Using Computational Thinking for Designing Resistance-Evading Drugs

Christopher James Langmead

August 22, 2008

Project Summary

The research objective of this proposal is to apply Computational Thinking to the task of designing drugs against infectious agents (i.e., viruses and bacteria). One of the key challenges in the treatment of viral and bacterial pathogens is the emergence of resistance-conferring mutations. The presence of a drug places a selective pressure on the pathogen's genome which, in turn, forces rapid evolution. The two primary risks associated with forcing evolution are (i) the reduction in the efficacy of the drug and, perhaps more importantly, (ii) the possible emergence of a significantly more harmful strain. We believe that these risks can be mitigated through the use of Computational Thinking in the design process. Specifically, we propose to develop a new design strategy that focuses on the design of a drug that will be effective against a given target molecule, **T**, *and* any variant of **T** that is likely to arise due to drug-induced evolution.

Overview

The goal of drug design is to find a drug which will disrupt a target molecule's function. For example, the inhibition of the protein HIV-1 protease disrupts HIV's ability to replicate and is therefore an important target for drug design. The traditional method for drug design can be interpreted as an optimization problem. Given a set of possible drugs, $\mathcal{D}$, a target molecule, **T**, and a function $G : \mathbf{T} \times \mathcal{D} \rightarrow \mathbb{R}$ that returns the *free energy* of binding between a drug and the target, the goal is to find a drug $d^* \in \mathcal{D}$ that minimizes G. There are a variety of ways to solve this optimization problem (at least approximately) using computational methods.

Unfortunately, by using drug d^*, target **T** is likely to evolve into a new target, $\mathbf{T}'$, which no longer optimally binds to d^*. This is the essence of drug resistance — nature responds to the drug by presenting a different target. However, if we use Computational Thinking and generalize the problem somewhat, we can imagine a multi-objective optimization problem where the goal is to find a drug (or drug cocktail) that minimizes the free energy of binding to a set of targets. That is, we will *anticipate* the set of likely mutations when designing the drug. More formally, let $\mathcal{T} = \{T_1, ..., T_k\}$ be the set of viable sequences for some target molecule **T**, and let $G_{\mathcal{T}} : \mathcal{T} \times \mathcal{D} \rightarrow \mathbb{R}$ be a function that returns the free energy of binding between a drug and each element of $\mathcal{T}$. In principle, if we can minimize $G_{\mathcal{T}}$, we can design drugs that will evade resistance.

Computationally, we can solve this problem by using advanced methods for modeling molecular systems. Previously, our work has focused on the use of Markov Random Fields for modeling protein structures and computing free energy calculations, including binding free energies (e.g., [2, 3, 4, 5]). A Markov Random Field (MRF) is a compact encoding of a multivariate probability distribution. In the context of molecular modeling, a MRF encodes the Boltzmann distribution over configurations, $P(\mathcal{C})$, where $\mathcal{C}$ is the configuration space of the target molecule. The partition function, Z, of this distribution can be approximated using Generalized Belief Propagation (GBP), which is an inference algorithm for probabilistic graphical models. Free energy is simply the negative log of Z; that is, $G = -\ln Z$.

We have pioneered the use of MRFs and GBP for free energy calculations and have demonstrated its practical advantages in a variety of settings. In particular, we have used it to compute $G : \mathbf{T} \times \mathcal{D} \rightarrow \mathbb{R}$ by

modeling the distribution $P(\mathcal{C}|d)$, where d is a drug. In a preliminary drug design project we computed the binding free energies of 36 putative drugs against a single target (unpublished results). The correlation coefficient between the predicted and experimental binding free energies was 0.714, with a standard error of 1.7 kcal/mol, outperforming a state-of-the art method for computer-aided drug design.

The goal of this year-long project will be to extend our method to compute multi-target free-energies (i.e., $G_{\mathcal{T}} : \mathcal{T} \times \mathcal{D} \rightarrow \mathbb{R}$). To do this, we will extend the MRF to model the probability distribution over sequence *and* structure, $P(\mathcal{T}, \mathcal{C}|d)$. That is, the MRF will model *every* possible mutation to a given target T. Binding free energies across sequence and structure can then be computed using GBP, as before.

The proposed method represents an elegant approach to the design of resistance-evading drugs because it models the conditional probability distribution over sequence and structure, given a drug. The change in the distribution over sequence and structure represents a thermodynamically based model of evolution. Indeed, we can use GBP to actually predict the set of sequences likely to arise in the presence of a drug. Additionally, we can use the computed binding free energies to rank a set of putative drugs. We note that the use of MRFs for simultaneously modeling sequence and structure was first proposed in [3] and subsequently demonstrated in [1]. Thus, the feasibility of the proposed method has been demonstrated.

Connections and Complementarity to ongoing Research at MSR

Microsoft Research (MSR) has conducted pioneering research in the application of Machine Learning and Formal Method to key problems in Biology and Medicine. Among the many contributions are cutting-edge Machine Learning techniques for studying the evolution of the HIV *genomic* sequence in the context of vaccine design [6]. The proposed research also uses techniques from Machine Learning, but studies the evolution of *proteomic* sequences from a biophysical perspective and in the context of drug design.

Expected Outcomes

We will implement the extended model and demonstrate its applicability on HIV 1 protease. This molecule has been selected as a target because a) mutations in this protein have been well-documented, and b) there are a number of well-characterized drugs targeting this molecule. Thus, we will be able to validate our method against a large-body of research in this important molecule.

References

[1] FROMER, M., AND YANNOVER, C. A computational framework to empower probabilistic protein design. *Proc. of the 16th Ann. Intl. Conf. on Intel. Systems. in Mol. Biol. (ISMB) Toronto, ON, July 17-23)* (2008), in press.

[2] KAMISETTY, H., AND LANGMEAD, C. Conformational Free Energy of Protein Structures: Computing Upper and Lower bounds. *Proc. of 3DSIG 2008 Struct. Bioinf. and Computat. Biophysics, Toronto, ON.* (2008), in press.

[3] KAMISETTY, H., XING, E., AND LANGMEAD, C. Free Energy Estimates of All-atom Protein Structures Using Generalized Belief Propagation. In *Proc. of the 7th Ann. Intl. Conf. on Research in Comput. Biol. (RECOMB)* (2007), pp. 366–380.

[4] KAMISETTY, H., XING, E., AND LANGMEAD, C. Free Energy Estimates of All-atom Protein Structures Using Generalized Belief Propagation. *J. Comp. Biol.* (2008), in press.

[5] LANGMEAD, C., AND KAMISETTY, H. Detecting Protein-Protein Interaction Decoys using Fast Free Energy Calculations. Tech. Rep. CMU-CS-07-156, Carnegie Mellon University, 2007.

[6] NICKLE, D. C., ROLLAND, M., JENSEN, M. A., POND, S. L. K., DENG, W., SELIGMAN, M., HECKERMAN, D., MULLINS, J. I., AND JOJIC, N. Coping with viral diversity in hiv vaccine design. *PLoS Comput Biol 3*, 4 (Apr 2007), e75.