Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Designing Vaccines that are Robust to Virus Escape

Swetasudha Panda and **Yevgeniy Vorobeychik**
Electrical Engineering and Computer Science
Vanderbilt University, Nashville, TN
swetasudha.panda@vanderbilt.edu, yevgeniy.vorobeychik@vanderbilt.edu

Abstract

Drug and vaccination therapies are important tools in the battle against infectious diseases such as HIV and influenza. However, many viruses, including HIV, can rapidly escape the therapeautic effect through a sequence of mutations. We propose to design vaccines, or, equivalently, antibody sequences that make such evasion difficult. We frame this as a bilevel combinatorial optimization problem of maximizing the escape cost, defined as the minimum number of virus mutations to evade binding an antibody. Binding strength can be evaluated by a protein modeling software, Rosetta, that serves as an oracle and computes a binding score for an input virus-antibody pair. However, score calculation for each possible such pair is intractable. We propose a three-pronged approach to address this: first, application of local search, using a native antibody sequence as leverage, second, machine learning to predict binding for antibody-virus pairs, and third, a poisson regression to predict escape costs as a function of antibody sequence assignment. We demonstrate the effectiveness of the proposed methods, and exhibit an antibody with a far higher escape cost (7) than the native (1).

We formulate antibody design as a formal bi-level optimization problem, where the "designer" chooses an antibody so as to maximize the shortest sequence of mutations that lead to escape. This formulation can be viewed as a Stackelberg game (Paruchuri et al. 2008) (Brückner and Scheffer 2011) between the designer and the virus in which the virus minimizes the cost of escaping the antibody chosen by the designer. The designer-virus game poses two challenges: 1) enormous search space for both the designer and the virus ($\geq 10^{50}$ in each case), and 2) determination whether an arbitrary antibody-virus pair bind. To tackle the former challenge, we propose, and compare the performance of, several stochastic local search heuristics (Hoos and Stutzle 2004), using the native antibody as a "springboard". Even for computing virus escape alone, this approach scales poorly. The major bottleneck is the second challenge: binding evaluation. For this purpose we make use of Rosetta, a premier computational protein modeling tool (Gray et al. 2003). Rosetta, however, can be extremely time consuming even for a single evaluation (which could take nearly an hour, as it makes use of its own sophisticated amalgam of local search techniques to simulate a binding complex). To significantly speed up the search, we use classification learning to predict whether or not an antibody-virus pair bind, limiting Rosetta evaluations only to cases in which the classifier predicts that they do not. While this makes the virus escape search practical, the bi-level nature of the problem means that antibody design is still quite time consuming. To address this, we make use of Poisson regression to predict virus escape cost. Making use of the resulting predictions now makes antibody design viable, with "inner loop" (virus escape) evaluations restricted to a small set of candidate antibodies predicted to be difficult to escape.

In summary, we make the following contributions:

1. A bi-level optimization (Stackelberg game) model of antibody design and virus escape interaction,
2. stochastic local search techniques to determine optimal virus escape, with classifier-in-the-loop used to speed up the evaluations, and
3. stochastic local search techniques for optimal antibody design, making use of Poisson regression to predict minimal virus escape time.

Our methods ultimately exhibit antibodies that are far more robust to mutations than the native antibody.

Related work include (Lathrop and Pazzani 1999), (Hernandez-Leal et al.), (Lathrop et al. 1999), (Richter, Augustin, and Kramer 2010).

Antibody Design as a Stackelberg Game

Let v^0 denote the native virus, which we treat simply as a sequence (vector) of amino acids, and v and a arbitrary virus and antibody sequences, respectively. Let $O(a, v)$ represent binding energy for the antibody-virus pair (a, v), which is computed by Rosetta. We stylize the "dilemma" faced by the virus as the following constrained optimization problem:

$$\min_{v \in V} \|v^0 - v\|_0 \tag{1a}$$

$$\text{s.t.}: O(a, v) \geq \theta, \tag{1b}$$

where V is the space of virus sequences under consideration, and θ is a threshold on binding energy which designates escape (that is, once binding energy is high enough,

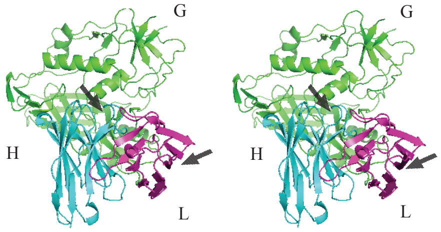

Figure 1: The native antibody, H and L, with the native virus, G (left) and antibody with escape cost=7 (right). The arrows point at some significant differences.

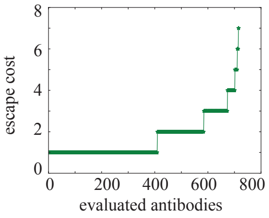

Figure 2: Evaluated antibodies for $\theta = 0$, ranked by escape cost. The native antibody escape cost is 1.

the proteins will no longer bind);[1] this threshold is typically domain-dependent. The l_0 norm simply computes the number of sequence positions in v that are different from v^0.

While in principle we could consider the space of all possible virus sequences in this subproblem, since virus structure and, consequently, its binding properties can be affected by a change in any residue (amino acid) in its sequence. However, first-order affect in regard to its antibody binding properties is determined by the sequence that is a part of the native virus binding site. Therefore, we only consider the problem of virus escape in terms of binding site mutations.

The optimization problem 1 can be viewed as a *best response* of the virus to a fixed antibody a. Now we consider the problem of designing an antibody, a, that is robust to virus escape. The target, virus escape, is now precisely defined by the virus optimization problem 1. Let $v(a)$ be the solution to this problem—naturally, a function of the antibody choice a. The designer's decision problem is then

$$\max_{a \in A} \|v^0 - v(a)\|_0, \tag{2}$$

where A is the antibody design space, which we restrict to the native binding site for the same reasons as for the virus. Alternatively, we can write this is a bi-level optimization problem composing 2 with 1.

Evaluation

To evaluate our approach we used a native antibody-virus interaction for HIV.

The native structure is the co-crystal structure of the antibody VRC01 complexed with the HIV envelope protein GP120.

The binding site on the virus is chain G with 45 residues, while the binding site on the antibody includes chains H and L with a total of 52 residues.

The visual representation of the native binding structure is shown in Figure 1 (left).

The actual set of antibodies we generated as a part of our search process, ranked in terms of evaluated escape cost (Figure 2). It is noteworthy that we found many antibodies which are much more robust to escape than the native when $\theta = 0$. In particular, our best has escape cost of 7, and the resulting antibody complexed with the native virus is shown in Figure 1 (right). Visually, the differences appear quite small, but make a significant difference in the ultimate breadth of binding, emphasizing the importance of a computational micro-level design approach.

Elaborated information on this research is at https://drive.google.com/file/d/0B3IRzPZ3ARZhdmh2SVBBVzhjQU0/edit?usp=sharing, and https://drive.google.com/file/d/0B3IRzPZ3ARZhX0FiR0U2Vjhzelk/edit?usp=sharing.

References

Brückner, M., and Scheffer, T. 2011. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 547–555. ACM.

Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; and Baker, D. 2003. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of molecular biology* 331(1):281–299.

Hernandez-Leal, P.; Fiedler-Cameras, L.; Rios-Flores, A.; González, J. A.; Sucar, L. E.; and Tonantzintla, S. M. Contrasting temporal bayesian network models for analyzing hiv mutations.

Hoos, H. H., and Stutzle, T. 2004. *Stochastic Local Search: Foundations & Applications*. Morgan Kaufmann.

Lathrop, R. H., and Pazzani, M. J. 1999. Combinatorial optimization in rapidly mutating drug-resistant viruses. *Journal of Combinatorial Optimization* 3(2-3):301–320.

Lathrop, R. H.; Steffen, N. R.; Raphael, M. P.; Deeds-Rubin, S.; Pazzani, M. J.; Cimoch, P. J.; See, D. M.; and Tilles, J. G. 1999. Knowledge-based avoidance of drug-resistant hiv mutants. *AI Magazine* 20(1):13.

Paruchuri, P.; Pearce, J. P.; Marecki, J.; Tambe, M.; Ordóñez, F.; and Kraus, S. 2008. Playing games with security: An efficient exact algorithm for Bayesian Stackelberg games. In *Proceedings of the Seventh International Conference on Autonomous Agents and Multiagent Systems*, 895–902.

Richter, L.; Augustin, R.; and Kramer, S. 2010. Finding relational associations in hiv resistance mutation data. In *Inductive Logic Programming*. Springer. 202–208.

[1] This idea may seem counterintuitive at first, but it is a reflection of the well-known tendency of chemical compounds towards low-energy states.