading to syncytia formation suggest gK plays a role in the negative regulation of fusion. Indeed, in transient assays expression of HSV-1 gB, gD, gH/gL along with UL20 and gK led to a decrease in the amount of fusion, while gK carrying a syncytial mutation did not have any effect on membrane fusion (Avitabile et al.,
It is possible that the N glycosylation moieties facilitate interaction with gB and are involved in gB-mediated fusion. Our mutagenesis analysis indicates that the second glycosylation site at residue 59 plays a crucial role, since when it is changed the resultant virus causes extensive cell fusion, while the N-glycosylation site at residue 48 did not appear to produce a substantial effect in either virus production or membrane fusion. Therefore, the N48 glycosylation site appears to play a secondary role, potentially involved in other unknown functions of gK. The apparent conservation of the distance between N-linked glycosylation sites in the amino terminus of alphaherpesvirus gKs suggest that this domain bracketed by the N glycosylation sites may play a role in binding a host protein. Alternatively, the use of N-linked glycosylation to evade humoral immunity is well documented (Wei et al., 2003; Zhou et al., 2017). This hypothesis is supported by the fact that gK glycosylation mutants grew to wild type levels. In addition, the VC2 live-attenuated vaccine strain, which has both N glycosylation sites deleted, has been shown to generate robust immune responses in guinea pigs and mice that are superior to those produced by its parental wild-type virus HSV-1 (F)((Liu et al., 2017; Stanfield et al., 2017; Stanfield et al., 2014), and unpublished observations). A question raised by the significant decrease in gK protein detected in cells infected with the double glycosylation mutant is how with such low levels of gK does the virus replicate essentially to wild type titers both in the supernatant and cell pellet fractions? We would suggest that in our protein preparations gK lacking glycosylation may not be as soluble in protein preparations as those from cells infected with virus in which gK is glycosylated. In support of this our attempts to synthesize the amino terminus of gK in bacterial systems yield very little soluble protein.
Role of cysteine residues in gK infectious virus production and virus-induced cell fusion

Syncytial cysteine mutants were defective for growth in cell pellet fractions and little infectious virus was recovered from the supernatant fraction. Therefore, we hypothesize that these cysteine residues are involved in the overall structure of the amino terminus of gK through disulfide bond formation. This allows the consideration that mutagenesis of the N58 glycosylation site that produces a syncytial phenotype may displace formation of the appropriate disulfide bond involving the cysteine residues that appear to also be involved in membrane fusion.

We have reported previously that virus with a deletion in amino acids 31-47 of HSV-1 gK, which includes the first cysteine, exhibits a gK null phenotype (Chouljenko et al., 2009). A virus in which the entire amino terminus of gK was deleted was similarly defective. However, virus with a deletion spanning amino acids 31-68 of HSV-1 gK, which includes the first cysteine as well as both N-linked glycosylation sites, exhibits only a one-log defect in peak titers and is not syncytial. These results suggest a potentially important relationship between glycosylation and disulfide bond formation in the amino terminus of gK. We hypothesize that C38 and C115 residues are critical to achieving proper structural conformation of the amino terminus of gK. The inability to achieve the proper structure may interfere with gK’s transit or function in some manner that results in a null phenotype. Interestingly mutation of the glycosylation sites leads to a syncytial phenotype, while deletion of amino acids 31-68 that contains both N-glycosylation sites does not. This suggest that the gK domain that is bracketed and includes both glycosylation sites may be dispensable for infectious virus production, but necessary for gK regulation of gB-mediated cell fusion, since syncytial mutations in gB fail to cause fusion in the presence of the gKD31-68 mutation (Chouljenko et al., 2009). Our syncytial gK mutants appear to be dominant
negative, since syncytial mutants in this study produced syncytial plaques on the gK complementing cell line VK302.

An intriguing possibility suggested by the role of disulfide bond reduction in the conversion of many viral fusion proteins from a pre-to-post fusion conformation is that the probable disulfide linkage between cysteines 82 and 243 of gK acts as a molecular switch for gB mediated fusion. The generation of free thiols by reduction of disulfide bonds has been shown to be important for fusion protein activity in multiple virus families (Stolf et al., 2011). Protein disulfide isomerases (PDI) are cellular enzymes that catalyze the breaking and formation of disulfide bonds (Wilkinson & Gilbert, 2004). For Newcastle disease virus (NDV) overexpression of PDI increased conversion of F protein to the post-fusion conformation and enhanced fusogenicity of NDV (Jain, McGinnes, & Morrison, 2008). Additionally, entry of HIV-1 requires disulfide exchange (Matthias et al., 2002) and inhibition of thiol–reactive proteins prevents entry of HIV (Ou & Silver, 2006; Ryser, Levy, Mandel, & DiSciullo, 1994). It will be interesting to pursue the role of thiol-reactive proteins in HSV-1 entry and gK function in future experiments.

7.5 Conclusions

In this chapter, we describe a role for the conserved N-linked glycosylation of the amino terminus of gK in the regulation of HSV-1 fusion. We have also described a role for the conserved cysteines in the HSV-1 fusion. The experimental results supporting the presence of disulfide bond formations are used to refine the computationally predicted model for gK (Figure 7.9). The role of these conserved structural features of gK in the entry of other alphaherpesvirus is currently under study. Specifically, how these mutations affect gB-mediated fusion, as well as the contribution of these structural features to the role of
alphaherpesvirus entry into neurons will be the focus of future research. This knowledge will enable the rational design of vaccination approaches to many significant human and non-human animal pathogens.

Figure 7.10. Refinement to gK model. Remodeled gK protein structure incorporates disulfide bonds between C37-C114 and C82-C243. The N-terminal domains of gK protein models before and after refinement are superimposed in gold and blue, respectively. Cysteine and N-glycosylation sites Asparagine residues are shown as sticks.
7.6 References


8. HSV-1 UL37 PROTEIN MODEL

8.1 Introduction

Herpes simplex virus type-1 (HSV-1) is the prototypic member of the alphaherpesvirinae (Davison et al., 2009). HSV-1 possesses a ~152 kb double-stranded DNA genome enclosed within an icosahedral capsid composed of 12 pentavalent and 150 hexavalent capsomeres (Zhou et al., 2000). The capsid is coated with a layer of viral proteins called the tegument, which is enclosed by a lipid envelope originating from host cellular membranes that is enriched with viral glycoproteins and other membrane-associated proteins. A variety of interactions between the viral proteins that make up the capsid, tegument and envelope are responsible for maintaining the structural integrity of mature virions and for shepherding the viral particle through the complex cytoplasmic envelopment process at trans-Golgi network-derived vesicles and endosomes (Johnson & Baines, 2011; Mettenleiter, 2004).

The UL37 gene of HSV-1 encodes a large 1123 aa (apprx. 120 kD) highly conserved tegument protein that is essential for viral growth in cell culture and is crucial for viral assembly and secondary envelopment in the cytoplasm (Bucks, Murphy, O'Regan, & Courtney, 2011; Kelly, Fraefel, Cunningham, & Diefenbach, 2009; Mettenleiter, 2002). The amount of UL37 protein (pUL37) incorporated onto mature virions is tightly regulated, since overexpression of pUL37 does not increase the pUL37 amount incorporated into virions (McLauchlan, 1997). pUL37 forms a complex with pUL36, and this interaction is conserved across all three

---

6 This chapter, published as Chouljenko DV, Jambunathan N, Chouljenko VN, Naderi M, Brylinski M, Caskey JR, Kousoulas KG. 2016. “Herpes simplex virus type 1 UL37 protein tyrosine residues conserved among all alphaherpesviruses are required for interactions with glycoprotein K (gK), cytoplasmic virion envelopment, and infectious virus production.” J Virol. 90(22):10351-10361. DOI: 10.1128/JVI.01202-16 Copyright © 2017, American Society for Microbiology
herpesviridae subfamilies. Interactions between proteins homologous to pUL36 and pUL37 have been documented in the alphaherpesviruses HSV-1, pseudorabies virus (PRV) and varicella zoster virus (VZV), in the betaherpesvirus human cytomegalovirus (HCMV) and in the gammaherpesvirus Kaposi’s sarcoma-associated herpesvirus (KSHV) (Bechtel & Shenk, 2002; Klupp, Fuchs, Granzow, Nixdorf, & Mettenleiter, 2002; Lee, Vittone, Diefenbach, Cunningham, & Diefenbach, 2008; Mijatov, Cunningham, & Diefenbach, 2007; Rozen, Sathish, Li, & Yuan, 2008; Uetz et al., 2006; Vittone et al., 2005). The presence of pUL36 is necessary for incorporation of pUL37 onto capsids (Ko, Cunningham, & Diefenbach, 2010). pUL37 is likely added to capsids after pUL36, since pUL36 is still detected on both HSV-1 and PRV capsids in mutants lacking pUL37 (P. Desai, Sexton, McCaffery, & Person, 2001; Klupp, Granzow, Mundt, & Mettenleiter, 2001). Along with pUL36, pUL37 may be involved in the organization of tegument structure. pUL37 attaches to capsid-bound pUL36 at the vertices, together forming thin flexible strands ranging from 15 to 70 nm in length that extend throughout the tegument, possibly providing a scaffold for the rest of the tegument (Newcomb & Brown, 2010). Deletion of either pUL36 or pUL37 prevents the acquisition of appreciable amounts of tegument in the cytoplasm and blocks cytoplasmic envelopment in HSV-1, resulting in cytoplasmic accumulation of unenveloped capsids (P. Desai et al., 2001; P. J. Desai, 2000; Roberts et al., 2009).

The UL37 protein encodes multiple functional domains. Co-immunoprecipitation experiments revealed that pUL37 domains spanning residues 1-300 and 568-1123 are involved in self-association in the absence of its binding partner pUL36 (Bucks et al., 2011). The pUL37 amino terminus contains an alanine-rich region (ARR) spanning residues 44-80, a leucine zipper motif covering residues 203-224 and a leucine-rich nuclear export signal (NES) encompassing
residues 263-272 (Bucks et al., 2011; Watanabe et al., 2000). The carboxyl terminus contains a domain spanning residues 1099-1104 involved in binding TNF receptor-associated factor 6 (TRAF6) to activate NF-κB pathway signaling (Bucks et al., 2011; Liu, Fitzgerald, Kurt-Jones, Finberg, & Knipe, 2008). The C-terminal 578-899 aa of pUL37 can interact with a spectraplakin protein called dystonin/BPAG1 known as a cytoskeletal cross-linker that is involved in microtubule stabilization and transport. Viral replication and cytoplasmic capsid mobility during egress from infected cells is impaired in dystonin-depleted cells, suggesting that pUL37 may play a role in capsid trafficking along microtubules (Pasdeloup, McElwee, Beilstein, Labetoulle, & Rixon, 2013). The C-terminus of pUL37 is also responsible for binding to pUL36 (Bucks et al., 2011; Kelly, Mijatov, Fraefel, Cunningham, & Diefenbach, 2012). Scanning alanine mutagenesis of pUL37 revealed that residue D631 of pUL37 mediates binding to pUL36. Altering this residue resulted in significantly decreased ability of the virus to replicate, with mutant viral titers approximately 2 logs lower than those of wild type virus (Kelly et al., 2012). A plasmid encoding the C-terminal portion of pUL37 spanning residues 568-1123 that includes the putative pUL36 interaction site partially rescued a UL37-null virus indicating that the carboxyl terminus of UL37 is particularly important for infectious virus production (Bucks et al., 2011).

The function of the central portion of pUL37 spanning aa 301-567 is not well defined. A mutant HSV-1 with a 12 aa protein C epitope tag inserted in-frame immediately after residue Y480 of pUL37 exhibited a severe defect in cytoplasmic envelopment, and surprisingly was partially complemented for replication and spread when grown on cells expressing pUL20 (Jambunathan et al., 2014). The inserted protein C (protC) epitope tag may directly disrupt protein-protein interactions mediated by adjacent residues, or even trigger a conformational
change in the pUL37 protein that could affect binding to other proteins (Jambunathan et al., 2014). Phosphorylation is a widespread form of posttranslational modification that can affect a multitude of protein functions, including modulation of protein-protein interactions and control of intracellular trafficking (Albright & Jenkins, 1993; Hunter & Karin, 1992). Phosphorylation of proteins such as the p53 tumor suppressor has been shown to mediate conformational changes that can affect protein function and regulation (Ashcroft, Kubbutat, & Vousden, 1999). Many viral proteins are also phosphorylated, by either viral or cellular kinases. The HSV-1 tegument includes at least three components that are protein kinases, encoded by the UL13, UL23 and US3 genes (Kelly et al., 2009; Mettenleiter, 2002). The UL37 protein is expressed late in the infection cycle and has been reportedly to be phosphorylated soon after translation of the UL37 gene. Phosphorylation of pUL37 is thought to be performed by a cellular kinase and is not dependent on the presence of any known HSV-1 binding partner because pUL37 expressed by a recombinant vaccinia virus has also been observed to be phosphorylated (Albright & Jenkins, 1993). It is not known whether the UL37 phosphorylation plays an important role in the virus lifecycle.

We have previously reported that pUL37 physically interacts with the membrane proteins pUL20 and gK, although the exact locations of the relevant gK and pUL37 binding sites remain unknown (Jambunathan et al., 2014). pUL20 and gK act as modulators of virus-induced fusion and interact with each other in addition to binding the major HSV-1 fusion protein gB, and pUL20 has also recently been shown to interact with gM (V. N. Chouljenko, Iyer, Chowdhury, Kim, & Kousoulas, 2010; Foster, Chouljenko, & Kousoulas, 2008; Foster, Melancon, Baines, & Kousoulas, 2004; Kim, Chouljenko, Walker, & Kousoulas, 2013; Melancon, Luna, Foster, & Kousoulas, 2005). In addition to their role in fusion, pUL20 and gK are required for secondary
envelopment, and their interactions with pUL37 may serve to facilitate the process of cytoplasmic virion envelopment (D. V. Chouljenko et al., 2012; Hutchinson & Johnson, 1995; Jambunathan et al., 2014; Jayachandra, Baghian, & Kousoulas, 1997).

We sought to better define the function of the central portion of HSV-1 pUL37 by performing scanning alanine mutagenesis of highly conserved amino acid residues. Specifically, we targeted residues likely to be exposed on the protein surface as predicted by computational 3-dimensional modeling based on the recently published crystal structure of the closely related PRV pUL37 amino terminus (Pitts, Klabis, Richards, Smith, & Heldwein, 2014). Previously, we reported that the mutant DC480 virus engineered to have a protC epitope tag inserted immediately after the Y480 residue exhibited a UL37-null-like defect for virus replication and cytoplasmic virion envelopment (Jambunathan et al., 2014). In addition, we constructed and characterized the DC447 virus having the 12-aa protC epitope tag inserted 100bp downstream of the Y480 residue, as well as mutant viruses that carried tyrosine-to-alanine changes within the central portion of the UL37 protein. Collectively, our results show that UL37 tyrosines conserved among alphaherpesviruses and especially tyrosines Y474 and Y480 are involved in UL37 interactions with gK and play crucial roles in cytoplasmic virion envelopment.

8.2 Materials and Methods

Cells and Viruses

African green monkey kidney (Vero) cells were obtained from the American Type Culture Collection (Rockville, MD). The Vero-based UL37-complementing cell line BD45 was a gift from Prashant Desai (Johns Hopkins University, Baltimore, MD). All cells were maintained in Dulbecco's modified Eagle's medium (Gibco-BRL, Grand Island, NY) supplemented with 10% fetal calf serum and antibiotics.
Plasmid and mutant virus construction

All mutations within the UL37 gene sequence were performed using the GeneTailor Site-Directed Mutagenesis System Kit from Invitrogen according to the kit manufacturers’ instructions. Construction of viral mutants with specific amino acid changes was accomplished in *Escherichia coli* by using the markerless two-step Red recombination mutagenesis system and synthetic oligonucleotides implemented on the bacterial artificial chromosome (BAC) plasmid pYEbac102 carrying the HSV-1(F) genome (a gift from Y. Kawaguchi, University of Tokyo, Japan). Construction of the HSV-1 mutant virus DC480 which has a 12-amino acid protein C epitope tag inserted immediately after amino acid 480 of HSV-1 UL37 was described previously (Jambunathan et al., 2014). The recombinant mutant virus YE102-VC1 was modified to express gK and UL20 genes containing V5 and 3×FLAG antigenic epitopes, respectively, and was described previously (Jambunathan et al., 2011). Double Red recombination was also used to construct the mutant DC474-480, which has the four conserved tyrosines at positions 474,476,477 and 480 of HSV-1 UL37 changed to alanine, as well as single (Y-to-A), amino acid changes at amino acid positions 476 and 477. All mutated DNA regions were sequenced to verify the presence of the desired mutations in BACs and the absence of any other spurious mutations on the viral genome.

Replication kinetics

Viral growth kinetics was done essentially as we have described earlier. Briefly, viruses were adsorbed on nearly confluent monolayers of Vero, or BD45 cell lines into each well of a six well plate at 4°C for 1hr. The cells were infected with multiplicity of infections (MOI) of 2. Thereafter the plates were incubated at 37°C with 5% CO2 for 1hr for viral penetration. Any unbound viruses were washed by treating with low pH buffer (pH 3.0) and the infection was
allowed to proceed for 0, 6, 12, 24 and 48 hpi. The virus titers were averaged and standard deviation was calculated for each time point.

Transfection-infection complementation assay

Vero cells were transfected with pUL37 constructs (genes cloned into pcDNA 3.3 vector, Invitrogen), UL37-PL262/263-AA, UL37-F294-A, UL37-P408-A, UL37-GF420/421-AA, UL37-Y474-A, UL37-Y480-A, UL37-P519-A, UL37-F620-A and UL37-P729-A with the noted residues altered to alanine. After 36 hours, transfected Vero cells were infected with the UL37-null virus at an MOI of 5. Infected cells were harvested at 24 hpi, and titration was performed using the UL37-complementing cell line BD45. A plasmid expressing pUL20 was used as the negative control. A plasmid expressing wild-type UL37 was used as a positive control. A second positive control consisted of mock transfection, followed by infection with the UL37-null or wild-type HSV-1(F) viruses.

Immunoprecipitation and immunoblot assays

Confluent Vero cells in T75 flasks were infected with the double-tagged recombinant virus YE102-VC1 (gK-V5 and UL20-FLAG), DC474-480 or F strain virus at an MOI of 2. At 24 hpi, the infected cells were lysed with NP-40 cell lysis buffer (Life Technologies) supplemented with protease inhibitor tablets (Roche). The samples were centrifuged at 13,000 rpm for 10 min at 4°C. The supernatants were then used for immunoprecipitation. The proteins from virus-infected cells were immunoprecipitated using protein G magnetic Dynabeads according to the manufacturer's instructions (Invitrogen). Briefly, the beads were bound to their respective antibodies and left on a nutator for 10 min, followed by the addition of cell lysates. The lysate-bead mixture was kept on the nutator for 10 min at room temperature and subsequently washed three times with phosphate-buffered saline (PBS). The protein was eluted from the magnetic
beads in 40 µl of elution buffer and used for immunoblot assays. Sample buffer containing 5% β-
mercaptoethanol was added to the protein and heated at 55°C for 15 min. Proteins were resolved
in a 4 to 20% SDS-PAGE gel and immobilized on nitrocellulose membranes. Immunoblot assays
were carried out using monoclonal mouse anti-FLAG antibody (Sigma-Aldrich, Inc., St. Louis,
MO), monoclonal mouse anti-V5 antibody (Invitrogen), mouse monoclonal anti-VP5 antibody,
horseradish peroxidase (HRP)-conjugated goat anti-mouse antibodies (Abcam, Inc., Cambridge,
MA), polyclonal rabbit anti-UL37 antibody (a gift from Frank J. Jenkins, University of
Pittsburgh Cancer Institute), and HRP-conjugated goat anti-rabbit antibody (Abcam, Inc.,
Cambridge, MA).

Structure modeling of the HSV-1 UL37 protein

Template-based modeling of the first 570 amino acids of the HSV-1 UL37 protein was
conducted by homology modeling with Clustal Omega (Thompson, Higgins, & Gibson, 1994)
and Modeller (Fiser & Sali, 2003). The X-ray crystal structure of PRV UL37 (PDB code:
4K70)(Pitts et al., 2014) is used as the modeling template. The amino acid sequence of HSV-1
UL37 protein (NCBI RefSeq: YP_009137112.1) is used for a multiple sequence alignment
(MSA) between 20 different strains of alphaherpesviruses including HSV-2 and PRV. The MSA
was constructed with Clustal Omega sequence profile alignment (Thompson et al., 1994) server
provided by Max-Planck Institute (http://toolkit.tuebingen.mpg.de/clustalw); The default
options with the modification of 5x HMM are used to run Clustal Omega. After removing the
first 29 residues, the first 541 amino acids from the sequence are used to build the homology
model using Modeller version 9.15. HSV-1 UL37 protein model is visualized using molecular
visualization package PyMol (The PyMol Molecular Graphics System, Version 1.2r3pre,
Schrödinger, LLC.). An alignment of HSV-2 UL37 (YP_009137112.1; AKC59563.1), with the
beta-herpesvirus HCMV UL37 (Merlin; YP_081505.1), and the gamma-herpesvirus EBV (YP_001129450.1) was performed using the same parameters and software described above for the alphaherpesviruses UL37 proteins.

Generating sequence logos

Sequence logos were generated using the web-based application WebLogo (Crooks, Hon, Chandonia, & Brenner, 2004) (http://weblogo.berkeley.edu/logo.cgi) with the default settings. Seven different Herpes Virus alignments from the MSA mentioned above are selected to generate the sequence logo. The names of the selected viruses and their accession numbers are as follows: (Human herpesvirus 1, YP_009137112), (Human herpesvirus 2, AKC59563), (Suid herpesvirus 1, AEM64124), (Cercopithecine herpesvirus 2, YP_164480), (Bovine herpesvirus 5, YP_003662487), (Bovine herpesvirus 1, NP_045321), (Anatid herpesvirus 1, AFC61845).

In silico Alanine scanning

Amino acid residues in the UL37 homology model are mutated into Alanine one at a time by AlaScan feature in the protein design package FoldX Suite downloaded in 2016 (http://foldxsuite.crg.eu/); the difference in the free Gibbs energy change ($\Delta G$) before and after performing each mutation is calculated by FoldX Suite (Schymkowitz et al., 2005). The numbers are reported as deviations from the median; a positive $\Delta G$ shows the importance of an amino acid residue in the structure stability or in contributing to protein interaction interface.

Electron microscopy

Sub-confluent layer of Vero cells in 6 well plates were infected with the wild type VC1 and the UL37 mutant 474-480 at an MOI of 3. After 1hr of infection the virus was removed and fresh medium was added and the infection was allowed for 16hrs at 37°C and 5% CO2. 16hpi,
cells were fixed with Glutaraldehyde fixative and processed for electron microscopy as described previously (Jambunathan et al., 2014).

8.3 Results

Delineation of UL37 functional domains

To delineate functional domains of the UL37 protein involved in infectious virus production, a cadre of UL37 gene mutations were generated by PCR-assisted mutagenesis targeting residues conserved among alphaherpesviruses. Alignment of a number of UL37 amino acid sequences specified by different alphaherpesviruses revealed that amino acids proline (262), leucine (263), phenylalanine (294), proline (408), glycine (420), phenyalananine (421), tyrosines (474 and 480), proline (519), phenylananine (620) and proline (729) exhibited a high degree of conservation (Figure 8.1). These residues were mutated to alanine residues as single or double mutations in instances where the two conserved amino acids were located adjacent to each other generating the mutant plasmids UL37-PL262/263-AA (adjacent proline and leucine residues changes to alanines), UL37-F294-A (phenylalanine at amino acid position 294 changed to alanine), UL37-P408-A (proline changed to alanine), UL37-GF420/421-AA (glycine and phenylalanine changed to alanines), UL37-P519-A (proline changed to alanine), UL37-F620-A (phenylalanine changed to alanine) and UL37-P729-A (proline changed to alanine), UL37-Y474-A and UL37 Y480-A (tyrosines changed to alanines). Plasmids expressing the mutant UL37 proteins were tested for their ability to complement a UL37-null virus by transfecting each plasmid into Vero cells followed by infection with the UL37-null virus grown in the UL37 complementing cell line BD45 (see Materials and Methods). All plasmids complemented the UL37-null virus replication and spread to variable extents except plasmids that specified the UL37-Y474-A, and UL37-Y480-A mutations, as evidenced by the appearance of substantially
larger viral plaques in comparison to the mock-transfected cells and the resultant accumulation of infectious virions (Figure 8.2: A, B, C).

Figure 8.1. Multiple sequence alignment of twenty alphaherpesvirus-encoded UL37 proteins. Sequence alignment of the UL37 protein from twenty different alphaherpesviruses encompassing the amino terminal half of UL37 corresponding to the portion of PRV UL37 with a published crystal structure (Pitts et al., 2014). The complete sequence alignment of the entire UL37 protein is included in the JVI published paper of this chapter (Jambunathan et al., 2016). Amino acid residues are colored according to their side chain chemistry. Aminoacids are colored by their sidechain properties; Yellow: Aromatic (F, W, Y), Red: Acidic (D, E), Blue: Basic (R, H, K), Orange: Nonpolar (A, G, I, L, M, P), Green: Polar (C, N, Q, S, T). Red arrowheads indicate the highly conserved residues that were mutated to alanines.

To visualize whether tyrosine residues identified to play important roles in the structure and function of the UL37 protein are exposed on the UL37 surface where they would be available to participate in protein-protein interactions, we generated the three-dimensional structure of the HSV-1 UL37 protein via homology modeling based on the x-ray crystal structure of the PRV UL37 protein (see Materials and Methods). The Y474 and Y480 residues are more conserved compared to Y476 and Y477, suggesting that they are important for UL37 structural
stability and function. A closer examination of the UL37 structure reveals that Y474 and Y480 are located in a pocket-like structure together with three other highly conserved residues, P408, F484 and P519, forming a helix-turn-helix motif. The position of aromatic rings in Y476, Y477 and F484 implies the possibility of noncovalent pi stacking interactions. Further, the presence of two highly conserved prolines, P408 and P519, within 6 Å of these aromatic residues suggests the possibility of C-H-pi interactions. Altogether, these putative interactions may have an essential role in stability of the protein, especially in preserving the pocket like structure which might be important in UL37 interaction with other proteins (Fig. 3B).

Figure 8.2. Delineation of UL37 amino acids involved in infectious virus production. A) Schematic of the UL37 protein depicting known functional domains and the approximate location of nine single and double amino acid replacements constructed to assess their role in infectious virus production. B) Images of viral plaques produced in the transfection-infection complementation of the UL37-null virus. Vero cells were transfected with each plasmid. After 36 hours, plasmid transfected cells were infected with UL37-null virus at an MOI of 5 and virus stocks were collected 24 hpi. Viral plaques were visualized by immunohistochemistry at 48 hpi. (see Materials and Methods). The plasmids where mutations in the amino acids of UL37 were able to complement the UL37 null virus is marked with green ‘+’ and the mutations which were not able to complement is marked by a red ‘-‘. C) Viral titers obtained in the transfection-
infection complementation assay. Virus stocks collected at 24 hpi with the UL37-null virus were titered on the BD45 cell line, which expresses the UL37 protein in trans. A plasmid expressing the wild-type UL37 protein was used as a positive control. A second positive control consisted of mock transfection, followed by infection with wild-type HSV-1, while a UL20 expressing plasmid was used as a negative control.

Structure modeling of the HSV-1 UL37 protein

To further determine the role of each tyrosine on the UL37 structure, the relative importance of the conserved tyrosines at positions 474, 476, 477 and 480 on the structural stability of the UL37 proteins were analyzed via computational alanine scanning. In this analysis the differences in the calculated free Gibbs energy change (DG) were measured between wild-type and each of the mutated proteins in order to quantify the effect of tyrosine replacement to alanine (see Materials and Methods). These differential energy calculations revealed that the Y276-A, Y474-A, Y480-A, and Y563-A mutations resulted in positive (DG) suggesting that they are important for the stability of the UL37 structure and/or the ability of the UL37 protein to interact with other proteins (Fig. 3C).

Construction and characterization of UL37 mutant viruses

The mutant virus DC480 contains a protein C (protC) epitope tag inserted into the HSV-1 UL37 gene immediately after the codon for amino acid 480 in the HSV-1(F) genetic background has been described previously (Jambunathan et al., 2014). To further assess the function of this central portion of the UL37 protein in virus replication, we constructed the mutant virus DC447 by inserting the ProtC epitope tag immediately after amino acid codon 447. In addition, we constructed mutant viruses DC476, DC477 and DC474-480 in which tyrosines at either individual positions 476, 477 or all four (474, 476, 477, 480) were changed to alanine residues,
respectively. These mutations were constructed in the VC1 genomic background expressing gK tagged with a V5 epitope and UL20 tagged with a FLAG epitope (Fig. 4A).

Figure 8.3. Predicted structure and dynamics of the HSV-1 UL37 protein. A) Space-filling predictive model of the UL37 amino terminal 570 amino acids of HSV-1 based on the known crystal structure of a corresponding amino terminal region of PRV UL37. Exposed residues are represented with unique colors: Y474, Y476, Y477, Y480 (blue), PL262-263 (orange), P408 (red), P519 (green), GF419-420 (magenta); Top portion of the panel (B). Sequence Logos show Y474 and Y480 are more conserved compared to Y476 and Y477, indicating their relative importance in protein stability and function. Other highly conserved residues within 6 Å of these tyrosines are P408 and P519. Within this range, C-H-pi interactions between a proline and an aromatic residue (Y, P or W) is expected. Sequence Logo also shows that in the other strains in position 484, the F residue is replaced by a Y, another aromatic residue. Bottom portion of the panel (B). The Y480 predicted location is within a helix-turn-helix amino acid domain. Approximate distances between the aromatic residues Y474, Y480 and F484 and the P408 and P519 are shown. C) In silico alanine-scanning mutagenesis. Relative energy changes of the protein after mutating each Y to A, one at a time in-silico. The difference between the free Gibbs energy before and after the mutation is reported on the Y-axis. Mutations that will disrupt the UL37 structural stability are expected to have a less negative ΔG. Y474 and Y480 exhibit the most positive ΔG, indicating that these Y-A mutations will have pronounced effects on the overall structure of the UL37 protein.

The DC480 mutant failed to replicate producing viral titers similar to the UL37-null virus
characterized by approximately 3 logs less infectious virus production than the parental HSV-1(F) virus, as we have reported previously (Jambunathan et al., 2014). In contrast, the DC447 virus replicated substantially better than the DC480 and UL37-null viruses, however final infectious virus production was more than one log less than the parental HSV-1(F) virus (Figure 8.4B). The DC476, DC477 viruses, having each tyrosine (476, 477) mutated to alanine, produced titers similar to the VC1 virus, and the DC474-480, having all four tyrosines (474, 476, 477, 480) mutated to alanines, produced viral titers similar to the UL37-null virus (Fig. 4D). All mutant viruses that exhibited a replication defect in Vero cells were efficiently complemented when grown in the BD45 cell line that expresses the UL37 gene in trans (Figure 8.4C, and 4E).

We have previously reported that the DC480 virus produced a UL37-null phenotype that resulted in accumulation of capsids in the cytoplasm of Vero Infected cells in a manner reminiscent to the gK-null and UL20-null virions indicating a defect in cytoplasmic virion envelopment (Jambunathan et al., 2014). Transmission electron microscopic examination of Vero cells infected with the DC474-480 virus revealed a similar defect characterized by accumulation of viral capsids in the cytoplasm and absence of infectious virions in the periphery of Vero infected cells in comparison to the parental VC1 virus (Figure 8.5).
Figure 8.4. Construction and characterization of UL37 mutant viruses. A) Schematic representation of the UL37 mutant viruses constructed on the viral genomic background via double-Red mutagenesis on the HSV-1 (F) and VC1 genomes cloned as a bacterial artificial chromosome. The approximate location of the constructed mutations is shown. B) Replication kinetics of wild-type and selected mutant viruses on Vero and BD45 cells. Viruses were adsorbed on nearly confluent monolayers of Vero, or BD45 cells. Cells were infected with each virus at multiplicity of infection (MOI) of 2. Virus stocks were collected at the indicated times and titered on both Vero and BD45 cells. Virus titers from three independent cultures were averaged and standard deviation was calculated for each time point.
Figure 8.5. Electron micrographs of Vero cells infected with VC1 or DC474-480 viruses. Vero cells were infected with either VC1 or DC474-480 viruses at an MOI of 5 and visualized by electron microscopy after 24 hpi. Arrowheads indicate the presence enveloped virions in the periphery of VC1-infected cells and capsids within the cytoplasm of DC474-480-infected cells. The cytoplasm (CYT) and extracellular space (ECS) are marked.

Role of conserved UL37 tyrosines in UL37-gK interactions

We have shown previously that the gK/UL20 protein complex interacts with UL37 (Jambunathan et al., 2014). Co-immunoprecipitation experiments were performed to evaluate whether the combined Y474-A, Y476-A, Y477-A and Y480-A mutations specified by the DC474-480 mutant virus affect the ability of UL37 to interact with gK. The VC1 virus was used as positive control and the HSV-1 (F) virus was used as negative control (the VC1 virus specifies gK and UL20 tagged with V5 and FLAG epitopes respectively, while the F strain is not epitope tagged). The UL37 protein was detected in all virus-infected lysates probed with anti-UL37 polyclonal rabbit antibody, although the amount of protein detected in the DC474-480 cell lysates was less than in the VC1 or the HSV-1(F) cell lysates. All cell lysates contained similar amounts of the major capsid protein VP5 indicating a similar overall viral protein content across samples. Immunoblot probing of immunoprecipitates from anti-gK (V5) immunoprecipitates
revealed the presence of both UL20 and gK as well as pUL37 when probed with anti-UL20 (FLAG), anti-gK(V5) and anti-UL37 antibodies, respectively, for all viruses except HSV-1(F), which does not have gK or UL20 proteins tagged with the V5 or FLAG epitope tags. Importantly, anti-gK(V5) immunoprecipitates contained very low amounts of UL37 appearing as very faint-labeled protein species for DC474-480 in comparison to the VC1 virus, while similar amounts of UL20 protein were immunoprecipitated from the same samples. The gK(V5) immunoprecipitates did not contain the VP5 protein (negative control) (Fig. 6).

8.4 Discussion

The HSV-1 UL37 protein exhibits a high level of conservation among all alphaherpesviruses. Previously, we and others have showed that the UL37 protein plays a critical role in cytoplasmic virion envelopment and egress of virions from infected cells and the UL37 protein interacts with glycoprotein K (gK) (Jambunathan et al., 2014). Herein, we show that specific tyrosine residues conserved among alphaherpesviruses, which are predicted to be exposed on the UL37 surface are critically involved in interaction with gK, cytoplasmic envelopment, egress and infectious virus production.

Both pUL37 and the UL20/gK proteins are conserved within the neurotropic alphaherpesvirinae subfamily, reflecting their important roles in the viral life cycle (Roizman & Campadelli-Fiume, 2007). Initial alignment of a number of UL37 homologs showed the presence of highly conserved residues scattered throughout the pUL37 ORF (Kelly et al., 2012). Alignment of UL37 homologs from 20 different alphaherpesviruses confirmed the conservation of a number of amino acid residues including tyrosines (Y474 and Y480). Transfection-complementation of a UL37-null virus using a cadre of plasmids carrying specific mutations in UL37 showed that both Y474 and Y480 serve important roles in the pUL37 structure and
function(s). This hypothesis was further supported by the UL37 predicted three-dimensional structure, which revealed that Y474 and Y480 were exposed on the UL37 surface and embedded within a “clef” domain reminiscent of a protein-binding site.

Figure 8.6. Effect of the DC474-480 mutations on UL37 interactions with gK. A) Lysates of Vero cells infected with either wild type HSV-1(F), VC1 or DC474-480 viruses were collected at 24 hpi, electrophoretically separated, blotted onto nitrocellulose membranes, and reacted with either anti-V5 (gK), anti-UL37, or anti-VP5 antibodies. B) Lysates prepared as in (A) were immunoprecipitated with the anti-V5 (gK). Electrophoretically separated proteins of the immunoprecipitates were blotted onto nitrocellulose membranes and the presence of UL37 gK, UL20 or UL37 protein was detected by reacting the blots with either anti-V5, anti-FLAG or anti-UL37 antibodies respectively. No VP5 was detected when the immunoprecipitate was probed with anti-VP5 antibody (negative control).

Additional evidence that tyrosines Y474 and Y480 were particularly important in UL37 functions was obtained from the characterization of the Y474-480 mutant virus in which all four tyrosines at positions 474, 476, 477 and 480 were mutated to alanines. This virus produced a UL37-null phenotype characterized by the accumulation of capsids in the cytoplasm of infected cells and inability to produce infectious virions. Since the mutant viruses Y476 and Y477 having only the Y476 or Y477 residues changed to alanine, replicated fairly efficiently, it can be concluded that Y474 and Y480 play crucial roles in the structure and function of the UL37
protein. This conclusion is supported by the inability of plasmids having either Y474-A or Y480-A mutations to complement the UL37-null virus replication suggesting that each of these two tyrosine residues play important role in the pUL37 structure and function. It was previously reported that the UL37 protein is phosphorylated (Albright & Jenkins, 1993). In contrast to these findings, multiple experiments failed to detect significant UL37 labeling with $^{32}$P in infected Vero cells (not shown) indicating that phosphorylation of both Y474 and Y480 residue does not play an important role in the UL37 structure and function.

We have shown previously that the UL37 protein interacts with gK and that UL37-null and gK-null viruses produce similar phenotypes characterized by drastic defects in cytoplasmic envelopment and egress (Jambunathan et al., 2014). Mutagenesis of the four conserved tyrosine residues to alanine revealed that they inhibited interactions of the UL37 protein with gK suggesting that they are involved in these interactions. The alanine scanning study with FoldX shows that Y474A and Y480A mutations lead to a more positive residual free energy ($\Delta\Delta G$) representing unfavorable or destabilizing change compared to Y476A and Y477A suggesting that both tyrosine residues play an important role in the pUL37 structure. The location of the Y474 and Y480 residues in a pocket like helix-turn-helix motif may play an important role in preserving the structure of this putative binding/interaction site, especially in preserving the geometry of the helix-urn-helix motif which is in general known to be contributory in protein-protein interactions (Khan & Vihinen, 2010; Thiltgen & Goldstein, 2012). Also, it is possible that Y474 and Y480 play important roles in preserving the 3-dimensional structure of the entire UL37 protein. However, the fact that Y480 is conserved only in alphaherpesviruses but not in betta or gamma herpesviruses (Fig. 7) argues that at least Y480 does not be play crucial functions in the overall UL37 structure, while it provides a specific residue that is utilized for
interactions with gK. It is worth noting that the precision of the energy calculation by alanine scanning methods can be assessed using experimental free energy measurements. In the current study, we have focused on the relative changes of ΔΔG rather than the absolute values in order to reduce any possible bias from energy calculations and to eliminate the need to examine the precision of such methods.

Figure 8.7. Sequence alignment of alpha, beta and gamma herpesvirus UL37 proteins. Sequence alignment of HSV-1, HSV-2, HCMV, and EBV UL37 proteins was performed as described in the Materials and Methods section. Orange is conserved proline, yellow is conserved phenylalanine.

Apparently, the drastic reduction in gK interactions with pUL37 did not affect interactions between the UL20 protein and pUL37. We have reported previously, that TGN localization of gK requires co-expression of the UL20 protein. However, UL20 assumes a cytoplasmic distribution in the absence of gK, while gK is trapped into rough endoplasmic reticulum (RER) membranes (Foster et al., 2004). Thus, it is conceivable that UL20 protein not bound to gK may interact with pUL37 independently of gK to facilitate cytoplasmic virion envelopment.

HSV-1 assembles in infected cells through a sequentially ordered process depending on precise protein-protein interactions that regulate intracellular transport and virion egress, as well as the amount and nature of each structural protein, which is ultimately incorporated within the mature virion. It has been reported that the UL36/UL37 protein complex can be transported to
TGN membranes where cytoplasmic envelopment occurs in the absence of capsid formation (P. Desai, Sexton, Huang, & Person, 2008). This result suggests that the UL36/UL37 complex may anchor to TGN membranes via binding to cytoplasmic domains of gK and UL20 proteins. Overall, our current and previous results indicate that physical interactions between the gK/UL20 protein complex and the UL37 protein serve important functions in cytoplasmic virion envelopment and infectious virus production.

8.5 Conclusions

We report here for the first time the HSV-1 UL37 structure model based on the overall conserved nature of the UL37 proteins encoded by alphaherpesviruses and the available x-ray structure of the PRV UL37. In addition, we utilized in silico alanine mutagenesis to analyze the relative importance of different amino acid residues in the UL37 structure and function providing significant complementary evidence of the relative role of conserved amino acid residues in the UL37 structure and function. This computational approach should be useful to explore the structure and function of other viral and cellular proteins.

8.6 References


9. CONCLUSIONS

Research and discoveries mentioned in chapters 2-8 of this PhD dissertation have lead to 4 manuscripts submitted for publication and more than 8 peer-reviewed papers published in Journal of Molecular Biology, Nature Partner Journals Systems Biology and Applications, Journal of Chemical Information Modeling, Journal of Cheminformatics, and Journal of Virology. The nature of the research discussed here is highly interdisciplinary and to preserve the context and scopes of this data-driven analysis of biomolecular interactions the collaborative efforts are presented with minimum alterations. Yet, it is necessary to acknowledge that the virology experiments were executed by members of Kousoulas lab (Chouljenko DV, Rider P, Jambunathan N, Chouljenko VN, Bergeron S, Charles AS, Subramanian R, Saied AA, and Kousoulas KG) and computer-programing efforts were performed in collaboration with the computer science department at LSU (Liu T, Alvin C, and Mukhopadhyay S). Computational Systems Biology group members (Ding Y, Govindaraj RG, Singha M, Lemoine J, and Brylinski M) contributed to the computational work.

Research highlights:

- Freely available eModel-BDB can be used to support structure-based drug discovery and repositioning, drug target identification, and protein structure determination.
  - The methodology to construct a pipeline for modeling drug-bound proteins based on protein threading and similarity-based drug docking is laid out in detail.
- eRepo-ORP, a comprehensive resource constructed by a large-scale repositioning of existing drugs to orphan diseases is freely available to support orphan drug discovery.
o Systematic drug repurposing for rare disease through holo-structure protein modeling, similarity based docking, and drug site matching is presented.
o Combination of sequence-order independent local binding site alignment and virtual screening is explored and implemented on drug bank and orphaned datasets.

• A novel software to build targeted virtual screening libraries, eSynth, is designed and distributed to assist with efficient lead generation.

• eSynth’s capabilities are examined and reported as:
o Construct chemically feasible molecules from molecular fragments
o Generate compounds that are distinctly different from parent molecules.
o Produce novel compounds that carry desired activity profiles.

• eMolFrag, a software to decompose chemical compounds to small molecular fragments, is produced and made available online.
o The robustness and computational performance of eMolFrag is assessed.
o Application of eMolFrag in de novo drug design are elucidated.

• Importance of protein modeling in biological research is showcased by applying state of the art homology modeling to inform virology experiments.

• The first protein structure model of glycoprotein-K in Herpes Simplex type-1 is computationally constructed.
o Functional domains in gK protein are predicted based on the structure model and are validated experimentally.
o Disulfide bond formations as well as N-Glycosylation sites are predicted computationally and further annotated with experiments.
Novel targetable sites are identified in gK to develop anti-HSV drugs.

- The first UL37 protein model in HSV-1 is generated.
  - Relative role of conserved amino acid residues in the UL37 structure and function are investigated computationally to complement experimental findings.

Summary

We created eModel-BDB, a database of 200,008 high-quality, comparative models of drug-bound proteins based on interaction data obtained from the Binding Database. Complex models in eModel-BDB were generated with a collection of the state-of-the-art techniques, including protein meta-threading, template-based structure modeling, refinement and binding site detection, and ligand similarity-based docking. In addition to a rigorous quality control maintained during dataset generation, a subset of weakly homologous models were selected for the retrospective validation against experimental structural data recently deposited to the Protein Data Bank. Validation results indicate that eModel-BDB contains high-quality models not only at the global protein structure level, but also with respect to the atomic details of the bound ligands (Chapter 2).

We present eRepo-ORP, a comprehensive resource constructed by a large-scale repositioning of existing drugs to orphan diseases with a collection of structural bioinformatics tools, including eThread, eFindSite and eMatchSite. Specifically, in order to find new targets for known drugs ultimately leading to drug repositioning, recently developed eMatchSite, a new computer program to compare drug-binding sites, is combined with virtual screening to systematically explore opportunities to reposition known drugs to proteins associated with rare diseases. A systematic exploration of 320,856 possible links between known drugs in DrugBank and orphan proteins obtained from Orphanet reveals as many as 18,145 candidates for
repurposing. The dataset comprises 31,142 putative drug-target complexes linked to 980 orphan diseases. The modeling accuracy is evaluated against the structural data recently released for tyrosine-protein kinase HCK. In order to illustrate how potential therapeutics for rare diseases can be identified with eRepo-ORP. The effectiveness of this integrated approach is demonstrated by case studies including a kinase inhibitor for Ras-associated autoimmune leukoproliferative disease, a kinase inhibitor candidate for repositioning to synapsin Ia, and a steroidal aromatase inhibitor to treat Niemann-Pick disease type C. Overall chapter 3 discusses this exhaustive exploration of the drug repositioning space which exposes new opportunities to combat orphan diseases with existing drugs. DrugBank/Orphanet repositioning data as well as the eRepo-ORP dataset are made freely available to research community at https://osf.io/qdjup/.

Computational research endeavors discussed in chapters 2 and 3 have far reaching consequences across a multitude of disciplines. Physiological responses and mechanisms of actions of therapies with multiple bioactive compounds lend themselves nicely to the philosophy of “one drug, multiple targets” and “multiple drugs, similar targets”. Research on cancer and neurodegenerative diseases has much to gain from insights into the human interactome. As such, it is imperative to scrutinize the intricate networks of drug-protein interactions at a systems level. We hope that our efforts translate into repurposed treatments for human diseases in the future.

In chapters 4 and 5, two new fragment-based cheminformatics computational tools were discussed. Due to exorbitant costs of high-throughput screening, many drug discovery projects commonly employ inexpensive virtual screening to support experimental efforts. However, the vast majority of compounds in widely used screening libraries, such as the ZINC database, will have a very low probability to exhibit the desired bioactivity for a given protein. Although combinatorial chemistry methods can be used to augment existing compound libraries with novel
drug-like compounds, the broad chemical space is often too large to be explored. Consequently, the trend in library design has shifted to produce screening collections specifically tailored to modulate the function of a particular target or a protein family. Assuming that organic compounds are composed of sets of rigid fragments connected by flexible linkers, a molecule can be decomposed into its building blocks tracking their atomic connectivity.

On this account, we developed eSynth, an exhaustive graph-based search algorithm to computationally synthesize new compounds by reconnecting these building blocks following their connectivity patterns. We conducted a series of benchmarking calculations against the Directory of Useful Decoys, Enhanced database. First, in a self-benchmarking test, the correctness of the algorithm was validated with the objective to recover a molecule from its building blocks. Encouragingly, eSynth could efficiently rebuild more than 80% of active molecules from their fragment components. Next, the capability to discover novel scaffolds was assessed in a cross-benchmarking test, where eSynth successfully reconstructed 40% of the target molecules using fragments extracted from chemically distinct compounds. Despite an enormous chemical space to be explored, eSynth was computationally efficient; half of the molecules were rebuilt in less than a second, whereas 90% took only about a minute to be generated.

Overall, eSynth could successfully reconstruct chemically feasible molecules from molecular fragments. Furthermore, in a procedure mimicking the real application, where one expects to discover novel compounds based on a small set of already developed bioactives, eSynth was capable of generating diverse collections of molecules with the desired activity profiles. Thus, we are very optimistic that our effort described in chapter 4 will contribute to
targeted drug discovery. eSynth is freely available to the academic community at www.brylinski.org/content/molecular-synthesis.

In chapter 5, we described eMolFrag, a new open-source software to decompose organic compounds into non-redundant fragments retaining molecular connectivity information. Given a collection of molecules, eMolFrag generates a set of unique fragments comprising larger moieties, bricks, and smaller linkers connecting bricks. These building blocks can subsequently be used to construct virtual screening libraries for targeted drug discovery by other software packages such as eSynth. The robustness and computational performance of eMolFrag was assessed against the Directory of Useful Decoys, Enhanced database conducted in serial and parallel modes with up to 16 computing cores. Further, the application of eMolFrag in de novo drug design was illustrated using the adenosine receptor. eMolFrag is implemented in Python and it is available as a stand-alone software and a webserver at www.brylinski.org/emolfrag and https://github.com/liutairan/eMolFrag.

In chapters 6-8, protein modeling as well as other strategies such as computational mutation analysis was explored in the context of Alphaherpesvirus research. Alphaherpesviruses, unlike beta- and gammaherpesviruses, have the unique ability to infect and establish latency in neurons. Glycoprotein K (gK) is a viral protein, which is conserved among all alphaherpesviruses. We have constructed the first structure model of gK. The processes involved predicting individual domains in gK and predict protein structure for each domain separately by threading.

We showed that a predicted β-sheet domain, which is conserved among alphaherpesviruses, functions in HSV-1 entry into neuronal axons, suggesting that it may serve similar functions for other herpesviruses based on structure conservation observed in partial gK
models that we constructed for other herpesvirus strains. These results, presented in chapter 6, are in agreement with previous observations that deletion of this gK domain prevents the virus from successfully infecting ganglionic neurons after ocular infection of mice.

We identified arginine-linked glycosylation sites in gK N-terminal domain. Based on the structure model, we were also able to recognize cysteine residues that could position in close proximity to each other providing the possibility for disulfide bond formations. The contribution of the two conserved N-linked glycosylation (N48 and N58) sites in the amino terminus of HSV-1 gK was determined experimentally. We found that N-linked glycosylation is important to the regulation of HSV-1-induced membrane fusion, since mutating N58 to alanine caused extensive virus-induced cell fusion. Due to the known contributions of N-linked glycosylation to protein processing and in particular, correct disulfide bond formation, we investigated whether the four, conserved extracellular cysteine residues within the amino terminus of gK contributed the regulation of HSV-1-induced membrane fusion. We found that mutation of C37 and C114 residues led to a gK-null phenotype characterized by very small plaque formation and drastic reduction in infectious virus production, while mutation of C82 and C243 caused extensive virus-induced cell fusion. Comparison of N-linked glycosylation and cysteine mutant replication kinetics identified disparate effects on infectious virion egress from infected cells. Specifically, cysteine mutations led to defects in the accumulation of infectious virus in both the cellular and supernatant fractions, while glycosylation site mutants did not adversely affect virion egress from infected cells. These results, discussed in chapter 7, demonstrate a critical role for the gK N glycosylation sites and cysteines for the structure and function of the amino terminus of gK.

To expand our findings in gK, in chapter 8, we investigated protein structure of HSV-1 UL37 that is known to interact with gK and its interacting partner UL20. It has been suggested
that UL37-gK/UL20 interactions facilitate cytoplasmic virion envelopment. In general, UL37 protein functions had been shown to be involved in virion envelopment at trans-Golgi membranes, as well as in retrograde and anterograde transport of virion capsids. Alignment of UL37 homologues encoded by different alphaherpesviruses revealed the presence of highly conserved residues in the central portion of the UL37 protein. To delineate UL37 domains and critical amino acids that are involved in virion envelopment, a cadre of nine UL37 site-specific mutations were produced and tested for their ability to inhibit virion envelopment and infectious virus production. Simultaneously, we developed the first structure model of UL37 to analyze the role of these amino acids in the structure stability and function of this protein. By computational alanine scanning we showed that UL37 protein structure stability was more sensitive to changes in two tyrosine residues, Y474 and Y480, compared to other amino acids in the region of interest. The recombinant virus DC474-480 constructed with tyrosines 474, 476, 477 and 480 mutated to alanine residues produced a gK-null-like phenotype characterized by the production of very small plaques and accumulation of capsids in the cytoplasm of infected cells. Recombinant viruses having either tyrosine 476 or 477 replaced with alanine residues produced a wild-type phenotype, supporting the computational findings. Immunoprecipitation assays revealed that replacement of all four tyrosines to alanines substantially reduced the ability of gK to interact with UL37. Moreover, alignment of herpes simplex virus (HSV) UL37 to the human cytomegalovirus (HCMV) and Epstein Barr Virus (EBV) UL37 homologues revealed that Y474 was conserved, while Y480 was not. Collectively, these results suggested that UL37 conserved tyrosine residue Y480 plays crucial role in interactions with gK to facilitate cytoplasmic virion envelopment and infectious virus production.
The overall goal of this dissertation was to explore proteins-drugs and protein-protein interactions by computational approaches. As a result of this data-driven approach in studying biomolecular interactions,

1. drug-bound protein models of BindingDB complexes and FDA approved drugs from DrugBank have been developed.

2. proteins that are associated with rare diseases were modeled, annotated with binding pockets, and candidate drugs were matched to their binding sites.

3. Cheminformatics tools were developed to deconstruct chemical compounds into fragments and reconstruct drug-like molecules by combinatorial chemistry following simple patterns to connect those fragments.

4. novel and important discoveries in the field of herpes virology were made by generating and analysing the first protein models for HSV-1 gK and UL37 proteins.

I hope that these findings, databases, software tools, and methods, made available here and in published scientific papers, can contribute to drug repurposing, fragment-based drug discovery, anti-viral drug design, as well as to further our understanding of biomolecular interactions.
APPENDIX A. DCIM PERMISSION FOR REPRINT

2/21/2018 Mail - mnader5@lsu.edu
https://outlook.office.com/owa/?realm=lsu.edu&vd=mail 1/3
RE: Fw: Re-use permission PhD dissertation

Dear Dr. Naderi,

Thank you for contacting JCIM with your request.

We grant you permission to republish all or part of your work as published in this journal, with the following proviso, as stated in the ACS Thesis/Dissertation Policy and the ACS Journal Publishing Agreement:

Appropriate citation of the Published Work must be made. If the thesis or dissertation to be published is in electronic format, a direct link to the Published Work must also be included using the ACS Articles on Request authordirected link – see http://pubs.acs.org/page/policy/articlesonrequest/index.html

Sincerely,

Letitia Glozer
Journal of Chemical Information and Modeling Editorial Office
Professor Kenneth M. Merz, Jr.
Editor in Chief,
Journal of Chemical Information and Modeling
Director, Institute for Cyber Enabled Research (iCER)
Joseph Zichis Chair in Chemistry
Department of Chemistry
Department of Biochemistry and Molecular Biology
Michigan State University
578 S. Shaw Lane
East Lansing, MI 48824132
Phone: 919.918.3912
Fax: 202.559.0834
Email: eic@jcim.acs.org

Original
APPENDIX B. ASM PERMISSION FOR REPRINT

Title: Cysteines and N-Glycosylation Sites Conserved among All Alphaherpesviruses Regulate Membrane Fusion in Herpes Simplex Virus 1 Infection

Author: Paul J. F. Rider, Misagh Naderi, Scott Bergeron et al.

Publication: Journal of Virology

Publisher: American Society for Microbiology

Date: Nov 1, 2017

Copyright © 2017, American Society for Microbiology

Permissions Request

Authors in ASM journals retain the right to republish discrete portions of his/her article in any other publication (including print, CD-ROM, and other electronic formats) of which he or she is author or editor, provided that proper credit is given to the original ASM publication. ASM authors also retain the right to reuse the full article in his/her dissertation or thesis. For a full list of author rights, please see: http://journals.asm.org/site/misc/ASM_Author_Statement.xhtml
VITA

Misagh Naderi, born in Tehran, Iran, attended Sharif University of Technology and received a Bachelor of Science degree in Chemical Engineering in 2008. In the next year he joined the Chemical Engineering graduate program at Louisiana State University, where he earned his Master’s degree in 2012. Misagh lost his grandfather to Parkinson’s and an aunt to cancer, so he was determined to establish his research in human health related topics. Consequently, he joined the Pathobiological Sciences department at the School of Veterinary Medicine at LSU. He researched the application of oncolytic viruses to treat cancers under the supervision of Prof. Konstantin Gus Kousoulas. Concurrently, Misagh joined the computational systems biology group lead by Dr. Michal Brylinski in 2013 to pursue his PhD research in data-driven approaches in analysis of biomolecular interactions. His research involved drug repurposing for rare and neglected diseases, fragment-based drug discovery, and protein modeling in viruses. His passion lies at the intersection of science, technology, and communication. He will continue his scientifically driven passion to discover cures for human diseases as well as educating the next generations.