379a Identifying the Interacting Positions of a Protein Using Boolean Learning and Support Vector Machines

Anshul Dubey, Bernard Loo, Matthew J. Realff, Jay H. Lee, and Andreas S. Bommarius

The mapping of a protein's sequence to its function is one of the most fundamental challenges in protein engineering and still remains an unsolved problem. This problem, if solved, can facilitate a rational design process in which the structure to function mapping is used to screen the potential designs. The enormous size of the sequence space or the state space of the problem, which can be defined as all the possible sequences that can be created for an enzyme, adds to the complexity of the problem. However, it is known that only a small fraction of the amino acids present in a protein contribute significantly to the protein's properties. Finding these amino acid positions can greatly improve our knowledge about proteins and also help us to design better experiments to alter them.

Apart from individual amino acids, it is known that in the three dimensional structure of a protein, certain amino acids can interact with each other in order to provide maintain structural integrity or aid in its catalytic function [1]. If these positions are mutated the loss of this interaction usually leads to a non-functional protein. Directed Evolution (DE) experiments [2], which probe the sequence space of a protein through mutations or recombination in search for an improved variant, frequently result in such inactive sequences. In this work, we extend our approach of using machine learning algorithms to find important amino acid residues [3] to interacting pairs. Boolean Learning and Support Vector Machines (SVMs) [4] are to identify pairs of interacting amino acid residues from the primary sequence of the variants that are generated during DE. It is shown this problem can be posed for Boolean Learning by transforming the sequences into Boolean vectors. The logical function that corresponds to the specific problem of finding amino acid residues with interactions is formulated in a Conjunctive Normal Form (CNF) [5]. A novel approach of combining SVMs with other algorithms like Boolean learning is also proposed and is applied in this problem. SVMs are unique in their ability to quantify the structural risk in terms of the generalization of the pattern learnt from the data. Most of the other learning algorithms like Boolean learning are solely based on empirical risk, which is their performance on the given data, without considering generalization or the performance on data not yet seen.

An extension of this approach to multiple rounds of evolution is proposed. The results obtained from one round can be used to design parent sequences for the subsequent rounds. The purpose of this strategy is to increase the average number of variants that retain catalytic activity as well as the average number of mutations obtained. Moreover, by using the variant sequences generated over multiple rounds, the identification of interacting pairs can also be improved when compared with using all sequences generated from the same round.

Simulations show that the pairs can be identified with a reasonably accuracy, which declines with increasing number of pairs per sequence and the length of the sequence. The combined approach with SVMs, in general gives better results that Boolean learning alone. The results from the multi-round strategy show that large improvements in the activity and the mutation levels of the libraries can be obtained. The identification of the interacting pairs is also significantly improved. The combined strategy again performs better, which justifies utilizing the structural nature of SVMs for an identification problem like this, which, because of its nature is well suited to Boolean learning.

To verify the strength of this approach, sequences from the recombination of mRFP and dsRED by using both RDA-PCR [6] and DNA-shuffling [7] will be used to identify the interactions that exist between different residues in their sequence.

1. Meyer, M.M., et al., Library analysis of SCHEMA-guided protein recombination. Protein Science: a Publication Of The Protein Society, 2003. 12(8): p. 1686-1693.

2. Petrounia, I.P. and F.H. Arnold, Designed evolution of enzymatic properties. Curr.Opin.Biotechnol., 2000. 11(4): p. 325-330. doi:10.1016/S0958-1669(00)00107-5.

3. Dubey, A., et al., Support vector machines for learning to identify the critical positions of a protein. Journal of Theoretical Biology, 2005. 234(3): p. 351-361.

4. Scholkopf, B. and A.J. Smola, Learning with Kernels. 2002, Cambridge: MIT Press.

5. Triantaphyllou, E., Inference of a Minimum Size Boolean Function from Examples by Using a New Efficient Branch-and-Bound Approach. Journal of Global Optimization, 1994. 5(1): p. 69-94.

6. Ikeuchi, A., et al., Chimeric Gene Library Construction by a Simple and Highly Versatile Method Using Recombination-Dependent Exponential Amplification. Biotechnology Progress, 2003. 19(5): p. 1460-1467.

7. Stemmer, W.P., Rapid evolution of a protein in vitro by DNA shuffling. Nature, 1994. 370(6488): p. 389-391.