Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Journal of Theoretical Biology 234 (2005) 187–199

Journal of Theoretical Biology

www.elsevier.com/locate/yjtbi

Directed molecular evolution by machine learning and the influence of nonlinear interactions

Richard Fox*

Codexis, Inc., 200 Penobscot Drive, Redwood City, CA 94063, USA

Received 30 August 2004; received in revised form 3 November 2004; accepted 22 November 2004
Available online 20 January 2005

Abstract

Alternative search strategies for the directed evolution of proteins are presented and compared with each other. In particular, two different machine learning strategies based on partial least-squares regression are developed: the first contains only linear terms that represent a given residue's independent contribution to fitness, the second contains additional nonlinear terms to account for potential epistatic coupling between residues. The nonlinear modeling strategy is further divided into two types, one that contains all possible nonlinear terms and another that makes use of a genetic algorithm to select a subset of important interaction terms. The performance of each modeling type as a function of training set size is analysed. Simulated molecular evolution on a synthetic protein landscape shows the use of machine learning techniques to guide library design can be a powerful addition to library generation methods such as DNA shuffling.

Keywords: Directed evolution; Genetic algorithm; DNA shuffling; NK landscape; Machine learning

1. Introduction

Technologies for protein engineering span a wide spectrum: from rational design based on molecular mechanics/dynamics at one end to random mutagenesis by error prone PCR at the other (van Regenmortel, 2000; Chen, 2001). Throughout this spectrum, recursive methods have been used to improve proteins through successive rounds of evolution. In particular, methods based on recombining beneficial diversity from improved variants have enjoyed great success. Such techniques are often able to obtain significantly improved proteins using fewer assays than required by ultra high throughput screening methods (Stemmer, 1994; Ness et al., 1999, 2002). Such recursive, recombination-based techniques are increasingly applied to protein engineering problems and are the focus of continued study (Kurtzman et al., 2001).

Given the widely observed success of these techniques, the next step in the evolution of protein engineering may be had in finding better ways to accelerate the recombination of beneficial diversity. For this to work, the ability to tease out important elements of diversity that contribute to improved function is key. Machine learning techniques used to analyse multivariate data sets are ideally suited to this task and have been used extensively in many engineering fields. Small molecule quantitative structure activity relationships (QSARs) have been used for many years in medicinal chemistry to improve the search for biologically active compounds (Kubinyi, 1997a, b; Byvatov et al., 2003; Byvatov and Schneider, 2004). Likewise peptide engineering has also made use of machine learning techniques to analyse and create improved molecules (Mee et al., 1997; Cho et al., 1998; Bucht et al., 1999; Lee et al., 2000).

Machine learning techniques are becoming more popular in biological engineering. Several researchers have looked at the fitness landscapes of a number of proteins and protein–protein interactions, building

*Tel.: +1 650 980 5616.
E-mail address: richard.fox@codexis.com (R. Fox).

0022-5193/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2004.11.031

statistical models for predicting function given the sequence alone (Lu et al., 2001; Aita et al., 2002; De Genst et al., 2002). To date, less work has been done using statistical models to optimize proteins (as opposed to the much smaller peptides). One example is a peptide-QSAR model that was used to optimize the glycosyl phosphatidylinositol (GPI) modification for a protein with a C-terminal signal peptide (Bucht et al., 1999).

The work presented here is similar to that found in earlier studies that promote the use of machine learning techniques for sequence-oriented evolution (Schneider et al., 1994, 1995a, b). Although these earlier studies were focused on peptides and protein precursor cleavage sites, they are conceptually similar to the current work in as much as proteins can be viewed mathematically and phenomenologically as large peptides. Though one study (Schneider et al., 1994) was criticized for possibly not using enough data to train a machine learning algorithm (Darius and Rojas, 1994), the criticism was later found to be unwarranted as highly active, novel peptides were found using the trained neural nets (Schneider et al., 1998; Wrede et al., 1998).

Two important advancements have taken place over the last 10 years to facilitate the use of machine learning techniques for protein engineering: (1) The ability to generate focused combinatorial protein libraries, (2) The availability of cheap, fast, and accurate DNA sequencing. Together these advancements make the use of machine learning techniques for protein engineering feasible today. For many practical applications of interest, only several dozens or hundreds of variants need to be analysed in order to generate meaningful statistical models that can be used to engineer improved variants, a condition well within today's capabilities.

The viability of using machine learning techniques for the directed evolution of proteins was evaluated earlier (Fox et al., 2003). The machine learning based algorithm begins with the creation of a combinatorial protein library, followed by physical assays that generate sequence/activity data. This information is then used to build statistical models that can then be interrogated to design new libraries. The cycle can be repeated, often adding new diversity at each cycle, by some combination of rational design or random mutagenesis. One perceived limitation of the earlier work is that it assumed a given residue's contribution to fitness was independent of context. Such a linear model may not be capable of identifying important nonlinear interactions between residues. The purpose of this work is to help ascertain the degree to which nonlinear interactions (and the use of models that attempt to capture such interactions) affect the efficiency and robustness of statistical, evolutionary search algorithms that are used to engineer improved proteins.

2. Methods

2.1. Problem coding

Protein variants can be created by any number of techniques that are currently available for use in the construction of combinatorial protein libraries. Such techniques include both classical and synthetic DNA shuffling (Stemmer, 1994; Ness et al., 1999, 2002). In both modes of shuffling the process begins with a fixed set of diversity found in either the homologs (for classical or family shuffling) or from rationally targeted diversity (for synthetic shuffling). This starting diversity is typically far less than the theoretically accessible space—even a small protein of 100 residues has 20^{100} possible sequences. Thus only a subset of the potential sequence space can be examined at any one round of evolution. In practice about 5–30% of a 300-residue chain may undergo some variation in a shuffled library. New diversity can be added in subsequent rounds of evolution (through random or site directed mutagenesis) in order to supply the evolutionary fuel required for further gains in fitness. In the present work, a fixed number of positions are assumed to undergo some degree of variation during one round of evolution. The task here is to examine how efficient the algorithm is for such a fixed search space. The inclusion of additional diversity in subsequent rounds of evolution is a separate, somewhat orthogonal consideration that should not detract from the general applicability and usefulness of the proposed method for optimizing proteins.

Let us assume that we have sequence and activity data for S protein variants. In practice, large segments of the sequence alignment may not contain any diversity between the variants and are excluded from further analysis. This does not mean that the variable positions in the alignment are not interacting physically with the fixed regions of the protein. In fact there is likely to be strong interactions with fixed parts of the protein, but since those parts are not varying they do not play a role in building a statistical model of the *local* fitness landscape. The local fitness landscape consists only of those residues that are undergoing variation. Our goal in building a statistical model is to generate an approximation to the local fitness landscape that can be used to extrapolate to nearby regions of sequence space. A *global* fitness landscape, though ideal to have, is well beyond the scope, capability and often the practical need of the engineering process.

We create dummy variables for the N variable positions within the sequence alignment. The example in Fig. 1a shows the first 27 positions from an alignment of protein variants. For brevity we only show the N-terminal portion of the protein alignment but in practice any number of positions along the length of the protein may undergo variation during a round of