Journal of Cellular Biochemistry Supplement 37:58–63 (2001)

Computationally Focusing the Directed Evolution of Proteins

Christopher A. Voigt,[1]* Stephen L. Mayo,[2] Frances H. Arnold,[3] and Zhen-Gang Wang[3]

[1]Biochemistry and Molecular Biophysics, California Institute of Technology, Pasadena, California
[2]Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, Pasadena, California
[3]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California

Abstract Directed evolution has proven to be a successful strategy for the modification of enzyme properties. To date, the preferred experimental procedure has been to apply mutations or crossovers randomly throughout the gene. With the emergence of powerful computational methods, it has become possible to develop focused combinatorial searches, guided by computer algorithms. Here, we describe several computational methods that have emerged to aid the optimization of mutant libraries, the targeting of specific residues for mutagenesis, and the design of recombination experiments. J. Cell. Biochem. Suppl. 37: 58–63, 2001. © 2002 Wiley-Liss, Inc.

Key words: directed evolution; recombination; mutagenesis; computational methods

Directed evolution has emerged as a powerful technique in protein engineering [Petrounia and Arnold, 2000; Arnold, 2001]. Diversity is created through mutagenesis or recombination and the resulting library is screened for improvements in interesting properties. Traditionally, the diversity generated by directed evolution has been applied randomly throughout the gene. On one hand, this allows for the discovery of beneficial mutations that may not have been predicted a priori to have a positive effect. However, the combinatorial explosion of possibilities, and some experimental biases (e.g., the genetic code), cause many opportunities to be missed.

Concurrent with the rise of directed evolution, computational techniques have improved dramatically [Dahiyat and Mayo, 1997; Street and Mayo, 1999; Voigt et al., 2001b]. In particular, inverse design algorithms, where an amino acid sequence is designed to fold into a predetermined structure, have shown tremendous success. These algorithms allow for the prediction of the effects of mutations on the stability of the protein. Further, a better understanding has emerged on the biases of PCR and the annealing of DNA fragments [Sun, 1999; Wang et al., 2000]. Together, these techniques will allow the optimization of directed evolution by focusing the diversity towards specific regions of the gene. In turn, this decreases the need for costly and time-consuming screening.

Here, we will describe several methods that have been proposed to merge computational algorithms with combinatorial experiments. First, a model is explored that describes the effect of mutations on the quality of the mutant library. These simulations have improved our understanding of the relationship between parameters describing the search space (e.g., interactions between amino acids) and experimental parameters such as the mutation rate and library size. Next, an algorithm is described that can calculate the tolerance of each residue to amino acid substitutions, based on the three-dimensional structure. Finally, a computational model to optimize recombination experiments is explored.

Optimizing the Mutant Library

A key constraint in the evolutionary search is the limited screening capacity. Typically, screening is restricted to 10^3–10^6 mutants [Daugherty et al., 2000; Petrounia and Arnold,

*Correspondence to: Christopher A. Voigt, Biochemistry and Molecular Biophysics, California Institute of Technology, Mail code 210-41, Pasadena, CA, 91125.

Received 18 September 2001; Accepted 19 September 2001

DOI 10.1002/jcb.10066

2000]. Even the state-of-the-art high throughput selection techniques, such as RNA–protein fusion, can handle on the order of 10^{12} mutants [Roberts and Szostak, 1997]. Despite these impressive experimental advances, the sampling ability remains tiny when compared with the vastness of sequence space. To reduce the project time and cost of an experiment, it is desirable to optimize the search parameters, such that the maximum fitness improvements can be found with the minimum screening effort. Towards this goal, this section is devoted to a model describing the properties of small libraries of mutants, generated by error-prone PCR.

For a given screening capacity, there is an optimal mutation rate, defined as the rate that produces the largest fitness improvement for a given library size. This is a consequence of two opposing effects. On the one hand, a large enough mutation rate is required to generate adequate diversity in the mutants. On the other hand, because the probability of an individual mutation demonstrating improvement is small, multiple mutations on the same sequence (the result of large mutation rate) are generally deleterious. Thus, in a limited screening pool, the probability of observing improvement decreases rapidly as the number of mutations increases.

Using a statistical mechanical model, we have investigated the effects of the finite screening size on libraries generated by mutagenesis [Voigt et al., 2001a,b]. Simulations using this model demonstrate the relationship between the number of mutants that can be screened and the optimal mutation rate. The optimal mutation rate is typically low (about one amino acid substitution per sequence) because the probability of an individual mutation demonstrating improvement is small (Fig. 1A). When multiple mutations are accumulated, it is likely that most are deleterious and these mutations quickly erode the improvement from the few beneficial mutations that may exist. This effect worsens as the number of mutants that can be screened decreases. In fact, as the mutation rate increases, the number of possible combinations increases exponentially. Therefore, to adequately sample higher mutation rates, exponentially larger libraries are required. Similarly, as the fitness of the parent sequence increases, the probability of improvement decreases, thus exaggerating the effect of deleterious mutations

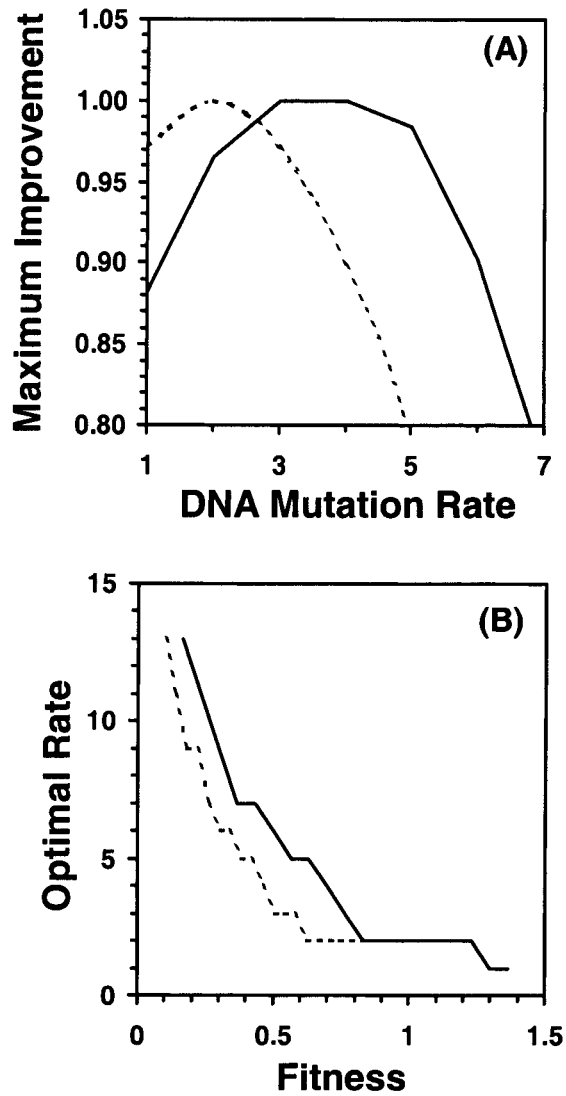

Fig. 1. The optimal DNA mutation rate as determined from a statistical model, similar to a spin-glass, that captures the effect of interactions between amino acids [Voigt et al., 2001]. The genetic code is included in the model. The data is for a $N = 50$ protein and 1,000 mutants are screened. The fitness improvement is the average maximum change in fitness as averaged over 10,000 landscapes. **A**: The optimal mutation rate as the total number of interactions between residues (the ''landscape ruggedness'') increases. The number of coupling interactions is 75 (dotted line) and 0 (solid line). As the landscape ruggedness increases, the optimal mutation rate decreases. To compare the relative location of the optima, the curves have been scaled so that the optima are at 1.0. **B**: The optimal mutation rate is shown as a function of the parental fitness for a smooth (solid line) and rugged (dotted line) landscape. As the parental fitness increases, the probability that a mutation is deleterious also increases, making a smaller mutation rate optimal. This effect is more rapid when there are more interactions between amino acids.

(Fig. 1B). Thus, as the generations of mutation and selection progress, an exponentially increasing screening size is required.

Further, the probability of improvement is affected by the ruggedness of the fitness landscape. The property of ruggedness is caused by interactions between amino acids, where two residues are considered interacting, or coupled, when the sum of individual effects from mutations at each residue is not equal to the combined effect of both mutations together [Kauffman and Levin, 1987; Matsuura et al., 1998; Juncovic and Poteete, 1999]. As the number of interactions increases, the probability that a mutation is deleterious also increases. When multiple mutations are accumulated on a gene, a larger fraction of these mutations will decrease the fitness. This effect quickly erodes the beneficial effect of any positive mutations. Therefore, to search rugged landscapes, a smaller mutation rate is optimal.

Within a protein, some residues are involved in more interactions than others. Single amino acid substitutions at a coupled residue are often not adequate to show improvement. Rather, it is necessary to simultaneously optimize all of the interacting residues, requiring a large mutation rate. However, the mutation rate is limited by the number of mutants that can be screened. Under this limitation, the probability that a single amino acid substitution will improve the fitness is lower for coupled residues. Simulations of the directed evolution algorithm on simple fitness landscapes have shown that the probability of discovering a beneficial mutation at a highly coupled residue decreases significantly as the sequence increases in fitness.

Targeted Mutagenesis Algorithms

Using error-prone PCR to mutate the entire gene has several disadvantages. First, the attainable amino acid substitutions are severely restricted by the genetic code because the probability of obtaining adjacent DNA mutations is small. This reduces the number of possible amino acid mutations at each residue from 19 to about 5.7. Further, the diversity is reduced by biases for transitions over transversions, and $A \leftrightarrow T$ mutations over $G \leftrightarrow C$ mutations. These limitations can be overcome by saturating a selected set of residues with all twenty amino acids. Towards this end, several non-computational strategies have been proposed to focus the mutations towards specific regions of the gene, including the selective saturation of residues near to the active site [Shinkai et al., 2001] or residues where improvements have been previously found by directed evolution [Miyazaki and Arnold, 1999]. Here, we describe a computational method, based on the simulations described in the above section, that determines which residues can be mutated without disturbing the stability of the protein.

Our statistical model has demonstrated that the largest fitness improvements are made at uncoupled residues when the fitness of the parent sequence is high and the number of mutants that can be screened is limited. We have extended these findings to real proteins, where data are available from directed evolution experiments [Voigt et al., 2001b]. To determine the coupling interactions between residues, the energetic interactions between amino acids are calculated using the ORBIT protein design software [Dahiyat and Mayo, 1997]. The effect of all amino acid substitutions was measured and the information condensed into a site entropy by the following equation:

$$s_i = \sum_{j=1}^{20} p_i(a) \ln p_i(a)$$

where $p_i(a)$ is the probability of amino acid a existing at residue i. A high site entropy indicates that many amino acids may be substituted at that residue. To circumvent the combinatorial difficulties in obtaining the probabilities required by the above equation, we applied mean-field theory, a technique borrowed from statistical mechanics [Saven and Wolynes, 1997]. A representative entropy profile for substilisin E is shown in Figure 2. Beneficial mutations accumulated in experimental directed evolution were found to be strongly correlated with the high entropy residues identified from our calculation [Chen and Arnold, 1993; You and Arnold, 1996; Zhao and Arnold, 1999]. Seven out of the nine mutations that improved the thermostability of subtilisin E occur at positions computed to be highly tolerant. Ten out of the thirteen mutations that improved the activity also occur at the calculated tolerant residues. Similarly strong correlations were found for T4 lysozyme [Pjura et al., 1993] and Antibody 4-4-20 [Boder et al., 2000].

Taken together, our results imply a structural and functional overlap in sequence space:

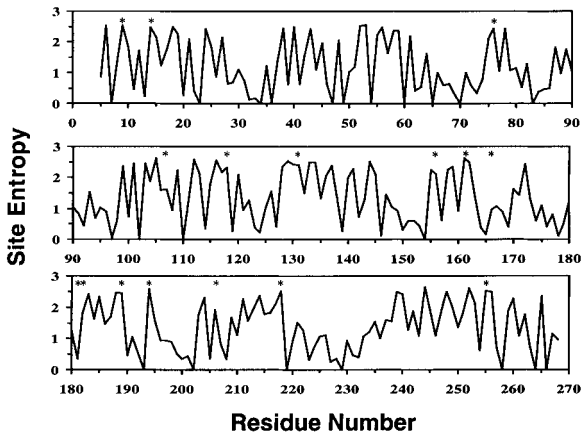

Fig. 2. The entropy profile for subtilisin E. Stars mark the positions where experimental directed evolution found mutations that improved thermostability or activity in organic solvent [Chen and Arnold, 1993; You and Arnold, 1996; Zhao and Arnold, 1999]. In these studies, the mutations were found by screening 2,000–5,000 mutants generated with an average mutation rate of 2–3 nucleotide substitutions per gene.

functional improvements tend to occur when the structure is least perturbed. This effect increases with the fitness of the parents, as it becomes more important to preserve the stability (Fig. 2). From the perspective of experimental design, targeting residues that are structurally tolerant will create a mutant library that is richer in folded, stable proteins. By focusing mutagenesis towards residues that preserve the stability, functional space can be more thoroughly explored [Huynen et al., 1996].

Recombination Strategies

Recombination is a powerful tool for in vitro protein evolution [Stemmer, 1994; Patten et al., 1997; Crameri et al., 1998; Ness et al., 1999; Ostermeier et al., 1999]. By exchanging genetic information from several parental genes, a library of recombinant mutants is generated. Screening or selection can identify those hybrid genes coding for proteins that are stable, functional, or have improved properties. Currently, crossovers are introduced at more or less random positions. There is little or no a priori input where the crossovers should occur, and many of the hybrid proteins are either unfolded or not functional. It is desirable to predict where crossovers will be accepted in functional protein hybrids in order to make focused shuffled libraries and minimize labor-intensive screening. In this section, we describe a computational algorithm to predict the location of crossovers, based on the assumption that the stability of the offspring need to be retained.

The dynamics of recombination have been explored extensively in computer science to study the effectiveness of genetic algorithms. Recombination is a powerful search strategy because it can combine beneficial clusters of interacting amino acids (schema or building blocks) that have previously survived selection onto a single offspring gene [Holland, 1998]. However, for recombination to be successful, the crossover locations in the gene have to correspond to those that allow the most schema to be retained. When a schemata is divided by recombination such that fractions of it are inherited from different parents, this is referred to as schema disruption [Vose and Liepens, 1991; Forrest and Mitchell, 1993]. When schema disruption is not controlled, genetic algorithms will often fail to converge on an optimal solution.

Applying this simple idea to protein structure, we postulate that recombination is most successful when the crossovers break the fewest stabilizing interactions between amino acids. Two residues are considered interacting if their side chains are within a cut-off distance. The schema disruption of a recombinant mutant counts the total number of interactions that are broken by the specific pattern of fragments inherited from each parent. The recombinants with the minimum schema disruption are most likely to retain the structure of the parents.

We have examined schema disruption in several shuffling experiments. To simulate the recombination of PurN and GART glycinamide ribonucleotide transformylase [Ostermeier et al., 1999; Lutz et al., 2001], the interacting residues was calculated from the structure of PurN. All single-crossover recombinant mutants were generated and the schema disruption of each was calculated. The crossovers found experimentally by screening the library for functional hybrid enzymes were strongly biased towards the two computed local minima (Fig. 3). We have found similar results when we compare the calculated schema disruption with recombination experiments with beta-lactamase, subtilisin, P450, and phytase.

CONCLUSIONS

Several new computational techniques to aid in the design of directed evolution experiments

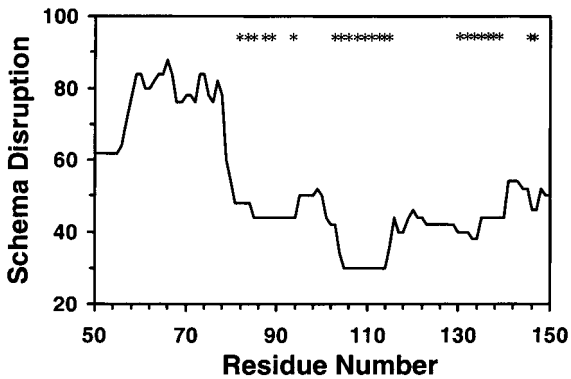

Fig. 3. To simulate the recombination of PurN and GART glycinamide ribonucleotide transformylase [Ostermeier et al., 1999; Lutz et al., 2001], the coupling matrix was calculated from the structure of PurN. All single crossover recombinant mutants were generated and the crossover disruption of each was calculated. The crossovers found experimentally by screening the library for functional hybrid enzymes (indicated by the stars) were strongly biased towards the computed local minima.

have been introduced. We employ a statistical model to study the dynamics of directed evolution as a search algorithm. Using this simplified model, we have explored the relationship between the optimal mutation rate, library size, fitness of the parents, and the interactions between amino acids. Further, a bias was discovered that mutations preferentially occur at uncoupled residues, when the mutation rate and number of mutants screened is small. This inspired the development of a more detailed structural model that measures the effect on stability of amino acid substitutions at each residue. Using this model, we find that mutations that were found to improve activity often occur at positions where the stability is least likely to be disturbed. Finally, a model is proposed that describes the optimal crossover locations as those that preserve structural schema. Together, these computational techniques represent a major step towards information-driven combinatorial protein design.

REFERENCES

Arnold FH. 2001. Combinatorial and computational challenges for biocatalyst design. Nature 409:253–257.

Boder ET, Midelfort KS, Wittrup KD. 2000. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. Proc Natl Acad Sci USA 97:10701–10705.

Chen K, Arnold FH. 1993. Tuning the activity of an enzyme for unusual environments: Sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. Proc Natl Acad Sci USA 90:5618–5622.

Crameri A, Raillard S-A, Bermudez E, Stemmer WPC. 1998. DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature 391:288–289.

Dahiyat BI, Mayo SL. 1997. De novo protein design: Fully automated sequence selection. Science 278:82–87.

Daugherty PS, Chen G, Iverson BI, Georgiou G. 2000. Quantitative analysis of the effect of the maturation of single chain Fv antibodies. Proc Natl Acad Sci USA 97: 2029–2034.

Forrest S, Mitchell M. 1993. Relative building-block fitness and the building-block hypothesis. In: Whitley LD, editor. Foundations of genetic algorithms 2. San Mateo, CA: Morgan Kauffman. p 109–126.

Holland J. 1998. Adaptation in natural and artificial systems. Boston, MA: The MIT Press.

Huynen MA, Stadler PF, Fontana W. 1996. Smoothness within ruggedness: The role of neutrality in adaptation. Proc Natl Acad Sci USA 93:397–401.

Juncovic M, Poteete AR. 1999. Protein salvage by directed evolution. Ann N Y Acad Sci 870:404–407.

Kauffman S, Levin S. 1987. Towards a general theory of adaptive walks on rugged landscapes. J Theor Biol 128: 11–45.

Lutz S, Ostermeier M, Benkovic SJ. 2001. Rapid generation of incremental truncation libraries for protein engineering using alpha-phosphothioate nucleotides. Nucleic Acids Res 29:e16.

Matsuura T, Yomo T, Trakulnaleamsai S, Ohashi Y, Yamamoto K, Urabe I. 1998. Nonadditivity of mutational effects and its application to directed evolution. Protein Eng 11:789–795.

Miyazaki K, Arnold FH. 1999. Exploring nonnatural evolutionary pathways by saturation mutagenesis: Rapid improvement of protein function. J Mol Evol 49:716–720.

Ness JE, Welch M, Giver L, Bueno M, Cherry JR, Borchert TV, Stemmer WPC, Minshull J. 1999. DNA shuffling of subgenomic sequences of subtilisin. Nat Biotech 17:893–896.

Ostermeier M, Shim JH, Benkovic SJ. 1999. A combinatorial approach to hybrid enzymes independent of DNA homology. Nat Biotech 17:1205–1209.

Patten PA, Howard RJ, Stemmer WPC. 1997. Application of DNA shuffling to pharmaceuticals and vaccines. Curr Opin Biotech 8:724–733.

Petrounia IP, Arnold FH. 2000. Designed evolution of enzymatic properties. Curr Opin Biotech 11:325–330.

Pjura P, Matsumura M, Baase WA, Matthews BW. 1993. Development of an in vivo method to identify mutants of phage T4 lysozyme of enhanced thermostability. Protein Sci 2:2217–2225.

Roberts RW, Szostak JW. 1997. RNA-peptide fusions for the in vitro selection of peptides and proteins. Proc Natl Acad Sci USA 94:12297–12302.

Saven JG, Wolynes PG. 1997. Statistical mechanics of the combintorial synthesis and analysis of folding macromolecules. J Phys Chem B 101:8375–8389.

Shinkai A, Patel PH, Loeb LW. 2001. The conserved active site motif a of *Escherichia coli* DNA polymerase I is highly mutable. J Biol Chem 276:18836–18842.

Stemmer WPC. 1994. Rapid evolution of a protein in vitro by DNA shuffling. Nature 370:389–391.

Street AG, Mayo SL. 1999. Computational protein design. Structure 7:R105–R109.

Sun F. 1999. Modeling DNA shuffling. J Comp Biol 6:77–90.

Voigt CA, Kauffman S, Wang Z-G. 2001a. Rational evolutionary design: The theory of in vitro protein evolution. Adv Protein Chem 55:79–160.

Voigt CA, Mayo SL, Arnold FH, Wang Z-G. 2001b. Computational method to reduce the search space in directed protein evolution. Proc Natl Acad Sci USA 98:3778–3783.

Vose MD, Liepens GE. 1991. Schema disruption. In: Belew RK, Booker LB. editors. Proceedings of the Fourth International Conference on Genetic Algorithms. San Mateo, CA: Morgan Kauffman. p 237–242.

Wang D, Zhao C, Cheng R, Sun F. 2000. Estimation of the mutation rate during e error-prone polymerase chain reaction. J Comp Biol 7:143–158.

You L, Arnold FH. 1996. Directed evolution of subtisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylformamide. Protein Eng 9:77–83.

Zhao H, Arnold FH. 1999. Directed evolution converts subtilisin E into a functional equivalent of thermitase. Protein Eng 12:47–53.