DARWINDOCK & GAG-DOCK: METHODS AND APPLICATIONS FOR SMALL MOLECULE DOCKING

Thesis by

Adam Reid Griffith

In Partial Fulfillment of the Requirements for
the degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2017
(Defended September 26, 2016)

ACKNOWLEDGEMENTS

I would first and foremost like to thank my advisor, William A. Goddard III, for his guidance, support, and patience during my time at Caltech. I continue to be in awe of his knowledge, wisdom, and passion for knowledge.

I would also like to thank my thesis committee: Dr. James Heath, Dr. Douglas Rees, and Dr. Linda Hsieh-Wilson. They have been extremely patient with me, much more than they should have been.

My interactions with Dr. Ravi Abrol as friend and mentor have been invaluable. He has provided guidance and direction on most of my projects, as well as providing encouragement and a friendly ear.

The countless members of the MSC and especially the Biogroup have been my friends and colleagues. Chief among them have been Dr. Soo-Kyung Kim, Dr. Ismet Caglar Tanrikulu, Dr. Heather Wiencko, Dr. Julius Su, Dr. John Keith, Dr. Jenelle Bray, Dr. Andrea Kirkpatrick, Dr. Caitlin Scott, Dr. Vaclav Cvicek, Dr. Fan Liu, and… many others. (I sincerely apologize to those I haven't mentioned.)

I have been blessed with many friends in Pasadena. Michael Torrice, Graef Allen, Michael D. Adams, Michelle Robbins, Heidi Privett, Dillon Ross, Ariele Hanek, Gavin and Jill Doughtie, Heather Wiencko, Ger Norton, Joel Louwsma, Greg ver Steeg, Jeff Krimmel, and Eve Stenson to name a few.

Last, certainly not least, I have had endless support and encouragement from my family, both immediate and extended. My cousins Margaret Griffith and Jamison Carter and Suzanna Griffith had never met me before I moved to California, yet they took me in during my first Thanksgiving out here and have been friends ever since. I've thoroughly enjoyed spending time discussing art and science with you. Thank you for providing an escape when I needed one. Nobody has been more supportive of me than my parents, Boyce and Diana Griffith, and my brother, Boyce E. Griffith. I wouldn't be here or doing what I'm doing without you. I can't even put into words how much you've done for me, or how much you mean to me.

ABSTRACT

Computational modeling is an effective tool in studying complex biological systems. Docking of small molecule ligands in particular is useful both in understanding the functioning of proteins as well as in the development of pharmaceuticals. Together with experiment, modeling can often provide a thorough picture of a given system. Computation can often provide details that are difficult or impossible to determine experimentally, while experiments provide guidance on what calculations are useful or interesting. Our goal is to extend computational modeling, specifically ligand docking, to systems not previously possible, such as the challenging glycosaminoglycan (GAG) systems. In order to do this it was first necessary to develop an automatic way of performing docking without extensive user input and experimental knowledge to narrow the list of candidate poses. DarwinDock represents our efforts in this respect. It is a method for small-molecule docking that separates pose generation and scoring into separate stages, which allows for complete binding site sampling followed by efficient, hierarchical sampling. Our convergence criteria for complete sampling allows for diverse systems to be studied without prior knowledge of how large a set of poses needs to be to span a given binding site, making the procedure more automatic. We also replace bulky, nonpolar residues with alanine, which we refer to as "alanization". This allows the ligand to interact more closely with polar sidechains, which help to orient the ligand. Additionally, alanization reduces the impact of incorrect sidechain placement on ligand placement, a concern that sometimes requires user intervention. With DarwinDock working for standard small molecules, it was then necessary to modify the procedure to work on challenging GAG ligands, which are large and have strong negative charges. A modification to DarwinDock – GAG-Dock – allows the method to be applied to GAGs and protein surface interactions. GAGs are large, linear polysaccharides with strong negative charge. They typically interact with the surfaces of proteins, rather than the cavities favored by most small-molecule drugs. GAG-Dock systematically samples the protein surface for unknown binding sites and modifies the pose generation to allow for large, surface-interacting ligands. GAG-Dock allowed us to study several systems important for neuronal development and answer interesting questions posed by experiment. Finally, we needed a

way to validate our predictions for GAG binding sites. We used a systematic approach to identify sets of beneficial mutations to the GAG binding sites by building up from individual *in silico* mutations. Standard mutation experiments typically employ large mutations, such as arginine to alanine, which decrease or destroy binding. However, such information is not always definitive, as large mutations can have wide-ranging effects beyond direct protein-ligand interactions. Mutations that *increase* binding, however, are less ambiguous because they must form new interactions with the ligand in order to affect binding energies or affinity. Therefore, we have identified and proposed sets of mutations for our GAG predictions for PTPs, NgR1, NgR3, and EphB3. We encourage our experimentalist colleagues to try these mutations and validate our predictions.

CONTRIBUTIONS

Chapter 2: Adam R. Griffith, Ismet Caglar Tanrikulu, Ravinder Abrol, William A. Goddard III, THE DARWINDOCK ALGORITHM FOR SMALL MOLECULE DOCKING

A.R.G. was involved in method development, wrote software to handle input/output of DOCK 6 calls, implemented the CNS clustering algorithm for the calculations reported, made modifications to the main DarwinDock program, prepared the DUD systems for calculations, performed extensive testing during method development, and performed all calculations reported

I.C.T. was involved in method development, wrote the main DarwinDock program, prepared DUD systems for testing during development, and performed testing on the CSN method

R.A. was involved in method development and wrote a development implementation of the CNS algorithm for testing

W.A.G. provided guidance for the project and valuable feedback

Chapter 3: Adam R. Griffith, Claude J. Rogers, Ravinder Abrol, Greg Miller, William A. Goddard III, Linda C. Hsieh-Wilson, PREDICTING GLYCOSAMINOGLYCAN-SURFACE PROTEIN INTERACTIONS: IMPLICATIONS FOR STUDYING AXONAL GROWTH

A.R.G. was involved in GAG-Dock method development, performed extensive testing, wrote software to automate docking calculations, developed the method for predicting mutations, and performed the mutation calculations

C.J.R. was involved in method development, prepared and evaluated the validation systems, prepared GAG ligands for docking, and performed calculations for the predicted systems

R.A. provided assistance with homology modeling and guidance

G.M. assisted with preparation of GAG ligands

W.A.G. provided guidance for the project and valuable feedback

L.C.H.-W. initiated the project and provided guidance on identifying relevant systems for calculation

Chapter 4: Adam R. Griffith, Claude J. Rogers, Ravinder Abrol, Greg Miller, William A. Goddard III, Linda C. Hsieh-Wilson, EXPLORING NOVEL INTERACTIONS BETWEEN CHONDROITIN SULFATE AND THE EPHB3 RECEPTOR

A.R.G. implemented APBS mapping of charges to *sphgen* spheres, performed all docking calculations, and performed mutation analysis

C.J.R. prepared GAG ligands for docking and provided extensive discussion and analysis relevant to the project

R.A. generated EphB2 and EphB3 homology models and provided guidance

G.M. assisted with preparation of GAG ligands

W.A.G. provided guidance for the project and valuable feedback

L.C.H.-W. initiated the project

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

List of Figures & Tables: Chapter 2

List of Figures & Tables: Chapter 3

List of Figures & Tables: Chapter 4

Chapter 1

SMALL MOLECULE DOCKING

Introduction

Computational modeling has become an important part of studying complex biological systems. It often provides details and information that either cannot or cannot easily be obtained via experiments. Rather, computation and experiment often work hand-in-hand, with insights from one driving questions for the other. A recent paper by Brian Shoichet[1] exemplifies this. G-protein coupled receptors (GPCRs) produce signaling via two pathways: a G protein or β-arrestin. The same GPCR can use both pathways to produce different effects. Experiments suggest that while the primary analgesic effect of the μ-opioid receptor (mOR) is carried out via G protein signaling, side effects such as the buildup of tolerance may be due to the β-arrestin pathway[2-4]. This suggests that a drug capable of activating the G protein pathway and *not* the β-arrestin pathway would be an improvement on current opioid drugs. While it may be possible to identify such a drug using an experimental approach, such an approach would be lengthy and expensive. Computation, on the other hand, is ideally suited to this sort of task. Shoichet, *et. al.*, used a virtual ligand screening (VLS) method to screen 3 million drug candidates against mOR to identify drug candidates that interact with the protein in a manner different than current opioid drugs. After refinement of the candidates, one drug was identified for testing. The result of the project was a drug that preferentially activates mOR G protein signaling over β-arrestin signaling. This drug will now provide a starting point for development of better mOR analgesics as well as providing a tool for the study of the differences of G protein and

β-arrestin signaling. Such a project likely would not have been successful without the use of computational methods for small-molecule docking.

Purpose of the Study

Our goal in this work is to tackle challenging computational problems, such as those related to glycosaminoglycan (GAG) docking, to answer questions raised by experiment, and to help guide new experiments. GAGs are large, linear polysaccharides that are heavily negatively charged, which makes them difficult to treat computationally. Furthermore, they typically interact with the surfaces of proteins, instead of the deeper cavities typically used by standard small-molecule ligands. These surface-protein interactions are less well defined, which requires much more extensive sampling to identify a binding site than a protein with a known binding cavity. We therefore need an automated docking method that is capable of identifying interesting poses and binding sites with little intervention by the user and little or no experimental information. We further need to be able to account for the size and charge of GAG ligands, as well as the ability to predict ligand binding to protein surfaces. And finally, we need a way to validate our predictions.

Our strategy for automated docking is DarwinDock. DarwinDock divides the problem of docking into two stages: geometry and scoring. The geometry or "completeness" stage thoroughly samples a putative binding site without performing any energy calls or using any scoring methods. Our method for generating a complete set of ligand poses differs from other docking methods in that the number of poses generated depends on the system, and not a default or assumed number of poses. Instead, poses are generated until a convergence threshold is met. This is important because each system – protein, ligand, or

binding site – is different. There is no default number of poses that will work in all situations. The scoring stage of DarwinDock efficiently and accurately evaluates the poses that are generated. Efficiency is derived from our hierarchical scoring method. Poses are clustered into families, and only the family head (centroid) pose is scored initially. This family head score is used to eliminate 90% of the families from consideration, significantly reducing computational cost. The children of the remaining 10% of families are then evaluated, and a final set of 120 best poses is output. The goal of the method is for the final set of 120 poses to contain correct poses for further, more detailed evaluations. We will describe the DarwinDock method and validate it against a set of known ligand-protein co-crystals.

We address the complexity of GAG ligand binding with GAG-Dock. GAG-Dock is a variation of DarwinDock that is designed to work for an extremely challenging and interesting problem: the study of glycosaminoglycan (GAG) binding to proteins. GAGs are linear polysaccharides that carry a strong negative charge. Their size and charge makes them particularly difficult candidates for docking. Furthermore, they typically interact with the surfaces of proteins. Typical small-molecule ligands tend to bury into cavities in a protein, which provides a clear, contained region to sample during docking. Interactions on the surface of a protein are less well defined and require much greater computational sampling. However, the interactions between GAGs and proteins are also very interesting. For instance, GAGs have been shown to be involved in directing neuronal development via interactions with the PTPs, NgR1, and NgR3 receptors. GAG-Dock is designed to study just these sorts of systems, for which structural data is often not known. We will show that

GAG-Dock is effective for these challenging systems by validating it against crystal structures with bound GAG polysaccharides. We will then apply it to the systems PTPs, NgR1, NgR3, and EphB3 in order to predict the binding sites of several GAG ligands and answer interesting questions posed by experimental evidence.

Finally, we use a systematic approach to identify sets of beneficial mutations to validate our GAG binding sites. Mutations that decrease or eliminate binding are often easy, but ambiguous. Mutating an arginine to an alanine could simply be removing a contact with the ligand, or it could be fundamentally altering the protein structure. Beneficial mutations that *increase* ligand binding energy/affinity, however, require that the mutation stabilize the binding site or provide new interactions with the ligand. They provide a much clearer signal. Therefore we identify sets of mutations for each of our predicted GAG cases where no crystal structure is known. This is particularly important for our CS-E/EphB3 predictions where even the general binding region of the protein was not known.

Overall, we show that we have developed methods capable of handling challenging, relevant systems that could not be approached before. Furthermore, we show that computational modeling and experiment can work together to guide and complement each other.

Outline

Chapter 2: Description of the DarwinDock method and validation against the DUD set, a set of protein-ligand co-crystals intended for testing small-molecule ligand docking methods.

Chapter 3: Description of GAG-Dock, validation against known GAG-protein co-crystals, and application to PTPs, NgR1, and NgR3 for the GAG ligands CS-A, CS-D, CS-E, and heparin.

Chapter 4: Application of GAG-Dock to the novel system of EphB3. We predict the binding site of CS-E to EphB3. We also explain why CS-E and not CS-A binds to EphB3, and why neither binds to EphB2.

References

1. Manglik, A., Lin, H., Aryal, D. K., McCorvy, J. D., Dengler, D., Corder, G., et al. (2016). Structure-based discovery of opioid analgesics with reduced side effects. *Nature*, *537*(7619), 185–190. http://doi.org/10.1038/nature19112
2. Bohn, L. M., Gainetdinov, R. R., Lin, F.-T., Lefkowitz, R. J., & Caron, M. G. (2000). μ-Opioid receptor desensitization by β-arrestin-2 determines morphine tolerance but not dependence. *Nature*, *408*(6813), 720–723. http://doi.org/10.1038/35047086
3. Bohn, L. M., Lefkowitz, R. J., Gainetdinov, R. R., Peppel, K., Caron, M. G., & Lin, F.-T. (1999). Enhanced Morphine Analgesia in Mice Lacking β-Arrestin 2. *Science*, *286*(5449), 2495–2498. http://doi.org/10.1126/science.286.5449.2495
4. Raehal, K. M., Walker, J. K. L., & Bohn, L. M. (2005). Morphine Side Effects in β-Arrestin 2 Knockout Mice. *Journal of Pharmacology and Experimental Therapeutics*, *314*(3), 1195–1201. http://doi.org/10.1124/jpet.105.087254
5. Shen, Y., Tenney, A. P., Busch, S. A., Horn, K. P., Cuascut, F. X., Liu, K., et al. (2009). PTPσ Is a Receptor for Chondroitin Sulfate Proteoglycan, an Inhibitor of Neural Regeneration. *Science*, *326*(5952), 592–596. http://doi.org/10.1126/science.1178310
6. Dickendesher, T. L., Baldwin, K. T., Mironova, Y. A., Koriyama, Y., Raiker, S. J., Askew, K. L., et al. (2012). NgR1 and NgR3 are receptors for chondroitin sulfate proteoglycans. *Nature Neuroscience*, *15*(5), 703–712. http://doi.org/10.1038/nn.3070