Chapter 54
Improving Prediction Accuracy via Subspace Modeling in a Statistical Geometry Based Computational Protein Mutagenesis

Majid Masso
George Mason University, USA

ABSTRACT

A computational mutagenesis is detailed whereby each single residue substitution in a protein chain of primary sequence length N is represented as a sparse N-dimensional feature vector, whose M << N non-zero components locally quantify environmental perturbations occurring at the mutated position and its neighbors in the protein structure. The methodology makes use of both the Delaunay tessellation algorithm for representing protein structures, as well as a four-body, knowledge based, statistical contact potential. Feature vectors for each subset of mutants due to all possible residue substitutions at a particular position cohabit the same M-dimensional subspace, where the value of M and the identities of the M nonzero components are similarly position dependent. The approach is used to characterize a large experimental dataset of single residue substitutions in bacteriophage T4 lysozyme, each categorized as either unaffected or affected based on the measured level of mutant activity relative to that of the native protein. Performance of a single classifier trained with the collective set of mutants in N-space is compared to that of an ensemble of position-specific classifiers trained using disjoint mutant subsets residing in significantly smaller subspaces. Results suggest that significant improvements can be achieved through subspace modeling.

DOI: 10.4018/978-1-4666-3604-0.ch054

INTRODUCTION

Protein engineering experiments involving the synthesis and analysis of new proteins, each differing from the wild type via a single amino acid replacement introduced into the native sequence, yield important insights into protein folding and activity by characterizing the structural and/or functional roles of constituent residue positions (Lehmann, Pasamontes, Lassen, & Wyss, 2000; Vieille & Zeikus, 2001; Yang, Wang, & Fitzgerald, 2004). However, the prohibitively time consuming and expensive nature of conducting comprehensive single residue mutagenesis studies leads researchers to prioritize these experiments based on information gleaned from a variety of sources, including predictions obtained from *in silico* models. Such models may utilize an evolutionary scoring function (Kumar, Henikoff, & Ng, 2009), apply physical (Kollman et al., 2000; Pitera & Kollman, 2000), statistical (Kwasigroch, Gilis, Dehouck, & Rooman, 2002; Parthiban, Gromiha, Hoppe, & Schomburg, 2007; Zhou & Zhou, 2002), or empirical (Bordner & Abagyan, 2004; Guerois, Nielsen, & Serrano, 2002) potentials, or implement machine learning tools (Capriotti, Fariselli, & Casadio, 2004; Cheng, Randall, & Baldi, 2006; Huang, Gromiha, & Ho, 2007; Verzilli, Whittaker, Stallard, & Chasman, 2005). In each case the model is designed to predict a specific outcome of single residue substitutions in a protein, for example any evidence of an effect on protein activity, the relative change in protein stability, or more broadly, any pathological consequence to the organism.

With machine learning models, the single residue protein mutants populating training sets are each typically represented as a feature vector consisting of attributes that characterize the sequence or structure, frequently combined with components that reflect evolutionary information. Recently, studies have begun to focus on combining potential functions with machine learning methods by using several of the energy terms associated with the mutant proteins as attributes in feature vectors (Dehouck et al., 2009; Lise, Archambeau, Pontil, & Jones, 2009; Masso & Vaisman, 2007, 2008). One such approach involves a computational mutagenesis procedure that utilizes a four-body, knowledge based, statistical contact potential and yields, for any mutation due to a single amino acid replacement in a protein structure, an N-dimensional vector of ensuing environmental perturbations occurring at each of the N constituent residue positions (Masso & Vaisman, 2007). Since the average sized polypeptide chain in a protein structure consists of $N \sim 200$ amino acids, protein mutants are represented by vectors in high-dimensional Euclidean space. For each mutant, the methodology captures only local effects at the $M << N$ residue positions that are structurally closest to the mutated residue (including the mutated position itself) identified via Delaunay tessellation of the protein structure, a classical computational geometry technique, where the value of M and the identities of the M nonzero vector components vary according to the residue position being mutated (Masso & Vaisman, 2007). Hence the vectors are sparse, with those representing the collective set of 19 mutants, obtained by introducing all possible alternative amino acid substitutions at one particular position in the protein, residing in the same particular M-dimensional subspace.

All possible single residue replacements in a protein of size N theoretically lead to $19N$ mutants; however, experimental comprehensive mutagenesis studies generally involve the synthesis and analysis of far fewer mutants, which subsequently can be classified based on observed changes relative to the native protein (e.g., the activity of each mutant is either "unaffected" or "affected" by the respective amino acid substitution). Such an experimental dataset for a protein, with each mutant represented as an N-dimensional feature vector of inputs along with a corresponding categorical output attribute identifying the experimental mutational consequence, can be used to train and evaluate the performance of models obtained via