Antibody Structural Modeling and Docking

Research Objectives

Significance of Antibody Structures

Therapeutic antibodies are powerful tools for the prevention and treatment of human and infectious disease because of their high affinity and specificity for target antigens. Antibody drugs are available to treat cancer, infectious and cardiovascular diseases, arthritis, inflammation and immune disorders,[1-7] and new therapeutic antibodies are driving further growth of the biotechnology industry.[8,9] Structures of antibodies in complex with their cognate antigens can yield insight into biological phenomena or drug and disease mechanisms, and the increasing sophistication of computational techniques makes it possible to increase antibody-antigen binding affinity[10,11] or deduce the structural origin of such affinity maturation.[12] However, a crystal structure may not be readily available for most newly developed antibody sequences, in which case an accurate antibody homology model and a model of the docked antibody-antigen complex is needed to perform structure-based antibody engineering *in silico*.

Although the basic structure of an antibody is conserved across the hundreds of solved x-ray crystal structures, the structure of the paratope varies, and it is the paratope that is critical for docking and antibody engineering. It is necessary for predictions to achieve near-Ångström accuracy for practical use of these models such as engineering of stability or affinity. For computational antibody approaches to be reliably useful, it is critical to improve the fidelity of the models through better prediction methods.

Much can be accomplished with the ability to predict antibody structure and antibody-antigen complexes. Immunomics has now provided single-cell analysis of human antibody repertoires[13,14] and high-throughput sequencing of antibody repertoires of zebrafish,[15,16] a human-derived combinatorial library,[17] murine bone marrow plasma cells,[18] and human CDR H3 repertoires.[19] Current technology limits sequencing to short read lengths on a single chain, or to small sets of sequences from single-cell analysis, but soon it will be possible to have a full set of heavy and light chain sequences for a biological antibody repertoire. Since antibody repertoire sizes for an individual range from 10^3 for zebrafish[15] to 10^{10} for humans,[17,20] computational approaches will be the only practical high-throughput method to create structures for the repertoire. Entire repertoires of structures will offer an unprecedented amount of information to answer questions about the biological process of creating a repertoire, how autoantibodies are filtered, and how antibodies are selected for recognizing pathogens. Such information will be useful in fighting immunological diseases and designing vaccines. Furthermore, understanding the biological process of rapidly evolving a protein-protein interface will suggest strategies for computational design of interfaces, a critical task for intercession in systems biology.

Our immediate application will be to determine complex structures for polyclonal repertoires isolated from bone marrow plasma cells after immunizations by two different antigens. This accomplishment will be significant in that it will show how the multitudes of antibodies in a polyclonal repertoire cover the surface of the antigen, revealing the immunodominant regions and the regions that escape antibody binding. Such information is critical, as vaccine design depends on choosing antigens that will elicit pathogen-neutralizing antibodies.[21] For example, HIV neutralization can be achieved by coordinating antibody binding to different evolutionarily constrained regions.[22] The particular systems we selected for study, ovalbumin and tetanus toxoid, are significant because of their roles in food allergies and as a model vaccine, respectively.

Besides the intrinsic impact of predicting antibody structures, antibodies serve as a model protein for developing broadly applicable computational homology modeling and docking methods.[23] Much is known about the genetic process of antibody formation and the conserved structure of antibodies. Thus, for studying loop building for homology modeling, we have a reliable framework to build upon. For prediction of binding and affinity, antibody paratopes can be superposed on non-cognate complexes to create cross-docking targets, thus allowing examination of differences between cognate and non-cognate partners. Loop modeling is an especially difficult problem that must be solved for homology modeling or docking of structures for which the loop is unresolved in the crystal.[24] It has long been a goal in computational biology for structure prediction methods to be applied on a genomic scale.[24,25] Our choice of antibodies allows us to isolate and test methods on a biologically and industrially relevant system.

Predictions of antibody and antibody-antigen complex structures

My lab developed a novel antibody modeling tool (RosettaAntibody).[26] Our approach uses a database of known antibody structures to select template V_H and V_L frameworks and canonical CDR loops (L1-L3, H1 and H2), which are grafted to the frameworks. The hypervariable CDR H3 loop is constructed using a combination of fragment assembly and high-resolution refinement in a Monte Carlo-plus-minimization scheme that simultaneously optimizes V_H-V_L orientation and the surrounding paratope loops. We tested the original implementation on a new benchmark of 54 antibody

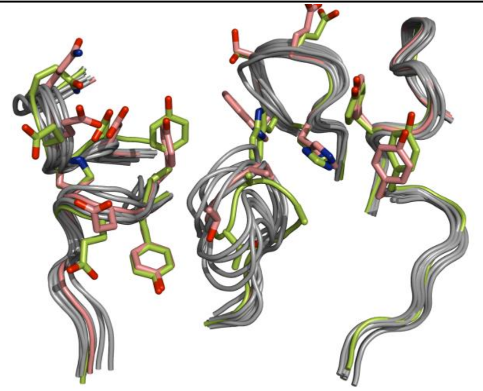

Figure 1: RosettaAntibody paratope models. Native structure of D3H44 Fab (1JPT), light-green; models, grey, low-rmsd model, salmon.

crystal structures. The median root mean square deviation (rmsd) of the antigen binding pocket comprised of all the CDR residues was 1.5 Å with 80% of the targets having an rmsd lower than 2.0 Å. The median backbone heavy atom global rmsd of the CDR H3 loop prediction was 1.6, 1.9, 2.4, 3.1, and 6.0 Å for very short (4-6 residues), short (7-9), medium (10-11), long (12-14) and very long (17-22) loops, respectively. Figure 1 shows the paratope from the ten top-scoring homology models of D3H44 Fab (PDB ID 1JPT,[27] H3 length 8 residues). The models show diversity sampled extensively in the vicinity of the native x-ray structure (light-green). The low-rmsd model (salmon) shows that many side chains are predicted near their native conformation. Most importantly, we showed that the models are sufficiently realistic to allow successful docking in some cases.[26]

We recently extended our method to the prediction of single-chain camelid antibodies, although accuracy is limited due to the long CDR H3 loops.[28] We provide access to the classic two-chain method via a web server;[29] over 2300 models have been built since the server's public release in 2009 leading to several publications from outside my lab.[30-36]

We advanced flexible docking by sampling multiple backbones simultaneously (EnsembleDock). The primary limitation in the docking field is the ability to dock successfully when the unbound backbone conformation differs from the bound. To address this challenge, we developed methods following the conformational selection and induced fit kinetic models of binding. In tests on docking targets with small amounts of conformational change, we showed that our method can improve both sampling and model quality.[37] The conformational selection method, EnsembleDock (Figure 2), can also be used to dock homology models that inherently have regions of uncertainty and can be represented as an ensemble of structures. Further, we demonstrated for the first time successful docking by using an entire ensemble of NMR models.

We created the first docking method tailored for antibody models (SnugDock). SnugDock leverages our ability to sample the degrees of freedom specific to an antibody.[38] CDR H3 and other CDR loop conformations and V_H-V_L relative orientation are sampled during the antibody-antigen docking process. SnugDock can be combined with the multiple-backbone method (EnsembleDock) for rapid docking of alternate low-energy homology models of the antibody. Together, the combined algorithm produced four medium and seven acceptable lowest-interface-energy predictions in a test set of fifteen complexes (using CAPRI rating criteria[39]). This work is the first to report robust docking of homology models.

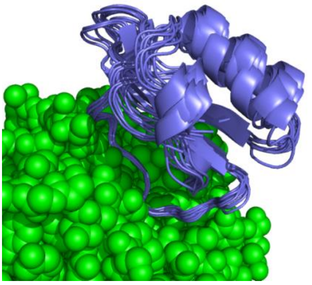

Figure 2: EnsembleDock selects potential new backbone conformations from a pregenerated ensemble.

Proposed Research

Our long-term goal is the accurate prediction of structures of antibodies and antibody-antigen complexes. Our specific aims are:

1. **Develop and test methods for accurate prediction of the structure of long, hypervariable CDR H3 loops.** Using kernel density estimates to gather statistics on CDR H3 subsequences, we will develop methods to identify short segments likely to form β turns (*i* to *i*+3 proximity). Subsequence and position information will restrict conformational sampling to effectively reduce the degrees of freedom to make long loop prediction feasible. Several loop modeling methods will be tested including kinematic closure[40] and a stepwise buildup method with configurational bias moves.
2. **Develop and test methods for docking with backbone flexibility using expanded ensemble approaches.** We will extend our EnsembleDock method to incorporate more extensive backbone flexibility by using large, pregenerated ensembles of clustered structures and a map to guide movements to neighboring structures. These methods will address currently intractable docking targets and also improve docking of homology models with backbone uncertainty.
3. **Develop and test methods for prediction of binding and affinity.** We will test reformulations and recalibrations of the energy function for (1) identifying binding and non-binding antibody-antigen interactions by docking unbound structures and (2) quantitative binding affinity prediction on both a recent benchmark set and antibodies for which point mutations have dramatic effects on affinity.
4. **Apply improved antibody modeling and docking methods to a large polyclonal repertoire.** From recent high-throughput sequencing studies, we will select 100 antibodies that bind the same antigen to predict structures of the antibodies alone and in complex with the antigen. We will examine the diversity of predicted epitopes and paratopes to explore the determinants of immunodominance. Data for two model antigens are currently available, including (1) ovalbumin, a culprit in food allergies, and (2) the tetanus toxoid fragment C, a vaccine.

These aims have recently been funded by the NIH (R01-GM078821, $204,800/year, 9/30/2012-8/31/2016).

Computational Methodology (application/codes)

Rosetta Modeling Platform

Rosetta is the best platform upon which to build our antibody methods because of its broadly tested energy function,[41-45] extensive library of interchangeable structure and sequence optimization tools,[46,47] and its success in folding,[48-52] docking,[37,53-55] design,[56-60] and numerous other applications.[26,38,40,41,61-63] As a member of the Rosetta Commons, my lab is one of the core contributors to the Rosetta project, leading development of the Rosetta applications in docking, antibody prediction, and protein-solid surface interactions. Together with Rhiju Das of Stanford, I lead the development of ROSIE, a web gateway to a variety of Rosetta protocols (see related proposal for a gateway allocation). My lab includes the lead test engineer for the Rosetta Commons, and we developed the PyRosetta interface to the code which allows facile and rapid development of new methods.[47] All developments will be incorporated and freely disseminated with the Rosetta and PyRosetta packages.

There are several underlying philosophies to the Rosetta approach. In most prediction problems, we seek the lowest free energy state among many possible conformations. Therefore, we use a variety of optimization methods to search for those conformations and a score function to identify the lowest energy state. The central optimization approach is a Monte Carlo-plus-minimization[64] algorithm that allows us to combine explicit, gradient-based minimization with discontinuous jumps through continuous spaces or discrete pre-generated libraries. In this way, we rapidly cover relevant conformation space to find local optima in the score function. Docking moves consist of rigid-body rotations and translations, and these are combined with a simulated annealing search for optimal side-chain conformations using a discrete rotamer library.[43,65,66] Antibody prediction also includes moves to explore loop conformations and to displace the heavy and light chains.

Rosetta uses a multi-scale approach, where a low-resolution representation is used to rapidly search large regions of conformation space, and a high-resolution, all-atom representation is used to provide accurate energetic discrimination between structures. There are scoring functions for each stage. The low-resolution stage is a combination of sterics and statistical potentials, and the high-resolution stage is a combination of physical potentials optimized to reproduce high-resolution x-ray structures and their features. Rosetta's object-oriented frameworks for score functions and optimization routines[46] allow us to rapidly prototype our new methods and combine them with previously established approaches.

Stampede is the best resource for our research needs

We chose Stampede as an ideal resource for our research for several reasons:

1. Our critical need is for CPU-hours. Stampede provides over 100,000 cores for these calculations.

2. Due to the underlying multi-start Monte Carlo algorithm, Rosetta's embarrassingly parallel scale-up for docking and antibody modeling means that the jobs will take advantage of Stampede's computational power with almost no penalty for scale-up to multiple cores.
3. The PI recently completed a sabbatical at the University of Texas at Austin where he ported the code to both Lonestar and Stampede. Stampede was more efficient and easier to manage. The PI also met with TACC staff regarding the installation of Rosetta and PyRosetta as standard modules on Lonestar and Stampede.
4. Through an ECS with our Startup grant (TG-MCB130133) and continued work proposed in a synergistic proposal for the ROSIE gateway, Rosetta is being optimized to run on Stampede's Intel Phi co-processors. These improvement have the potential to increase ROSIE's efficiency on Stampede by an order of magnitude or more.
5. Stampede has ample storage and memory for the antibody and docking applications.

Performance and Scaling

The antibody and docking protocols are embarrassingly parallel, *i.e.*, they compute multiple trajectories where each core calculates one trajectory without need to communicate to other cores besides job and initial seed assignment. Therefore these jobs will use multiple CPU cores for its computations with extremely little or no penalties.

Typical antibody jobs require between 300-350 CPU-hours. On Stampede, we will assign appropriate job sizes for each protocol, using MPI to manage the job across cores. Figure 3 below shows the scaling performance of the antibody CDR H3 modeling application. As the number of CPUs are increased from 16 to 512, there is very little penalty in efficiency. At 1024 CPUs, the efficiency drops off to about 80%. This drop in efficiency can be easily explained and gives insight into the best practice for choosing the number of cores for each job. The antibody job creates 1000 candidate structures using 1000 independent trajectories. The 1000 trajectories are each assigned to one core; in the case where 1024 cores are requested, 24 cores are idle immediately. As each trajectory finishes, since there are no remaining trajectories to compute, the assigned core will become idle. Under MPI, all cores wait to exit until all trajectories are complete. For jobs with smaller numbers of cores, there will be less time wasted waiting for all jobs to complete. Thus, the number of CPUs should be smaller than the number of trajectories, at least by a factor of two and higher if rapid results are not needed. The final decision is a trade-off between computational efficiency and rapid job completion time. We will typically choose cores equal to about one quarter of the number of independent trajectories to compute. For the antibody application, which averages 361 CPU-h per job, this corresponds to 97% efficiency and job completion in around 1.5 hours on 256 cores.

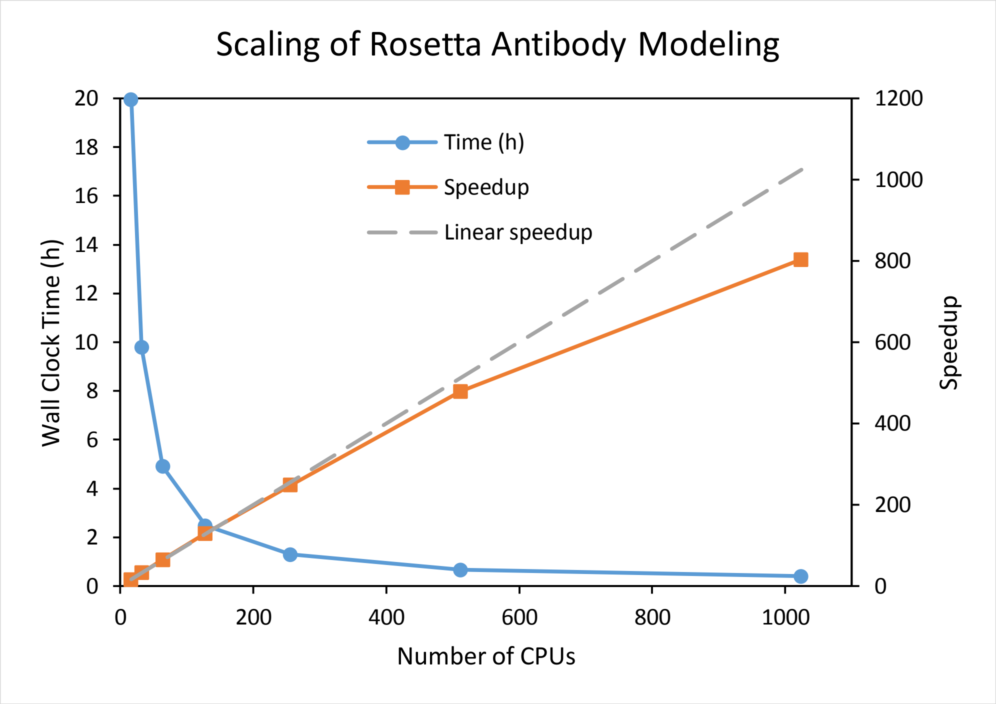

Figure 3: Scaling of Rosetta's antibody modeling application. Test case is a naïve repertoire antibody (Ab ND2-1005) with a CDR H3 loop of 12 residues. Calculation performed on Stampede using standard run options. Code was compiled with Intel C++ compilers using MKL libraries for automatic offloading of appropriate computations to the Intel Xeon Phi co-processors.

Computational Research Plan

We propose to complete the following runs on Stampede in the coming year:

1. **Antibody H3 loop modeling.** We will test KIC and CCD methods on a small test set of 20 antibodies and then a larger test set of 100 antibodies.
2. **Flexible docking conformer generation.** We will test several methods (Rosetta relax, backrub, etc.) of generation of ensembles for structures for conformer selection docking. We will then test conformer selection docking on a small test set of targets and then on a large benchmark set.
3. **Score function calibration and affinity prediction.** Using local docking data of near-native and non-native protein-protein interfaces, we will calibrate score functions for docking. Score functions tested will include both (a) pH-based score functions based on our recent work on pK_a prediction and variable protonation state side-chain packing and (b) Talaris-2013 based score functions using the functional form agreed upon at the April 2013 Talaris conference, and combinations of the two.

4. **Repertoire docking.** We will test the docking of a repertoire of antibodies against the model antigen hemagglutinin (comparing against known antibody-antigen complex structures) as well as novel antigens including the tetanus toxoid vaccine.

Justification for SUs Requested

Table 2 summarizes the computational needs for the runs for the four aims in our study. The largest needs are for runs to generate alternate conformations to capture protein flexibility upon docking. For most aims, we will first use a small test set of 10-20 proteins, or targets, and later scale up to a large benchmark set as is the standard in the field. Although we have over 1,500,000 CPU-h of work in these aims for the next year, our request is for 1,000,000 since we will continue to perform some of these runs on our lab cluster as we transition to the shared XSEDE resources.

Table 1: Major runs planned for the year.

Aim	Run type	Protocol	CPU-h /run/target	Targets	Runs	Total CPU-h	Storage (GB)
1	H3 testing	KIC, CCD	250	15	5	18,750	1
1	H3 production	KIC	250	50	3	37,500	5
1	Ab homology testing	antibody_H3	300	20	10	60,000	5
1	Ab homology production	antibody_H3	300	100	2	60,000	5
2	Conformer generation	Relax	100	70	10	70,000	8
2	Conformer generation	Backrub	30	70	10	21,000	16
2	Conformer generation	KIC	600	70	10	420,000	8
2	Flexible docking testing	Expanded EnsembleDock	1000	10	10	100,000	50
2	Flexible docking production	Expanded EnsembleDock	1000	70	5	350,000	200
3	Score function calibration	Docking	200	100	5	100,000	250
3	Docking benchmarking	Docking	100	180	10	180,000	500
4	Hgg docking	SnugDock/Ensemble Dock, ReplicaDock	2000	20	2	80,000	500
4	Tetanus docking	SnugDock/Ensemble Dock, ReplicaDock	2000	20	1	40,000	500
	TOTAL					**1,537,250**	**2,048**

Storage Needs

Our work fits within the limits of Stampede (5 GB on $HOME, 400 GB on $WORK). Most data only need transient storage. For example, candidate predicted docked structures are retained during a large run, and at the end the lowest-energy structures are retained. Thus, temporary storage on $SCRATCH will require 2 TB throughout the project, with peak loads around 0.5 TB at any one time. Distilled final data will be copied back to the Gray lab servers. 1 TB of space is requested for data backup on Ranch.

Additional Considerations

HPC Facilities in the Gray lab

For the past decade, we have maintained our own computing cluster in our lab. The last major upgrade to this equipment was in 2010, and we are in the process of transitioning to more efficient shared resources. Our computing cluster is a Dell PowerEdge high-performance computing cluster composed of 8 dual hex-core nodes, 13 dual quad-core nodes, and 32 quad-core nodes for a total of 328 CPUs at up to 2.66 GHz with 1-2 GB of memory per core. The cluster has 11 TB of storage and gigabit networking and is located in Garland Hall on the Johns Hopkins University Homewood campus with appropriate cooling and power supply. The cluster is capable of creating, evaluating and storing hundreds of thousands of protein structures per day, and it is for the exclusive use of the Gray laboratory.

Desktop and cluster servers participate in a shared central data backup system (10 TB disk cache backed up by 0.5 PB LTO-4 based tape library).

The Gray lab also hosts the Rosetta Commons Testing Cluster of 13 dual quad-core Dell PowerEdge nodes (2.5 GHz, 8 GB memory per node) and one 3 TB server. This system runs and reports the results of continual tests of the Rosetta source code (build tests, unit tests, integration tests, performance tests, and scientific tests) as it is updated by ~100 developers at over 15 participating labs worldwide. The Rosetta cluster is not available for scientific use by the Gray laboratory.

Other XSEDE Allocations

The Gray lab has an XSEDE startup allocation for testing these protocols and various projects related to Rosetta and PyRosetta dissemination. The startup allocation was awarded from May 10, 2013 through May 9, 2014. We used our initial allocation rather rapidly and subsequently requested and received a supplemental allocation, bringing our total allocation to 50,000 SU on Lonestar and 200,000 on Stampede.

Along with co-PI Rhiju Das at Stanford, the PI is submitting a second application to support a Rosetta gateway server named ROSIE. Das and Gray developed an online server to interface to a variety of Rosetta protocols including antibody modeling, docking and RNA folding. The goal of the ROSIE proposal is to use Stampede as a back-end to the web server to complete biomolecular modeling calculations for the scientific community via a simple web interface.