Therapeutic antibodies and other protein-based drugs are playing an increasingly important role in the pharmaceutical industry and in basic research. The main advantage of protein-based pharmaceuticals is that, unlike small organic molecules, proteins can be manufactured to bind to virtually any target with high specificity. Furthermore, advances in recombinant DNA technology and cell-surface display techniques have made it possible to experimentally screen billions of candidates in order to find proteins that bind to the desired target with the highest affinity [(Levin and Weiss 2006)](http://wizfolio.com/?citation=1&ver=3&ItemID=717&UserID=14332&AccessCode=10A85EC21D0D4F24B1461DADDA20A7A7&CitationSuffix=). Nevertheless, several challenges in the design of effective protein-based pharmaceuticals remain unsolved. Most naturally-occurring proteins have evolved a closely-matched energy profile of folding and unfolding, which makes them vulnerable to degradation and limits their shelf-life and stability in a pharmaceutical preparation [(Korkegian et al. 2005)](http://wizfolio.com/?citation=1&ver=3&ItemID=699&UserID=14332&AccessCode=5143F69ADCCD4D8CA3CE3274F7B4DF06&CitationSuffix=). Furthermore, while the goal generally is to find proteins that bind to a specific region on the target, high-throughput screening techniques isolate all molecules that are able to bind to the target at any location. Topological specificity can be created by using DNA libraries that code only for the proteins that can bind to the target at the desired location, but the design of those libraries requires specific knowledge about the protein-protein interaction interface and may compromise the ability of the screen to produce high-quality “hits”.

Since thermal stability and intermolecular affinity both are determined primarily by a small set of closely-packed residues in the hydrophobic core of a protein or protein-protein interface, respectively, a model that could predict correctly the stabilising or destabilising effect of different mutations at those regions would be of important practical use in addressing both problems [(Chen et al. 2012;](http://wizfolio.com/?citation=1&ver=3&ItemID=720&UserID=14332&AccessCode=DAF25C7D2D514810920616FFE03092E8&CitationSuffix=) [Park and Cochran 2009)](http://wizfolio.com/?citation=1&ver=3&ItemID=693&UserID=14332&AccessCode=93FF69E09D1E42BC9E87BC1759318A9F&CitationSuffix=). Previously-described approaches used either structure-based protein redesign or data mining techniques to predict the effect of a given mutation on the stability of the protein.

Structure-based techniques involve making successive mutations to the protein and evaluating the effect of those mutations using molecular force fields [(Shifman and Mayo 2003)](http://wizfolio.com/?citation=1&ver=3&ItemID=671&UserID=14332&AccessCode=AF66C60C4AE24D0A8D56228BBF5DE56E&CitationSuffix=). Sampling all possible combinations of mutations is not tractable, and different heuristics or machine learning algorithms have to be used to predict the regions on the protein where the mutations are most likely to produce favourable results [(Yu et al. 2012)](http://wizfolio.com/?citation=1&ver=3&ItemID=721&UserID=14332&AccessCode=0F5DD1860BEA4657B2D5E1154B29B67B&CitationSuffix=). The backbone of the protein usually is held fixed during the redesign process, and the side-chains are fixed at discrete angles called rotamers, both of which decrease the predictive power of the approach but often are necessary to reduce the complexity of the search space [(Hong et al. 2009)](http://wizfolio.com/?citation=1&ver=3&ItemID=670&UserID=14332&AccessCode=9C1E8261E05B4508923BAB079B1CAA64&CitationSuffix=). Energy functions used to evaluate the quality of the structure often are inaccurate, further complicating the interpretation of the results [(Lippow et al. 2007)](http://wizfolio.com/?citation=1&ver=3&ItemID=722&UserID=14332&AccessCode=1074064BE0FA4C3E9BBA636E0C9BB479&CitationSuffix=).

Data-mining techniques involve training machine learning algorithms on available experimental data using features designed to capture the effect of a particular amino acid on the thermal stability of the protein. For instance, Li *et al.* used 41 features to train a random forest algorithm on data sets reporting the changes in free energy caused by a variety of single and multiple point mutations, and using that model they were able to predict accurately the free energy changes associated with most mutations of Staphylococcal nuclease [(Li and Fang 2012)](http://wizfolio.com/?citation=1&ver=3&ItemID=702&UserID=14332&AccessCode=80CEB0ABA76845A09DBBD60CE64DFF9D&CitationSuffix=). Some of the mutations that were predicted less accurately involved a tryptophan residue at position 140, which is known to be critical for Staphylococcal nuclease structure, stability and function. A possible explanation for this is that a highly conserved amino acid is likely to make strong contacts with one or more other amino acids, and those contacts will affect free energy change associated with a given mutation

In a recently published set of papers it has been shown that if a global maximum entropy model is used to measure evolutionary covariance between different amino acids in a large sequence alignment, then the obtained evolutionary constraints contain enough information to predict the 3-dimentional structure of a protein to within 4 Å accuracy [(Marks et al. 2012;](http://wizfolio.com/?citation=1&ver=3&ItemID=651&UserID=14332&AccessCode=4870F096D12A4BFAB925C026D374371B&CitationSuffix=) [Hopf et al. 2012)](http://wizfolio.com/?citation=1&ver=3&ItemID=650&UserID=14332&AccessCode=AD6AE867FBE7486DBBD02CFAFDD92694&CitationSuffix=). We propose to unite the evolutionary approach with PROTS-RF in order to improve the performance of PROTS-RF on evolutionarily constrained residues, such as tryptophan 140 in Staphylococcal nuclease. This could be done by splitting the amino acids of each protein in each data set into two groups based on whether or not those amino acids shows strong evidence of evolutionary constraint. Residues that do not show any evidence of evolutionary constraint will be used to train a random forest classification algorithm using the same set of features that were used by Lee *et al.* [(Li and Fang 2012)](http://wizfolio.com/?citation=1&ver=3&ItemID=702&UserID=14332&AccessCode=80CEB0ABA76845A09DBBD60CE64DFF9D&CitationSuffix=). However the residues that do show evidence of evolutionary constraint will be classified as pairs using a separate set of features designed to capture the energy profile of the pairwise interaction. We would then use the two models to predict the net free energy change associated with each possible mutation in our scaffold protein, with the first model used to predict the effect of mutations for non-constrained amino acids and the second model used to predict the effect of mutations for evolutionarily constrained pairs of amino acids. Mutations that are predicted to be highly stabilising would be selected for experimental validation. An analogous approach would be applied to the refinement of protein-protein interfaces.

[**References**](https://wizfolio.com/?style=1&ver=3&UserID=14332&StyleName=Medline-AuthorYear%20(Alexey%20mod%20Final))

Chen CT, Peng HP, Jian JW, Tsai KC, Chang JY, Yang EW, Chen JB, Ho SY, Hsu WL, Yang AS. Protein-protein interaction site predictions with three-dimensional probability distributions of interacting atoms on protein surfaces. *PloS one* 2012;7(6):e37706.

Hong EJ, Lippow SM, Tidor B, Lozano-Pérez T. Rotamer optimization for protein design through MAP estimation and problem-size reduction. *J Comput Chem* 2009;30(12):1923-1945.

Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 2012;149(7):1607-1621.

Korkegian A, Black ME, Baker D, Stoddard BL. Computational thermostabilization of an enzyme. *Science* 2005;308(5723):857-860.

Levin AM, Weiss GA. Optimizing the affinity and specificity of proteins with molecular display. *Mol Biosyst* 2006;2(1):49-57.

Li Y, Fang J. PROTS-RF: a robust model for predicting mutation-induced protein stability changes. *PloS one* 2012;7(10):e47247.

Lippow SM, Wittrup KD, Tidor B. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat Biotechnol* 2007;25(10):1171-1176.

Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. *Nat Biotechnol* 2012;30(11):1072-1080.

Park SJ, Cochran JR. Protein Engineering and Design. CRC Press; 2009. 416 p.

Shifman JM, Mayo SL. Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc Natl Acad Sci U S A* 2003;100(23):13274-13279.

Yu CM, Peng HP, Chen IC, Lee YC, Chen JB, Tsai KC, Chen CT, Chang JY, Yang EW, Hsu PC, Jian JW, Hsu HJ, Chang HJ, Hsu WL, Huang KF, Ma AC, Yang AS. Rationalization and design of the complementarity determining region sequences in an antibody-antigen recognition interface. *PloS one* 2012;7(3):e33340.