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)

© 2016

Adam Reid Griffith

All Rights Reserved

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

Acknowledgements	iii
Abstract	iv
Contributions	V
Table of Contents	vii
List of Figures & Tables	ix
Chapter 2	ix
Chapter 3	. ???
Chapter 4	.???
Chapter 1: Small Molecule Docking	
Introduction	1
Purpose of the Study	2
Outline	3
References	4
Chapter 2: The DarwinDock Algorithm For Small Molecule Docking	
Introduction	6
DarwinDock	9
The DUD Set	13
System Preparation	15
Predicted Sidechain Sets	19
"Best Case" Predicted Sidechain Results	21
Standard Predicted Sidechains	26
Crystal Sidechains	30
Conclusions	32
References	34
Chapter 3:	
Abstract	36
Introduction	37
Summary of the GAG-Dock Method	39
DarwinDock/GenDock	40
GAG-Dock Modifications	42
System Preparation	43
GAG Ligand Preparation	43
Results and Discussion	44
Case 1: Validation of systems for which there are	
x-ray structures of the co-crystal	45
FGF1	45
FGF2	47
FGF2-FGFR1	47
α-Antithrombin III	48
Case 2: Predictions for systems for which no co-crystal structure is available	50

RPTPs	50
NgR	53
Suggested Post-Prediction Validations	56
Conclusions	60
Acknowledgements	61
References	61
Supplementary Information	64
System Preparation	64
DarwinDock	65
Closest-Neighbor Seeded Ligand Clustering	67
Forcefield	68
Sidechain Optimization	68
Sphere Generation	68
Sphere Clustering	69
Ligand Preparation	70
Molecular Dynamics (MD)	71
Supplemental References	72
Supplemental Figures & Tables	75
Chapter 4:	
Introduction	
GAG-Dock Overview	
Modifications to GAG-Dock	
EphB2 and EphB3 Models	109
Results	
Suggested Post-Prediction Validations	113
Conclusions	117
Acknowledgements	119
References	119
Figures & Tables	121
Supplemental Information	
Mutation Methodology	
Supplemental Figures & Tables	

LIST OF FIGURES AND TABLES

List of Figures & Tables: Chapter 2

- Figure 2-1 Diagram of the DarwinDock algorithm. Pose generation and scoring are partitioned into two completely separate stages. The geometry or "completeness" stage generates 5000 poses, clusters them into 2Å families, and adds 5000 additional poses until the completeness threshold has been reached. The scoring stage initially only evaluates the 2Å family heads. 90% of the families are eliminated based on the family head energy, and the children of the remaining 10% are scored fully. Typically, 120 interesting candidate poses are output from this final list.
- Figure 2-3 Analysis of docking to the "best" sidechain predictions. The series show either flat or distance-dependent dielectric constants for varying percentages of families scored. The left half of the chart shows scoring energies with 100% polar and the phobic energy scaled from 0 to 90%. The right half of the chart shows scoring energies with 100% phobic energy and the polar energy scaled from 90 to 0%. The center point has 100% polar and 100% phobic scaling. There is a clear preference for 100% polar and 10% phobic regardless of dielectric and percent of families scored.
- Figure 2-4 Analysis of docking to the "best" sidechain predictions with the scoring energy set to 100% polar, 10% phobic. The red columns show a flat dielectric of 2.5 with the percent of families scored set to 10, 25, 33, 50, and 100%. The blue columns show the distance-dependent dielectric for the same percent of families scored. There is a clear, modest preference for the distance-dependent dielectric. There is also a slight, but insignificant improvement from increasing the percent of families scored. Due to the added computational cost, there is no reason to score more than 10% of the families for such insignificant improvement.
- Figure 2-5 Analysis of docking to the crystal, non-alanized sidechains. Unlike with the "best" predicted sidechains, these results show a preference for a higher phobic content, although not 100%. The best result is for distance-dependent dielectric and 100% polar, 50% phobic when scoring 100% of the families. This drops to 100% polar, 30% phobic when only scoring 10% of the families.

 bumps with the ligand. Even small numbers of bumps can make it impossible for the ligand Table 2-4 – Summary of results for various completeness thresholds and clustering diversities. Smaller completeness thresholds and clustering diversities yield improved results, but at Table 2-5 – Comparison of docking results for flat-dielectric and distance-dependent dielectric predicted sidechains with and without alanization. The average number of 2Å hits per system looks encouraging for the flat-dielectric set without alanization, but *half* of the systems produce Table 2-6 – Average number of 2Å hits for flat dielectric and distance-dependent dielectric predicted sidechains, with and without alanization. There are some systems where alanization is essential (e.g. "ppar"), some where it is detrimental (e.g. "ppp"), and some where it doesn't make an impact (e.g. "trypsin"). Similarly, some systems do better with the flat dielectric (e.g. "dhfr") and some do better with the distance-dependent dielectric (e.g. "rxr"). Note: "A" refers to a histidine replaced with alanine, "B" refers to a neutral histidine changed to

List of Figures & Tables: Chapter 3

xi

- Figure S3-9 Structure of FGF2-FGFR1 [PDB: 1FQ9, res. 3.00 Å] chain B with predicted heparin hexamer ligand (magenta) and 5 Å binding site shown (cyan). Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein. The predicted ligand has excellent agreement with the crystal ligand, RMSD: 1.51/0.75 Å.
- Figure S3-10 Structure of FGF2-FGFR1 [PDB: 1FQ9, res. 3.00 Å] chain C with predicted heparin hexamer and octamer ligands (magenta) and 5 Å binding site shown (cyan). Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein. The predicted ligand has excellent agreement with the crystal ligand, RMSD: 1.51/0.75 Å. 80
- Figure S3-11 Structure of FGF2-FGFR1 [PDB: 1FQ9, res. 3.00 Å] chain C with predicted heparin hexamer and octamer ligands (magenta) and 5 Å binding site shown (cyan). Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein. The predicted ligand has excellent agreement with the crystal ligand, RMSD: 1.51/0.75 Å. 81

xii

Figure S3-22 – Detail of predicted NgR1/heparin structure after docking and dynamics with heparin hexamer (magenta) and 5 Å binding site (cyan) shown. Dashed lines indicate Figure S3-23 – NgR3. (A) Structure of NgR3. (B) Electrostatic potential surface. (C-F) Predicted structures of CS-A, CS-D, CS-E, and heparin after docking and molecular Figure S3-24 – Detail of predicted NgR3/CS-A structure after docking and dynamics with CS-A hexamer (magenta) and 5 Å binding site (cyan) shown. Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein......100 Figure S3-25 – Detail of predicted NgR3/CS-D structure after docking and dynamics with CS-D hexamer (magenta) and 5 Å binding site (cyan) shown. Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein......101 Figure S3-26 – Detail of predicted NgR3/CS-E structure after docking and dynamics with CS-E hexamer (magenta) and 5 Å binding site (cyan) shown. Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein......102 Figure S3-27 – Detail of predicted NgR3/heparin structure after docking and dynamics with heparin hexamer (magenta) and 5 Å binding site (cyan) shown. Dashed lines indicate Table 3-1 – Summary of docking validations. The resolution of the x-ray structure is given along with the heavy-atom RMSD between the predicted and x-ray position of the ligand.49 Table 3-2 – Predicted sets of mutations to either increase (left) or decrease (right) binding of ligands to RPTPs. Note that none of the sets show improved binding for heparin. Changes in binding energy are shown relative to the wildtype structures in both absolute Table 3-3 – Predicted sets of mutations to either increase (left) or decrease (right) binding of ligands to NgR1. Note that none of the sets show improved binding for CS-E. Changes in binding energy are shown relative to the wildtype structures in both absolute change in Table 3-4 – Predicted sets of mutations to either increase (left) or decrease (right) binding of ligands to NgR3. Changes in binding energy are shown relative to the wildtype structure in Table S3-5 – Per-residue energetic contributions in the FGF1/heparin predicted (left) and Table S3-6 – Per-residue energetic contributions in the FGF2/heparin predicted (left) and Table S3-7 – Per-residue energetic contributions in the FGF2-FGFR1/heparin predicted (left) and crystal (right) structures for chains A and B. [PDB: 1FQ9, res. 3.00 Å, RMSD: Table S3-8 – Per-residue energetic contributions in the FGF2-FGFR1/heparin-A predicted (left) and crystal (right) structures for chain C. [PDB: 1FQ9, res. 3.00 Å, RMSD: 1.51/0.75 Table S3-9 – Per-residue energetic contributions in the FGF2-FGFR1/heparin-A predicted (left) and crystal (right) structures for chain D. [PDB: 1FQ9, res. 3.00 Å, RMSD: 1.51/0.75

x111

Table S3-10 – Per-residue energetic contributions in the FGF2-FGFR1/heparin-B predicted (left) and crystal (right) structures for chain C. [PDB: 1FQ9, res. 3.00 Å, RMSD: Table S3-11 – Per-residue energetic contributions in the FGF2-FGFR1/heparin-B predicted (left) and crystal (right) structures for chain D. [PDB: 1FQ9, res. 3.00 Å, RMSD: 1.51/0.75 Table S3-12 – Per-residue energetic contributions in the Antithrombin-III/heparin analog predicted (left) and crystal (right) structures. [PDB: 1E03, res. 2.90 Å, RMSD: 0.60 Å]. 89 Table S3-13 – Per-residue energetic contributions in the predicted RPTP σ /CS-A (left) and Table S3-14 – Per-residue energetic contributions in the predicted RPTP σ /CS-E (left) and Table S3-15 – Per-residue energetic contributions in the predicted NgR1 structures for CS-A, Table S3-16 - Per-residue energetic contributions in the predicted NgR3 structures for CS-A, Table S3-17 – Single residue mutation data for RPTPs. Values show change in binding energy (kcal/mol) relative to wildtype structures. Values are shown both for the change in hydrogen bonding for the specific mutated residue as well as the overall change in the full cavity binding energy. The cavity binding energy is further separated into hydrogen bonding + van der Waals or Coulomb energy......105 Table S3-18 – Single residue mutation data for NgR1. Values show change in binding energy (kcal/mol) relative to wildtype structures. Values are shown both for the change in hydrogen bonding for the specific mutated residue as well as the overall change in the full cavity binding energy. The cavity binding energy is further separated into hydrogen Table S3-19 – Single residue mutation data for NgR3. Values show change in binding energy (kcal/mol) relative to wildtype structures. Values are shown both for the change in hydrogen bonding for the specific mutated residue as well as the overall change in the full cavity binding energy. The cavity binding energy is further separated into hydrogen

List of Figures & Tables: Chapter 4

	xiv
Figure 4-4 – Pharmacophore for best pose in EphB3/CS-E mode 1.	
Figure 4-5 – Pharmacophore for best pose in EphB3/CS-E mode 2.	. 123
Figure 4-6 – Pharmacophore for best pose in EphB3/CS-E mode 3.	. 124
Figure 4-7 – Pharmacophore for best pose in EphB3/CS-E mode 4.	.124
Figure 4-8 – Pharmacophore for best pose in EphB3/CS-E mode 5.	.125
Figure 4-9 – Mutations to Gln predicted to increase EphB3/CS-E binding. Mutated resi	dues
are colored orange. Red hydrogen bond markers denote new hydrogen bonds with	1 the
ligand due to mutations and blue markers denote hydrogen bonds to the ligand that	t are
common to both mutant and wild type. (A) Mutations for binding mode 1: T448Q, V33	39Q,
I446Q, A442Q, T319Q, A443N, N322Q. Binding energy improved by 66.0 kcal/mo	ol or
16.5% over wild type. (B) Mutations for binding mode 2: V339Q, T422Q, T338N, N44	45Q,
N323Q, S341Q. Binding energy improved by 46.6 kcal/mol or 11.4% over wild type	126
Figure S4-10 - Schematic showing placement of CS-E binding modes bound to EphB3	.128
Figure S4-11 – Electrostatic surfaces of (A) EphB3 and (B) EphB2. Sphere regions used	d for
coarse docking are shown in green for (C) EphB3 and (D) EphB2. Note that the reg	gions
sampled cover the positively charged regions of the proteins	. 129
Figure S4-12 – Detailed view of the best CS-E/EphB3 Mode 1 binding pose.	.130
Figure S4-13 - The best CS-E/EphB3 Mode 1 binding pose, with the VDW surface of	f the
protein shown to illustrate how well the ligand fits to the protein	.131
Figure S4-14 – The placement of all CS-E/EphB3 Mode 1 poses. The top pose in this n	node
is the #3 pose overall (-377.46 kcal/mol), but this mode shows the most consistent	cy in
placement	.132
Figure S4-15 – (A) Detailed view of CS-E docked to EphB3 in top pose from binding mod	de 2.
(B) Placement of all CS-E poses docked to EphB3 in binding mode 2. The top pose in	ı this
mode is the #1 pose overall (-380.80 kcal/mol), but shows less consistency in	pose
placement than Mode 1. (C) Top view of the best Mode 2 pose appears to fit closely to	o the
protein surface, but the rotated view (D) shows that the middle section of	the
octasaccharide is separated from the surface.	.133
Figure S4-16 – (A) Detailed view of CS-E docked to EphB3 in top pose from binding mod	de 3.
(B) Placement of all CS-E poses docked to EphB3 in binding mode 3. The top pose in	i this
mode is the $\#2$ pose overall (-380.53). This mode shows less contact with the surface	te of
the protein. (C) Top view of the best Mode 5 pose appears to fit closely to the pro-	otein
surface, but the rotated view (D) shows that the much of the octasaccharide is separ	
Figure S4.17 Discompany of only CS E page dorked to EnhB2 in hinding mode 4. This p	. 134
rigure 54-17 – Placement of only CS-E pose docked to EpnB5 in binding mode 4. This is	125
Eigure S4.18 Diagement of the only CS E pose decked to EphB3 in binding mode 4.	. 155 with
the protein surface shown	136
Figure S4 19 Discement of the only CS E pose docked to EphB3 in binding mode 5	This
mode contains only one pose in the top 25 poses. This pose is $\#22$ overall (-31	8 65
$\frac{1}{100}$ kcal/mol	137
Figure S4-20 – Placement of the only CS-E pose docked to EphB3 in binding mode 5	with
the protein surface shown.	.138
Figure S4-21 – Structure with proposed mutations for Mode 1: T319O. N322O. V32	390.
A4420 A443N 14460 and T4480. We predict that this set of mutations for Mo	de 1

A442Q, A443N, I446Q, and T448Q. We predict that this set of mutations for Mode 1 improves binding energy by 66.03 kcal/mol, or 16.5% better than binding to the wild-type.141

Figure S4-26 – Structure and pharmacophore for Mode 4 mutations: S341Q, A388Q, I446Q. This improves binding energy by 25.71 kcal/mol or 6.43% better than the wild-type. 145

- Table S4-5 Binding energies for all mutations to glutamine that improve binding energy for the best pose in mode 2. Note that most mutations do not make new hydrogen bonds (highlighted in red). The increase in binding energy for those mutations can be attributed

xvi

ΔV1
to Coulomb energy. We only wish to use mutants that make new contacts with the
ligand (highlighted in green)150
Table S4-6 - Binding energies in kcal/mol for different sets of mutations to best pose from
binding mode 2. Sets are ranked by binding energy. Set 4 (163.5 kcal/mol or 40%
improvement in binding energy) and Set 11 (46.6 kcal/mol or 11.4% improvement in
binding energy) are both selected for mode 2 due to the presence of residue E424. It is
interesting to find E424 in such close proximity to a negatively charged ligand like CS-E
and we are unsure of what other role E424 may be playing in the protein. Thus, mutations
to E424 may have unexpected consequences, even with a relatively close mutant such as
glutamine
Table S4-7 - Binding energies for all mutations to glutamine that improve binding energy for
the best pose in mode 3. Note that only <i>one</i> residue was able to make a new contact with
the ligand. This is a strong indicator that binding mode 3 is not a reliable result. The
increase in binding energy for those mutations can be attributed to Coulomb energy 152
Table S4.8 - Binding energies for all mutations to glutamine that improve binding energy for
the only pose in mode 4. Note that most mutations do not make new hydrogen bonds
(highlighted in red). The increase in binding energy for those mutations can be attributed
to Coulomb energy. We only wish to use mutants that make new contacts with the ligend
(highlighted in group)
(Ingingined in green)
Table S4-9 - Binding energy in Kcal/mol for the only set of mutations to the only pose from $1 \cdot 1 $
binding mode 4. Mutation results in 25.7 Kcal/mol or 6.4% improvement in binding energy.15.
Table S4-10 - Binding energies for all mutations to glutamine that improve binding energy for
the only pose in mode 5. Note that most mutations do not make new hydrogen bonds
(highlighted in red). The increase in binding energy for those mutations can be attributed
to Coulomb energy. We only wish to use mutants that make new contacts with the ligand
(highlighted in green)
Table S4-11 - Binding energy in kcal/mol for mutations to the only pose from binding mode
5. Set 1, which includes a mutation to E361, improves binding energy by 133.8 kcal/mol or
37%. Set 2, which skips the mutation to E361, improves binding by 29.1 kcal/mol or 8.1%.
Similarly to binding mode 2, it is interesting to find E361 (a different glutamic acid) in such
close proximity to a negatively charged ligand like CS-E and we are unsure of what other
role E361 may be playing in the protein. Thus, mutations to E361 may have unexpected
consequences, even with a relatively close mutant such as glutamine154

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 Shoichet¹ 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 uopioid receptor (mOR) is carried out via G protein signaling, side effects such as the buildup of tolerance may be due to the β -arrestin pathway²⁻⁴. 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

- 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
- 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
- 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
- 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
- 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
- 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