

# David F. Green SeaWulf Research Program

## Scientific goals

We are using a range of computational and theoretical methods to study problems of specificity in protein–protein interactions. We are interested both in understanding underlying chemical and physical principles and in developing novel methods, with our primary focus of study being various components of the heterotrimeric G-protein signal transduction pathway, a system of significant medical and biotechnological relevance. Systems involving HIV-1 viral–cell recognition and entry are also a target of study.

**Understanding specificity of protein–protein interactions.** The chemistry of biology takes place within an incredibly complex, heterogeneous environment, including a rich diversity of many classes of molecules. However, the proper function of biological systems requires the specific association of individual molecules from this vast array of molecules into functional complexes. Understanding the basic principles behind this specificity of interaction has important consequences, not only for understanding the natural complexity of biology, but also for the design of novel biomolecular complexes. Beginning with subunit association in the heterotrimeric G-proteins, we are studying the basic biophysical principles behind both promiscuous and highly specific binding. Starting with experimentally determined structures, we are using computational methods to develop detailed maps of the various interactions contributing to affinity in complexes of varying sequence. In addition, the complete network of “allowed” and “disallowed” interactions between the three components of this system is being elucidated using techniques from the field of computational protein design, with a focus on developing an understanding suitable for application to the design of specifically interacting molecules. Ultimately, this analysis will be extended to both upstream and downstream components of the signalling network.

**Effects of glycosylation on protein interactions.** Proteins can be chemically modified in numerous ways that alter the properties of the affected protein, and may have profound effects on molecular association. These effects can be due to gross physical changes (*e.g.* large scale conformational change, blocking of a binding site, or significant change in charge at the binding interface), or to more subtle changes (*e.g.* local conformational change, altered dynamics of the molecule or binding site, or long-range electrostatic effects). Understanding how the chemical modification of proteins affects their association properties will help in furthering our understanding of regulation of cellular processes. While known to be an essential component in modulating essential biomedical characteristics such as immune system recognition and viral entry into cells, the effects of glycosylation on protein structure, dynamics, and function are particularly poorly understood. We are extending traditional methods for the study of protein systems to investigate the effects of glycosylation on protein–protein interactions. As an initial system, we are focusing on the interactions made by the HIV-1 surface glycoprotein gp120 as part of the cell recognition and entry process. Future work will extend the methods and understanding gained in this system to the effects of glycosylation on the function of additional systems, including additional viral cell-entry mechanisms.

## Computational Tools

**Analysis of biomolecular structural energetics.** The energetics of biomolecules are typically modeled using a Molecular Mechanics (MM) approach, in which an energetic model based on classical physics is parameterized to reproduce data from both experiments and higher-level quantum mechanical models. Within this model, the energy of a single microstate is very fast to evaluate, but accuracy in calculated results requires evaluation of the entire ensemble of accessible microstates. Molecular Dynamics (MD) provides a means to obtain this ensemble, through propagation of the laws of motion over the MM energy surface, and provides the additional benefit of describing molecular motions in exquisite detail. These calculations are implemented with use of the CHARMM software package.

**Continuum electrostatic models of solvation.** A major challenge of molecular modeling is obtaining a reasonably complete sampling of the microstates of the highly mobile solvent molecules. Proteins and

protein complexes typically have well-defined structures, with dynamical motions around a single average, and thus a MD approach can provide a reasonable sampling of possible structures without excessive computational cost. However, any near-complete sampling of solvent degrees of freedom requires immense effort. One solution to this is the use of another approximation from classical physics — the dielectric continuum model. Explicit solvent molecules are replaced with a continuum of high dielectric constant, and the Poisson–Boltzmann (PB) equation is solved to obtain solute-solvent electrostatic interaction energies; a microscopic surface tension is used in addition to account for non-polar effects in a surface-area dependent fashion (SA). These PB/SA computations are somewhat more expensive for a single state than a single-state molecular mechanics calculation, taking on the order of minutes, but only need to be performed on a relatively small number of structural states. These calculations, and algorithms to decompose the energies into contributions from different components, are implemented in the ICE software package, co-written by Dr. Green.

**Protein design and structural model building** A recurring problem in biophysical modeling is the prediction of the structure of a protein or protein complex. While a solution to the general protein-folding problem remains elusive, an important subset of this problem has become tractable. This is the side-chain packing problem — the prediction of the preferred conformations of amino-acid side chains given a fixed, or nearly fixed, conformation of the protein backbone. This provides a means to build reasonable structural models using the known structure of related systems (homology modeling), as well as to design novel sequences that will stabilize a given target structure (protein design). The problem is one of a combinatorial search; each position in the sequence is allowed many conformations and, in protein design, many amino acid choices. The Dead-End Elimination (DEE) and A\* algorithms provide a highly efficient procedure to solve these problems, representing the problem as one of a tree search. An implementation of these search algorithms has been graciously provided by the Tidor group at MIT.

**Hierarchical models for protein design** While algorithms such as DEE/A\* are incredibly efficient, they rely on approximations to the underlying energetics that limits the accuracy of the final results. A solution to this is the use of hierarchical methods, where low energy ensembles of states are passed through successively more accurate, but more costly, energetic models; at each stage of the hierarchy, a decreasing number of states is evaluated. This allows for the efficiency of the DEE/A\* algorithms to be harnessed, while maintaining accuracy in the final results. These methods can also be extended to consider more complex properties at various stages in the hierarchy. These hierarchical schemes involve use of all of the previously described software, integrated through locally written scripts.

## Computational Needs

The MD simulations and DEE/A\* searches are the single most involved computations used in these studies; both can be parallelized, but with some loss of efficiency. However, the bulk of the calculational effort in both cases is not in the initial calculation, but rather is in the individual calculations of structural energetics. While the requirements for individual calculations of the energetics of any given state are modest, both the analytical methods (MM/PBSA and energetic component analysis) and protein design problems (propagation through a hierarchy of models) require the evaluation of the energies of a large number of states. Due to the independent nature of each of these calculations, they are perfectly suited to a cluster computing environment. The simulation and detailed analysis of a complex requires about 2,000 CPU hours at minimum, and 10,000 CPU hours or more for larger systems or more complete analyses, generating 5-25 GB of data. A typical protein design calculation, from the initial search through a final stage ranking of the low energy states, may be estimated at roughly 7,500 CPU hours, and will generate roughly 5 GB of stored data (compressed). A total of 50,000 CPU hours per month would allow for reasonable progress on the G-protein association studies (including both analysis and design), and an additional 25,000 CPU hours per month are required for the studies of HIV-1 cell-entry proteins. These studies will generate data at an expected rate of 100–200 GB per month, with 2–4 TB of total storage space likely adequate over an extended period.