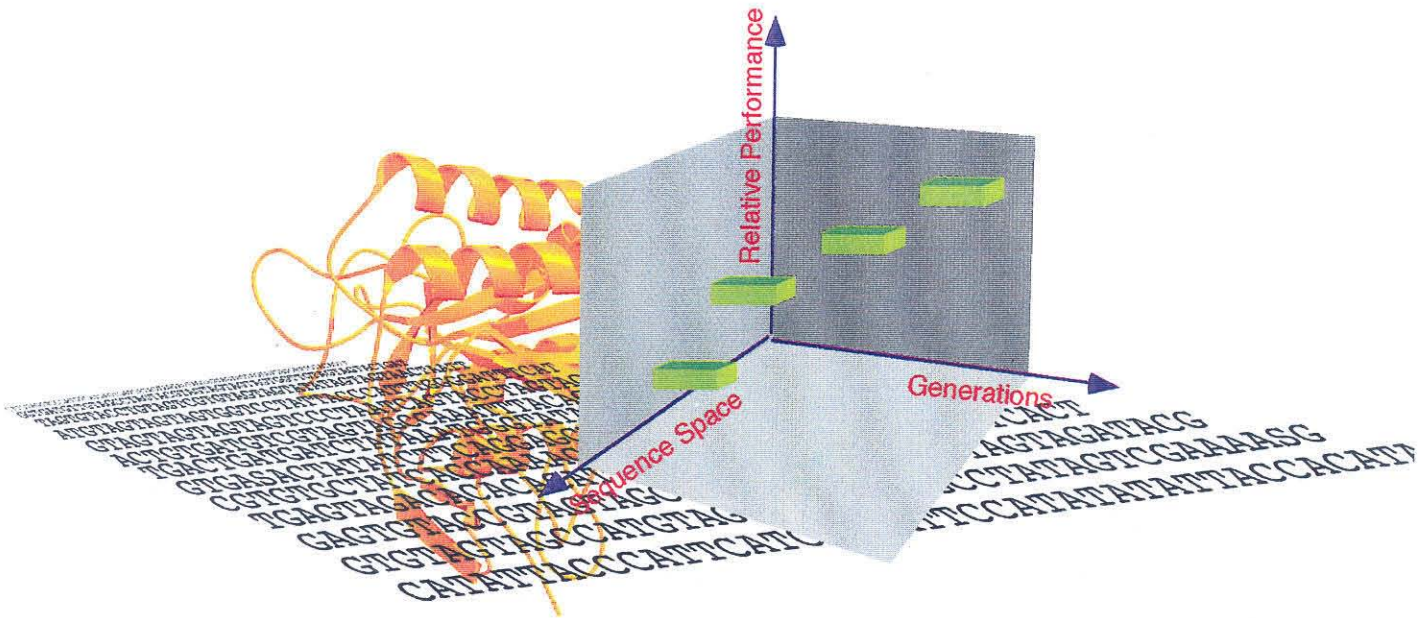
Enzyme Design By Directed Evolution

Thesis by
Huimin Zhao



In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1998
(submitted March 2, 1998)

Acknowledgments

An old Chinese proverb states: "Don't forget the diggers of the well after you drink the water." I am fully aware that I am deeply indebted to many people in my life. Without them I would not have become what I am now. Thus, upon finishing my thesis, I would like to take this opportunity to thank all these people.

First and foremost, I must thank my advisor, Dr. Frances Arnold, for her enthusiastic support during my graduate work at Caltech. Her insight and encouragement kept my research heading in directions. She has given me an extraordinary amount of freedom to pursue the research topic which interested me most. In my culture, a teacher once is a father forever. I will always keep this kind of feeling deep in my heart.

I thank my thesis committee members: Drs. Harry Gray, John Richards, Rich Roberts and Douglas Rees for helpful comments and suggestions. I would also like to thank Drs. Zhengang Wang and Harvey Blanch for their advice on my career decision.

Life is full of surprises. I feel very lucky to know Dr. Joseph Affholter from Dow chemical. Although he stayed at Caltech for less than two months, he made significant contributions to the progress of my research project. I feel privileged to become his friend and exchange our thoughts about life, work, science and love. Most importantly, he helped me get my "dream job."

Past and present Arnold group members have given me tremendous support in numerous ways. Among them, Dr. Li You, Dr. Rob Johnson, Dr. Jeff Moore, Dr. Huamin Jin, Dr. Jihu Zhang, Dr. Zhixin Shao, Dr. Lori Giver, Dr. Sean Plunkett, Dr. Tsao-chen Chen, Dr. Zhibin Guan, Dr. Zhanglin Lin, Dr. Per-Ola Freskgard and Dr. Kentaro Miyazagi deserve special thanks. Research discussions with them have always been stimulating and helpful. I am also grateful to Dr. Guohua Chen and Dr. Weigong Zheng for being my friends and Ping-Pong buddies. The time I spent with them playing Ping-Pong helped keep me sane, and made my research and my thesis writing less suffering. In addition, I sincerely thank Noah Malmstadt, Yongkai Ow, and Jessie Kim, three brilliant Caltech undergraduates, for their assistance and fondly wish them success in the future. Finally, special appreciations to Donna Johnson for keeping the Arnold group functioning smoothly and keeping track of the whereabouts of Frances.

Six years ago, as a college student in China, it was one of my wildest dreams to one day obtain a Ph.D. from one of the finest institutes in the world. I owe my thanks to Dean Yunyu Shi of University of Science and Technology of Science of China for taking me into her lab. She helped me develop independence and an appreciation for how seemingly unrelated disciplines such as biology, chemistry, statistical mechanics and

physics can be integrated and used to solve biological problems. I will never forget the days when she helped me to publish my first paper in an international journal. To my great happiness, the Arnold group emphasizes the similar integration of scientific disciplines and carry it into a new high level.

At Caltech, my student life would be less colorful and memorable without my good friends from Caltech C (Chinese Association). Special thanks to Wenge Zhong, Dr. Li Pu for helping me solve some chemical synthesis problems that I had in my early experiments; Xiangdong Fang for various help offered; Guangyang Wang for once being my enjoyable roommate; Dr. Yibin Cao, Dr. Shuyan Qi, Qing Yang, Dr. Yu Cao, Dr. Wei Lin, Jianzhong Li, Cao Ku, Dr. Xiaotian Zhu, Dr. Guanghua Gao and Dr. Danny Koh for trustful personal friends. Although we have taken different career paths, my friends, keep in mind "All roads lead to Rome." We will meet each other somewhere sometime in the future.

Finally, my deepest thanks to my parents for their love and support throughout the years. I was also immensely fortunate to meet and marry Minqin at Caltech. I believe this was my greatest achievement I have ever had in my entire life. Thank you, Minqin, for your love and encouragement during times of ups and downs. Minqin, do you still remember an ancient Chinese saying, "A lady puts makeup for who is fond of her. A gentleman dies for who knows him the best" ? Sure, I do.

Abstract

Directed evolution, inspired by Darwinian evolution in Nature, is an effective approach for protein design. An industrially-important enzyme, subtilisin E, has been chosen as the research target. Important methodologies for directed evolution have been developed, including optimizing the error-prone polymerase chain reaction (PCR) to allow easy and precise control of the mutation rate, optimizing DNA shuffling for high fidelity recombination, and developing three new *in vitro* recombination methods: random priming recombination (RPR), defined primer recombination (DPR) and staggered extension process (StEP) recombination.

Using these techniques, subtilisin E isolated from the mesophilic organism *Bacillus subtilis* has been rapidly converted into its thermophilic counterpart (without compromising its activity). After five generations of directed evolution, the resulting variant 5-3H5 is as stable as its naturally-occurring thermostable homolog, thermitase, isolated from the thermophilic organism *Thermoactinomyces vulgaris*. The half-lives of thermal inactivation at 83 °C of both 5-3H5 and thermitase are 3.5 min. Their temperature optima are 76 °C, 18 °C higher than that of wild type subtilisin E. In addition, 5-3H5 is more active than wild type subtilisin E over the whole range of temperatures. The mutations responsible for the enhanced thermostability were identified and mapped into the structure of subtilisin E. Our findings strongly supports the notion that thermal stability is achieved by the cumulative effect of small improvements at many locations within the protein molecule. Thus, not surprisingly, the pursuit of a 'holy grail' of rules for protein thermostabilization was deemed unsuccessful. However, as demonstrated here, directed evolution is a generally applicable, highly effective approach to increase protein thermostability.

The concepts and techniques developed for directed evolution may also be applied to solving problems associated with molecular evolution in Nature. For example, due to

significant sequence divergence, identification of the adaptive mutations, neutral mutations and deleterious mutations in evolutionarily-related proteins is a difficult task. We developed a convenient method to identify functional mutations by gene recombination and sequence analysis of a small sampling of the recombined library exhibiting the evolved behavior. As a demonstration, this approach was used to identify the two thermostable mutations out of ten mutations in a laboratory-evolved thermostable subtilisin E variant.

Table of Contents

Chapter 1

Introduction

Biocatalysis and Biocatalysts

Enzymes exhibit exquisite catalytic power unmatched by conventional catalysts. Their many applications range from serving as catalysts for chemical synthesis to use in diagnostic testing, foods and pharmaceuticals. Compared to conventional catalysts, enzymes as biocatalysts are advantageous for several reasons: (1) Enzymes are highly efficient. Under identical conditions, the rate of an enzymatic reaction may be as much as 10-14 orders of magnitude faster than the rate of the reaction without a catalyst. (2) Enzymes often promote highly chemoselective, regioselective, and stereoselective reactions which are difficult or impossible to emulate using other techniques of synthetic organic chemistry. (3) Enzymes catalyze reactions under relatively mild conditions with regard to temperature (ca. 37 °C), pressure (1 atm), and pH (ca. 7.0). This makes biocatalysis remarkably energy-efficient compared to the corresponding chemical processes. Furthermore, this also minimizes the problems of side-reactions, such as decomposition, isomerization and rearrangements. (4) Enzymes are natural catalysts which are generally environmentally-benign and produce less hazardous waste, such as toxic organic solvents or metals.

There are also some disadvantages associated with enzymes. The most important ones are: (1) Many potentially useful enzymes are relatively unstable. (2) Due to their natural origin, enzymes usually function best in natural environments, i.e., aqueous media, room temperature, neutral pH, etc. This may cause difficulties when the substrate or product is poorly soluble in or sensitive to water, or when the biocatalytic process needs to be operated at high temperature, or pressure. (3) A given enzyme is often capable of transforming only a very narrow selection of substrates. The ideal biocatalysts for synthetic applications are those enzymes having broad specificity towards both natural and nonnatural substrates while at the same time maintaining high degree of selectivity where stereoselectivity is needed.

These problems of using enzymes as biocatalysts may be regarded as consequences of nature's five-billion-year-evolutionary project. Proteins in nature have evolved, through selective pressure, to perform specific biological tasks. From another angle of view, these problems are in fact rooted in the lack of knowledge of two more fundamental issues. They are, 1) how does a protein fold from primary sequence into a well-defined three-dimensional structure? and 2) how does the sequence and structure of a protein determine its function(s) ?

Protein Structures and Functions

The principal component of all known enzymes is protein. Proteins are linear polymers of 20 naturally occurring amino acids. In theory, the three-dimensional structure arises from the sequence of the amino acids (the primary structure). The major driving force is energy minimization of the interactions of the side-chains of amino acids. In reality, in spite of considerable efforts over the decades, this folding problem still remains a major unsolved intellectual challenge. The fundamental reason lies in the fact that, with 20 different amino acids, there are a vast number of ways in which similar structures can be generated by different amino acid sequences. Enormous computing power is required to search through the protein configuration space.

Each enzyme contains an active site responsible for acting on substrate. This active site is composed of a set of amino acids (residues) finely positioned in the protein. The interactions among the substrate, the side chains of these active site residues and solvents determine the enzyme function, including rate of catalysis, specificity and selectivity. The basic principle underlining enzyme catalysis proposed by Pauling half a century ago is that enzymes increase the rate of a chemical reaction by preferentially binding the transition state of the substrate [1]. However, these interactions are very subtle and complex and depend on their precise locations. This kind of precision requirement is not only below the limit of resolution of X-ray crystallography but also poses a supreme technical difficulty for

crystallographers to capture the interactions in the transition state. Furthermore, enzyme catalysis is a dynamic process, far from a lock-and-key model, in which both enzyme and substrate structures adjust to accommodate each other. Consequently, the static enzyme structures determined by X-ray crystallography are of limited utility in exposing these dynamic processes. Take triosephosphate isomerase (TIM) as an example. This enzyme catalyzes the interconversion of dihydroxyacetone phosphate and D-glyceraldehyde 3-phosphate through the formation of a *cis*-enediol(ate) [2]. The catalytic center is composed of a protein loop (residue 166 to 176) that binds and stabilizes the reaction intermediate, and a catalytic base (Glu165) and a catalytic acid (His94) to mediate the enolizations. Extensive studies showed that this protein loop has two catalytic functions: it ensures an efficient throughput of substrate to product, and it stabilizes the reaction intermediate. However, the exact catalytic mechanism is still unknown due to the lack of high-resolution structures of the enzyme complexed with substrate and with the intermediate. Furthermore, the full catalytic power relies on the very precise positioning of Glu 165. When this group is moved by as little as ~1 Å (as in the mutant E165D), the catalytic efficiency of the enzyme decreases nearly a thousand fold, though its crystal structure shows no other significant alterations. A single message coming from all of the mechanistic and structural studies of TIM is one of precision, which means not only the exact fit between substrate and active site, but also other elements of precision which have not been fully appreciated [3].

A widely used approach to probe this structure-function relationship is "rational" design in which the presumed importance of a particular amino acid or a set of amino acids is probed by changing or deleting them and then examining the functional consequences [4,5]. Extensive structural and mechanistic information is required to guide such efforts. Identifying the amino acids responsible for existing protein functions and those which might give rise to new functions remains an often-overwhelming challenge. It is clear that our present understanding of protein structure and function does not yet guarantee that

rationally designed changes will yield the predicted outcomes. In fact, protein engineers frequently have been surprised by the range of effects caused by single mutations designed to change only one specific and simple property in an enzyme. This situation becomes even worse now that mutations far away from active sites have been found to affect catalysis [6,7, 8, 9]. For example, in a subtilisin E variant functioning in 60% dimethylformamide (DMF), several functional mutations are located more than 20 Å away from the active site and bound substrate [6]. Another example is isocitrate dehydrogenase of *E. coli*. Six mutations in the adenosine binding pocket have been engineered to shift coenzyme preference toward NAD. Additional mutations in the binding pocket impaired performance with NAD. Two mutations outside the pocket, however, promote the binding and catalysis by a net eight-fold increase in performance [8]. The implication from these studies is that the combined effects of many substitutions outside binding sites and catalytic sites may be considerable in aggregate. The often surprising results reveal how little we know of what is required to effect specific small changes.

Directed Evolution

An alternative and highly efficient approach to address this dilemma is random mutagenesis or directed evolution. This approach involves the generation and selection or screening of molecular repertoires with sufficient diversity for the altered function to be represented. This approach avoids preconceived ideas about what is important, and since (at least in principle) all possible single amino acid changes and some fraction of double and higher order changes can be made, it is possible to identify those structural changes that produce a particular functional effect with no bias. This may give rise to new insights and result in a deeper understanding of the relationship between protein sequence, structure and function.

Random mutagenesis differs from directed evolution in key features. Random mutagenesis involves mutating a single gene randomly followed by selection or screening

in an attempt to alter its function. Typically only a single round of point mutagenesis and selection or screening is performed. In contrast, directed evolution has a more ambitious goal of evolving novel protein functions in the test tube. Directed evolution may start from several homologous genes and involves multiple rounds of mutation, recombination and screening or selection. Interest in protein design by directed evolution has grown significantly in the past several years [6,10]. Directed evolution has been successfully used to engineer enzymes with increased thermostability [11,12,13], altered substrate specificity [14, 15, 16, 17], and enhanced catalytic activity in organic solvents [14,18,19].

Protein evolution is protein design as it occurs in nature. Directed evolution is protein evolution as it occurs in the laboratory. The term 'evolution' implies a gradual alteration in contrast to a sudden change. Evolution in nature occurs spontaneously and constantly during reproduction and survival such that organisms enable to adapt to ever changing environments. Generally associated with the name of Charles Darwin, natural evolution usually consists of two processes: generation of diversity and natural selection. Several theories, including the neo-Darwinian theory of evolution and the neutral theory of evolution, were postulated to explain the origin and nature of diversity [20]. Strictly speaking these theories aim to deal with evolution of species, while our concern in directed evolution is to deal with evolution at the molecular level (molecular evolution). Nevertheless, the patterns of relationship between species include their chemical components at the level of macromolecules, that is, the genetic material (e.g., DNA sequences) and its products (e.g., proteins). By mimicking the evolutionary processes in the test tube, directed evolution may give new insights to the driving force behind the evolutionary processes in nature as well as to the relationship between protein structure and function. After all, proteins are not the products of rational design, but rather arose from a combination of random mutation and natural selection. Thus not all protein behaviors are rational and understandable [21].

Evolutionary changes at the molecular level in nature are dynamic processes, which are the origin of molecular diversity. Such processes include gene duplication, shuffling of DNA (exon shuffling), random mutation, transposition, gene recombination, and gene conversion [22]. These processes have created and are creating the stunning variety of living things, cell types, and biological molecules existing in the world, each with its own highly specialized talents. However, these *in vivo* mechanisms operate at very low efficiency, eliciting insignificant changes of gene structures or functions even after millions of years. For example, random changes (neutral substitutions) of one residue only occur at a rate of roughly one per 10^8 years, and highly conserved residues less than one per 10^{11} years [23]. In order to harness the power of natural evolution for practical applications, this variation-selection scenario must occur very quickly, preferably in the order of weeks or days. The solution relies on two capabilities: rapid generation of diversity at the molecular level and rapid identification of the fittest-among-survivors. Thus one of the major challenges of directed evolution is to develop the technological tools necessary to accelerate these two processes.

Two natural evolutionary processes have been mimicked so far by directed evolution in the test tube: random point mutagenesis and *in vitro* recombination. These two techniques are fundamentally different. As shown in Fig. 1.1, random point mutagenesis usually starts from a single parent gene and introduce new mutations randomly in the progeny genes, while recombination can start from a pool of closely-related parent genes and generate combinations of existing mutations. Due to the use of polymerases, recombination methods can also introduce new point mutations.

The first step in directed evolution is to create molecular diversity starting from a target gene or a family of related genes (Fig. 1.2). The diversity can be created by introducing mutations and/or by recombination. The gene products are sorted by screening or selection, and those genes encoding improved products can be returned for further generations of evolution. This evolutionary process can be repeated until the goal is

achieved (or until there is no further improvement). The two major requirements for successful directed evolution are (1) functional expression of the target protein in a suitable microbial host; (2) developing an efficient screen (or selection) sensitive to the target property.

Project Overview

My research has focused on two aspects of directed evolution: (1) developing methods and strategies for design by evolution (Chapter 2 and Chapter 3) and (2) demonstrating them by engineering a novel, industrially useful enzyme (Chapter 4) and addressing a long-standing puzzle existing in both laboratory-evolved and naturally-evolved proteins, identification of the functional mutations (Chapter 5). All these efforts have been illustrated with the serine protease subtilisin E and its evolved variants.

Serine proteases are extremely widespread and exhibit diverse functions. They have been grouped into six clans, of which the two largest are the (chymo)trypsin-like and subtilisin-like clans. More than 140 members of subtilisin-like serine proteases have been discovered in Archaea, Bacteria, fungi, yeast, and high eukaryotes [24]. The mature enzymes range from 266 to 1775 residues. From several known crystal structures and a multiple alignment of known amino acids sequences, a core structure was predicted for the catalytic domain of all subtilisin-like proteases (Fig. 1.3) [24]. Only 19 of these core residues are highly conserved.

Subtilisin E is an alkaline serine protease produced in *Bacillus subtilis* [25]. Its gene has been cloned and sequenced [26]. Recently, its X-ray crystal structure has also been solved [27]. Subtilisins are produced from pre-pro-subtilisins consisting of the pre-sequence of 29 residues, the prosequence of 77 residues, and the mature protease of 275 residues [26]. The pre-sequence functions as the signal peptide for protein secretion across the membrane [28], while the pro-sequence acts as a "foldase" to guide the appropriate folding of the subtilisin molecule [29,30].

Over the past several years, fueled by the development and use of new technologies, directed evolution has emerged as a powerful tool for engineering new enzymes as well as addressing fundamental questions about structure, function and evolution of proteins. Chapters 2 and 3 describe the theoretical models and major techniques developed for directed evolution, including optimization of existing techniques as well as new techniques. To facilitate the DNA manipulation and protein expression, especially increasing the size of variant library, the subtilisin E working system has been optimized. This includes the construction of a shuttle vector between *E. coli* and *B. subtilis* and the establishment of a sensitive and efficient screen for thermostability. Based on this system, important methodologies for directed evolution have been developed. These include optimizing the error-prone polymerase chain reaction (PCR) to allow easy and precise control of the mutation rate, optimizing a recently-developed *in vitro* recombination method -- DNA shuffling -- for high fidelity recombination, and developing three new *in vitro* recombination methods: random priming recombination (RPR), defined primer recombination (DPR) and the staggered extension process (StEP).

With these techniques, I have attempted to address two fundamental questions in protein structure-function. The first question is, how can we increase protein thermostability efficiently and what is the molecular basis of this thermostability? Stability here includes thermodynamic stability, as measured by reversible denaturation, and kinetic stability, as measured by the unfolding rate for enzymes that are subject to irreversible denaturation [31]. Denaturation and stability are interconnected, since perturbing the native structure of a protein is the only way to quantify its stability. *Ab initio* calculations of the free energy of stabilization of proteins are not feasible [32]. As mentioned earlier, one of the major problems of using biocatalysts for industrial applications is the low stability of enzymes. In many industrial applications, stability is defined as having a sufficient lifetime under specified conditions to complete a reaction. As such, the factors affecting kinetic stability are as important as those affecting the folding-unfolding equilibrium [31].

The three-dimensional structure of proteins is determined by two classes of non-covalent interactions, electrostatic and hydrophobic. The electrostatic interactions include ion pairs, hydrogen bonds, weakly polar interactions and van der Waals forces [32,33]. Hydrophobic interactions imply van der Waals forces and hydration effects of non-polar groups [34,35,36]. Protein stability represents the cumulative effect of these interactions at many locations. However, interpretation of stabilization in terms of these interactions is extremely difficult since the free energy of stabilization of proteins represents a marginal difference of large numbers as a consequence of the delicate balance of attractive and repulsive forces. In fact, the overall free energy of stabilization is equivalent to the energy required to break a maximum of five hydrogen bonds, corresponding to about 1% of the total number of H-bonds in the folded structure [32]. The difference of free energy of stabilization between thermophilic and mesophilic enzymes is of the same order of magnitude (estimated ~5-7 kcal/mol), which is equivalent to a few hydrogen bonds or two ion pairs [38, 39]. Thus, no 'holy grail' in terms of predicting protein stabilization from specific amino acid changes is expected to exist. However, a 'holy approach' does exist. As shown in Chapter 4, directed evolution is an extremely efficient and powerful approach to increase thermostability. By mimicking the natural process of molecular adaptation to thermophilic conditions, the important contributions towards thermostability coming from delicate balance of stabilizing and destabilizing interactions can be probed. Furthermore, since thermophiles were primordial and subsequent organisms were derived from them [40], in other words, mesophilic enzymes evolved from their thermophilic ancestors, it is interesting to know whether we can reverse this evolution process. Can we convert a mesophilic enzyme (descendent) into its thermophilic counterpart (ancestor)?

The second fundamental issue in protein structure-function that I would like to address is how to rapidly identify functional mutations in both laboratory-evolved and naturally-evolved proteins. According to neutral theory of evolution, the great majority of mutant substitutions are caused by random fixation through sampling drift of selectively

neutral mutants [41]. As a consequence, the sequences of evolutionarily-related proteins usually have diverged significantly, as we have seen for the subtilisin-like proteases. Thus, identification of those adaptive mutations (i.e., those affecting the growth and survival of the organism), neutral mutations and deleterious mutations or even identification of the important determinants in a specific case becomes an overwhelming task. This problem exists for enzymes evolved *in vitro* as well. While *in vitro* evolution can lead to the development of useful new protein functions, the responsible mutations almost always occur in a background of mutations which are neutral or even deleterious to the behavior(s) of interest. To address these problems, I developed a convenient method based on what we have learned from *in vitro* evolution. As shown in Chapter 5, this approach has been used to identify two thermostable mutations out of ten mutations in a laboratory-evolved thermostable subtilisin E variant. This method involves the random recombination of homologous sequences followed by screening for the altered behavior. A similar approach, coupled with selection rather than screening, could be used to distinguish adaptive from neutral mutations.

Finally, two previously written papers have been included in the end of this thesis. The first one is a chapter written for a book --"ASM Manual of Industrial Microbiology and Biotechnology," which describes the methods for optimizing industrial enzymes by directed evolution (Appendix A). The second one is a review paper describing different kinds of screening methods developed for random mutagenesis or directed evolution experiments (Appendix B).

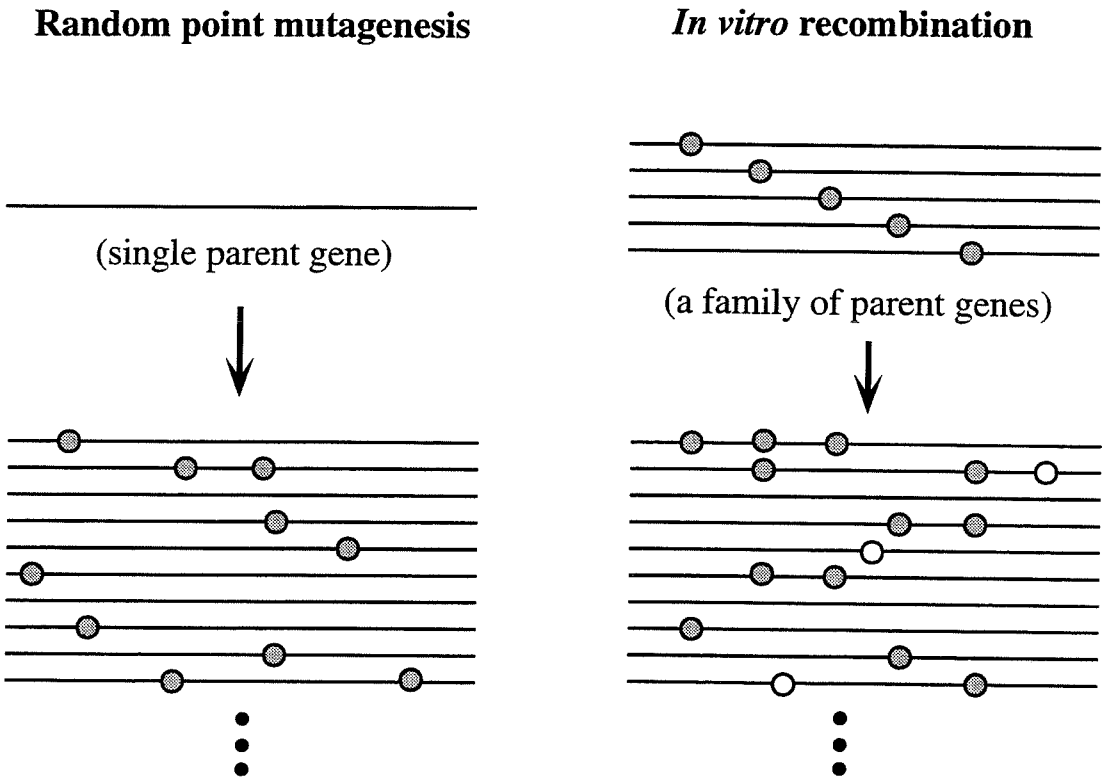


Fig. 1.1. Two major gene diversification processes mimicked in directed evolution. DNA sequences are shown in lines and mutations in circles. New mutations introduced during *in vitro* recombination are represented by empty circles.

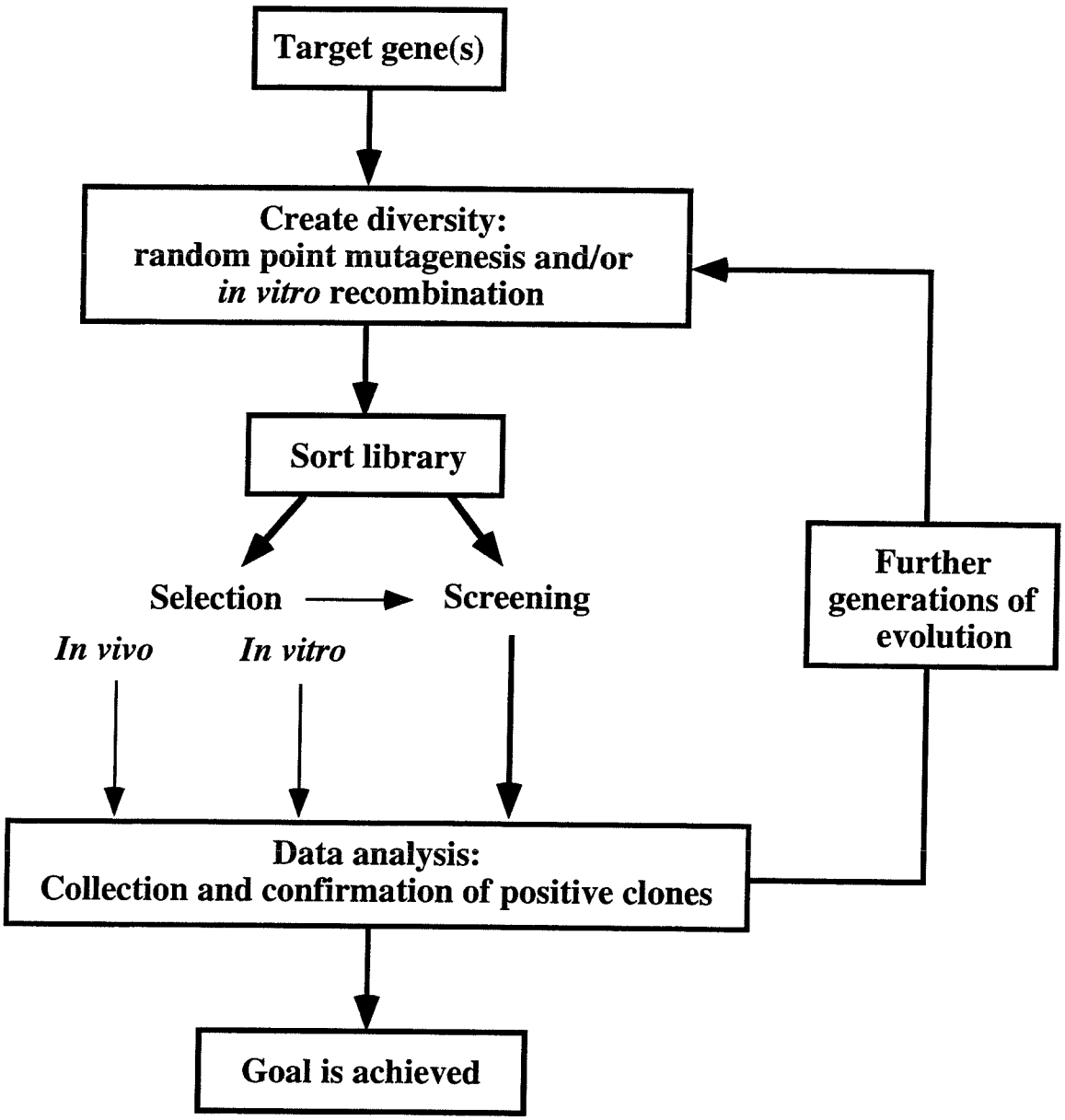


Fig. 1.2. Flowchart for directed evolution process.

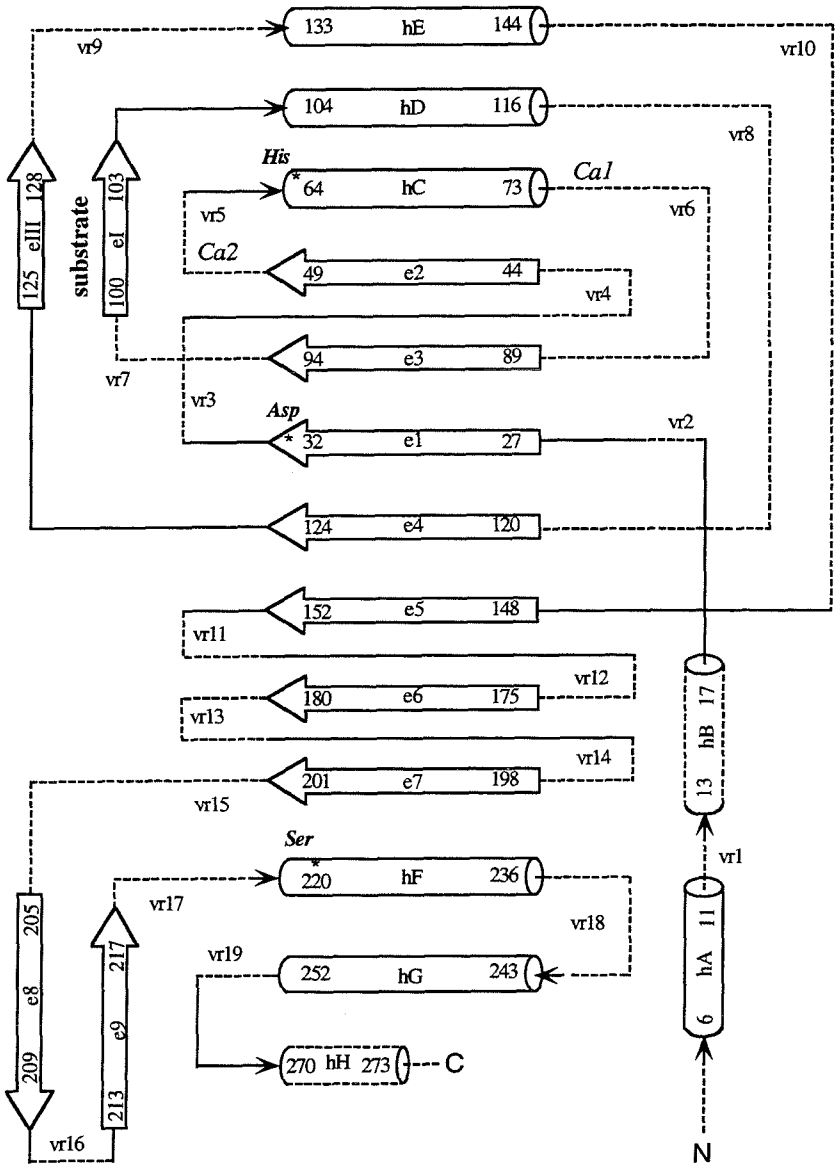


Fig. 1.3. Schematic representation of the secondary structure topology of subtilisins, with α-helices shown as cylinders and β-sheet strands as arrows. Solid lines indicate the conserved regions and dashed lines the variable regions. Approximate location is indicated of the main Ca^{2+}-binding sites (Ca1 and Ca2), catalytic triad residues D32, H64, S221 (by *) and substrate-binding region (between strands eI and eIII). (redrawn from reference 23).

References

1. Pauling, L. (1948) Nature of forces between large molecules of biological interest. *Nature* **161**, 707-709.

2. Maister, S. G., Pett, C. P., Albany, W. J. and Knowles, J. R. (1976) Energetics of triosephosphate isomerase: the appearance of solvent tritium in substrate dihydroxyacetone phosphate and in product. *Biochemistry* **15**, 5607-5612.

3. Knowles, J. R. (1991) Enzyme catalysis: not different, just better. *Nature* **350**, 121-124.

4. Leatherbarrow, R. J. and Fersht, A. R. (1986) Protein engineering. *Protein Eng.* **1**, 7-16.

5. Knowles, J. R. (1987) Tinkering with enzymes - what are we learning. *Science* **236**, 1252-1258.

6. Arnold, F. H. (1998) Design by directed evolution. *Accounts of Chemical Research* **31**, 125-131.

7. Mace, J. E. and Agard, D. A. (1995) Kinetic and structural characterization of mutations of glycine 216 in α-lytic protease: a new target for engineering substrate specificity. *J. Mol. Biol.* **254**, 720-736.

8. Chen, R., Greer, A. and Dean, A. M. (1995) A highly active decarboxylating dehydrogenase with rationally inverted coenzyme specificity. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 11666-11670.

9. Liu, D. R., Magliery, T. J., Pasternak, M. and Schultz, P. G. (1997) Engineering a transfer-RNA and aminoacyl-transfer-RNA synthetase for the site-specific incorporation of unnatural amino acids into proteins *in vivo*. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 10092-10097.

10. Kuchner, O. and Arnold, F. H. (1997) Directed evolution of enzyme catalysts. *Trends in Biotechnology* **15**, 523-530.

11. Liao, H., Mckenzie, T. and Hagemen, R. (1986) Isolation of a thermostable enzyme variant by cloning and selection in a thermophile. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 576-580.

12. Giver, L., Gershenson, A., Freskgard, P. O. and Arnold, F. H. Laboratory evolution of a thermostable enzyme. submitted.

13. Zhao, H. and Arnold, F. H. manuscript in preparation.

14. Moore, J. C. and Arnold, F. H. (1996) Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. *Nature Biotechnol.* **14**, 458-467.

15. Black, M. E., Newcomb, T.G., Wilson, H. M. P. and Loeb, L.A. (1996) Creation of drug-specific herpes-simplex virus type-1 thymidine kinase mutants for gene therapy. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 3525-3529.

16. Zhang, J.-H., Dawes, G. and Stemmer, W. P. C. (1997) Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 4504-4509.

17. Stemmer, W.P.C. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389-391.

18. Chen, K. and Arnold, F.H. (1993) Tuning the activity of an enzyme for unusual environments - sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 5618-5622.

19. You, L. and Arnold, F. H. (1996) Directed evolution of subtilisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylformamide. *Protein Eng.* **9**, 77-83.

20. Maynard-Smith, J. (1975) The theory of evolution 3rd edition (Penguin, New York).

21. Benner, S. and Ellington, A. D. (1990) Evolution and structural theory: the Frontier between chemistry and biology. in *Bioorganic Chemistry Frontiers*, **1** (Springer-verlag, Berlin), p2-69.

22. Li, W.-H. and Graur, D. (1991) Fundamentals of Molecular Evolution (Sinauer Associates, Inc., Massachusetts).

23. Robson B. and Garnier, J. (1986) Introduction to Proteins and Protein Engineering. (Elsevier, Amsterdam / New York / Oxford) p321.

24. Siezen, R.J. and Leunissen J. A. M. (1997) Subtilases: the superfamily of subtilisin-like serine proteases. *Protein Science* **6**, 501-523.

25. Boyer, H.W. and Carton, B.C. (1968) Production of two proteolytic enzymes by a transformable strain of *Bacillus subtilis*. *Arch. Biochem. Biophys.* **128**, 442-455.

26. Stahl, M.L. and Ferrari, E. (1984) Replacemnet of the *Bacillus subtilis* subtilisin structural gene with an *in vitro*-derived deletion mutation. *J. Bacteriol.* **158**, 411-418.

27. Chu, N. M., Chao, Y. and Bi, R. C. (1995) The 2Å crystal structure of subtilisin E with PMSF inhibitor. *Protein Eng.* **8**, 211-215.

28. Wong, S. L. and Doi, R. H. (1986) Determination of the signal peptidase cleavage site in the preprosubtilisin of *Bacillus subtilis*. *J. Bio. Chem.* **261**, 10176-10181.

29. Ohta, Y. and Inouye, M. (1990) Pro-subtilisin E: purification and characterization of its autoprocessing to active subtilisin E *in vitro*. *Molecular Microbiology* **4**, 295-304.

30. Gallagher, T., Gilliland, G. and Bryan, P. (1995) The prosegment-subtilisin BPN' complex- crystal structure of a specific foldase. *Structure* **3**, 907-914.

31. Shaw, A. and Bott, R. (1996) Engineering enzymes for stability. *Curr. Opin. Struct. Bio.* **6**, 546-550.

32. Jaenicke, R., Schurig, H., Beaucamp, N. and Ostendorp, R. (1996) Structure and Stability of hyperstable proteins: glycolytic enzymes from hyperthermophilic bacterium *Thermotoga Maritima*. *Advances in protein chemistry* **48**, 181-269.

33. Dill, K. A. (1990) Dominant forces in protein folding. *Biochemistry* **29**, 7133-7155.

34. Burley, S. K. and Petsko, G. A. (1988) Weakly polar interactions in proteins. *Adv. Protein Chem.* **39**, 125-186.

35. Tanford, C. (1973) *The hydrophobic effect* (Wiley, New York).

36. Privalov, P. L. and Gill, S. J. (1988) stability of protein structure and hydrophobic interaction. *Adv. Protein Chem.* **39**, 191-234.

37. Privalov, P. L. and Gill, S. J. (1989) The hydrophobic effect - a reappraisal. *Pure Appl. Chem.* **61**, 1097-1104.

38. Perutz, M. F. (1978) Electrostatic effects in proteins. *Science* **201**, 1187-1189.

39. Fersht, A. R. and Serrano, L. (1993) Principles of protein stability derived from protein engineering experiments. *Curr. Opin. Struct. Biol.* **3**, 75-83.

40. Woese, C. R. (1993) The archaea: their history and significance In *New Comprehensive Biochemistry* v. 26. eds. Kates, M., Kushner, D. J. and Matheson, A. T. (Elsevier, Amsterdam) vii - xxvii.

41. Kimura, M. (1991) Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. *Proc. Natl. Acad. Sci., U. S. A.* **88**, 5969-5973.

Chapter 2

Directed Evolution I : Implementation of the System and Error-prone Polymerase Chain Reaction (PCR)

INTRODUCTION

Setting up the Working System

As mentioned in Chapter 1, successful application of directed evolution requires sorting large numbers of variants for the desired feature(s). In an ideal situation, variants representing all possible changes responsible for a specific property should be sorted rapidly (in a few generations) in order to reach the optimum (at least locally) of that specific property. Thus, the target protein should be functionally expressed in a suitable microbial host, and an efficient screen (or selection) sensitive to that specific property is required. Based on these considerations, my initial efforts were devoted to constructing a new shuttle vector with higher transformation efficiency both in *E. coli* and *B. subtilis* and developing an efficient thermostability screening method.

An early directed evolution experiment in this laboratory was to make the serine protease subtilisin E more active in high concentrations of aqueous organic solvent by sequential random mutagenesis [1,2]. An enzyme that tolerates organic solvents is useful for synthetic reactions on substrates poorly soluble in water. Organic solvents can also shift the direction of equilibrium: in the case of proteases, addition of organic solvent promotes peptide synthesis over hydrolysis [3]. Most useful (polar) cosolvents, however, dramatically reduce enzyme activity. After multiple generations of random mutagenesis and screening, the best subtilisin E variant works almost as well in 60% dimethylformamide (DMF) as the wild type enzyme did in aqueous solution, a nearly 500-fold increase in total enzyme activity.

The system to evolve the activity of subtilisin E was not very efficient. Due to the low transformation efficiency of *B. subtilis* (a few hundred transformants per μg of plasmid DNA as compared to up to 10^9-10^{10} transformants/μg plasmid DNA in *E. coli.*), the size of the variant libraries was severely limited. This is at least partially due to the fact that only multimeric linear or circular forms of plasmids can efficiently transform *B.*

subtilis [4,5,6]. To overcome this problem, a *B. subtilis-E. coli* shuttle vector was constructed to facilitate establishment of the variant library in *B. subtilis* [2]. The transformation efficiency of *Bacillus subtilis* by this vector was improved to yield 10^3 transformants/μg plasmid DNA. However, the transformation efficiency of *E. coli* by this shuttle vector was very low and the yield of plasmid DNA was also poor. To some extent, previous workers were lucky in that mutations that improved the specific activity of subtilisin E in aqueous DMF were surprisingly abundant (generally one for every few hundred clones screened).

The particular feature of subtilisin E that my research focused on is thermostability. As outlined in Chapter 1, proteins are only marginally stable and it is difficult to identify the stabilizing interactions. Directed evolution, however, can efficiently fine-tune these subtle interactions by sorting through all available solutions and accumulating the effective ones. Since the chemical nature of DNA molecules does not vary much from one sequence to another, the techniques to generate molecular diversity are almost universal, requiring only minor modifications towards individual genes. However, due to the different chemical and physical nature of protein properties, individual fitness tests (screen or selection) have to be devised to pick out enzyme variant(s) with a certain desirable property.

Screening is the most flexible sorting method for directed evolution [7]. Two major criteria for a good screening method are the efficiency and sensitivity. In other words, the assay should be rapid and the screening conditions should ensure that the expected small improvements brought by single amino acid substitutions can be measured. In accordance with these considerations, the thermostability screen was developed by modifying a similar method performed in petri dishes [8]. The method is based on the retention of activity after incubation at an elevated temperature and is not designed to distinguish various mechanisms of inactivation. The ratio between residual activity and initial activity (normalized residual activity) is taken as index of thermostability. Variants with higher index of thermostability than wild type subtilisin E and with initial activity still comparable

to wild type are chosen as positives. These positives are then subject to a more rigorous analysis: measurement of thermal inactivation kinetics. Variants with half-lives longer than their parents were regarded as "true" positives.

Error-prone PCR

The errors occurring during DNA replication are called mutations. Mutations can be classified into four types: (1) substitutions (the replacement of one nucleotide by another), (2) deletions (the removal of one or more nucleotides), (3) insertions (the addition of one or more nucleotides), and (4) inversions (the rotation by 180° of a double-stranded DNA segment comprised of two or more base pairs) [9]. Among these, mutation by nucleotide substitutions is the most common process and can be easily mimicked in directed evolution. Effective techniques to introduce mutations over the whole length of a gene include chemical mutagens [10], UV radiation [11], mutator strains [12] and error-prone PCR [13]. In theory, since mutations occur independently (randomly), and no two mutations can occur at the same position simultaneously, the distribution of mutations generated by all of these techniques should follow a Poisson distribution (see Appendix 2 for detailed mathematical analysis). In practice, however, error-prone PCR is the most attractive for directed evolution, since it is very simple, robust, efficient, and most importantly, the mutation rate (the average number of mutations per gene) can be easily and precisely controlled. It is critical for the success of directed evolution that the mutation rate of random mutation be precisely controlled. Higher mutation rates will result in too many inactive clones, while lower mutation rate will yield too many parent-like (unmutated) clones.

Error-prone PCR was originally developed by Leung and coworkers [13] and further modified by several other groups [14-19] as well as this group. In principle, error-prone PCR is analogous to nature's own design -- mutations occur due to the inaccurate copying of a DNA template by a polymerase. The inaccuracy is the result of two factors:

the intrinsic low fidelity of *Taq* DNA polymerase (In fact, the fidelity (error-rate) of *Taq* polymerase is the lowest among commercially available thermostable polymerases [20]), and the extrinsic factors of the reaction conditions, including the addition of $MnCl_2$, increased concentration of $MgCl_2$, increased and unbalanced concentration of the four dNTPs, increased concentration of *Taq* polymerase, increased PCR cycles, etc. In addition, the polymerase fidelity also depends on the nature of target sequences. Different mutation rates will be observed on different sequences, even when the exact same polymerases and reaction conditions are used. Reagents from different manufacturers may also cause some small differences. Thus, the mutation rates of error-prone PCR should be determined for each gene. To simplify this, a simple and efficient approach has been developed to estimate the mutation rate of a specific error-prone PCR protocol. In practice, this approach can serve as a diagnostic check for the successful creation of a randomly-mutated library.

RESULTS AND DISCUSSION

Setting up the working system

Construction of an E. coli - B. subtilis Shuttle Vector pBE2 for Directed Evolution

Shuttle vector pBE2 was constructed by ligation of fragments from two widely-used plasmids: *B. subtilis* cloning vector pUB110 and *E. coli* cloning vector pGEM3 (Promega) following published protocols [21]. As illustrated in Fig. 2.1, these two purified plasmids were first digested with restriction enzyme *Eco*RI. In order to avoid self-ligation of vectors in the following step, the linearized pGEM3 plasmid was further de-phosphorylated. Equal amounts of these two linearized vectors were ligated with T4 DNA ligase and transformed into *E. coli* HB101. Several transformants appeared on an LB agar

plate supplemented with 20 μg/ml kanamycin and 100 μg/ml ampicillin after overnight incubation. One transformant was randomly picked and inoculated. Its plasmid was extracted, purified and checked by agarose gel electrophoresis. This plasmid turned out to have the right size, containing both pUB110 and pGEM3 fragments. The correct orientation of the ligation reaction was also confirmed by digestion with restriction enzyme *Pvu*II. In order to reduce the size and remove the redundant restriction sites (e.g., *Bam*HI, *Eco*RI and *Pvu*II), this vector was further digested with *Pvu*II. The large fragment was isolated, ligated and transformed into *E. coli* HB101 again. This resulted in shuttle vector pBE2. The transformation efficiency of pBE2 is $2x10^4$ / μg plasmid DNA (to *B. subtilis* DB428) and 10^8 /μg plasmid DNA (to *E. coli* HB101 by electroporation). The stability of pBE2 was found to be similar to that of pUB110 in *B. subtilis*. The plasmid copy-number was also found to be similar to that of pGEM3 in *E. coli* (several hundred copies) which results in high yield of plasmid DNA.

Subcloning of Wild Type Subtilisin E Gene: Construction of Plasmid pBE3

Directed evolution has been carried out on the DNA fragment containing the complete gene encoding mature subtilisin E. To facilitate the experiment, the DNA fragment to be manipulated should contain two unique restriction sites flanking the fragment. Thus, *Nde*I (located within prosequence, 10 amino acids ahead of the first amino acid of mature subtilisin E) was chosen as the upstream restriction site and *Bam*HI (located 113 nt after the stop codon of mature subtilisin E gene) as the downstream restriction site. Since there is another *Nde*I site within the vector (located within the fragment from pUB110), site-directed mutagenesis was carried out to remove that site. The successful modification of pBE2 vector was confirmed by the digestion with *Nde*I and *Eco*RI (only one fragment appeared after agarose gel electrophoresis). The wild-type subtilisin E gene was then removed from pKWZ by cutting with *Eco*RI and *Bam*HI and subcloned into *Eco*RI-*Bam*HI-digested modified pBE2 to form pBE3 (Fig. 2.2). Using

pBE3, all DNA manipulation can be performed conveniently in *E. coli*, while the enzymes are still expressed in *B. subtilis*. Since the plasmid DNA from the variant library can be obtained in large quantities in *E. coli*, the transformation efficiency of *B. subtilis* is no longer a limiting factor for the library size. For example, the transformation efficiency of *E. coli* by ligation mixture is usually 10^5-10^6 /μg DNA, the DNA from this plasmid library can be amplified to produce as much as several hundred micrograms in *E. coli*. Since the transformation efficiency of *B. subtilis* is $2x10^4$ /μg plasmid DNA, this means the size of the enzyme library (expressed) can be larger than 10^6, comparable to the library size that could be obtained if the gene could be directly expressed in *E. coli*.

Finally, it is worth pointing out that a significant fraction of plasmids isolated from *E. coli* HB101 are in the form of supercoiled dimer or multimers, which transform *B. subtilis* competent cells with high efficiency. Plasmids from other *E. coli* strains, such as JM109, XL1-blue, ER1648, and DH5α' act differently and transform *B. subtilis* with much lower efficiency. No explanation has been found for this observation.

Establishment of the Thermostability Screening Method

As illustrated in Fig. 2.3, after transformation and incubation on petri dishes, colonies were usually picked by toothpicks and grown in 96-well plates. After overnight incubation, a small portion of the supernatant from each well was transferred into two 96-well plates. One plate is used to measure the initial activity (after brief incubation at an elevated temperature, such as 65 °C), and the other is used to measure the residual activity (after prolonged incubation at the same temperature). Both initial activity and residual activity are measured at 37 °C spectrophotometrically by 96-well plate reader.

The activity assay is based on a synthetic substrate for subtilisin E - succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide, which provides a colorimetric signal (yellow color) when hydrolyzed by the enzyme. The activity is determined by the initial rate of the reaction. This assay is very rapid (~1 min) and sensitive (as low as ~μM of enzyme). To determine

the reproducibility of the assay and whether there is any positional variability in the plate, a control plate containing only wild type clones was evaluated. This caution allows us to identify systematic errors due to pipetting, heat transfer, etc., and to determine the values of the thermostability index that should be associated with true positives. In the control plate, wild type subtilisin E clones grew in 200 µl of SG medium supplemented with 30 µg/ml kanamycin and incubated for 20 h. The cells were spun down and 5 µl of supernatant from each well was transferred into two fresh empty plates, into which we added 15 µl of SG medium (to avoid a problem with evaporation during incubation at high temperatures). Initial activity was measured after 5 min of incubation at 65 °C, and residual activity was measured after 20 min of incubation at 65 °C. The 65 °C incubation was performed in an oven on an aluminum block machined to closely contact standard multi-well plates for uniform heating. As shown in Fig. 2.4, a small amount of systematic variability is observed, possibly due to uneven heating of the wells. In general, however, the activity values vary within the range of 30-40% normalized residual activity. This provides confidence that variants with more than 40% normalized residual activity are likely to be more thermostable variants of subtilisin E. In practice, I usually chose variants exhibiting about twice this activity level (normalized residual activity, 80% vs. 30-40%).

These conditions were used to screen first generation variants only (i.e., wild type is their parent). For the second generation and thereafter, the incubation times for initial and residual activity remained the same, but the temperature was elevated gradually. For each generation, this temperature was determined by the point where the parent shows 30-40% normalized residual activity after incubation.

Kinetics of Thermal Inactivation

Proteases, including subtilisins and neutral proteases, differ from other proteins by the fact that autolysis is the main cause of inactivation at elevated temperatures. It is generally believed that the rate of thermal inactivation is determined by the rate of local

unfolding processes that render proteases susceptible to autolysis [21-25]. In accordance with this model, the process of thermal inactivation should follow first-order kinetics, and the rate of thermoinactivation (or thermostability index defined in my thermostability screen) should also be independent of the enzyme concentration (see Appendix 1 for mathematical description).

Several experiments have been performed to check whether this model also holds true for subtilisin E. First, the kinetics of thermal inactivation of purified wild type subtilisin E in 10 mM Tris-HCl (pH 8.0), 1 mM $CaCl_2$, was investigated by following the time dependence of the enzyme activity (see Materials and Methods) at 60 °C and 65 °C. As shown in Fig. 2.5, plots of the logarithm of residual activity versus time were linear up to the point where greater than 80% of the enzyme activity has been lost. The coefficients of least-squares fit were all larger than 0.995. Thus, the rate of thermal inactivation of wild type subtilisin E at high temperatures followed first-order kinetics. Second, the inactivation experiments were carried out at 65 °C at different enzyme concentrations within the range of 40 - 800 μg/ml. As shown in Table 2.1, under these conditions, the rate of thermal inactivation was virtually independent of its concentration. Both findings strongly suggests that the rate-limiting step of thermal inactivation of wild type subtilisin E is also the local unfolding processes.

Error-prone PCR

A Modified Error-Prone PCR Protocol

The gene encoding subtilisin E was subjected to mutagenic PCR with varying concentrations of $MnCl_2$. The mutagenic PCR was based on the protocols outlined by Cadwell and Joyce [15] and Shafikhani *et al.* [18] with slight modifications. After mutagenesis, the PCR products were purified and subcloned into *E. coli-B. subtilis* shuttle

vector pBE3. The ligation mixtures were then transformed into *E. coli* HB101 by electroporation. More than 2000 colonies for each library (containing PCR product generated at a given concentration of $MnCl_2$) were obtained. These libraries were further collected and their plasmids were isolated by mini-prep and transformed into *B. subtilis* DB428 for expression. About 300 clones from each library were picked by toothpicks and grown overnight in SG medium supplemented with 30 μg/ml kanamycin in 96-well plates. Enzyme activities of all variants from each library were measured using a plate reader. Data were sorted and plotted in a descending order to generate activity profiles. Fig. 2.6 shows the activity profiles of each library. Clones exhibiting at least 10% of wild type subtilisin activity are scored as active. The fraction of active clones varied from 90% at 0 mM $MnCl_2$ to only 30% at 0.5 mM. Even a concentration difference as small as 0.05 mM $MnCl_2$ makes a clear difference in the activity profile.

To assess the mutation rate as well as mutation type, I randomly picked eight clones from the library with ~50% active clones (0.2 mM $MnCl_2$) and sequenced. As shown in Table 2.2, twenty-three single-base substitutions were identified in a total of 7500 nt sequenced (an overall mutation rate of 0.3% per position) and no deletions or insertions were found. The frequency of mutating each of the four bases is approximately equal except for a 1.5-fold enhanced probability of mutating G bases. The mutation probability at AT base pairs is almost equal to the probability at GC pairs (the ratio of the sum of A→N and T→N over the sum of G→N and C→N is 11/12 or ~1.0, N=G,A,C,T). This is consistent with the observation by Cadwell and Joyce [15] but inconsistent with the observation by Shafikhani *et al.* [18]. Another difference from both these protocols [15,18] is that our sequencing results clearly show that transitions (substitutions between A and G (purines) and between C and T (pyrimidines)) occurs almost twice as often as transversions (substitutions between a purine and a pyrimidine). Based on all the published results of existing protocols [13-19], error-prone PCR does not create random DNA substitutions, i.e., the frequency of all possible 12 mutation types is not equal. The

least common mutation types are G→C and C→G changes, which presumably reflects the difficulty in forming and extending G•G and C•C mismatches. Although Table 2.1 shows the absence of these two types as well as A→C and G→T changes, these four types did all occur in my later studies (Chapter 4). Thus, the absence of these types of mutation is probably due to the small sampling size for sequencing (7500 nt). Systematic studies have been conducted to define the rules relating the frequencies of each mutation type to the set of the dNTP concentrations in the PCR experiment, but no conditions have been obtained where more balanced mutations can be generated [17]. In all published studies, as well as our experiments, no mutational hot spots have been detected.

Since point mutation is one of the most important factors in the natural evolution of DNA sequences, molecular evolutionists have long been interested in determining the pattern of spontaneous mutation. Pseudogenes are usually chosen as the research targets due to their selective neutrality. Data from 13 mammalian pseudogene sequences showed that (1) the direction of mutation is nonrandom [13]. For example, A changes more often to G than to either T or C. (2) All transitions, and in particular C→T and G→A, occur more often than transversions. The sum of the frequency of transitions is 59.2%. Under random mutation the expected frequency of transition is only 33%, for there are only four types of transitions but eight types of transversions (4 / (4+8) = 33%). The observed frequency of transitions is almost twice the value expected under random mutation. (3) G→C, C→G, A→C, G→T occur much less frequently [13]. These observations are similar to our results with error-prone PCR and those of others. It seems that the diversity generated by nature is also somewhat limited.

Guidelines for Performing Error-Prone PCR on Other Genes

In error-prone PCR, the two most important factors to consider are mutation frequency and mutation pattern. The overall mutation frequency is the product of three parameters: the error rate (fidelity) of the polymerase under the specific reaction conditions,

the length of the mutagenized gene and the number of effective doubling cycles. The first parameter is the most complex. It can be affected by the reaction conditions as well as the nature of the target sequences. Varying the concentration of $MnCl_2$ may be one of the simplest methods to control the error rate. The second parameter, the gene length, is usually fixed. The third, the effective number of cycles, is easily adjustable by altering the PCR cycle number or the ratio of template to primers. However, the adjustable range is usually very narrow (less than two-fold).

In comparison with mutation frequency, the pattern of point mutations is almost unpredictable. Significant variability among protocols has been observed. However, as evidenced by natural evolution as well as the successful results from random mutagenesis and directed evolution experiments, some bias in the pattern of point mutation can be tolerated, while large variations in mutation frequency usually cannot.

The actual overall mutation frequency of error-prone PCR may vary from gene to gene or protocol to protocol. This means we had to sequence a few random clones (random sampling) every time we created a library by error-prone PCR in order to ensure the optimum overall mutation frequency. Unfortunately, this is time-consuming and very expensive. To alleviate this problem, we developed a simple and efficient approach to estimate the mutation frequency by analyzing the activity profile for each library. This approach is based on the statistical relationship between the mutation frequency and activity profile, mainly the fraction of active clones. A simple linear model for this relationship has been proposed by Shafikhani *et al.* [18]. However, this model is not consistent with our experimental data. This relationship may be hard to interpret mathematically, but it does exist since the experiments can be reproduced very reliably. Thus, for a given sequence, even with lack of a mathematical model, this fraction of active clones can serve as a diagnostic check for the successful creation of randomly-mutated library with defined overall mutation frequency. For example, the error-prone PCR of subtilisin E gene with a

mutation frequency of two base changes per gene, or ~0.2%, always produces ~65% active clones.

In summary, for any given sequence, general guidelines to create the desired mutation level for directed evolution are: (1) Calculate the desired mutation frequency. A good target mutation level is ~2-3 base substitutions per gene [26], which is the product of the length of the DNA coding sequence and the mutation frequency. Thus, for a 1 kb sequence, the mutation frequency should be ~0.2-0.3%. (2) Use the outlined mutagenic PCR protocol as a starting point and set up several reactions at a series of $MnCl_2$ concentrations. Choose the conditions producing 50% active clones and determine its mutation frequency by sequencing ten random clones. (3) Adjust the mutation frequency by slightly varying the $MnCl_2$ concentration. (4) Always use the activity profile as a diagnostic check for successful creation of the desired library.

MATERIALS AND METHODS

Reagents and Kits

Plasmids from *E. coli* were purified with QIAprep spin plasmid miniprep kit (Qiagen, Chatsworth, CA). The DNA fragments after restriction-digestion are separated by agarose gel electrophoresis and purified with QIAEX II gel purification kit (Qiagen, Chatsworth, CA). Restriction enzymes were purchased from New England Biolabs. Succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide (s-AAPF-*p*Na) was from Sigma. Enzyme activity was measured as described elsewhere [27] in 96-well plates using a Thermomax microplate reader (Molecular Devices, Sunnyvale, CA). Preparation and electroporation of *E. coli* HB101 competent cells were performed as described in the manual accompanying Bio-Rad Gene Pulse II Electroporator. *Bacillus subtilis* strain DB428 and *Bacillus* cloning vector

pKWZ containing the subtilisin E gene were kindly provided by Dr. R. Doi of University of California, Davis.

Elimination of NdeI Site Within pBE2 Vector

The extra *Nde*I site is located at position of 1076 of pUB110 (numbering based on pUB110 alone). ExSite^{TM} PCR-based site directed mutagenesis kit (Stratagene, San Diego) was used to remove this site. Two primers: 5'-CTGTA AATCG CTCCT TTTTA G-3' (underlined letter is a change from A to C which eliminates the restriction site) and 5'-ATGAG TTATG CAGTT TGTAG A-3' were designed according to the instructions accompanying the kit. After mutagenesis, six plasmids (pBE2) were randomly picked and digested with *Nde*I and *Eco*RI. The vector without *Nde*I site should only have a single band after agarose gel electrophoresis. Two out of these six plasmids had single bands and the rest had double bands.

DNA Transformation of E. coli HB101

Plasmid DNAs or ligation mixtures are transformed into *E. coli* HB101 by electroporation on the Bio-Rad Gene Pulser II electroporator. Electrocompetent *E. coli* HB101 cells are prepared according to the protocol supplied by the Bio-Rad.

DNA Transformation of B. subtilis DB428

The following transformation protocol is modified after published protocols [28, 29]. *Materials*: SPI salts (1x): 0.2% $(NH_4)_2SO_4$, 1.4% K_2HPO_4, 0.6% KH_2PO4, 0.1% Na-citrate, 0.02% $MgSO_4$; 50 mM $CaCl_2$ (100x); 250 mM $MgCl_2$; 50% glucose; CAYE (100x): 2% casamino acids, 10% yeast extract; 100 mM ethylene glycol-bis(β-aminoethylether)-N,N,N'N'-tetraacetic acid (EGTA). SPI medium: any volume SPI salts plus 1/100 volume glucose and CAYE; SPII medium: any volume SPI medium plus 1/100 volume each of $CaCl_2$ and $MgCl_2$. *Procedures*: (1) Inoculate 1 ml of SPI medium with a

single colony of DB428, and shake at 30 °C overnight. (2) Add 10 ml of fresh SPI medium (which usually gives an initial turbidity of 20 ± 5 Klett units (54, green filter)). Shake at 200 rpm at 37 °C (about 3.5 h). (3) Follow the growth by measuring turbidity with the Klett colorimeter and plot the growth curve on semilog paper until the cells have just entered the stationary phase. (4) Add 10 ml of the early stationary culture to 90 ml of prewarmed SPII medium plus 1 μM $MnCl_2$ (final concentration) and incubate at 37 °C for 90 min with slow shaking (50-100 rpm). The turbidity should increase from around 35 Klett units to 80-100 Klett units, and cells grow exponentially. (5) Harvest cells by centrifugation (6500 xg) at 4 °C for 10 min. Gently resuspend the pellet in 10 ml of the **same** culture supernatant to which glycerol has been added to a final concentration of 10% (v/v). (6) Take 0.5 ml aliquots into Eppendorf tubes, quickly freeze in liquid nitrogen and store at -70 °C (transformation efficiency will drop several fold after 6 months). (7) To use, quickly thaw the 0.5 ml aliquot by shaking in a 42 °C water bath, and then add 2 ml of prewarmed SPII medium to bring the volume to one-half of its original culture volume. (8) Add 1/100 volume of EGTA and continue the incubation for 10 min. (9) Add 0.2-0.5 ml of the competent cells to a prewarmed glass tube containing 0.05 - 0.1 μg of DNA and incubate at 37 °C for 90 min with moderate shaking (200 rpm). (10) Spread aliquots onto appropriate plates for selection of desired transformants.

Error-prone PCR

The target is the wild type subtilisin E gene, including 45 nt of pro-sequence, the complete mature subtilisin E sequence and 113 nt after the stop codon. Primers P5N (5'-CCGAG CGTTG CATAT GTGGA AG-3', underlined sequence is *NdeI* restriction site) and P3B (5'-CGACT CTAGA GGATC CGATT C-3', underlined sequence is *BamHI* restriction site) were used to amplify this ~1 kb DNA fragment. The PCR reaction contained (100 μl final volume): 10 mM Tris-HCl (pH 8.3 at 25 °C), 50 mM KCl, 7 mM $MgCl_2$, 0.01% (wt/vol) gelatin, 0.2 mM dGTP, 0.2 mM dATP, 1 mM dCTP, 1 mM TTP,

varying concentration of $MnCl_2$, 0.3 μM of both primers, 5 ng of template and 5U *Taq* DNA polymerase (Promega). No mineral oil was overlaid since the lid of thin PCR tube was pre-heated. PCR was performed in a MJ Research (Watertown, MA) PTC-200 thermocycler for 13 cycles : 1 min 94 °C, 1 min 5 0 °C and 1 min 72 °C. The PCR products were purified using Wizard PCR Preps (Promega), followed by restriction digestion by *Nde*I and *Bam*HI. These digestion products were purified again using Wizard PCR Preps (Promega), subcloned into *E. coli* - *B. subtilis* shuttle vector pBE2 and transformed into *E. coli* HB101 by electroporation. The transformants were incubated on LB + ampicillin (100 μg/ml) agar plates at 37 °C for 15 h and then harvested. Plasmid DNAs were isolated from these transformants and further transformed into competent *B. subtilis* DB428 cells for protein expression.

DNA Sequencing

Genes were individually purified from *B. subtilis* DB428 using a QIAprep spin plasmid miniprep kit (Qiagen, Chatsworth, CA) with the modification that 2 mg/ml lysozyme was added to P1 buffer and the cells were incubated for 5 min at 37 °C, retransformed into competent *E. coli* HB101 and then purified again using QIAprep spin plasmid miniprep kit to obtain sequencing quality DNA. Sequencing was done on an ABI 373 DNA Sequencing System using the Dye Terminator Cycle Sequencing kit (Perkin-Elmer, Branchburg, NJ).

Enzyme Activity Assay in 96-well Plates

Clones in *B. subtilis* were picked and grown in 200 μl of SG medium supplemented with 30 μg/ml kanamycin. After 20 h incubation, cells were spun down. 5 μl of supernatant from each well was transferred into a fresh 96-well plate. 100 μl of prewarmed (37 °C) enzyme activity assay solution (100 mM Tris-HCl, 10 mM $CaCl_2$, 0.2 mM suc-AAPF-*p*NA, pH 8.0) was added into each well. The initial rates were measured at

405 nm on Thermomax 96-well plate reader (Molecular Device, Sunnyvale, CA). For each library, data from different plates were combined, sorted and plotted in descending order.

Enzyme Purification

(A). Growth and Expression of Subtilisin. Day 1: Grow a fresh LB+ kanamycin (50ug/ml) plate of *Bacillus* strain containing subtilisin gene of interest at 37 °C overnight. Day 2: Innoculate 3 ml cultures with single colonies from this plate. Use SG media and grow at 37 °C on shaker overnight (250 rpm). Day 3: Innoculate 2x500 ml of SG medium (in 2000 ml flasks) with one tube for each flask. Grow for 36h, occasionally checking the media for subtilisin activity. (B) Purification by Chromatography. (1) Remove cells from 36h culture by centrifugation in the JA10 rotor at 5000 xg for 15 min. Combine supernatants and check for activity. (2) Ammonium Sulfate Precipitation. Stir the 1 liter of supernatant in the coldroom, slowly add 656 g ammonium sulfate (take about 15 min). Allow the mixture to stir in the coldroom for another 15 min. Finally, take mixture out of coldroom and stir at room temperature for 20 min more. Centrifuge for 20 min at 5000 xg in JA10 rotor. Resuspend pellet in 100 ml of 10 mM sodium phosphate buffer pH 6.2, dialyze in this same buffer at 4 °C (2x4 liters). Measure the volume and activity of the solution. (3) Acetone Precipitation. Keeping all the solutions cold during this procedure is very important. Chill acetone to -20 °C. Add NaCl to the enzyme sample to give a final concentration of 50 mM (this helps the precipitation). Have the enzyme stirring in the water bath. Add acetone slowly to solution to give a final concentration of 50%. Stir solution in saltwater ice bath for 15 min. Put rotor in centrifuge and turn the centrifuge on so it is cold when you want to use it. Keep sample on ice when you are transferring it to the centrifuge. Centrifuge the sample at 20,000 xg for 15 min and then take the sample back to the coldroom. Activity should still be in the supernatant, throw away the pellet. Add additional acetone to bring sample to 65% acetone. To obtain a precipitate, follow the same procedure for the 50% solution described above. The subtilisin is now in the pellet.

Resuspend the pellet in (about 30ml) 10mM sodium phosphate buffer pH 6.2. Dialyze in this same buffer at 4 °C (2x4 liters). (4) CM Sepharose Chromatography. For an original 1 liter sample of Bacillus, prepare a 2.5x20 cm column of CM sepharose CL6B (Pharmacia). Pour out the correct amount of CM sepharose (ml) and do several washes with 100mM sodium phosphate buffer pH 6.2. Using the equilibrated sepharose pour the column, allow it to pack to correct height by "pulling" from the bottom at a rate of 2.6 ml/min. When you have packed the column to the correct height, put the plunger on top of the column and begin "pushing" 10 mM sodium phosphate buffer pH 6.2 through the column. Equilibrate at a rate of 2.6 ml/ml for about 2 hours. Check to make sure buffer coming off column is close to pH 6.2. Reduce speed of pump to 1.3 ml/min. and load the enzyme extract that has been acetone precipitated. Monitor the column protein output using a UV detector. Collect column output in 8 ml fractions. Wash the column with 10 mM sodium phosphate buffer pH 6.2 until all proteins that do not stick to the column have been washed off. Lower plunger as close to the top of sepharose as possible and begin gradient, 300 ml, 0-0.4 M NaCl buffer, stir 0 M NaCl buffer. Run column at 1.3 ml/min and mix gradient at 1/2 that rate. Subtilisin peak will come off about 1/2 way through gradient as indicated by a large peak on the monitor. Measure activity of fractions in peak and pool those with the highest amount of activity. Dialyze into an appropriate buffer (for example, 10 mM phosphate buffer, pH 6.2, or 10 mM Tris pH 8.0, 1 mM $CaCl_2$).

Thermal Inactivation

Kinetics measurements were made using a thermostatted Milton-Roy Spectronic 3000 Array spectrophotometer. Purified enzymes were dialyzed in 10 mM Tris-HCl, 1 mM $CaCl_2$, pH 8.0 at 4 °C overnight before characterization. The enzyme concentration was determined by absorbance at 280 nm ($\varepsilon = 35886\ M^{-1}cm^{-1}$). Subtilisin activity was determined by the initial rates of hydrolysis of s-AAPF-pNa substrate in 0.99 ml of 10 mM Tris-HCl, 1 mM $CaCl_2$, pH 8.0 at 37 °C as described previously [1]. A MJ Research

(Watertown, MA) PTC-200 thermocycler was used as an incubator for precisely controlling the temperatures. Purified enzymes were incubated in 10 mM Tris-HCl (pH 8.0), 1 mM $CaCl_2$ at specific temperatures. Aliquots, taken at various time intervals, were removed and diluted into 1.0 ml of activity assay solution (0.2 mM s-AAPF-*p*Na, 10 mM Tris-HCl (pH 8.0), 1 mM $CaCl_2$) equilibrated at 37 °C. Typically, inactivation was followed until greater than 80% of the enzyme activity has been lost and plots of the logarithm of residual activity versus time were linear. The rate of thermal inactivation (k_{inact}) was the slope of the straight line fitted by linear regression algorithm and values of half-lives were calculated by the equation: $t_{1/2} = \ln 2 / k_{inact}$.

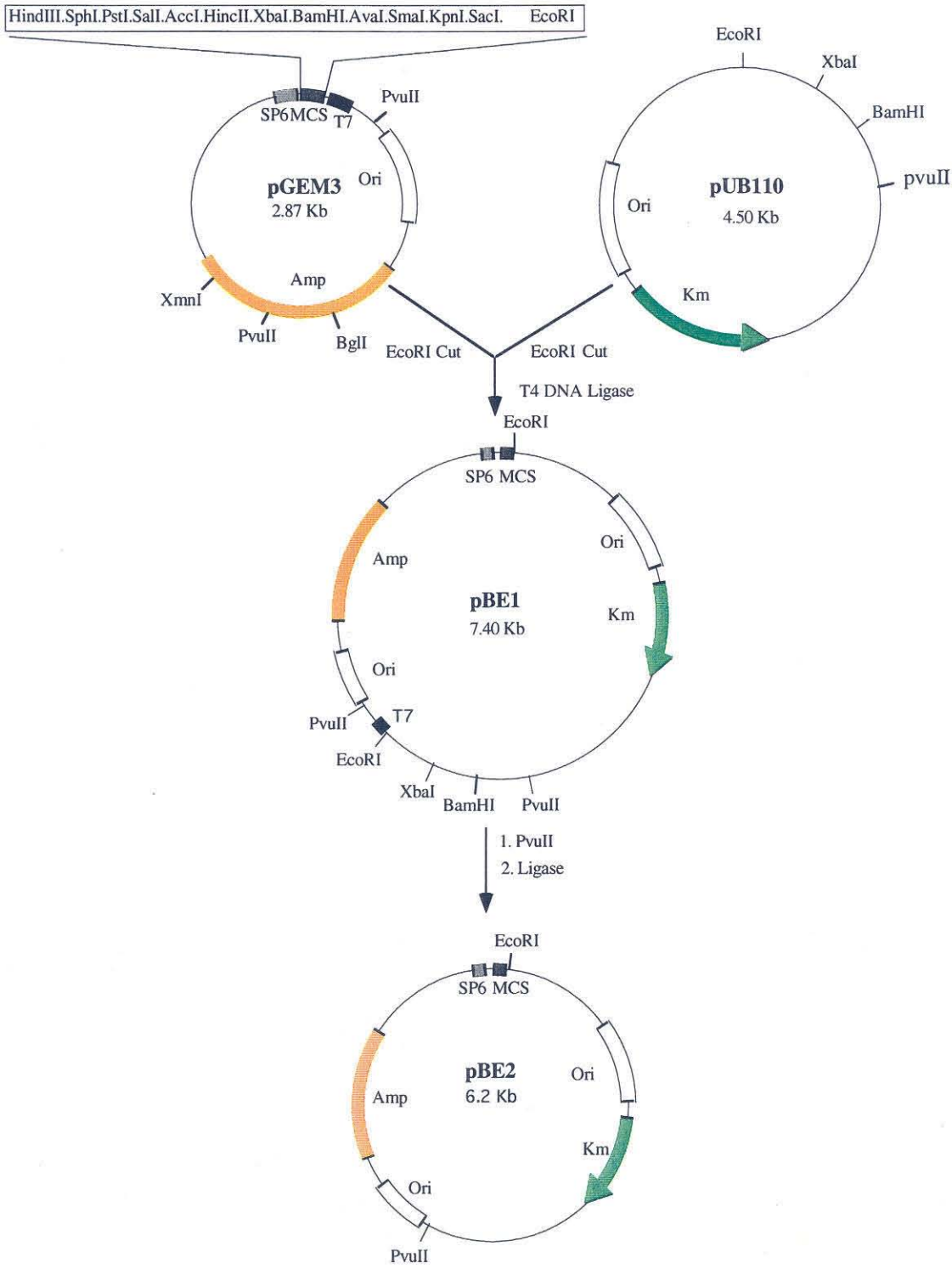


Fig. 2.1. Construction of *E. coli /B. subtilis* shuttle vector pBE2.

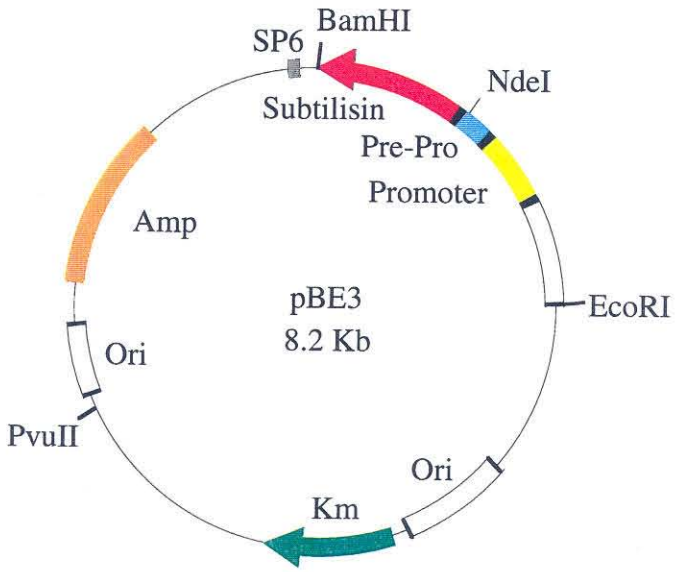
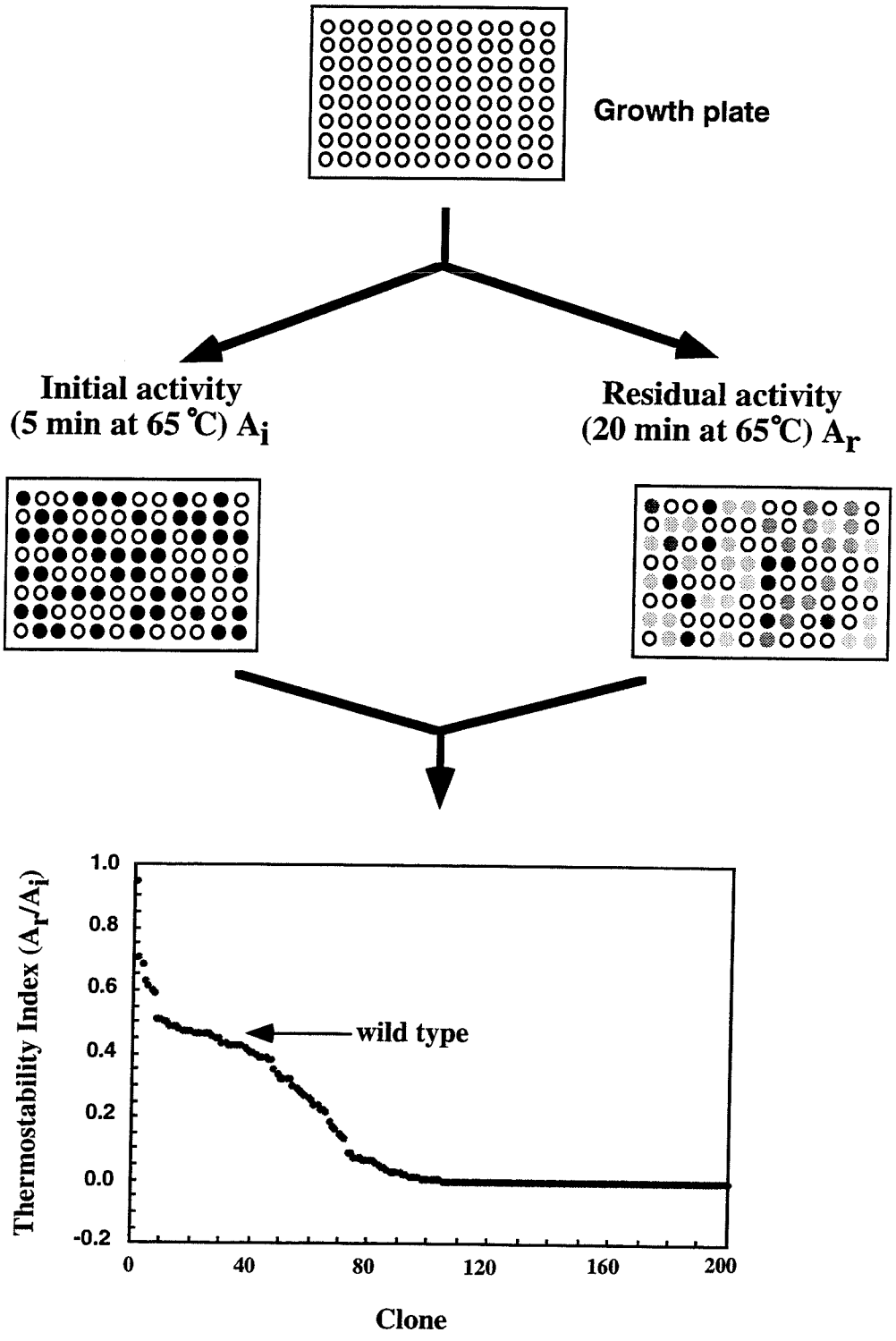


Fig. 2.2. The plasmid map of pBE3. The subtilisin E gene including its natural promoter was subcloned into *E. coli-B. subtilis* shuttle vector pBE2 by *Bam*HI and *Eco*RI restriction sites.

Fig. 2.3. The scheme for thermostability screen. Variants are usually picked by toothpicks and grown in 96-well plates (growth plates). After overnight incubation, a small portion of the supernatant from each well is transferred into two 96-well plates (assay plates). One plate is used to measure the initial activity (after brief incubation at an elevated temperature. For first generation variants, the condition is 5 min at 65 °C), and the other is used to measure the residual activity (after prolonged incubation at the same temperature. For generation variants, the condition is 20 min at 65 °C). Both initial activity and residual activity towards s-AAPF-*p*NA are measured at 37 °C spectrophotometrically by 96-well plate reader. The ratio between residual activity and initial activity (normalized residual activity) is taken as index of thermostability. Data are sorted and plotted in descending order. Variants with higher index of thermostability than the parent and with initial activity still comparable to the parent are chosen as positives.


Growth plate
Initial activity
(5 min at 65 ℃) A_i
Residual activity
(20 min at 65℃) A_r
Thermostability Index (A_r/A_i)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
wild type
0
40
80
120
160
200
Clone

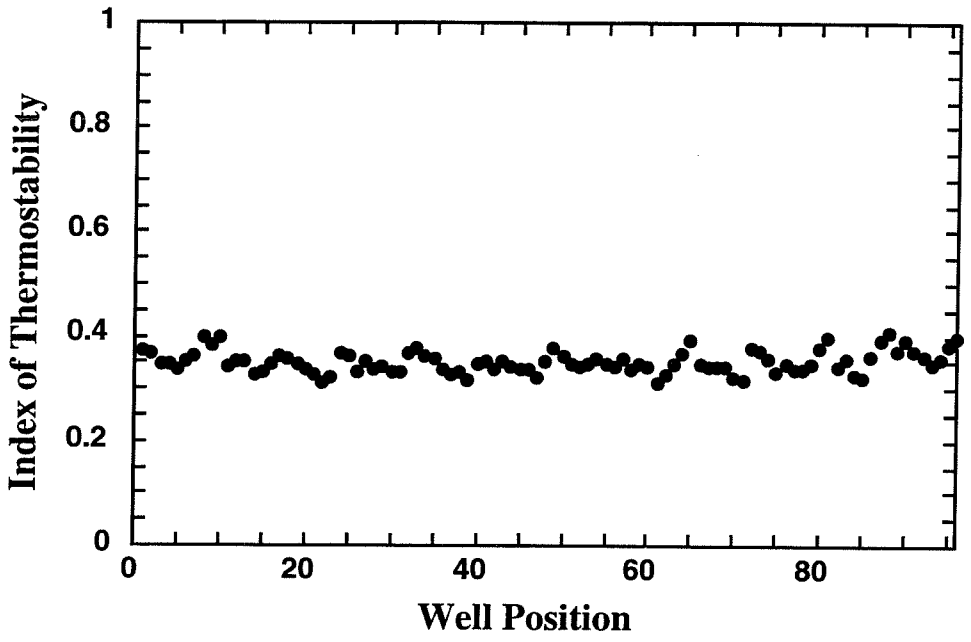


Fig. 2.4. Variations in the thermostability index of wild type subtilisin E clones in a 96-well plate assay.

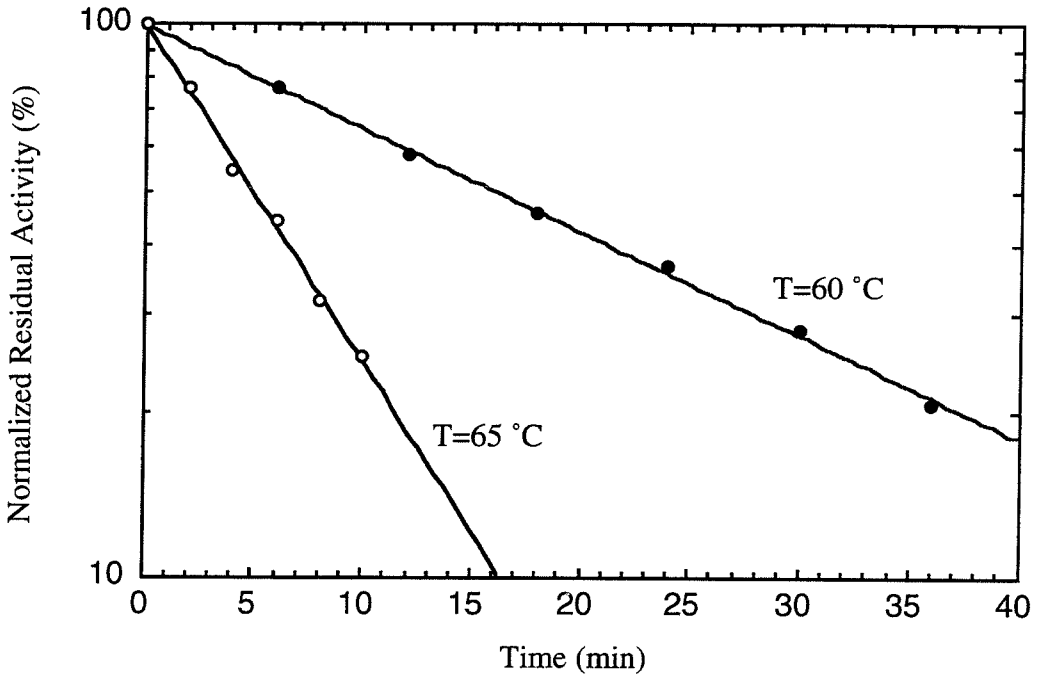


Fig. 2.5. The kinetics of thermal inactivation of wild type subtilisin E at 60 and 65 °C.

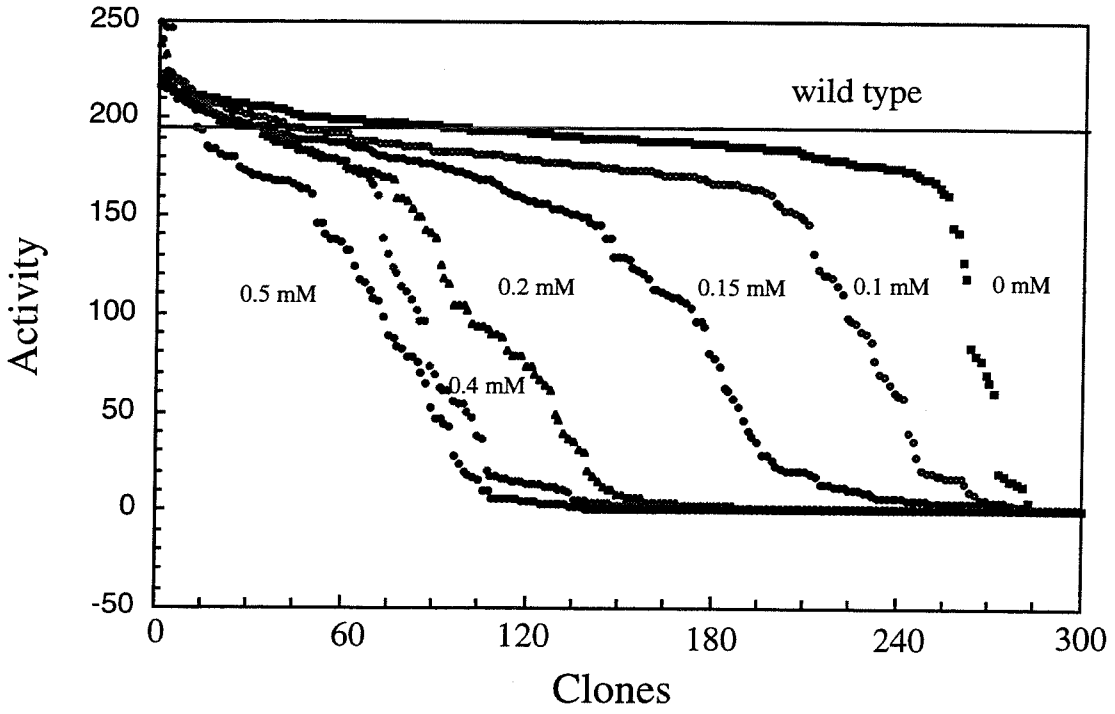


Fig. 2.6. Activity profiles of subtilisin E variant libraries generated at varying concentrations of $MnCl_2$ (0 - 0.5 mM). About 300 variants from each library were taken for initial activity measurements. Data were sorted and plotted in descending order. The activity of wild type is shown by the horizontal line.

Table 2.1. Concentration independence for the half-life of irreversible thermal inactivation of wild type subtilisin E.

concentration of wild type subtilisin E (ug/ml)	half-life (min) for thermal inactivation [a,b]
800	4.88
400	5.01
200	4.92
100	4.87
40	4.93

a. Inactivation half-lives are determined as described in Materials and Methods, in 1 mM $CaCl_2$ and 10 mM Tris buffer (pH 8.0) at 65 °C.

b. Values of half-lives are the average of duplicate experiments. Deviations are less than or equal to ± 5%.

Table 2.2. Mutations introduced by error-prone PCR at 0.2 mM $MnCl_2$ (about 7500 nt sequenced).

Transition	Occurrences	Transversion	Occurrences
G → A	8	A → T	2
A → G	3	A → C	0
C → T	2	C → A	2
T → C	2	C → G	0
		G → C	0
		G → T	0
		T → A	3
		T → G	1

Total : 23

Appendix 1. Rationale Behind the Thermostability Screen

A. The Process of Thermoinactivation Follows First-order Kinetics.

At elevated temperatures most subtilisins and neutral proteases are irreversibly inactivated as a result of autolysis. It is generally believed that reversible unfolding processes which render the protease susceptible towards the irreversible process of autolysis determines its rate of inactivation. This can be described by a simple model:

$$\mathrm{N} \underset{k_r}{\overset{k_f}{\rightleftharpoons}} \mathrm{U} \xrightarrow{k_i} \mathrm{I},$$

where N = folded (intact) protease, U = unfolded protease, I = inactivated protease.

Since

$$\frac{d[\mathrm{I}]}{dt} = -\frac{d[\mathrm{N}]}{dt} = k_i[\mathrm{U}] \qquad (1)$$

and according to principle of steady-state theory, we have

$$\frac{d[\mathrm{U}]}{dt} = 0 = k_f[\mathrm{N}] - k_i[\mathrm{U}] - k_r[\mathrm{U}]$$

$$k_f[\mathrm{N}] = (k_i + k_r)[\mathrm{U}]$$

$$[\mathrm{U}] = \frac{k_i + k_r}{k_f}[\mathrm{N}] \qquad (2)$$

Substitute (2) into (1), we have

$$\frac{d[\mathrm{U}]}{dt} = -\frac{d[\mathrm{N}]}{dt} = \frac{k_i + k_r}{k_f}[\mathrm{N}]$$

or

$$-\frac{d[\mathrm{N}]}{dt} = k[\mathrm{N}] \quad \text{where } k = \frac{k_i + k_r}{k_f}.$$

Further we have

$$-\frac{d[\mathrm{N}]}{[\mathrm{N}]} = kdt$$

Integration of both sides of the equation,

$$-\int_{\mathrm{N}_0}^{\mathrm{N}_t} \frac{d[\mathrm{N}]}{[\mathrm{N}]} = \int_0^t kdt$$

$$\ln[\mathrm{N}]\Big|_{[\mathrm{N}_0]}^{[\mathrm{N}_t]} = -kt$$

$$\frac{[\mathrm{N}_t]}{[\mathrm{N}_0]} = e^{-kt} \quad \text{or } \mathrm{N}_t = \mathrm{N}_0\, e^{-kt}$$

Since enzyme activity (total activity) is correlated to the concentration of folded enzyme, we have

$$\mathrm{A}_t = \mathrm{A}_0 e^{-kt}$$

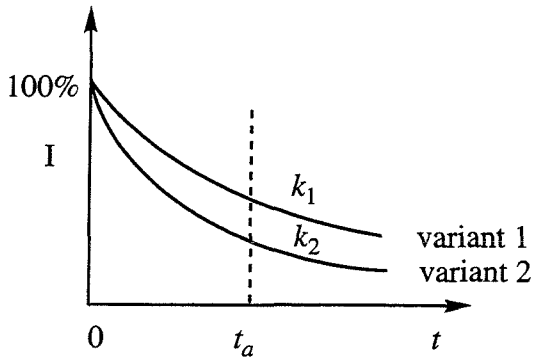
where A_t = enzyme activity after incubation at elevated temperatures for a time period of t, A_0 = initial activity, k = rate of thermoinactivation.

B. The Thermostability Screen for Subtilisin E is Independent of Enzyme Concentration and Activity.

As shown above, the loss of activity for subtilisin E at high temperatures follows first-order kinetics. By definition, the index of thermostability (I) is

$$\mathrm{I} = \frac{\mathrm{A}_{res}}{\mathrm{A}_{ini}} = \frac{\mathrm{A}_{t_a}}{\mathrm{A}_0} = e^{-k_1 t_a}$$

Suppose we have two variants with different rate of thermoinactivation k_1 and k_2 as shown in the following figure:



then for

$$\text{variant 1} \quad I_1 = \frac{A_{t_a,1}}{A_{0,1}} = e^{-k_1 t_a}$$

$$\text{variant 2} \quad I_2 = \frac{A_{t_a,2}}{A_{0,2}} = e^{-k_2 t_a}$$

thus

$$\ln\left(\frac{I_1}{I_2}\right) = \ln\left(e^{-k_1 t_a + k_2 t_a}\right) = (k_2 - k_1)t_a$$

If $I_1 > I_2$, then $\ln\left(\frac{I_1}{I_2}\right) > 0$, thus $k_1 < k_2$

This result means that at an elevated temperature, variant 1, deactivates more slowly than variant 2 (variant 1 is more thermostable than variant 2). In other words, the variant with higher index of thermostability is more thermostable.

Appendix 2. The Distribution of Random Mutations Follows Poisson Distribution

For generality, suppose a gene is composed of n nucleotides (the length of the gene = n), the probability of having a mutation is p_n (the mutation rate = p_n), then the average number of mutations per gene (λ) is $n \cdot p_n$ ($\lambda = n \cdot p_n$ or $p_n = \lambda/n$).

Assume (1) each mutation is independent (random) and (2) no two mutations occur at the same position simultaneously, then the probability of having exact k mutations in n nucleotides (positions) is

$$p\{x = k\} = C_k^n \, p_n^k (1 - p_n)^{n-k} \qquad (1)$$

To simplify equation (1), we have

$$p\{x = k\} = \frac{n!}{(n-k)!k!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n=k}$$

$$= \frac{1}{k!} \times \frac{n(n-1)\cdots(n-k+1)}{n^k} \times \lambda^k \times \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^k}$$

Since $k << n$, we have $\frac{n-i+1}{n} \to 1 (i = 1, 2, \cdots, k)$, and

$$\frac{n(n-1)\cdots(n-k+1)}{n^k} \to 1, \left(1 - \frac{\lambda}{n}\right)^k \to 1, \left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}$$

thus

$$C_k^n \, p_n^k (1 - p_n)^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}$$

or

$$p\{x = k\} = \frac{\lambda^k}{k!} e^{-\lambda} (k = 0, 1, 2, \cdots) \text{ (Poisson distribution)}$$

Because assumptions (1) and (2) can be applied to all currently existing random mutagenesis techniques, the above deduction is general or in other words the distribution of random mutations generated by existing random mutagenesis techniques should all follow Poisson distribution.

The importance of this finding can be understood in the following. First, mutations generated by any existing technique are not uniform, instead, are scattered over a wide range. Thus, variants with any defined number of mutations only occupy a small fraction of the whole library of variants generated by random mutagenesis. Second, this finding makes it possible to calculate the optimal mutation rate at which the number of single mutants (variant with one amino acid substitution) is maximized.

References

1. Chen, K. and Arnold, F. H. (1993) Tuning the activity of an enzyme for unusual environments - sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 5618-5622.

2. You, L. and Arnold, F. H. (1996) Directed evolution of subtilisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylforamide. *Protein Eng.* **9**, 77-83.

3. Wong, C.-H. and Wang, K.-T. (1991) New developments in enzymatic peptide-synthesis. *Experientia* **47**, 1123-1129.

4. Canosi, U., Morelli, G. and Trautner, T. A. (1978) The relationship between molecular structure and transformation efficiency of some *S. aureus* plasmids ioslated from *B. subtilis*. *Mol. Gen. Genet.* **166**, 259-267.

5. Mottes, M., Grandi, G., Sgaramella, V., Canosi, U., Morelli, G. and Trautner, T. A. (1979) Different specific activities of the monomeric and oligomeric forms of plasmid DNA in transformation of *B. subtilis* and *E. coli*. *Mol. Gen. Genet.* **174**, 281-286.

6. de Vos, W. M., Venema, G., Canosi, U. and Trautner, T. A. (1981) Plasmid transformation in *Bacillus subtilis*: fate of plasmid DNA. *Mol. Gen. Genet.* **181**, 424-433.

7. Zhao, H. and Arnold, F. H. (1997) Combinatorial protein design: strategies for screening protein libraries. *Curr. Opin. Struct. Biol.* **7**, 480-485.

8. Bryan, P. N., Rollence, M. L., Pantoliano, M. W., Wood, J., Finzel, B. C., Gilliland, G. L., Howard, A. J. and Poulos, T. L. (1986) Proteases of enhanced stability:characterization of a thermostable variant of subtilisin. *Proteins Struct. Funct. Genet.* **1**, 326-334.

9. Li, W.-H. and Graur, D. (1991) Fundamentals of Molecular Evolution (Sinauer Associates, Inc., Massachusetts).

10. Myers, R. M., Lerman, L. S. and Maniatis, T. (1985) A general method for saturation mutagenesis of cloned DNA fragments. *Science* **229**, 242-247.

11. Botstein, D. and Shortle, D. (1985) Strategies and applications of *in vitro* mutagenesis. *Science* **229**, 1193-1201.

12. Greener, A., Callahan, M. and Jerpseth, B. (1997) An efficient random mutagenesis technique using an *Escherichia coli* mutator strain. *Mol. Biotech.* **7**, 189-195.

13. Leung, D. W., Chen, E. and Goeddel, D. V. (1989) A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Technique,* **1**, 11-15.

14. Zhou, Y., Zhang, X. and Ebright, R. H. (1991) Random mutagenesis of gene-sized DNA molecules by use of PCR with *Taq* DNA polymerase. *Nucleic Acids Res.* **19**, 6052.

15. Cadwell, R. C. and Joyce, G. F. (1992) Randomization of genes by PCR mutagenesis. *PCR Methods and Appli.* **2**, 28-33.

16. Cadwell, R. C. and Joyce, G. F. (1994) Mutagenic PCR. *PCR Methods and Appli.* **3**, S136-S140.

17. Fromant, M., Blanquet, S. and Plateau, P. (1995) Direct random mutagenesis of gene-sized DNA fragments using polymerase chain reaction. *Anal. Biochem.* **224**, 347-353.

18. Shafikhani, S., Siegel, R. A., Ferrari, E. and Schellenberger, V. (1997) Generation of large libraries of random mutants in *Bacillus subtilis* by PCR-based plasmid multimerization. *Biotechniques* **23**, 304-310.

19. Lin-Goerke, J. L., Robbins, D. J. and Burczak, J. D. (1997) PCR-based random mutagenesis using manganese and reduced dNTP concentration. *Biotechniques* **23**, 409-412.

20. Eckert, K. A. and Kunkel, T. A. (1991) DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Applic.* **1**, 17-24.

21. Guo, X.-H., Xiong, Z., Zhou, M., Jia, S.-F. and Xu, Y. (1991) The construction of multifunctional shuttle vectors of *Bacillus subtilis-Escherichia coli. Chinese J. of Biotechnology* **7**, 224-229.

22. Dahlquist, F. W., Long, J. W. and Bigbee, W. L. (1976) Role of calcium in the thermal stability of thermolysin. *Biochemistry*, **15**, 1103-1111.

23 Vriend, G. and Eijsink, V. G. H. (1993) Prediction and analysis of structure, stability and unfolding of thermolysin-like proteases. *J. Comput.-Aided Mol. Design*, **7**, 367-396.

24. Kidokoro, S., Miki, Y., Endo, K., Wada, A., Nagao, H., Miyake, T., Aoyama, A., Yoneya, T. and Kai, K. (1995) Remarkable activity enhancement of thermolysin mutants. *FEBS Letters* **367**, 73-76.

25. Braxton, S. and Wells, J. A. (1992) Incorporation of a stabilizing Ca^{2+}-binding loop into subtilisin BPN'. *Biochemistry* **31**, 7796-7801.

26. Moore, J. C. and Arnold, F. H. (1996) Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. *Nature Biotechnol.* **14,** 458-467.

27. Zhao, H. and Arnold, F. H. (1997) Functional and non-functional mutations distinguished by random recombination of homologous genes. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 7997-8000.

28. Gryczan, T. J., Contente, S. and Dubnau, D. (1978) Characterization of *Staphylococcus aureus* plasmids introduced by transformation into *Bacillus subtilis*. *J. Bacteriol.* **134**, 318-329.

29. Dubnau, D. and Davidoff-Abelson, R. (1971) Fate of transforming DNA following uptake by competent *Bacillus subtilis*. *J. Mol. Biol.* **56**, 209-221.

Chapter 3

Directed Evolution II : *In vitro* Recombination

INTRODUCTION

Genetic recombination is the reassortment of a series of nucleotides along a nucleic acid molecule, usually of double stranded DNA, in exceptional cases also of RNA. Genetic recombination can be classified into several types: (1) reciprocal recombination, which is a symmetrical exchange between two DNA double helices; (2) homologous recombination, which is recombination occurring between two DNA duplexes with the same or nearly the same base sequences; (3) site-specific recombination, which occurs at highly preferred sites; and (4) illegitimate recombination between two DNA duplexes with little or no DNA homology [1]. Such events make it possible to (1) increase the variation in a population by combining different variants, (2) combine positive traits of two parent into one offspring, (3) repair damaged genes, and (4) maintain a large potential genetic variability in a population which contains only a small amount of variability in each individual, as evolutionary insurance against a changing environment [2,3]. More importantly, this natural process reduces the time required for a population to become evolutionarily "fit" to a few generations as opposed to the large number of generations required if such positive trait information was not shared [4,5]. However, natural *in vivo* recombination mechanisms usually operate at low efficiencies, eliciting insignificant changes in gene structures or functions even after tens of generations. Furthermore, *in vivo* recombination in most organisms is cumbersome and difficult to adapt to the redesign of genes, operons or pathways.

Various approaches have been developed to mimic and accelerate nature's recombination strategy in order to direct the evolution of protein functions. *In vitro* recombination methods offer higher recombination efficiencies and greater experimental flexibility than *in vivo* approaches [6]. Stemmer was the first to develop an *in vitro* recombination method -- "DNA shuffling" -- for directed evolution [7,8]. As shown in Fig. 3.1, this technique usually consists of four steps: (1) preparation (amplification) of

genes to be shuffled by conventional PCR, (2) random fragmentation of these double-stranded DNA templates by an endonuclease DNase I, (3) reassembly of these short fragments into full-length genes by thermocycling in the presence of polymerase, and (4) amplification of the reassembled product of correct size by a conventional PCR. During the reassembly step (step 3), the short fragments prime each other based on homology (equivalent to multiple mini-overlap-extension PCR), and recombination occurs when fragments from one copy of a gene prime on another copy, causing a template switch. This technique has been successfully used to alter substrate specificity [8, 9], increase the functional expression of green fluorescent protein [10], increase catalytic activity in aqueous organic solvents [11], and increase arsenic resistance of an arsenate detoxification pathway [12] . Several limitations associated with the original protocol prompted us to optimize DNA shuffling for high fidelity recombination as well as develop several new approaches which provide useful features and advantages for different applications. The new approaches are random-priming *in vitro* recombination (RPR), defined-primer *in vitro* recombination (DPR) and staggered extension process (StEP) *in vitro* recombination.

While recombining the DNA sequences, the original DNA shuffling technique also introduces new point mutations at a relatively high rate (0.7%) [7]. Though these point mutations may provide useful diversity for some *in vitro* evolution applications, they are problematic for others, especially when the mutation rate is this high [13]. On the technical side, it is time-consuming and labor-intensive to monitor the digestion process of DNA templates by electrophoresis and further purify the digested product (smear) of certain sizes. Concerned by these issues, I made some modifications to Stemmer's original protocol. The new protocol is simpler, more efficient and, most importantly, the mutation rate can also be controlled over a wide range, 0.05% - 0.7%. The high-fidelity version has been successfully used to identify functional mutations in a laboratory-evolved thermostable subtilisin E variant (Chapter 5).

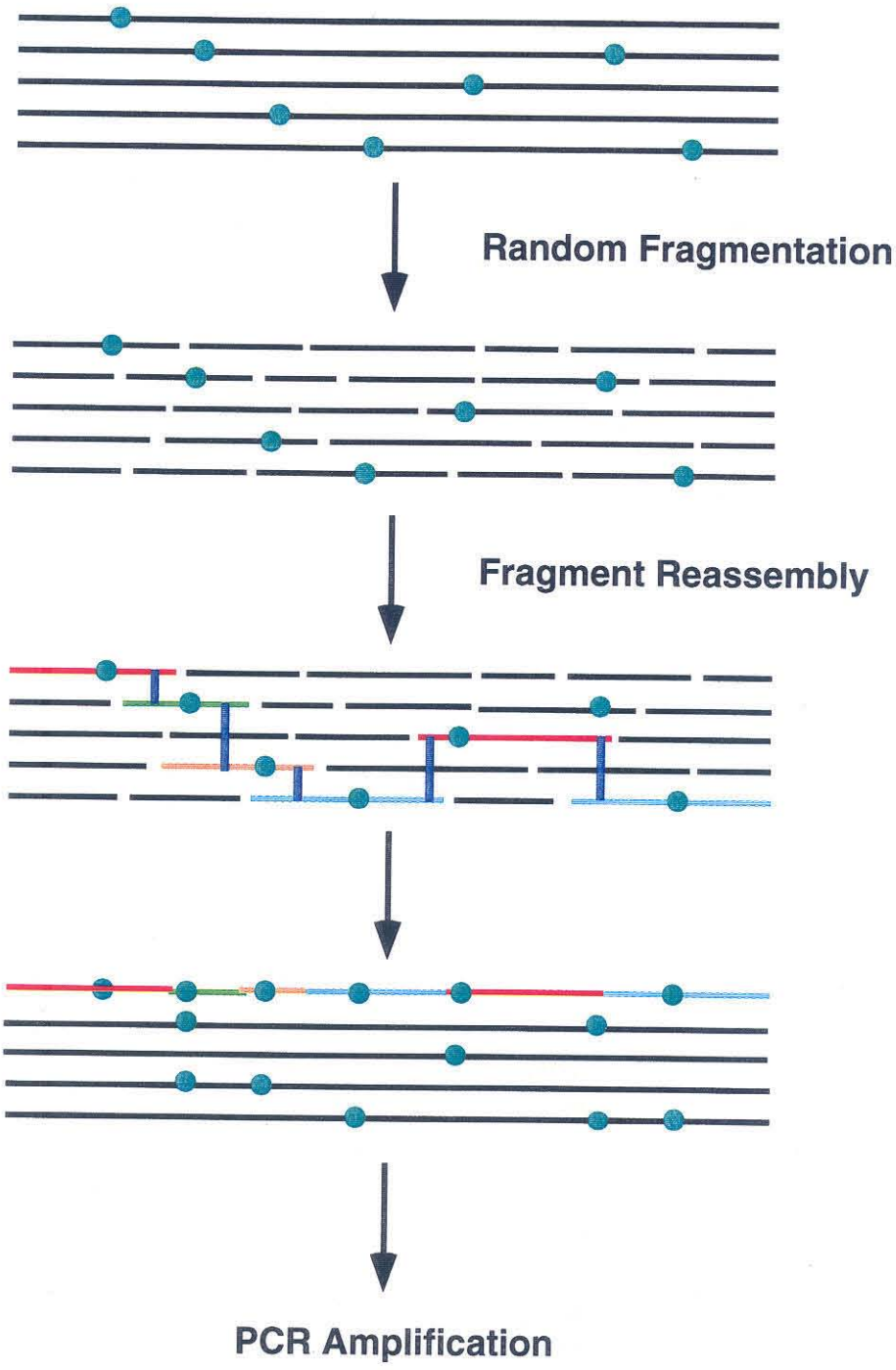
Other limitations of DNA shuffling include the following: (1) The non-random DNA fragmentation associated with DNase I and other endonucleases may introduce bias and limit the recombination diversity. (2) DNA shuffling is limited to recombination of double-stranded polynucleotides and cannot be used on single-stranded templates. (3) DNA shuffling is somewhat labor-intensive, requiring significant quantities of DNA and several DNA purification steps. The first two issues can be addressed by using extension of random sequence primers to generate the pool of short DNA fragments for reassembly instead of using digestion by DNase I [14]. Furthermore, all three issues can be addressed by using a conceptually novel and technically simple approach called staggered extension process or StEP [6]. Instead of reassembling the recombined genes from a pool of fragments generated by DNase I fragmentation of parent genes [7,8] or random-priming synthesis from parent templates [14], full-length, recombined genes are prepared directly in the presence of the template(s). This process mimics nature's own design for retroviruses. Retroviruses are a family of RNA viruses that replicate through a DNA intermediate. A unique feature of retroviruses is that two retroviral RNA molecules are packaged into one virion. During the process of reverse transcription, reverse transcriptases (RTs) frequently switch templates, which results in a high rate of homologous recombination. In addition, RT is error-prone which results in point mutations. This mechanism greatly contributes to the high variation and evolutionary potential of the retrovirus populations [15,16,17].

All the above *in vitro* recombination methods require no knowledge of the template sequence(s). However, in some cases, if structural and functional information for the template sequences is available, the important determinants in the linear sequences (segments of sequences) for a particular feature are known, recombination can be conducted in a more defined fashion. Such considerations have resulted in the development of another recombination method called defined primer *in vitro* recombination (DPR). Instead of randomly generating recombination cassettes (the short fragments as

mentioned above), this technique uses multiple defined primers to generate recombination cassettes in the presence of the templates by StEP followed by reassembly into full length genes. As expected, recombination occurs more frequently within defined region(s).

In this chapter, I will present four *in vitro* recombination methods: (1) optimized DNA shuffling for high-fidelity *in vitro* recombination, (2) random-priming *in vitro* recombination (RPR), (3) defined-primer *in vitro* recombination (DPR), and (4) staggered extension process (StEP) *in vitro* recombination.

Fig. 3.1. The scheme of DNA shuffling [7]. A pool of homologous genes with different point mutations (green solid circles) is randomly fragmented with DNase I. These short fragments are reassembled into full-length genes by repeated cycles of annealing in the presence of DNA polymerase. Recombination occurs when fragments from one copy of gene prime on another one, causing a template switch (crossovers). A recombinant gene containing five crossovers (blue lines) is shown. Fragments from different copy of genes are coded in color. The products are amplified by a conventional PCR, followed by cloning into proper expression vector.


Random Fragmentation
Fragment Reassembly
PCR Amplification

References

1. Wurgler, F. E. (1992) Recombination and gene conversion. *Mutation Research* **284**, 3-14.

2. Stahl, F. W. (1987) Genetic recombination. *Sci. Am.* **256**, 90-101.

3. Crow, J. F. (1988) The importance of recombination. In *The Evolution of Sex: An Examination of Current Ideas.* (Sinauer Associates, Inc., Sunderland, Mass.) 56-73.

4. Maynard Smith, J. (1988) The evolution of recombination. In *The Evolution of Sex: An Examination of Current Ideas.* (Sinauer Associates, Inc., Sunderland, Mass.) 56-73.

5. Muller, H. J. (1932) Some genetic aspects of sex. *Amer. Nature* **66**, 118-138.

6. Zhao, H., Giver, L., Shao, Z., Affholter, J. A. and Arnold, F. H. (1998) Molecular evolution by staggered extension process (StEP) *in vitro* recombination. *Nature Biotechnol.* **16**, 258-262

7. Stemmer, W. P. C. (1994) DNA shuffling by random fragmentation and reassembly: *in vitro* recombination. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 10747-10751.

8. Stemmer, W. P. C. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389-391.

9. Zhang, J. H., Dawes, G. and Stemmer, W. P. C. (1997) Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 4504-4509.

10. Crameri, A., Whitehorn, E., Tate, E. and Stemmer, W. P. C. (1996) Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nature Biotechnology* **14**, 315-319.

11. Moore, J. C., Jin, H. M., Kuchner, O. and Arnold, F. H. (1997) Strategies for the *in vitro* evolution of protein function: enzyme evolution by random recombination of improved sequences. *J. Mol. Biol.* **272**, 336-347.

12. Crameri, A., Dawes, G., Rodriguez, E., Silver, S., and Stemmer, W. P. C. (1996) Molecular evolution of an arsenate detoxification plasmid by DNA shuffling. *Nature Biotechnol.* **15**, 436-438.

13. Zhao, H. and Arnold, F. H. (1997) Optimization of DNA shuffling for high fidelity recombination. *Nucleic Acids Res.* **25**, 1307-1308.

14. Shao, Z., Zhao, H., Giver, L. and Arnold, F. H. (1998) Random-priming *in vitro* recombination: an effective tool for directed evolution. *Nucleic Acids Res.* **26**, 681-683.

15. Hu, W.-S., Bowman, E.H., Delviks, K.A., and Pathak, V.K. (1997) Homologous recombination occurs in a distinct retroviral subpopulation and exhibits high negative interference. *J. Virol.* **71**, 6028-6036.

16. Hu, W.-S. and Temin, H.M. (1990) Retroviral recombination and reverse transcription. *Science* **250**, 1227-1233.

17. Hu, W.-S. and Temin, H.M. (1990) Genetic consequences of packaging 2 RNA genomes in one retroviral particle - pseudodiploidy and high rate of genetic recombination. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 1556-1660.

Technique 1

Optimization of DNA Shuffling for High Fidelity Recombination

(Huimin Zhao and Frances H. Arnold)

(appeared in *Nucleic Acids Research*, 1997, **25**, 1307-1308)

Nucleic Acids Research, 1997, Vol. 25, No. 6 1307–1308

Optimization of DNA shuffling for high fidelity recombination

Huimin Zhao and Frances H. Arnold*

Division of Chemistry and Chemical Engineering 210-41, California Institute of Technology, Pasadena, CA 91125, USA

Received December 19, 1996; Accepted January 27, 1997

ABSTRACT

A convenient 'DNA shuffling' protocol for random recombination of homologous genes *in vitro* with a very low rate of associated point mutagenesis (0.05%) is described. In addition, the mutagenesis rate can be controlled over a wide range by the inclusion of Mn^{2+} or Mg^{2+} during DNase I digestion, by choice of DNA polymerase used during gene reassembly as well as how the genes are prepared for shuffling (PCR amplification versus restriction enzyme digestion of plasmid DNA). These protocols should be useful for *in vitro* protein evolution, for DNA based computing and for structure–function studies of evolutionarily related genes.

The method of DNA shuffling, or 'sexual PCR', is used to recombine homologous DNA sequences during *in vitro* molecular evolution (1,2). While randomly recombining the DNA sequences, the technique also introduces new point mutations at a relatively high rate (0.7%; 3). Though these point mutations may provide useful diversity for some *in vitro* evolution applications, they are problematic for others, especially when the mutation rate is this high. Much lower mutagenesis rates are desired, for example, during the *in vitro* evolution of long genes or whole operons (4), during recombination of beneficial mutations already identified previously (5), for DNA-based computing, or when the method is used to differentiate adaptive from neutral or deleterious mutations in evolutionarily-related sequences (6).

In order to optimize the DNA shuffling technique with regard to fidelity, we have attempted to minimize the number of point mutations introduced in each step. Here we describe a convenient DNA shuffling protocol which randomly recombines genes with a very low rate of associated point mutagenesis. In addition, we show how the mutagenesis rate associated with DNA shuffling can be controlled over a practically useful range by appropriate changes in the protocol.

DNA shuffling consists of four steps: (i) preparation of genes to be shuffled, (ii) fragmentation with DNase I, (iii) reassembly by thermocycling in the presence of a DNA polymerase, and (iv) amplification of reassembled products by a conventional PCR. Point mutations may be generated during each of these steps. Lorimer and Pastan reported that use of Mn^{2+} instead of Mg^{2+} during the DNase I fragmentation step improves the fidelity of DNA shuffling ~3-fold (7). Our protocols include this improvement. Conventional PCR with *Taq* polymerase is usually used to prepare the genes to be shuffled (step 1) as well as to amplify the reassembled products (step 4) (1,7). However, the fidelity of *Taq* polymerase is the lowest among commercially available thermostable DNA polymerases. In fact, between 33 and 98% of the amplification products will contain mutation(s) when a 1 kb fragment is amplified for 20 effective cycles (one million-fold amplification) using *Taq* polymerase. Extensive studies have revealed that fidelity during PCR depends on the specific conditions and DNA polymerase used (8–10). Avoiding PCR where possible and using higher fidelity DNA polymerase during amplification and reassembly should further reduce the point mutagenesis rate associated with DNA shuffling.

Wild-type subtilisin E and its thermostable mutant 1E2A genes were randomly recombined by DNA shuffling using the conditions summarized in Table 1. Gene 1E2A, obtained by directed evolution of wild-type subtilisin E, differs by 10 base changes (6; Fig. 1). The enzyme encoded by this gene retains wild-type activity. The ~1 kb fragments encoding mature subtilisin E from residue −15 (from the prosequence) to the C-terminus (including 113 nt after the stop codon) were obtained by restriction digestion of plasmid DNA and purified from a 0.8% agarose gel using the QIAEX II gel extraction kit (QIAGEN, Chatsworth, CA). After DNA shuffling, the gene library was amplified in *E.coli* HB101 and transferred into *B.subtilis* DB428 competent cells for expression and screening, as described elsewhere (6). Screening for protease activity was carried out at 37 °C in 96-well plates using suc-Ala-Ala-Pro-Phe-*p*-nitroanilide (0.2 mM) as substrate, as described previously (11). All PCR reactions were done on a MJ Research (Watertown, MA) PTC200 thermocycler. Sequencing was done on an ABI 373 DNA Sequencing System using the Dye Terminator Cycle Sequencing kit (Perkin-Elmer, Branchburg, NJ).

The frequency of clones exhibiting ≥ 10% of wild-type subtilisin activity was used as a convenient index of the fidelity of DNA shuffling under different conditions. Table 1 shows that only 20% of the clones retained subtilisin activity after DNA shuffling using the protocol of Lorimer and Pastan (conditions A) (7). Replacing *Taq* with *Pwo* polymerase in the reassembly step (conditions B) increased this to 46%.

Improvements in fidelity could also be achieved by preparing the genes directly from plasmids by restriction enzyme digestion (conditions C–E). This, together with PCR amplification using a 1:1 mixture of *Taq* and *Pfu*, yielded 58% active clones (conditions C). Finally, the use of proofreading enzymes (*Pfu* or *Pwo*) in the reassembly step dramatically increased the frequency of active clones to as high as 95% (conditions E). These data are consistent with the known fidelities of the various DNA polymerases.

To determine the mutagenesis rate of our optimized DNA shuffling protocol (conditions E) as well as to analyze the efficiency

*To whom correspondence should be addressed. Tel: +1 818 395 4162; Fax: +1 818 568 8743; Email: frances@cheme.caltech.edu

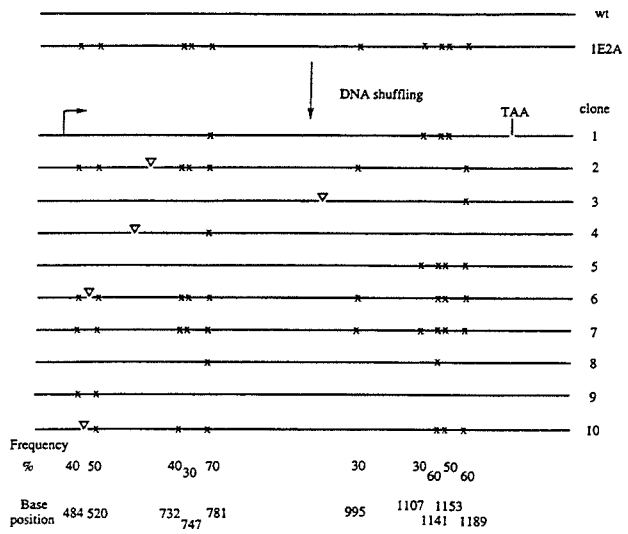


Figure 1. Results of sequencing genes from 10 randomly-selected unscreened clones from DNA shuffled library. Lines represent 986 bp of subtilisin E gene including 45 nt of its prosequence, the entire mature sequence and 113 nt after the stop codon. Crosses indicate positions of mutations from 1E2A, while triangles indicate positions of new point mutations introduced during the DNA shuffling procedure.

Table 1. Frequency of active clones obtained after DNA shuffling under different conditions

Step / Condition	1 Gene preparation	2 DNase I digestion[a]	3 Reassembly	4 PCR amplification	Active clones[b] %
A	PCR / *Taq*	Mn^{2+}	*Taq*	*Taq*	20
B	PCR / *Taq*	Mn^{2+}	*Pwo*	*Taq*	46
C	plasmid digestion	Mn^{2+}	*Taq*	*Taq:Pfu* (1:1)	58
D	plasmid digestion	Mn^{2+}	*Pwo*	*Taq:Pfu* (1:1)	87
E	plasmid digestion	Mn^{2+}	*Pfu*	*Taq:Pfu* (1:1)	95

[a]As suggested in ref. 7.
[b]Clones exhibiting >10% wild-type subtilisin E activity.

of recombination, 10 unscreened clones were selected randomly for sequencing. The plasmids from these clones were purified and the inserts were sequenced in both directions. Comparison of these 10 sequences (Fig. 1) shows that all except clone 7 result from different recombination events. (Clone 7 is the intact 1E2A parent sequence.) The frequency of occurrence of a particular point mutation from parent 1E2A in the shuffled genes ranged from 30 to 70%, fluctuating around the expected value of 50%. No pair of point mutations was found to be inseparable, even those as little as 12 bp apart. However, a certain degree of linkage of nearby mutations is apparent. In 9860 total bases sequenced, no insertion or deletions (frameshifts) were found. The overall mutagenic rate for this protocol is only 0.05% (5/9860), which is the lowest rate thus far reported for DNA shuffling.

High-fidelity DNA shuffling protocol

Preparation of genes to be shuffled. About 10 µg plasmids containing wild-type subtilisin E and the thermostable mutant 1E2A gene were digested at 37°C for 1 h with *Nde*I and *Bam*HI (30 U each) in 50 µl 1× buffer B (Boehringer Mannheim, Indianapolis, IN). Fragments of ~1 kb were purified from 0.8% preparative agarose gels using QIAEX II gel extraction kit (QIAGEN, Chatsworth, CA). The DNA fragments were dissolved in 10 mM Tris–HCl (pH 7.4). The DNA concentrations were estimated, and the fragments were mixed 1:1 for a total of ~2 µg.

DNase I digestion in the presence of Mn^{2+}. The mixture was diluted to 45 µl in 10 mM Tris–HCl (pH 7.4) and 5 µl 10× digestion buffer (500 mM Tris–HCl pH 7.4, 100 mM $MnCl_2$) was added. This mixture was equilibrated at 15°C for 5 min on a thermocycler before 0.30 U DNase I (10 U/µl; Boehringer Mannheim) was added. The digestion was done at 15°C and terminated after 2 min by heating at 90°C for 10 min. That the fragments were <50 bp was confirmed on a 2% agarose gel before purification on a Centri-Sep column (Princeton Separations, Inc., Adelphia, NJ).

Fragment reassembly. Spin-column purified fragments (10 µl) were added to 10 µl 2× PCR premix [5-fold diluted cloned *Pfu* buffer, 0.4 mM each dNTP, 0.06 U/µl cloned *Pfu* polymerase (Stratagene, La Jolla, CA)]. The reaction mixture was overlaid with 30 µl mineral oil. PCR program: 3 min 96°C followed by 40 cycles of 1 min 94°C, 1 min 55°C, 1 min + 5 s/cycle 72°C, followed by 7 min at 72°C.

PCR amplification of reassembled products. One microlitre of this reaction was used as template in a 25-cycle PCR reaction. PCR conditions (100 µl final volume): 30 µmol each primer, 1× *Taq* buffer, 0.2 mM each dNTP and 2.5 U *Taq/Pfu* (1:1) mixture. PCR program: 2 min 96°C, 10 cycles of 30 s 94°C, 30 s 55°C, 45 s 72°C, followed by another 14 cycles of 30 s 94°C, 30 s 55°C, 45 s + 20 s/cycle 72°C, and finally 7 min 72°C. This program gives a single band at the correct size.

We have developed a high-fidelity DNA shuffling protocol with mutagenic rate of only 0.05%. The mutagenic rate can be controlled over a wide range, 0.05–0.7%, by the inclusion of Mn^{2+} or Mg^{2+}, by the choice of DNA polymerase and/or using restriction enzyme digestion, as indicated in Table 1. Further improvements in fidelity could be achieved by reducing the PCR cycle number in the reassembly step and using *Pfu* only in the final amplification step. The current high fidelity protocol has been used successfully to distinguish the functional and non-functional mutations in an evolved gene by random recombination and sequence analysis of genes conferring the evolved phenotype (6). This approach should prove extremely useful for structure–function studies of homologous genes.

REFERENCES

1 Stemmer,W.P.C. (1994) *Nature* **370**, 389–391.
2 Crameri,A., Whitehorn,E.A., Tate,E. and Stemmer,W.P.C. (1995) *Nature Biotech.* **14**, 315–319.
3 Stemmer,W.P.C. (1994) *Proc. Natl Acad. Sci. USA* **91**, 10747–10751.
4 Moore,J.C. and Arnold,F.H. (1996) *Nature Biotech.* **14**, 458–467.
5 Moore,J.C. and Arnold,F.H. *(1997) Adv. Biochem. Engng*, **58**, in press.
6 Zhao,H. and Arnold,F.H., submitted.
7 Lorimer,I.A.J. and Pastan,I. (1995) *Nucleic Acids Res.* **23**, 3067–3068.
8 Kunkel,T.A., Loeb,L.A. and Goodman,M.F. (1984) *J. Biol. Chem.* **259**, 1539–1545.
9 Keohavong,P. and Thilly,W.G. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 9253–9257.
10 Eckert,K.A. and Kunkel,T.A. (1990) *Nucleic Acids Res.* **18**, 3739–3744.
11 Chen,K. and Arnold,F.H. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 5618–5622.

Technique 2

Random-priming *in vitro* Recombination: an Effective Tool for Directed Evolution

(Zhixin Shao, Huimin Zhao, Lori Giver and Frances H. Arnold)

(*Nucleic Acids Research*, 1998, **26**, 681-683)

Nucleic Acids Research, 1998, Vol. 26, No. 2 681–683

Random-priming *in vitro* recombination: an effective tool for directed evolution

Zhixin Shao, Huimin Zhao, Lori Giver and Frances H. Arnold*

Division of Chemistry and Chemical Engineering 210-41, California Institute of Technology, Pasadena, CA 91125, USA

Received September 15, 1997; Revised and Accepted November 21, 1997

ABSTRACT

A simple and efficient method for *in vitro* mutagenesis and recombination of polynucleotide sequences is reported. The method involves priming template polynucleotide(s) with random-sequence primers and extending to generate a pool of short DNA fragments which contain a controllable level of point mutations. The fragments are reassembled during cycles of denaturation, annealing and further enzyme-catalyzed DNA polymerization to produce a library of full-length sequences. Screening or selecting the expressed gene products leads to new variants with improved functions, as demonstrated by the recombination of genes encoding different thermostable subtilisins in order to obtain enzymes more stable than either parent.

Directed evolution (1) has proven particularly effective for exploring and optimizing enzyme functions (2–5). Including recombination in the creation of gene libraries allows the rapid accumulation of beneficial mutations and removal of deleterious ones (2–4). The sequence diversity upon which recombination acts can be obtained by mutating the wild-type gene or, alternatively, by using homologous genes isolated from nature or obtained in parallel laboratory evolution experiments (e.g. genes optimized for different properties). Stemmer (2,3) recently introduced the 'DNA shuffling' technique for random *in vitro* mutagenesis and recombination. Here we present a new method, random-priming recombination (RPR), as an effective alternative to DNA shuffling. As shown in Figure 1, random sequence primers are used to generate a large number of short DNA fragments complementary to different sections of the template sequence(s). Due to base misincorporation and mispriming, these short DNA fragments also contain a low level of point mutations. The short DNA fragments can prime one another based on homology, and be recombined and reassembled into full-length genes by repeated thermocycling in the presence of thermostable DNA polymerase. These sequences can be further amplified by conventional PCR and cloned into a vector for expression, followed by screening or selection. RPR and screening or selection can be repeated over multiple cycles in order to evolve the desired properties.

Compared to DNA shuffling (2,3), the RPR technique has several advantages: (i) RPR can use single-stranded polynucleotide templates without an intermediate step of synthesizing the whole second strand. Potential mutations and/or crossovers can be introduced at the DNA level from single- or double-stranded DNA template by using DNA polymerases, or directly from mRNA by using RNA-dependent DNA polymerases. (ii) DNA shuffling requires fragmentation of the double-stranded DNA template (generally done with DNase I). The DNase I must be removed completely before the fragments can be reassembled into full length sequences. Gene reassembly is generally easier with the RPR technique, which employs random priming synthesis to obtain the short DNA fragments. Furthermore, since DNase I hydrolyzes double-stranded DNA preferentially at sites adjacent to pyrimidine nucleotides (6), its use in template digestion may introduce a sequence bias into the recombination. (iii) The synthetic random primers are uniform in their length and lack sequence bias. The sequence heterogeneity allows them to form hybrids with the template DNA strands at many positions, so that, at least in principle, every nucleotide of the template should be copied or mutated at a similar frequency during extension. The random distribution of the short, nascent DNA fragments along the template(s) and the random distribution of point mutations within each nascent DNA fragment should guarantee the randomness of crossovers and mutations in the full length progeny genes. (iv) The random-priming DNA synthesis is independent of the length of the DNA template. DNA fragments as small as 200 bases can be primed equally well as large DNA molecules such as linearized plasmids or λ DNA (7). This offers unique potential for engineering small peptides. Hodgson and Fisk (8) found that the average size of synthesized single-stranded DNA is an inverse function of primer concentration. Based on this guideline, proper conditions for random-priming synthesis can be easily set for a given gene. (v) Since the parent polynucleotide serves solely as the template for the synthesis of nascent, single-stranded DNA, 10–20 times less parent DNA is needed as compared to DNA shuffling.

We demonstrated the RPR method by the mutagenesis and recombination of genes RC1 and RC2 encoding thermostable *Bacillus subtilis* subtilisin E variants (Fig. 2). The mutations at base positions 1107 in RC1 and 995 in RC2, giving rise to amino acid substitutions Asn_{218}→Ser (N218S) and Asn_{181}→Asp (N181D), lead to improvements in subtilisin E thermostability; the remaining mutations, both synonymous and non-synonymous, have no detectable effects on thermostability. At 65°C, the single variants

*To whom correspondence should be addressed. Tel: +1 626 395 4162; Fax: +1 626 568 8743; Email: frances@cheme.caltech.edu

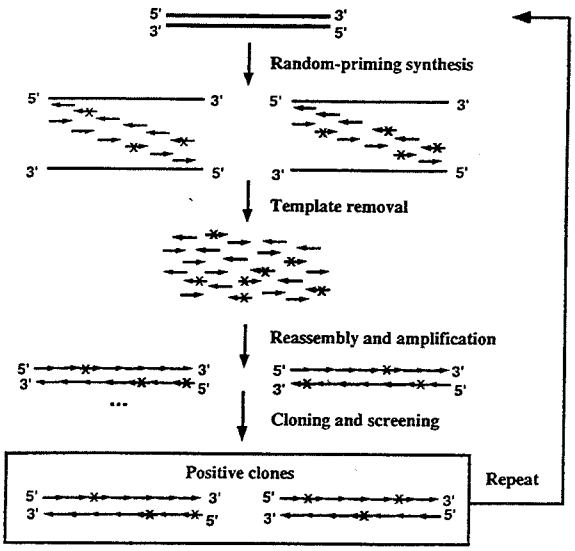


Figure 1. Random-priming *in vitro* recombination (RPR). (i) Synthesis of short single-stranded DNA fragments from random-sequence primers. X's indicate newly introduced mutations. (ii) Template removal. (iii) Reassembly and amplification. (iv) Cloning and screening (or selection). (v) Cycle is repeated until desired functional improvement is achieved.

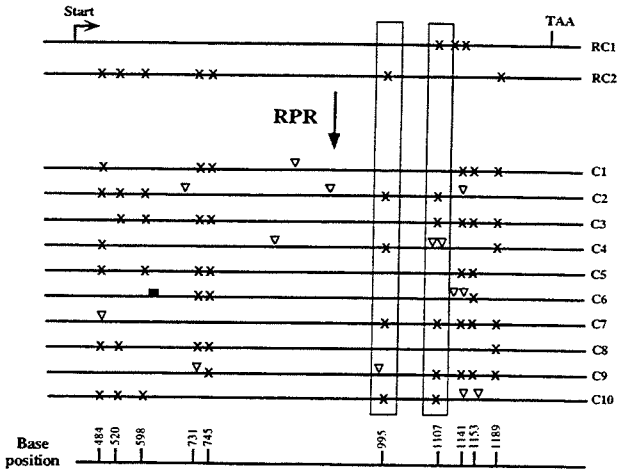


Figure 2. Sequence analysis of 10 randomly-picked clones from (unscreened) *B.subtilis* subtilisin E RPR library. Each line represents 986-bp subtilisin E gene including the 45 bp prosequence, the entire mature sequence and 113 bp after the stop codon. Triangles (▽) indicate positions of new point mutations introduced during RPR. Box (■) in C6 indicates the short stretch containing four mutations between two repeats (5′-CGGTACG-3′). Positions and frequency of mutations in the 10 clones are indicated.

N181D and N218S have ~3-fold and 2-fold longer half-lives, respectively, than wild-type subtilisin E, and variants containing both mutations have half-lives that are 8-fold longer (9). The half-lives of a population of subtilisin E variants can therefore be used to estimate the recombination efficiency. In particular, random recombination between these two mutations (in the absence of point mutations affecting thermostability) should generate a library in which 25% of the population exhibits the thermostability of the double mutant. Similarly, 25% of the population should exhibit wild-type like stability, as N181D and N218S are eliminated at equal frequency.

For random-priming synthesis using the Klenow fragment, 200 ng (0.7 pmol) of genes RC1 and RC2 (1:1) was mixed with 13.25 µg (6.7 nmol) of $dp(N)_6$ random primers (Pharmacia Biotech Inc., Piscataway, NJ). After denaturation at 100°C for 5 min, 10 µl of 10 × reaction buffer (900 mM HEPES, pH 6.6, 0.1 M $MgCl_2$, 20 mM dithiothreitol and 5 mM each dATP, dCTP, dGTP and dTTP) was added, and the total volume of the reaction mixture was brought to 95 µl with H_2O. Ten units (in 5 µl) of the Klenow fragment of *Escherichia coli* DNA polymerase I (Boehringer Mannheim, Indianapolis, IN) was added. The synthesis reaction was carried out at 22°C for 3 h and terminated by cooling the sample to 0°C. The high primer concentration facilitates production of shorter fragments, presumably by blocking extension. Under the reaction conditions described here (10 000-fold excess primer over template) the strand displacement activity of the Klenow is much reduced, and the random priming products are ~50–500 bases as determined by alkaline agarose gel electrophoresis. After the addition of 100 µl of ice-cold H_2O to the reaction mixture, the random primed products were purified by passing the whole reaction mixture through a Microcon-100 filter (Amicon, Inc., Beverly, MA) to remove the template, proteins and large nascent DNA fragments, followed by a Microcon-10 filter to remove the primers and fragments <30 bases. The retentate fraction (~65 µl) recovered from the Microcon-10 step was buffer-exchanged against PCR reaction buffer. Of this fraction, 10 µl was used for the whole gene reassembly as described (9). Subtilisin E mutant gene amplification, cloning, expression and thermostability screening and analysis were carried out as described (9).

We measured the thermoinactivation rate at 65°C of subtilisins E from ~400 randomly-picked clones. Approximately 26% of the active clones exhibited thermostability comparable to the N181D + N218S double mutant, indicating that the RPR had efficiently recombined the N181D mutation from RC2 and the N218S mutation from RC1. Sequence analysis of the clone exhibiting the highest thermostability showed that mutations N181D and N218S were both present (data not shown). To further characterize the recombination efficiency and point mutagenesis associated with the RPR process, 10 clones from the mutant library were selected at random and sequenced. As summarized in Figure 2, all 10 clones were novel recombinants, different from the parent genes. The frequency of occurrence of any particular point mutation from parent RC1 or RC2 in the recombined genes (C1–C10) ranged from 40 to 70%, fluctuating around the expected value of 50%. The minimum number of crossovers ranged from two (C4) to six (C1). Figure 2 also shows that all 10 mutations in the parent genes can be recombined, including mutations only 12 bp apart (positions 1141 and 1153 in C6).

As shown in Figure 2, 18 new point mutations were found in the 10 random clones. This error rate of ~0.18% (1–2 new point mutations per gene) is close to ideal for directed evolution (4). The new mutations appear randomly distributed along the gene, except for a cluster of four mutations within a very short stretch of DNA

in clone C6 (see Fig. 2). The direction of mutation, however, is clearly non-random: A changes more often to G than to T or C. Transitions (in particular T→C and A→G) occur more often than transversions. One G→C, one C→G and one C→A transversion were found within the 10 sequenced clones. These mutations were generated much more rarely during the error-prone PCR mutagenesis of subtilisin BPN′ (10). This result suggests that the RPR technique may allow access to a greater range of amino acid substitutions than PCR-based point mutagenesis.

RPR is a very flexible and easy to implement technique for generating mutant libraries for directed evolution. Other polymerases with different fidelities, including bacteriophage T4 DNA polymerase, T7 Sequenase® version 2.0 DNA polymerase, the Stoffel fragment of *Taq* polymerase and *Pfu* polymerase, have been used successfully for the random priming DNA fragment synthesis. The length and concentration of random primer, as well as the time, temperature and other reaction conditions can also be manipulated in order to achieve the desired mutagenic rate and recombination frequency.

ACKNOWLEDGMENTS

This research is supported by Eli Lilly and Co., the US Office of Naval Research (N0014-96-1-340 and N00014-97-1-0433), the US Department of Energy's program in Biological and Chemical Technologies Research within the Office of Industrial Technologies, Energy Efficiency and Renewables. Z.S. is grateful to Drs Peggy Arps and Kentaro Miyazaki for helpful discussions and to Michaeleen Callahan for excellent technical assistance.

REFERENCES

1 Kuchner, O. and Arnold, F.H. (1997) *Trends Biotech.*, **15**, 523–530.
2 Stemmer, W.P.C. (1994) *Nature*, **370**, 389–391.
3 Stemmer, W.P.C. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 10747–10751.
4 Moore, J.C., Jin, H.-M., Kuchner, O. and Arnold, F.H. (1997) *J. Mol. Biol.*, **272**, 336–347.
5 Moore, J.C. and Arnold, F.H. (1996) *Nature Biotech.*, **14**, 458–467.
6 Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*. 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
7 Feinberg, A.P. and Vogelstein, B. (1983) *Anal. Biochem.*, **132**, 6–13.
8 Hodgson, C.P. and Fisk, R.Z. (1987) *Nucleic Acids Res.*, **15**, 6296.
9 Zhao, H. and Arnold, F.H. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 7997–8000.
10 Shafikhani, S., Siegel, R.A., Ferrari, E. and Schellenberger, V. (1997) *Biotechniques*, **23**, 304–310.

RELATED PAPERS RECENTLY PUBLISHED IN *NUCLEIC ACIDS RESEARCH*

He,M. and Taussig,M.J. (1997) Antibody-ribosome-mRNA (ARM) complexes as efficient selection particles for *in vitro* display and evolution of antibody combining sites. *Nucleic Acids Res.* **25**, 5132–5134.

Zhao,H. and Arnold,F.H. (1997) Optimization of DNA shuffling for high fidelity recombination. *Nucleic Acids Res.* **25**, 1307–1308.

Morl,M. and Schmelzer,C. (1990) Group II intron RNA-catalyzed recombination of RNA *in vitro*. *Nucleic Acids Res.* **18**, 6545–6551.

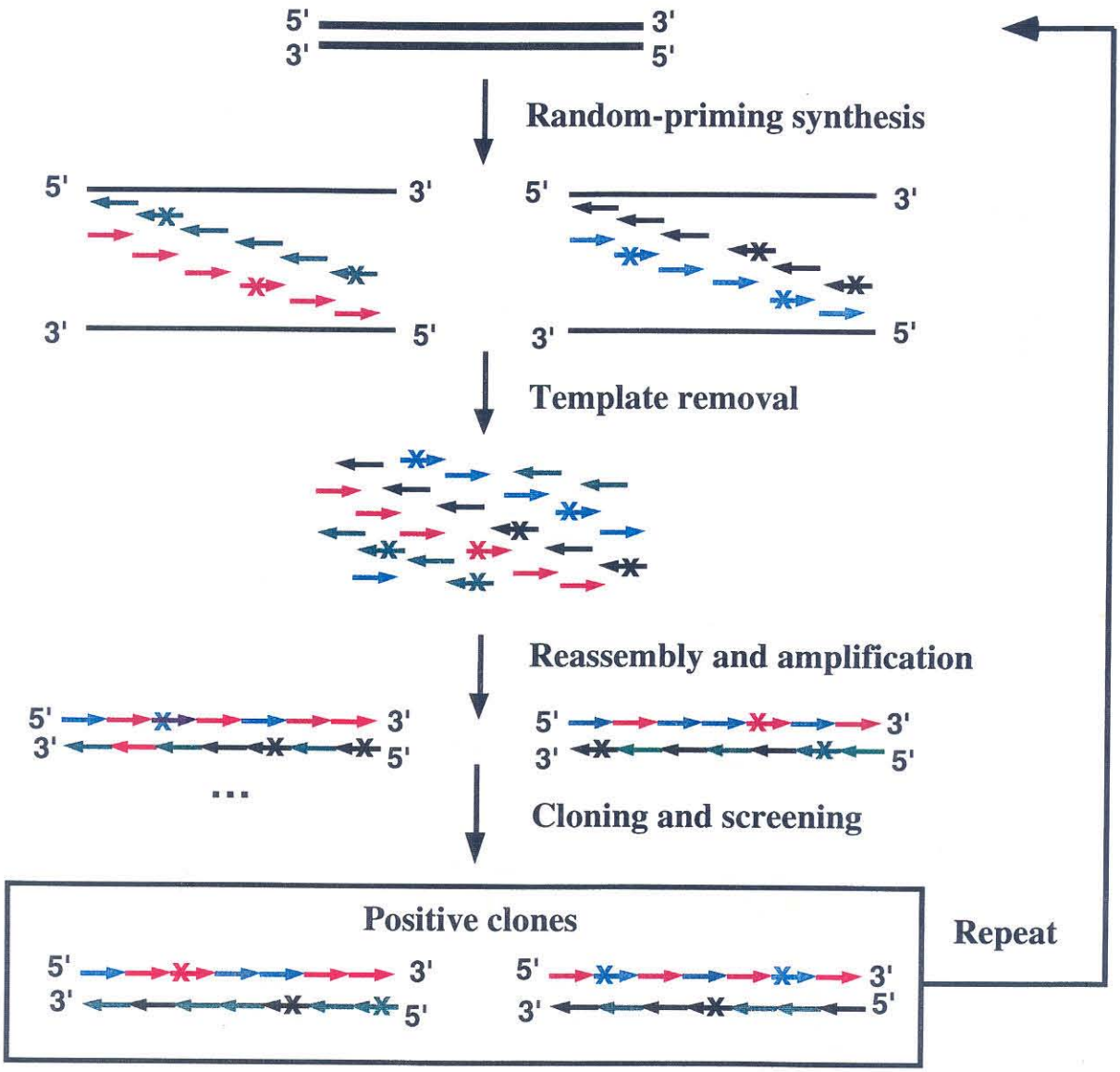


Fig. 1. Random-priming *in vitro* recombination (shown in color).

Technique 3

'Defined Primer' Based *in vitro* Recombination for Directed Evolution

Directed evolution, inspired by natural evolution, has emerged as a powerful tool for engineering new proteins as well as addressing fundamental questions about protein structure, function and evolution in recent years [1,2]. Various protein properties have been targeted by directed evolution, such as stability, catalytic activity, substrate specificity, activity in non-natural environments, expression level in a heterologous host, and others [2]. This approach involves the generation and selection or screening of a library of mutated genes which has sufficient diversity for a gene encoding a protein with altered or enhanced function to be present therein. The sequence diversity can be generated by random point mutagenesis and/or random recombination of a pool of homologous genes. In contrast to random point mutagenesis, random recombination allows the rapid accumulation of beneficial mutations and removal of deleterious ones.

Several *in vivo* and *in vitro* recombination methods have been developed [3,4,5,6,7]. Aiming at a high degree of randomness in the recombination process, these methods were designed on purpose not to use any structural and functional information on the genes. In many scenarios, however, this knowledge should help molecular evolutionists identify the important region(s) responsible for a particular protein feature. Thus it may be more efficient to evolve a protein feature by biasing the recombination events such that recombination appears most often within these regions. To this end, a new technique, we call defined primer *in vitro* recombination (DPR), has been developed. This approach uses multiple internal defined primers to generate recombination cassettes by staggered extension process (StEP) [7] followed by reassembly into full length gene products. After removal of templates, these products were further amplified by a conventional PCR and cloned into a proper vector for expression, followed by screening or selection. DPR and screening or selection can be repeated over multiple cycles in order to evolve the desired properties. In this paper, we demonstrated the DPR method by the mutagenesis and recombination of two genes encoding thermostable *B. subtilis* subtilisin E variants, each of which carries a single thermostable mutation along with several other

neutral mutations. In addition, we also showed that specific mutations can be introduced into the recombined sequences by using appropriate defined primer sequence(s) containing the desired mutation(s).

Genes RC1 and RC2 encoding thermostable subtilisin E variants have been used as a model system to develop DPR method as well as other new *in vitro* recombination methods [6,7]. Two phenotypic markers (N218S and N181D) and as many as eight non-phenotypic markers (neutral) can be used to estimate the recombination efficiency (Fig. 1). Recombination between the two phenotypic markers only 113 bp apart will generate much more thermostable variants, provided no new deleterious mutations are introduced into these variants during the recombination process [6,7,8]. Four internal defined primers have been designed (Fig. 1). In particular, primer P50F contains a mutation (A→T at base position 598) which eliminates a *Hind*III restriction site and simultaneously adds a new unique *Nhe*I site. This primer is used to demonstrate that specific mutations can also be introduced in the population of recombined sequences by specific design of the defined primer. Gene RC2 also contains a mutation A→G at the same base position, which eliminates the *Hind*III site as well. Thus restriction analysis (cutting by *Nhe*I and *Hind*III) of random clones sampled from the recombined library will indicate the efficiency of the recombination and of the introduction of a specific mutation via the mutagenic primer. Sequence analysis of randomly-picked (unscreened) clones provides further information on the recombination and mutagenesis events occurring during DPR.

DPR consists of four steps: (1) Preparation of genes to be recombined. About 10 μg of plasmids containing RC1 and RC2 gene were digested at 37 °C for 1 hour with *Nde*I and *Bam*HI (30 U each) in 50 μl of 1x buffer B (Boehringer Mannheim, Indianapolis, IN). Inserts of ~1 kb were purified from 0.8% preparative agarose gels using QIAEX II gel extraction kit (Qiagen, Chatsworth, CA). The DNA inserts were dissolved in 10 mM Tris-HCl (pH 7.4). The DNA concentrations were estimated, and the inserts were mixed 1:1 for a concentration of 50 ng/μl. (2) Staggered extension process (StEP) and reassembly.

Conditions (100 µl final volume): about 100 ng inserts were used as templates, 50 ng of each of 4 internal primers, 1x *Taq* buffer, 0.2 mM of each dNTP, 1.5 mM $MgCl_2$ and 0.25 U *Taq* polymerase. Program: 7 cycles of 30 sec at 94 °C, 15 sec at 55 °C, followed by another 10 cycles of 30 sec at 94 °C, 15 sec at 55 °C, 5 sec at 72 °C (staggered extension), followed by 53 cycles of 30 sec at 94 °C, 15 sec at 55 °C, 1 min at 72 °C (gene assembly). (3) *Dpn*I digestion of the templates. 1 µl of this reaction was diluted up to 9.5 µl with dH_2O and 0.5 µl of *Dpn*I restriction enzyme was added to digest the DNA templates for 45 min, followed by incubation at 70 °C for 10 min and then this 10 µl was used as template in a 10-cycle PCR reaction. (4) PCR amplification of reassembled products. PCR conditions (100 µl final volume): 30 pmol of each flanking primers P5N (5'-CCGAG CGTTG CATAT GTGGA AG-3', underlined sequence is *Nde*I restriction site) and P3B (5'-CGACT CTAGA GGATC CGATT C-3', underlined sequence is *Bam*HI restriction site), 1x *Taq* buffer, 0.2 mM of each dNTP, 1.5 mM $MgCl_2$ and 2.5 U of *Taq* polymerase. PCR program: 10 cycles of 30 sec at 94 °C, 30 sec at 55 °C, 1 min at 72 °C. This program gave a single band at the correct size. The product was purified and subcloned into pBE3 *E. coli-B. subtilis* shuttle vector. This gene library was amplified in *E. coli* HB101 and transferred into *B. subtilis* DB428 competent cells for expression and screening, as described elsewhere [9]. Thermostability of enzyme variants was determined in the 96-well plate format described previously [10]. All PCR reactions were done on a MJ Research (Watertown, MA) PTC-200 thermocycler. Sequencing was done on an ABI 373 DNA Sequencing System using the Dye Terminator Cycle Sequencing kit (Perkin-Elmer, Branchburg, NJ).

Forty clones were randomly picked from the recombinant library. Their plasmids were isolated and digested with restriction enzymes *Nhe*I and *Bam*HI. In a separate experiment the same 40 plasmids were digested with *Hind*III and *Bam*HI. These reaction products were analyzed by agarose gel electrophoresis. 19 out of 40 clones (~50%) contain the *Hind*III site. 8 out of 40 clones (20%) contain the newly introduced *Nhe*I

restriction site, demonstrating that the mutagenic primer has indeed been able to introduce the specified mutation into the population.

The first ten randomly picked clones were further subjected into sequence analysis, and the results were summarized in Fig. 2. A minimum of six out of ten genes have undergone recombination. Among these six genes, the minimal crossover events (recombination) between genes RC1 and RC2 vary from 1 to 4. All visible crossovers occur within the region defined by the four primers. Mutations outside this region are rarely, if ever, recombined, as shown by the fact that there is no recombination between the two mutations at base positions 484 and 520. These results show that the defined primers can bias recombination so that it appears most often in the portion of sequence defined by the primers (inside the primers). Mutation very close tend to remain together (for example, base substitutions 731 and 745 and base substitutions 1141 and 1153 always remain as a pair). However, the sequence of clone 7 shows that two mutations as close as 34 bases apart can be recombined (base position at 1107 and 1141). In a total, 23 new point mutations were introduced in the ten genes during the recombination process. This overall mutation frequency of 0.23% corresponds to two to three new point mutations per gene, which is the frequency that has been determined optimal for generating mutant libraries for directed enzyme evolution (Chapter 2). The mutation types are listed in Table 1. Mutations are mainly transitions and are evenly distributed along the gene.

Phenotypic analysis of approximately 450 clones were performed as described previously [9]. Approximately 56% of the clones expressed active enzymes, which indicates a mutation rate on the order of two to three mutations per gene (see Chapter 2.4). Approximately 5% of clones showed double mutant (N181D+N218S)-like phenotypes, which is below the expected 25% value for random recombination alone. This is primarily due to the high level of point mutagenesis since even in a random sampling as small as ten clones, two out ten (20%) clones contain both N181D and N218S (as shown in clone 7 and 8 (Fig. 2)).

DPR is a very flexible and easy to implement technique for controlling random recombination events within defined region(s). Polymerases other than *Taq* polymerase may be used. Multiple defined primers or defined primers exhibiting limited randomness may also be used. The extreme extension of this approach is to only use two defined primers flanking the whole length of gene, as reported elsewhere [7].

Table 1. New mutations introduced during the DPR process. Ten randomly picked clones from (unscreened) *B. subtilis* subtilisin E DPR library were sequenced (a total of 9860 bases). The overall mutation rate was ~0.23%.

Transition	Frequency	Transversion	Frequency
G → A	4	A → T	1
A → G	4	A → C	1
C → T	3	C → A	1
T → C	5	C → G	0
		G → C	1
		G → T	0
		T → A	3
		T → G	0

Total: 23

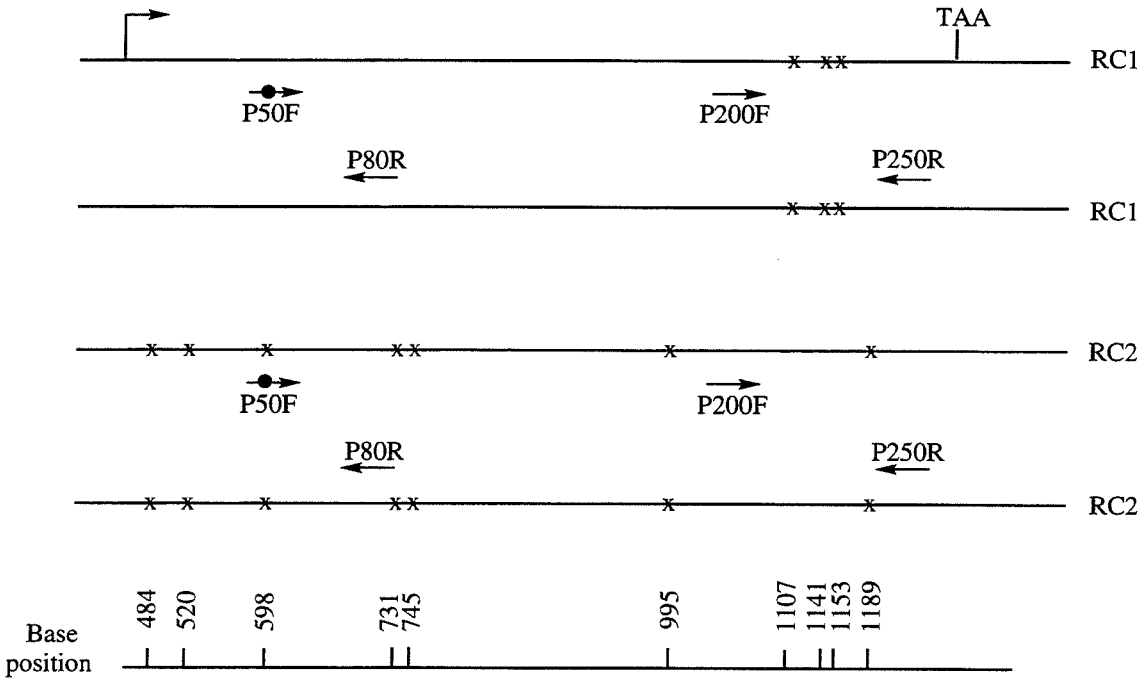
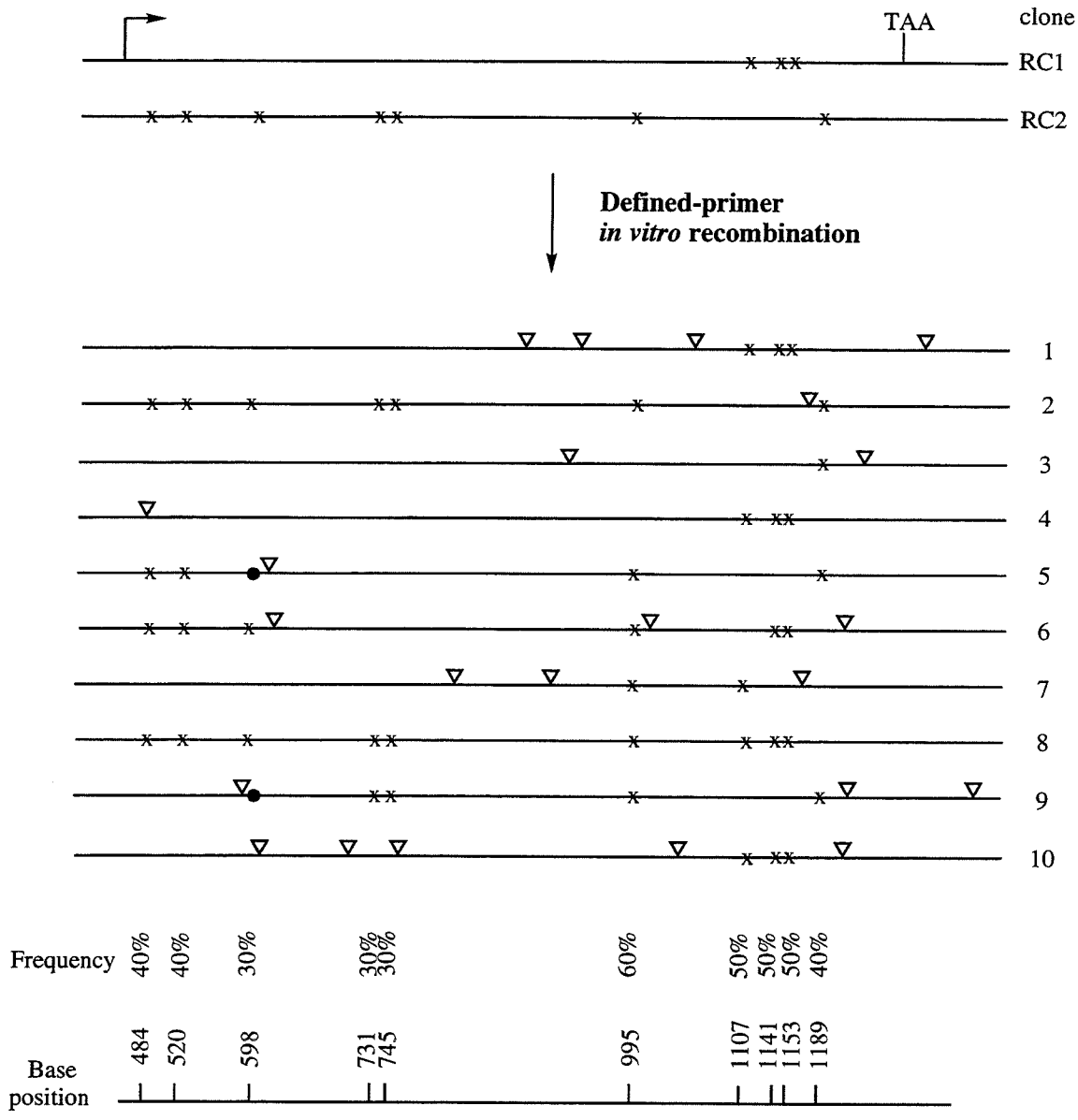


Fig. 1. Four internal primers used as defined primers to recombine genes RC1 and RC2. Two forward primers: P50F: 5'-GGCGG AGCTA GCTTC GTA-3' (mutagenic primer, underlined base is the mutagenized base at position 598) and P200F: 5'-GATGT GATGG CTCCT GGC-3'. Two reverse primers: P80R: 5'-CAGAA CACCG ATTGA GTT-3' and P250R: 5'-AGTGC TTTCT AAACG ATC-3'. Positions of mutation in RC1 and RC2 are indicated.

Fig. 2. Sequence analysis of ten randomly picked clones from (unscreened) *B. subtilis* subtilisin E DPR library. Each line represents 986-bp of subtilisin E gene including 45 nt of its prosequence, the entire mature sequence and 113 nt after the stop codon. Crosses indicate positions of mutations from RC1 and RC2, filled circles indicate the designed specific mutation within primer P50F, and triangles indicate positions of new point mutations introduced during DPR. Positions and frequency of mutations in the ten clones are indicated.


TAA
clone
RC1
RC2
Defined-primer
in vitro recombination
1
2
3
4
5
6
7
8
9
10
Frequency
40%
40%
30%
30%
30%
60%
50%
50%
50%
40%
Base position
484
520
598
731
745
995
1107
1141
1153
1189

REFERENCES

1. Arnold, F.H. (1998) Design by directed evolution. *Accounts of Chemical Research* **31**, 125-131.

2. Kuchner, O. and Arnold, F. H. (1997) Directed evolution of enzyme catalysts. *Trends in Biotechnology* **15**, 523-530.

3. US Patent. No. 5,093,257.

4. Pompon, D. and Nicolas, A. (1989) Protein engineering by cDNA recombination in yeasts: shuffling of mammalian cytochrome P-450. Gene **83**, 15-24.

5. Stemmer, W. P. C. (1994) DNA shuffling by random fragmentation and reassembly - *in vitro* recombination for molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 10747-10751.

6. Shao, Z., Zhao, H., Giver, L. and Arnold, F. H. (1998) Random-priming *in vitro* recombination: an effective tool for directed evolution. *Nucleic Acids Res.* **26**, 681-683

7. Zhao, H., Giver, L., Shao, Z., Affholter, J. A. and Arnold, F. H. (1998) Molecular evolution by staggered extension process (StEP) *in vitro* recombination. *Nature Biotechnol.* **16**, 258-262

8. Zhao, H. and Arnold, F. H. (1997) Functional and nonfunctional mutations distinguished by random recombination of homologous genes. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 7997-8000.

9. Zhao, H. and Arnold, F. H. (1997) Optimization of DNA shuffling for high fidelity recombination. *Nucleic Acids Res.* **25**, 1307-1308.

Technique 4

Molecular Evolution by Staggered Extension Process (StEP) *in vitro* Recombination

(Huimin Zhao, Lori Giver, Zhixin Shao, Joseph A. Affholter and Frances H. Arnold)

(*Nature Biotechnology*, 1998, **16**, 258-262)

Molecular evolution by staggered extension process (StEP) in vitro recombination

Huimin Zhao, Lori Giver, Zhixin Shao, Joseph A. Affholter[1], and Frances H. Arnold*

*Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125. [1]Materials R&D–Biocatalysis, The Dow Chemical Company, 1707 Building, Midland, MI 48674. *Corresponding author (e-mail: frances@cheme.caltech.edu).*

Received 5 November 1997; accepted 23 January 1998

We have developed a simple and efficient method for in vitro mutagenesis and recombination of polynucleotide sequences. The staggered extension process (StEP) consists of priming the template sequence(s) followed by repeated cycles of denaturation and extremely abbreviated annealing/polymerase-catalyzed extension. In each cycle the growing fragments anneal to different templates based on sequence complementarity and extend further. This is repeated until full-length sequences form. Due to template switching, most of the polynucleotides contain sequence information from different parental sequences. The method is demonstrated by the recombination of two genes encoding thermostable subtilisins carrying two phenotypic markers separated by 113 base pairs and eight other point mutation markers. To demonstrate its utility for directed evolution, we have used StEP to recombine a set of five thermostabilized subtilisin E variants identified during a single round of error-prone PCR mutagenesis and screening. Screening the StEP-recombined library yielded an enzyme whose half-life at 65°C is 50 times that of wild-type subtilisin E.

Key words: directed evolution, random mutagenesis, subtilisin

Homologous recombination is a ubiquitous process that plays an important role in species adaptation and survival. Its importance is illustrated by its duality of functions—increasing genetic diversity in populations by reshuffling genes and preserving genetic integrity by aiding in the repair of damaged genes[1,2]. Computer simulations have shown that recombination with a low level of point mutation is efficient for the evolution of complex linear sequences[3–5]. Natural in vivo recombination mechanisms, however, usually operate at low efficiencies, eliciting insignificant changes in gene structures or functions even after tens of generations. Furthermore, in vivo recombination in most organisms is cumbersome and difficult to adapt to the redesign of genes, operons, or pathways.

Various approaches have been developed to mimic and accelerate nature's recombination strategy to direct the evolution of protein function[6]. In vitro recombination methods generally offer higher recombination efficiencies and greater experimental flexibility than in vivo approaches. In the widely used "DNA shuffling" method developed by Stemmer[7,8], a set of parent genes is digested with DNase I to create a pool of short DNA fragments that are reassembled into full-length genes by repeated thermocycling in the presence of DNA polymerase.

We describe a new approach to in vitro recombination that is both technically simple and conceptually novel. Rather than reassembling recombined genes from a fragment pool, our method prepares full-length recombined genes in the presence of the template(s) by what we call the "staggered extension" process (StEP). StEP consists of priming the template sequences followed by repeated cycles of denaturation and extremely abbreviated annealing/polymerase-catalyzed extension (Fig. 1). In each cycle the growing fragments can anneal to different templates based on sequence complementarity and extend further to create "recombination cassettes." Due to the template switching, the growing polynucleotides contain sequence information from different parental genes. StEP is continued until full-length genes are formed. It can be followed by a gene amplification step, if desired. The whole process can be performed using flanking universal primers.

We have assessed the degree and efficiency of StEP recombination by recombining two genes encoding thermostable *Bacillus subtilis* subtilisin E variants, each of which carries a single thermostable mutation along with several other neutral mutations. We have also demonstrated the utility of the StEP method for directed evolution of a mesophilic subtilisin into its thermophilic counterpart.

Results and discussion

StEP recombination between two thermostable subtilisin E genes. Two thermostable subtilisin E mutants, *RC1* and *RC2*, were used to test the recombination efficiency of the StEP method. The 10 positions at which these genes differ from one another are shown in Table 1. Only those mutations leading to amino acid substitutions Asn181 → Asp (N181D) and Asn218 → Ser (N218S) confer thermostability; the remaining mutations are neutral[9]. Single variants N181D and N218S have half-lives approximately threefold and twofold longer, respectively, than wild type subtilisin E at 65°C. The variant containing both mutations has an eightfold longer half-life. Thus N218S and N181D are convenient phenotypic markers for recombination events. Recombination between these positions only 113 bp apart can be measured easily by phenotypic analysis of a small sampling from the recombined variant library[9,10]. If the point mutagenesis rate is very low, 25% of the recombined population should exhibit wild-type–like stability, 25% of the population should have double mutant (N181D+N218S)-like stability and the remaining 50% should have single mutant (N181D or N218S)-like stability. Additional information on recombination efficiency can be obtained by sequencing a small sampling of the recombined library.

An equimolar mixture of plasmid DNAs containing the *RC1* and *RC2* genes was subjected to StEP recombination with two flanking primers located 45 nucleotides before the first codon and 113 nucleotides after the stop codon of the mature sequence. The progress of staggered extension was monitored by removing 10 µl aliquots from the reaction tube at various time points and separating the DNA fragments by agarose gel electrophoresis. The average

size of the smear increased gradually with increasing cycle number (Fig. 2). The front of the smear approached 100 bp after 20 cycles, 400 bp after 40 cycles, and 800 bp after 60 cycles. Finally, a clear, discrete approximately 1 kb band appeared within the smear after 80 cycles. This band containing a mixture of recombined products was gel purified, digested with restriction enzymes BamHI and NdeI, and ligated with vector generated by BamHI-NdeI digestion of the *Escherichia coli/B. subtilis* pBE3 shuttle vector[9]. This gene library was amplified in *E. coli* HB101 and transferred into *B. subtilis* DB428 competent cells for expression and screening.

The overall point mutagenesis rate associated with StEP recombination can be estimated from the catalytic activity profile of a small sampling of the recombined variant library[11,12]. The relationship between the point mutagenesis rate and the fraction of the library that encodes enzymes with activity significantly lower than wild type (<10%) is known[12]. Catalytic activities of enzyme variants were measured in a 96-well plate format[9]. Of the 368 clones screened, about 84% retained subtilisin activity. This level of inactivation corresponds to a point mutation rate of roughly 0.1% (a rate commonly observed when using *Taq* polymerase).

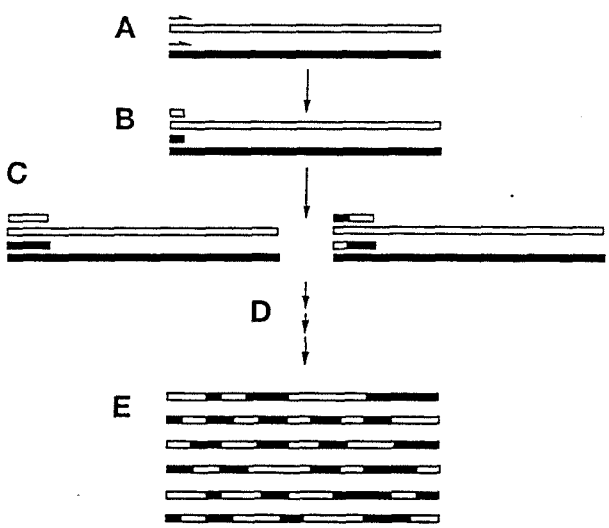


Figure 1. StEP recombination. Only one primer and single strands from two parent genes (templates) are shown. (A) Denatured template genes are primed with one defined primer. (B) Short fragments are produced by brief polymerase-catalyzed primer extension. (C) Through another cycle of StEP, fragments randomly prime the templates (template switching) and extend further. (D) This process is repeated until full-length genes are produced. (E) Full-length genes are purified and (optionally) amplified in a PCR reaction with external primers.

Table 1. DNA and amino acid substitutions in thermostable subtilisin E variants *RC1* and *RC2*.

Gene	Base	Base substitution	Position in codon	Amino acid substitution	Amino acid
	1107	A→G	2	218	Asn→Ser
RC1	*1141*	A→T	3	229	synonymous
	1153	A→G	3	233	synonymous
	484	A→G	3	10	synonymous
	520	A→T	3	22	synonymous
	598	A→G	3	48	synonymous
RC2	731	G→A	1	93	Val→Ile
	745	T→C	3	97	synonymous
	995	A→G	1	181	Asn→Asp
	1189	A→G	3	245	synonymous

Genes also contain base substitution A→G at position 780 relative to wild type.

The thermostabilities of the active clones are shown in Figure 3. With regards to thermostability, approximately 18% were similar to wild-type subtilisin E, 21% were comparable to the N181D+N218S double mutant, and 61% were comparable to the single mutants (N181D or N218S). This distribution was similar to that expected for random recombination of the two phenotypic markers separated by 113 bp.

The recombination efficiency was further analyzed by sequencing genes from 10 randomly selected clones. All 10 genes were novel recombinants, different from the parent genes (Fig. 4). The frequency of occurrence of any particular point mutation (marker) from parent *RC1* or *RC2* in the recombined genes ranged from 20% to 70% (fluctuating around the expected value of 50%), which indicated that the template-switching events during StEP were reasonably random. The minimum number of crossovers required to generate each chimeric gene ranged from one to four. A certain degree of linkage was apparent for mutations that are close together. The two closest, yet separable markers among this small sampling are 34 bp apart (positions 1107 and 1141 in clone 7). No significant differences in recombination efficiency were seen, compared with other in vitro recombination methods. When these two genes were recombined using the DNA shuffling method, for example, the minimum number of crossovers ranged from one to four[9]; it ranged from one to six for recombination using the random-priming technique[10]. None of the three methods was able to efficiently recombine the most closely spaced mutations. Six new point mutations were identified in the 10 genes. This mutation rate (0.06%) is close to that estimated from the frequency of active clones (*vide supra*).

Recombination is also observed when a pool of homologous templates is amplified by the PCR[13–16]. However, our PCR control experiments indicated that the efficiency of recombination by this route is very low. Ten nanograms of an equimolar mixture of plasmids containing genes *RC1* and *RC2* were used as templates in two separate conventional PCRs, one with a low annealing temperature (45°C) and the

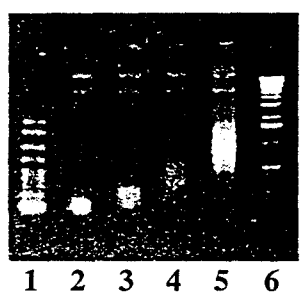


Figure 2. Agarose electrophoresis gel showing the progress of recombining two thermostable subtilisin E genes *RC1* and *RC2* by StEP. Lane 1: AmpliSize DNA Size standards (Bio-Rad, Hercules, CA), from top to bottom: 2000, 1500, 1000, 700, 500, 400, 300, 200, 100, and 50 bp; lane 2: after 20 cycles ; lane 3: after 40 cycles; lane 4: after 60 cycles; lane 5: after 80 cycles; lane 6: 1 kb ladder from Boehringer Mannheim.

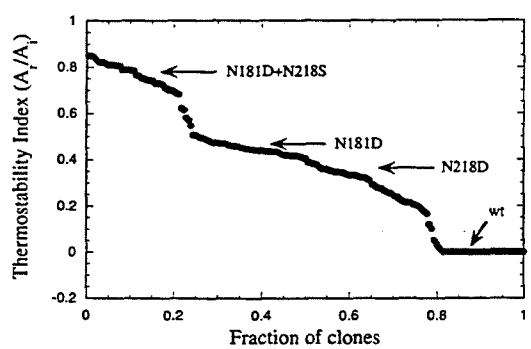


Figure 3. Results of screening 368 variants from the recombined library for activity after incubation at 65°C for 10 min (initial activity, A_i) and 40 min (residual activity, A_f). The ratio A_f:A_i (thermostability index) was used to estimate thermostability[9]. Data from active variants are sorted and plotted in descending order.

other with an annealing temperature of 60°C. Gene amplifications were conducted with *Taq* polymerase for 25 cycles: 1 min at 94°C, 1 min at 45°C or 60°C, and 1 min at 72°C. One hundred eighty-four clones from the two libraries were screened for thermostability. The fraction of clones expressing active subtilisin E increased from 53% to 80% when the annealing temperature was raised from 45°C to 60°C. This significant change most likely reflects an improvement in polymerase specificity at the higher temperature. However, among these active variants, the frequency of recombination of the two phenotypic markers in *RC1* and *RC2* (variants with wild-type–like stability plus those with double-mutant–like stability) decreased from 11% at 45°C (4% for double-mutant–like variants plus 7% for wild-type–like variants) to 5% at 60°C (2% for double-mutant–like variants plus 3% for wild-type–like variants). From the library generated at an annealing temperature of 45°C, genes from one variant exhibiting double-mutant-like stability and one variant exhibiting wild-type like stability were sequenced. The gene with wild-type–like stability had neither the N218S nor N181D mutation, while the thermostable gene had both. No new point mutations were found (data not shown). Recombinants are postulated to arise from incomplete extension of the annealed primer during one cycle when the polymerase pauses or prematurely disassociates from the template strand for unknown reasons[13–16]. This phenomenon shares the same recombination mechanism as StEP. In StEP recombination, however, the partial extension of the annealed primer deliberately controlled by the very brief annealing/extension steps greatly enhances recombination efficiency.

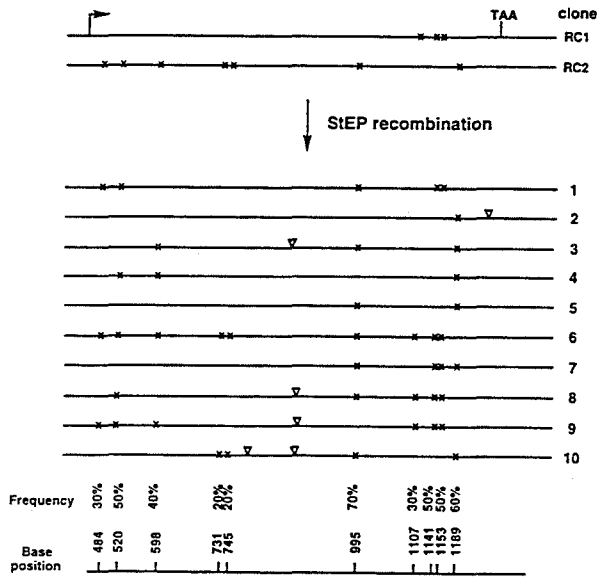


Figure 4. Sequence analysis of unscreened StEP-recombined gene libraries. Lines represent 986 bp of subtilisin E gene including 45 nucleotides of its prosequence, the entire mature sequence, and 113 nucleotides after the stop codon. X: positions of mutations from parent genes *RC1* and *RC2*; ▽: positions of new point mutations introduced during StEP. The four new point mutations in the center of genes 3, 8, 9, and 10 are located at different positions.

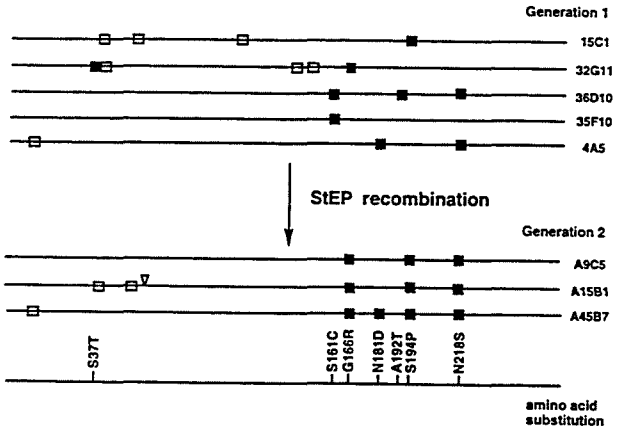


Figure 5. Sequence analysis of first- and second-generation thermostable subtilisin E variants. Filled squares indicate positions of nonsynonymous mutations, empty squares indicate positions of synonymous mutations, and triangles indicate positions of new point mutations introduced during StEP recombination.

The key to successful recombination by StEP is to tightly control the polymerase-catalyzed DNA extension. Denaturation is followed by random annealing of the extended fragments to template sequences and continued partial extension. This process is repeated multiple times, depending on the concentration of primer and template, until full-length sequences are made. Too much extension during each cycle severely limits recombination events. The annealing/extension step is therefore carried out under conditions that allow high fidelity primer annealing ($T_{annealing} > T_m$-25°C) but limit polymerization/extension (no more than a few seconds). Thermostable DNA polymerases typically exhibit maximal polymerization rates of 100–150 nucleotides/s at optimal temperatures and follow approximate Arrhenius kinetics at temperatures approaching the optimum temperature. Thus, at 55°C, a thermostable polymerase may exhibit only 20–25% of its steady-state polymerization rate at 72°C, or approximately 24 nucleotides/s. At 37°C and 22°C, *Taq* polymerase is reported to have extension activities of only 1.5 and 0.25 nucleotides/s, respectively[17]. The time and temperature of DNA polymerization must be optimized, depending on the template genes, polymerase, and particular reaction conditions (e.g., thermocycler used), to obtain the desired degree of recombination.

Unlike gene amplification (which generates new DNA exponentially), StEP generates new DNA linearly in its early cycles. In StEP, the ratio of primer to template is usually between 100 and 500, as compared with 10^6 in a typical gene amplification process. When significant numbers of primer-extended DNA molecules begin to reach sizes of more than one-half the length of the full-length gene, a rapid jump in molecular weight occurs, as half-extended forward and reverse strands begin to cross-hybridize to generate fragments nearly twice the size of those encountered to that point. Rapid consolidation of the smear into a discrete band of the appropriate molecular weight occurs either by continuing StEP or by altering the thermocycling program (by increasing the extension time or optimizing the extension temperature) to allow complete extension of the primed DNA and drive exponential gene amplification.

Directed evolution of a thermostable subtilisin E. Directed evolution involves the generation and selection or screening of molecular repertoires with sufficient diversity for the altered function to be represented. This "irrational" design approach has proven particularly effective for exploring and optimizing enzyme functions[6]. An effective directed-evolution strategy is to generate molecular diversity by error-prone PCR at a low error rate (two to three mutations per gene) and select or screen variants that show improvement with respect to the desired feature[18]. It is usually the case that several positive variants are identified after one round of selection or screening. Using in vitro recombination, beneficial mutations from these variants can be accumulated rapidly while the deleterious mutations are removed[7,8,19].

Subtilisin E is a protease produced by the mesophile *B. subtilis*. At 65°C, pH 8.0, and in the presence of 1 mM $CaCl_2$, the half-life of wild-type subtilisin E is approximately 5 min[9]. Our long-term goal was to convert this enzyme into its equivalent hyper-thermostable counterpart by directing its evolution in vitro. Mutagenesis by error-prone PCR on the mature subtilisin E gene and screening yielded five variants

with half-lives three to eight times greater than wild type at 65°C. Equal amounts of the genes from the five variants were recombined by StEP. Approximately 8000 clones expressing genes from the recombined library were screened for thermostability at 75°C. Three variants with the highest thermostabilities were identified and sequenced. Their half-lives at 65°C were 25–50 times that of wild type.

The DNA sequences of the five first-generation thermostable variants and the three thermostable variants obtained by StEP recombination are summarized in Figure 5. From inspection of the sequences of the first-generation variants, we can conclude that amino acid substitutions S194P and S161C are responsible for the enhanced thermostability of 15C1 and 35F10, respectively, as these are the only nonsynonymous mutations in their genes. Although synonymous mutations may affect expression, they are not expected to influence thermostability. Mutations leading to N181D (4A5) and N218S (36D10 and 4A5) have been confirmed previously to be thermostabilizing[9]. The remaining two variants, 32G11 and 36D10, each have more than one nonsynonymous mutation; the effects of S37T, G166R, and A192T are therefore not clear from the first generation sequences.

Among the second-generation variants, the most thermostable is A45B7, with a half-life 50 times that of wild-type subtilisin E. A45B7 contains four nonsynonymous mutations: G166R, N181D, S194P, and N218S, and the two less-thermostable variants A9C5 and A15B1 (half-life approximately 25–30 times that of wild type) contain only G166R, N181D, and N218S. The presence of G166R in all three indicates that it contributes to the observed thermostabilization. The absence of S37T and A192T in all three sequences supports the contention that these mutations are neutral, if not slightly deleterious. There are two possible explanations for our failure to identify a second-generation variant with known beneficial mutation S161C. The first possibility is that S161C does not contribute sufficiently to thermostability in the background of the other mutations. The second is that S161C is too close to G166R (only 15 bp apart) to recombine, and therefore that recombinant thermostable genes will contain one or the other, but not both. If G166R contributes more to thermostability than S161C, it will appear preferentially in the screened population. Based on these arguments, we conclude that mutations S161C, G166R, S194P, N181D, and N218S are thermostabilizing.

The efficiency of StEP recombination is similar to other in vitro recombination methods. However, the StEP recombination reaction can be carried out in a single tube; separation of parent templates from the recombined products is not necessary. StEP is in some ways reminiscent of the template-switching recombination mechanism that contributes to the evolutionary potential of retroviral populations[20]. The simple and efficient StEP recombination method provides a powerful new tool that can be applied to directed evolution of genes, operons, pathways, and even whole bacterial or viral genomes for specific applications.

Experimental protocol

Enzymes. Restriction enzymes were purchased from Boehringer Mannheim (Indianapolis, IN). Succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide (s-AAPF-pNa) was from Sigma (St. Louis, MO).

Staggered extension process (StEP). 5′ and 3′ flanking primers' P5N (5′-CCGAG CGTTG CATAT GTGGA AG-3′, underlined sequence is NdeI restriction site) and P3B (5′-CGACT CTAGA GGATC CGATT C-3′, underlined sequence is BamHI restriction site) were used for recombination. StEP conditions (100 µl final volume): 0.15 pmol (total) plasmid DNAs (pBE3 containing the subtilisin E genes[9]) were used as templates, 30 pmol of each primer, 1× *Taq* buffer, 0.2 mM of each dNTP, 1.5 mM $MgCl_2$, and 2.5 U *Taq* polymerase (Promega, Madison, WI). Program: 5 min of 95°C, 80 cycles of 30 sec 94°C, and 5 s 55°C. StEP was performed in an MJ Research (Watertown, MA) PTC-200 thermocycler. A clear, discrete electrophoretic band of the correct size (about 1 kb) among smears is typically obtained. All plasmids were isolated and purified from *E. coli* HB101 using QIAprep spin plasmid miniprep kit (Qiagen, Chatsworth, CA).

Cloning, expression, and thermostability screening. The product of correct size (about 1 kb) was cut from a 0.8% agarose gel after electrophoresis of the whole reaction mixture and purified using QIAEX II gel extraction kit (Qiagen). This purified product (approximately 300 ng) was subjected to a standard restriction-digestion reaction by Nde I and BamHI (a total of 20 µl of reaction volume), followed by electrophoresis. The DNA of correct size (about 1 kb) was again cut from the 0.8% agarose gel and purified using QIAEX II gel extraction kit (Qiagen). Subsequent cloning, expression, and thermostability screening were carried out as described[9]. For recombination of thermostable subtilisins E genes *RC1* and *RC2*, initial and residual activities were measured after incubation at 65°C for 5 and 20 min, respectively. Recombinants generated by StEP from five thermostable first-generation subtilisin E mutants were screened by measuring their activities after incubation at 75°C for 5 and 15 min.

Error-prone PCR. Random mutagenesis of subtilisin E genes was performed under conditions similar to those described[11,12]. Primers P5N and P3B were used to amplify approximately 1 kb fragments including a partial prosequence (a length of 15 residues), mature subtilisin E gene and 113 nucleotides after the stop codon. The PCR reaction contained (100 µl final volume): 10 mM Tris (pH 8.3 at 25°C), 50 mM KCl, 7 mM $MgCl_2$, 0.01% (wt/vol) gelatin, 0.2 mM dGTP, 0.2 mM dATP, 1mM dCTP, 1 mM TTP, 0.15 mM $MnCl_2$, 0.3 µM of both primers, 5 ng of template, and 5 U *Taq* DNA polymerase (Promega). No mineral oil was overlaid because the lid of thin PCR tube was pre-heated. PCR was performed in an MJ Research PTC-200 thermocycler for 13 cycles:1 min at 94°C, 1 min at 50°C, and 1 min at 72°C. The PCR products were purified using Wizards PCR Preps (Promega), followed by restriction digestion by Nde I and BamHI. These digestion products were purified again using Wizards PCR Preps. Cloning, expression, thermostability screening, enzyme purification, and DNA sequencing were carried out as described[9].

Acknowledgments

This work was supported by the US Department of Energy's program in Biological and Chemical Technologies Research within the Office of Industrial Technologies, Energy Efficiency and Renewables. The authors are grateful to Kentaro Miyazaki for helpful discussions and Yongkai Ow for her technical assistance.

1. Stahl, F.W. 1987. Genetic recombination. *Sci. Am.* 256:91–101.
2. Kucherlapati, R. and Smith, G.R. 1988. *Genetic recombination*. American Society of Microbiology, Washington, DC.
3. Holland, J.H. 1992. Genetic algorithms. *Sci. Am.* 267:66–72.
4. Holland, J.H. 1992. *Adaptation in natural and artificial systems*. 2nd ed. MIT Press, Cambridge, MA.
5. Forrest, S. 1993. Genetic algorithms: Principles of natural selection applied to computation. *Science* 261:872–878.
6. Kuchner, O. and Arnold, F.H. 1997. Directed evolution of enzyme catalysts. *Trends Biotech.* 15:523–530.
7. Stemmer, W.P.C. 1994. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* 370:389–391.
8. Stemmer, W.P.C. 1994. DNA shuffling by random fragmentation and reassembly - In vitro recombination for molecular evolution. *Proc. Natl. Acad. Sci. USA* 91:10747–10751.
9. Zhao, H. and Arnold, F.H. 1997. Functional and nonfunctional mutations distinguished by random recombination of homologous genes. *Proc. Natl. Acad. Sci. USA* 94:7997–8000.
10. Shao, Z., Zhao, H., Giver, L., and Arnold, F.H. 1998. Random-priming in vitro recombination: an effective tool for directed evolution. *Nucl. Acids Res.* 26:681–683.
11. Shafikhani, S., Siegel, R.A., Ferrari, E., and Schellenberger, V. 1997. Generation of large libraries of random mutants in *Bacillus subtilis* by PCR-based plasmid multimerization. *Biotechniques* 23:304–310.
12. Zhao, H. and Arnold, F.H. 1997. Optimization of DNA shuffling for high-fidelity recombination. *Nucl. Acids Res.* 25:1307–1308.
13. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T. et al. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239:487–491.
14. Scharf, S.J., Friedmann, A., Brautbar, C., Szafer, F., Steinman, L., Horn, G. et al. 1988. HLA class II allelic variation and susceptibility to pemphigus vulgaris. *Proc. Natl. Acad. Sci. USA* 85:3504–3508.
15. Scharf, S.J., Long, C.M., and Erlich H.A. 1988. Sequence analysis of the HLA-DRβ and HLA-DQβ loci from 3 pemphigus vulgaris patients. *Hum. Immunol.* 22:61–69.
16. Bradley, R.D. and Hillis, D.M. 1997. Recombinant DNA sequences generated by PCR amplification. *Mol. Biol. Evol.* 14:592–593.
17. Gelfand, H.A. 1989. *Taq* DNA polymerase, p. 18 in *PCR technology*. Ehrlich, H.A. (ed.). Stockton Press, New York.
18. Moore, J.C. and Arnold, F.H. 1996. Directed evolution of a para-nitrobenzyl esterase. *Nature Biotechnology* 14:458–467.
19. Moore, J.C., Jin, H.M., Kuchner, O., and Arnold, F.H. 1997. Strategies for the in vitro evolution of protein function: enzyme evolution by random recombination of improved sequences. *J. Mol. Biol.* 272:336–347.
20. Hu, W.-S., Bowman, E.H., Delviks, K.A., and Pathak, V.K. 1997. Homologous recombination occurs in a distinct retroviral subpopulation and exhibits high negative interference. *J. Virol.* 71:6028–6036.
21. Cadwell, R.C. and Joyce, G.F. 1994. Mutagenic PCR. *PCR Methods and Applications* 2:28–33.

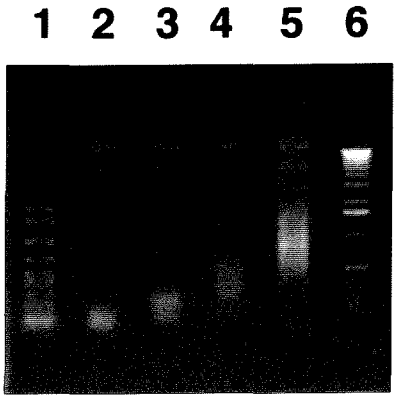


Fig. 2. Agarose electrophoresis gel showing the progress of recombining two thermostable subtilisin E genes RC1 and RC2 by StEP. Lane 1: AmpliSize^{TM} DNA Size standards (Bio-Rad Laboratories), from top to bottom: 2000, 1500, 1000, 700, 500, 400, 300, 200, 100 and 50 bp; Lane 2: after 20 cylces ; Lane 3: after 40 cylces; Lane 4: after 60 cycles; Lane 5: after 80 cycles; Lane 6: 1 kb ladder from Boehringer Mannheim.

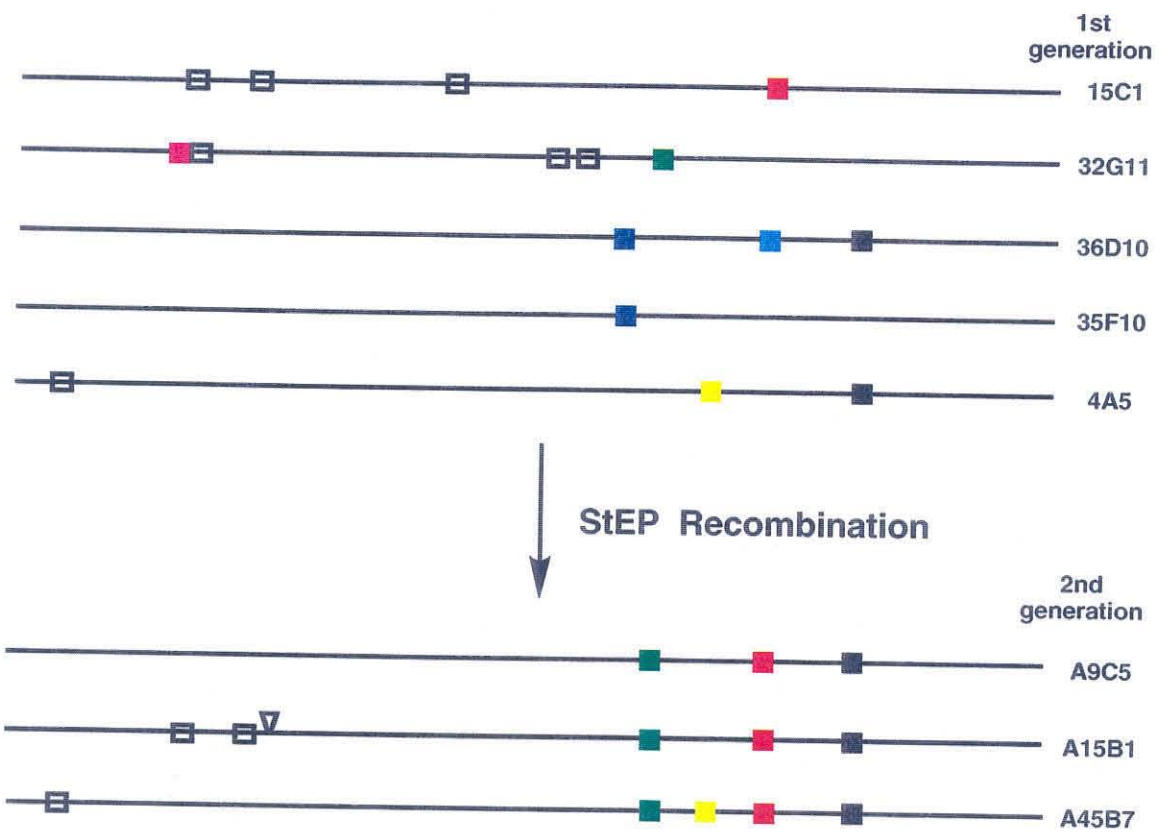


Fig. 5 (in color)

could form the basis of a simple bioassay for signaling molecules. Specific cells of the early frog embryo have the potential to develop along any of several pathways of differentiation depending upon the type of signals they receive[2]. Cells from a transgenic embryo containing DNA that encodes a particular tissue-specific promoter driving a reporter gene could be used to identify molecules that promote or inhibit differentiation. In addition, frog transgenic technology is ideally suited for identifying the temporal and tissue-specific transcriptional elements of developmentally important genes.

Until now, the identification of DNA regulatory sequences has been hampered by the problems of unreliable and inappropriate transgene expression[4]. The incorporation of ITRs into plasmids that contain specific promoters should alleviate these problems and make it feasible to reliably define regulatory sequences. Microinjecting hundreds of frog embryos with different permutations of a specific promoter may become the fastest and least expensive way to map transcriptional regulatory elements that function in vertebrate organisms. Given the evolutionary conservation of many developmental control mechanisms, it may prove to be more effective, at least initially, to analyze promoters of other vertebrate organisms in transgenic frogs.

Developmental biologists who study frogs will benefit the most from this technology in the short term. Currently, proteins are expressed in frog embryos via injection of mRNA. The problems with this strategy are that expression lasts only as long as the mRNA persists, and there is little control over the time and place of expression. Plasmids containing ITRs represent a significantly more reliable method for controlling the expression of molecules in early embryos.

We now have two methods for generating transgenic frogs: the introduction of plasmids containing ITRs, and REMI. The advantages of the ITR technology are its efficiency and its simplicity. The work of Sylvia Evans and colleagues is a most promising advance for all biologists who relish frog embryos for their many experimental advantages. It is worth speculating that, with further improvements, the ITR strategy in frogs may rival the use of mouse transgenics for many applications.

1. Kroll, K.L. and Amaya, E. 1996. *Development* 122:3173–3183.
2. Kroll, K.L. and Gerhart, J.C. 1994. *Science* 266:650–653.
3. Fu, Y., Wang, Y., and Evans, S.M. 1998. *Nature Biotechnology* 16:253–257.
4. Vize, P.D. et al. 1991. *Methods Cell Biol.* 36:367–387.
5. Philip, R. et al. 1994. *Mol. Cell. Biol.* 14:2411–2418.
6. Mohun, T.J., Garrett, N., and Gurdon, J.B. 1986. *EMBO J.* 5:3185–3193.
7. Dawid, I.B. 1991. *Methods Cell. Biol.* 36:311–328.

One small StEP in molecular evolution. . .

Lance P. Encell and Lawrence A. Loeb

Nature has evolved some remarkably precise and efficient enzymes through a dynamic equilibrium between mutagenesis and Darwinian selection. While natural evolution takes millions of years, methods are currently being developed to evolve enzymes in vitro in a matter of days. These new methods take advantage of the ease of genetic recombination and allow rearrangements of an expanded set of sequences for a particular protein. In this issue, Frances Arnold and colleagues[1] present a novel approach for creating molecular diversity, termed the staggered extension process (StEP). Their approach complements, extends, and perhaps simplifies methods previously used for in vitro DNA shuffling.

Evolution in the test tube is not subject to the web of intradependent pathways that restricts evolution in cells. In addition, an improved enzyme in the mind of a scientist may not offer selective advantages during natural evolution, but may be of great potential value to industry and medicine[2].

Until recently, in vitro evolution involved substituting random nucleotide sequences for designated portions of genes[3] or mutating larger regions by either chemical modification[4] or error-prone PCR[5]. Using large libraries containing random nucleotide substitutions in DNA, it is possible to select rare mutant proteins with desired phenotypes or screen a limited number of variants for altered properties[6]. Sequential applications of these methods mimic Darwinian evolution, but only sparsely sample all of the possible sequences for a particular protein (sequence space)[7].

The use of random mutagenesis for applied molecular evolution is based on the premise that we lack sufficient information about how amino acids interact within proteins to predict the phenotypes resulting from most single substitutions, saying nothing about multiple substitutions. As a result, the rational design of novel enzymes by site-specific mutagenesis has been restricted to a few well-studied situations. In contrast, applied molecular evolution circumvents these problems; proteins are evolved to possess the features that we desire them to have. The use of random mutagenesis does not require a priori knowledge of the effects of amino-acid changes; direction is achieved by screening or selecting interesting mutants from among large populations containing randomized sequences.

Currently, only a handful of methods are being used for the generation of gene libraries containing random mutations and for the selection of active mutants. These methods can be divided into two categories (Fig. 1): Those that emphasize individual substitutions for introducing mutations, and those that emphasize recombination. With random oligonucleotide mutagenesis, one can target and saturate multiple sites in a region of a plasmid-encoded gene, and this approach has been used extensively for altering substrate specificities and for examining structure–function relationships[4,8–10]. With chemical modification and error-prone PCR, one can scan larger gene segments to produce predominantly single mutations. In contrast to these methods, recombination in vitro is more global in scope and facilitates the exchange between different gene segments.

DNA shuffling methods developed by Stemmer[11] and the new approach (StEP) developed by Zhao and Arnold[12] are particularly suited for exploring multiple distant domains and larger sequence space. Shuffling involves randomly fragmenting single or homologous genes by hydrolysis with DNase I followed by PCR amplification in which the digested fragments serve as primers for DNA synthesis. During each round of annealing, template switching results in crossovers in regions of sequence homology. Mutants exhibiting desired properties can be selected and reshuffled against each other or against the wild-type sequence (backcross) to remove deleterious recombination products from the pool. Initial studies involved DNA shuffling between different mutants of the same gene. However, more recent work has involved shuffling of homologous genes from different species. In the later studies, the shuffling of four microbial class C cephalosporinase genes was more robust and resulted in mutants exhibiting as high as a 540-fold improvement in moxalactam resistance[13].

StEP, like DNA shuffling, can be used to promote in vitro recombination among mutant genes. In PCR-like reactions, mixed templates containing different mutations are

Lance P. Encell is a postdoctoral fellow and Lawrence A. Loeb is director of the Gottstein Memorial Cancer Research Laboratory, departments of pathology and biochemistry, University of Washington School of Medicine, Seattle, WA 98195-7705

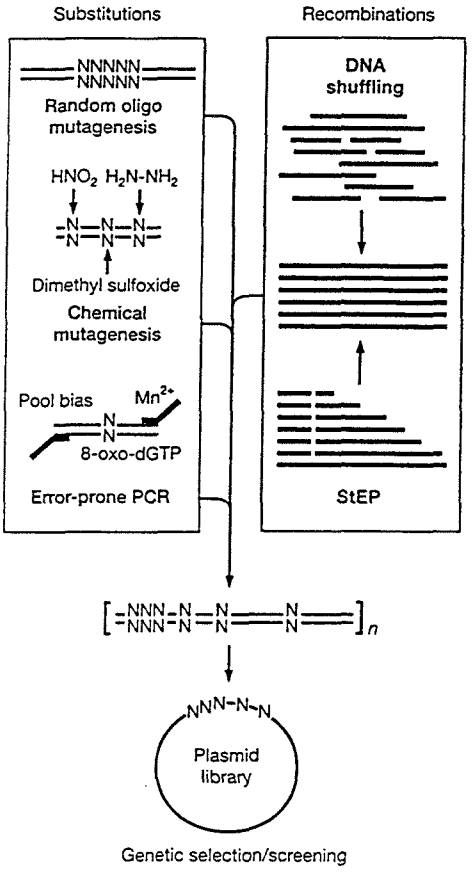


Figure 1. Substitution and recombinant approaches for creating molecular diversity. Techniques can be used independently, in combination, or sequentially to evolve enzymes in vitro that can be genetically screened or selected.

primed and extended for very short periods of time, followed by cycles of denaturation and annealing. Using conditions that allow primer annealing but decrease polymerization rates, the partially extended fragments hybridize randomly onto templates harboring different mutations.

In the present paper, Zhao and Arnold use StEP to combine two *Bacillus subtilis* subtilisin E mutants previously shown to be thermostable[12] and recover mutants in a Mendelian-like fashion (1:2:1). They go on to generate five thermostable mutants of subtilisin E by error-prone PCR, recombining them by StEP. Among the resulting recombinant enzymes, they recovered a mutant with a half-life at 65°C that is 50-fold longer than that of wild-type subtilisin E. Their long-term goal is to convert this enzyme into a hyperthermostable variant; along the way they are likely to obtain information about protein folding. The authors point out the similarity between StEP and natural recombination processes that occur in viruses such as HIV containing two encapsulated RNA template strands.

StEP is similar in efficiency to shuffling, but it is a simpler process. Both methods harness the power of recombination in mixing different domains and in sampling large regions of sequence space. In either approach, the extent of diversity (recombination) can be controlled by altering the time and temperature of the annealing steps.

The authors state in their paper that a drawback to DNA shuffling is that DNase I treatment does not result in completely random fragments. With StEP, the fragments may also not be totally random due to sequence-specific pause sites during DNA synthesis. In fact, StEP shows no improvement over shuffling as far as the minimum distance that can exist between mutations (34 bp), suggesting that this approach also does not yield a completely random library of clones before screening.

StEP, DNA shuffling, and substitution-oriented approaches are all being used to create enzymes with multiple substitutions that are highly efficient and exhibit altered substrate specificities. Such mutant proteins are currently beyond the reach of rational design. With each of these techniques, the diversity of functional mutants that one can create is astronomic. Moreover, these techniques can be used in combination or sequentially, and it is likely that newer methods for applied molecular evolution will be rapidly forthcoming. Our limitation is not the ability to create, but the imagination of the investigator in devising powerful techniques for the selection or screening of mutant proteins with desired properties.

1. Zhao, H. et al. 1998. *Nature Biotechnology* 16:258–261.
2. Skandalis, A., Encell, L.P., and Loeb, L.A. 1997. *Chem. Biol.* 4:889–898.
3. Horwitz, M.S. and Loeb, L.A. 1986. *Proc. Natl. Acad. Sci. USA* 83:7405–7409.
4. Sweasy, J.B. and Loeb, L.A. 1993. *Proc. Natl. Acad. Sci. USA* 90:4626–4630.
5. Cadwell, R.C. and Joyce, G.F. 1992. *Cold Spring Harbor Laboratory* 2:28–33.
6. Black, M.E. et al. 1996. *Proc. Natl. Acad. Sci. USA* 93:3525–3529.
7. Eigen, M. and Winkler-Oswatitsch, R. 1990. pp. 505–530 in *Methods in enzymology*, Vol. 183. Doolittle, R.F. (ed.). Academic Press, London.
8. Suzuki, M. et al. 1996. *Proc. Natl. Acad. Sci. USA* 93:9670–9675.
9. Gulick, A.M. and Fahl, W.E. 1995. *Proc. Natl. Acad. Sci. USA* 92:8140–8144.
10. Encell, L.P., Coates, M.M., and Loeb, L.A. 1998. *Cancer Res.* In press.
11. Stemmer, W.P.C. 1994. *Nature* 370:389–391.
12. Zhao, H. and Arnold, F.H. 1997. *Proc. Natl. Acad. Sci. USA* 94:7997–8000.
13. Crameri, A. et al. 1998. *Nature* 391:288–291.

Chapter 4

Engineering Highly Thermostable and Active Subtilisins by Directed Evolution

(Rapid Conversion of a Mesophilic Enzyme into its Thermophilic Counterpart by Directed Evolution)

ABSTRACT

In the order of months we have evolved a mesophilic enzyme, wild type subtilisin E isolated from mesophilic microorganism, *Bacillus subtilis* into its thermophilic counterpart (without compromising its activity) in the test tube. Our approach, 'directed evolution', consisted of repeated cycles of random mutagenesis, gene recombination and screening. After five generations of directed evolution, the resulting variant 5-3H5 is as stable as its homolog, thermitase isolated from thermophilic microorganism *Thermoactinomyces vulgaris*. Their half-lives at 83 °C were 3.5 min and their temperature optima were 76 °C, 18 °C higher than that of wild type subtilisin E. In addition, this variant was more active towards succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide than wild type subtilisin E over the whole range of temperatures. It was also shown that directed evolution provided unique tools to unveil the mechanism of thermal adaptation and the molecular basis of thermostabilization. Our work demonstrates that directed evolution is a generally applicable, highly effective approach to increase protein thermostability.

INTRODUCTION

The enhancement of enzyme thermostability is of central importance for biomedical, chemical and industrial enzyme applications. Despite many successful examples of protein thermostabilization, however, the mechanism of thermostabilization is still far from understood. A major reason for this is that proteins are only marginally stable as a result of the delicate balance between numerous stabilizing and destabilizing interactions; protein stability represents the cumulative effect of small improvements involving all levels of the hierarchy of the protein structure [1,2,3].

One of the experimental approaches that have been widely followed to alter the stability of proteins is rational design, either by comparison between homologous proteins from thermophiles and mesophiles [4,5,6] or by analysis of the contributions of the different interactions in a protein [7] . However, identifying the important sequence determinants of thermostability involved in a specific case remains an overwhelming task since protein sequences have usually diverged significantly. Further structural comparisons between thermostable and non-thermostable proteins failed to identify simple patterns that characterize stabilizing interactions [1,6]; minute and subtle local structural alterations are sufficient to confer the small difference of free energy of stabilization between them. The marginal stability of proteins and the complex effects of temperature on the different interactions also makes it very difficult to assess the contributions of these interactions on an energetic scale [1,3]. Thus, it is not surprising that no general rules of thermostabilization have been defined.

An alternatiive to rational design is to couple random mutagenesis with a suitable screen or selection to isolate point mutants that are more thermostable [8,9,10,11,12]. It has also been demonstrated that such mutations can be combined and the stability increases can be additive [13,14]. However, this approach usually involves only a single round and has the limitation that further improvements rely on identifying the stabilizing mutations out

of other accompanying neutral and/or deleterious mutations followed by combining them by site-directed mutagenesis. Sequence information is required to guide all the efforts, and additivity has to be assumed for all the stabilizing mutations.

Directed evolution provides an effective alternative to these two approaches. Inspired by Darwinian evolution, directed evolution involves multiple rounds of mutation, recombination and screening or selection. New, powerful tools to implement evolution in the test tube have been developed rapidly [16]. Various protein properties have been designed by directed evolution, including stability, substrate specificity, activity in non-natural environments and protein expression [16, 17]. Apart from engineering novel protein functions, directed evolution can also probe the molecular mechanisms underlying those improvements. Unlike in natural occurring proteins, functional mutations of laboratory-evolved proteins are much easier identified [16].

To study the relationship between enzyme activity and thermostability, we have directed the evolution of an alkaline serine protease - subtilisin E isolated from *Bacillus subtilis*, to create its thermostable counterpart. The sequence of subtilisin E shares 80% identity with that of subtilisin BPN'. Both subtilisins are produced from pre-pro-subtilisins consisting of the pre-sequence of 29 residues, the prosequence of 77 residues, and the mature protease of 275 residues [18]. The pre-sequence functions as the signal peptide for protein secretion across the membrane [19], while the pro-sequence acts as a "foldase" to guide the appropriate folding of the subtilisin molecule [20,21].

Subtilisin has been widely used as an additive in laundry detergents. The enzyme loses its function by irreversible inactivation with a half-life in the order of weeks at 20°C and in the order of minutes at 60 °C (pH 7-9) [22]. There is considerable practical interest in engineering highly stable and active subtilisins. The observation that thermostable enzymes from thermophilic bacteria are often less active at low temperature than their mesophilic counterparts [23,24], and extensive studies on the relationship between protein stability and function by site-directed mutagenesis [25] have been used to support the

hypothesis that there is a balance between stability and function. Previous studies on subtilisin E [26,27] also seemed to support this hypothesis. Thus, engineering highly stable and active enzymes poses a difficult challenge to the rational protein design.

Here we describes a general approach, directed evolution, to increase enzyme thermostability. Starting from a wild type subtilisin E isolated from mesophile *Bacillus subtilis*, we obtained a subtilisin E variant 5-3H5 after five generations of directed evolution. This variant is more thermostable than one of its thermostable homologs, thermitase isolated from the thermophile *Thermoactinomyces vulgaris*, and at the same time, is more active than wild type subtilisin E over the whole range of temperatures. We also discuss the underlying mechanism of thermal adaptation and molecular basis of thermostabilization.

Experimental Design and Strategy

In directed evolution experiments, the key processes--mutation, recombination and screen or selection -- are carefully controlled by the experimenter. Because the vast protein sequence space can not be completely searched by screen or selection, and most mutations are deleterious while beneficial mutations are rare, a good evolutionary strategy has to be carefully designed [16].

Our strategy for directing the evolution of subtilisin E is first to make a library of subtilisin E gene, each with 2-3 base changes on average (which maximizes the population of single mutants). The gene library is then cloned into *E. coli* and further expressed in *B. subtilis*. Several thousand subtilisin variants from the library are screened for their thermostability. Typically, a few variants are much more stable for their parent. The genes encoding these variants are then recombined in order to find the best combination of beneficial mutations and at the same time remove potential deleterious mutations. The most

thermostable variant identified from this library then serves as a template for the next round of mutagenesis and screening. This sequence of events is repeated until the goal is achieved. In addition, since recombination of relatively few mutations can lead to very large screening requirements, the number of the genes to be recombined is controlled to be small (usually less than five) in order to explore all the possible combinations [28]. As a result, even though there are more than five positive variants identified from the mutant library, only the top four or five are selected.

Screening is the most flexible sorting method for directed evolution [29]. Two major criteria for a good screening method are the efficiency and sensitivity. In other words, the assay should be rapid and the screening conditions should ensure that the expected small improvements brought by single amino acid substitutions can be measured. In accordance with these considerations, the thermostability screen was developed by modifying a similar method performed in petri dishes [11, Chapter 2]. The method is based on the retention of activity after incubation at an elevated temperature for a fixed period of time and is not designed to distinguish various mechanisms of inactivation. The ratio between residual activity and initial activity (normalized residual activity, both are measured at 37 °C) is taken as index of thermostability. Variants with higher index of thermostability than wild type subtilisin E and with initial activity still comparable to wild type at 37 °C are chosen as positives. These positives are then subjected to a more rigorous analysis: measurement of thermal inactivation kinetics. Variants with half-lives longer than their parents are regarded as "true" positives. With each successive generation, the incubation temperature is gradually elevated concomitantly. For each generation, this temperature is determined by the point where the parent shows 30-40% normalized residual activity after incubation.

RESULTS

The Process of Directed Evolution

An initial round of mutagenic PCR was performed on the mature wild type subtilisin E gene, followed by cloning, expression and thermostability screening. Initial and residual activities were measured after incubation at 65°C for 5 and 20 minutes, respectively (the incubation time was fixed for the subsequent generations). Of the ~5000 clones screened, five variants (1-4A5, 1-15C1, 1-32G11, 1-35F10, 1-36D10; the initial x- indicates generation number) were identified with higher thermostability index and longer half-lives at 65 °C than wild type. Equal amounts of the genes from the five first generation variants were recombined by staggered extension process (StEP) [30]. Approximately 8000 clones expressing genes from the recombined library were screened for thermostability at 75°C. Many variants showed improved thermostability and one particular variant, 2-45B7, exhibited the highest thermostability index and the longest half-life at 75 °C. This variant was therefore used as parent for the second round of mutagenesis.

The third generation of the evolution process began with a mutagenic PCR reaction on the gene isolated from 2-45B7. Approximately 2000 clones were screened for thermostability at 76 °C. Three variants (3-5H2, 3-16D11, 3-20E8) were identified with higher thermostability index and longer half-lives than 2-45B7. The genes isolated from these three variants were recombined again by StEP. Another ~3000 clones were screened at 78 °C, which yielded variant 4-8B3 with the highest thermostability index and the longest half-life. This variant was used as the parent for the third round of mutagenesis. Among ~2000 clones screened at 80 °C, one variant 5-3H5 was identified with higher thermostability index as well as longer half-life than 4-8B3.

Thermostability of Evolved Thermostable Subtilisin Variants

With other proteases, subtilisins differ from other proteins by the fact that autolysis is the main cause of irreversible inactivation at elevated temperatures [31,32,33]. Preventing autolysis during stability measurement is highly complicated, if possible at all, and a rigorous thermodynamic analysis of protease stability therefore seems to be impossible [31,33]. Thus, the thermostability of subtilisins is usually measured by kinetic analysis of the inactivation process, either by rates of thermal inactivation (more commonly by half-lives, $t_{1/2}$, which are the time needed to reach 50% of the initial activity) [31] or T50, which is the temperature causing 50% loss of protease activity during a fixed incubation period [33,34]. The difference of T50 between evolved thermostable variants and wild type subtilisin E is shown by dT50.

Twelve subtilisins consisting of 0-WT, 1-4A5, 1-15C1, 1-32G11, 1-35F10, 1-36D10, 2-45B7, 3-5H2, 3-16D11, 3-20E8, 4-8B3, and 5-3H5 were grown in half liter cultures and purified. In all cases, purified protein appeared as a single major band upon sodium dodecyl sulfate-polyacryamide gel electrophoresis (SDS-PAGE). Subtilisin activity was assayed by monitoring the hydrolysis of 0.2 mM solutions of the tetrapeptide substrate succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide (sAAPF-*p*Na) at 37 °C. As shown in Table 4.1, at 65 °C, in the presence of 10 mM Tris buffer (pH 8.0) and 1 mM $CaCl_2$, the half-life and T50 of wild type subtilisin E are ~5 min and 59.2 °C, respectively. The first generation variants (1-4A5, 1-15C1, 1-32G11, 1-35F10, 1-36D10) show half-lives of 2-8 times greater than wild type at 65 °C and dT50 of 2.3 °C - 8.2 °C. The half-life of their recombined product 2-45B7 is ~50 times that of wild type at 65 °C and its T50 increases by 13.1°C compared to wild type. The third generation variants (3-5H2, 3-16D11, 3-20E8) prolong their half-lives by ~50% at 75 °C and each of them exhibits ~1°C increase of T50. Their subsequently recombined product 4-8B3 shows a half-life at 75 °C greater than any of them. T50 of 4-8B3 is increased by 15.2 °C. The half-life of the final product 5-3H5 is

two times that of 4-8B3 at 80 °C or at least 200 times that of wild type at 65 °C (Fig. 4.1). T50 of 5-3H5 reaches at 76.4 °C or is increased by 17.2 °C, compared to wild type.

Activity of Evolved Thermostable Subtilisin Variants

The specific activities of the evolved thermostable subtilisin variants and wild type are listed in Table 4.2. None of the first generation variants decreased its specific activity even slightly relative to wild type. Instead, 1-4A5, 1-32G11 and 1-35F10 each showed a significant increase of specific activity. Beginning from second generation, all the thermostable variants exhibited 3.5 times higher activity than wild type.

More detailed kinetic analyses were carried out to determine their kinetic constants, k_{cat} and K_M. As shown in Table 4.2, among the first generation variants, k_{cat} of 1-4A5 and 1-35F10 was two times greater than that of wild type, while their K_M remained unchanged. Both k_{cat} and K_M remained unchanged for 1-36D10, while 1-15C1 showed a small increase (20%) only in k_{cat}. K_M of 1-32G11 was lowered more than two fold but its k_{cat} was increased by 20%, thus its catalytic efficiency was three times that of wild type. Starting from second generation, all the subsequent generation variants showed similar k_{cat} and K_M with more than 2-fold increase in k_{cat} and more than 2.5-fold decrease in K_M.

Comparison With Thermitase

Thermitase, a naturally-occurring thermophilic homologue of subtilisin E isolated from *Thermoactinomyces vulgaris*, consists of 279 amino acid residues [35,36]. Its sequence alignment with wild type subtilisin E shows 43% identity (Fig. 4.2).

To make a side-by-side comparison between 5-3H5 and thermitase, the half-lives of thermitase and 5-3H5 were measured under the same conditions (at 83 °C in 10 mM Tris buffer, pH 8.0, and 1 mM Ca^{2+}). As shown in Fig. 4.3, the rates of thermal inactivation of both enzymes follow first-order kinetics and both exhibited a half life of 3.5 min. The

half-life of thermitase was 1.2 min under a similar assay condition (83°C in 0.1 M HEPES buffer, pH 7.6, and 1 mM Ca^{2+} [37]). In addition, the temperature optima for wild type subtilisin E, 5-3H5 and thermitase were also determined, as described in Materials and Methods. As shown in Fig. 4.4, the temperature optima of 5-3H5 and thermitase are all at 76 °C, 18 °C higher than that of wild type subtilisin E, which is in good agreement with the temperature increase determined by T50. Interestingly, compared to that of wild type, the temperature profile of 5-3H5 is not only broadened towards higher temperatures but also elevated at all temperatures.

The specific activity and kinetic constants of thermitase were also determined at the same consition as used for subtilisins. As shown in Table 4.2, the specific activity of thermitase at 37 °C is ~15 times of that of wild type subtilisin E. The k_{cat} of thermitase is five times that of wild type subtilisin E while its K_M is ~27 times that of wild type subtilisin E. Thus, the higher catalytic efficiency of thermitase with respect to wild type subtilisin E is mostly due to the difference of substrate specificity; the synthetic peptide substrate binds the active site of thermitase much more tightly.

Sequence Analysis

To further probe the molecular basis of thermostabilization and the evolutionary pathway, the DNA sequences of the above eleven evolved thermostable subtilisin E variants were determined. As shown in Fig. 4.3, all together the sequences contain 21 unique base substitutions. Eleven out of 21 substitutions lead to amino acid changes. The locations of substitutions are well distributed throughout the mature subtilisin E gene. No two base changes occurred simultaneously in a single codon. The types of substitutions are also more or less balanced with regard to transitions and transversions; the number of transitions (purine to purine changes - A to G or C to T) is almost equal to that of transversions (purine to pyrimidine changes) (10 vs. 11).

Among the first five generation variants, S161C, G166R, S194P, N181D and N218S have been identified as thermostable mutations, S37T and A192T being neutral [30, Technique 4 in Chapter 3]. Among the three third generation variants, S9F, P14L, and N76D are responsible for the enhanced thermostability, since each is the only nonsynonymous mutation in its gene. Two of the three thermostable mutations, P14L and N76D, are recombined in the fourth generation variant 4-8B3. The failure to identify a variant with known beneficial mutation S9F could be attributed to the same reasons accounted for the lack of S161C in the recombined product 2-45B7 [30]. One is that S9F does not contribute sufficiently to thermostability in the background of the other mutations. The other is that S9F is too close to P14L (only 15 bp apart) to recombine, and therefore that recombinant thermostable genes will contain one or the other, but not both. Interestingly, thermostable mutation S161C reappears in the fifth generation variant 3H5. N118S in 5-3H5 is thought to be a stabilizing mutation since it increases the thermostability of wild type subtilisin E [38]. However, it is also possible that N118S may have little or no effect on thermostability in the background of 5-3H5. It is noteworthy that N218S occurred both in 1-4A5 and 1-35F10, while S161C occurred in three different genes.

Sequence analysis also revealed the mutations affecting activity. As shown in Table 4.2, single mutant 1-36D10 (S161C) has a k_{cat} and K_M similar to those of wild type, thus S161C does not affect activity. In contrast, single mutant 1-15C1 (S194P) shows a slight increase in k_{cat} but not in K_M, thus S194P slightly affects activity. In accordance with the finding that N181D does not affect specific activity while N218S increases specific activity 2-fold [39], the double mutant 1-4A5 (N181D+N218S) has the same K_M as wild type but a k_{cat} two times that of wild type. Furthermore, since the kinetic constants of triple mutant 1-35F10 (S161C+A192T+N218S) are the same as those of 1-4A5, A192T is indeed neutral in terms of activity. In addition, the finding that K_M of the second generation variant 2-45B7 (G166R+N181D+S194P+N218S) is similar to double mutant 1-32G11 (S37T+G166R) while its k_{cat} is similar to double mutants containing N218S indicates that

G166R decreases K_M by more than 2-fold and slightly increases k_{cat} while S37T does not affect activity. Data from subsequent generation variants showed that the remaining mutations S9F, P14L, N76D and N118 clearly do not affect activity.

DISCUSSION

Kinetics of Thermal Inactivation

Thermal inactivation of many proteins is irreversible, and involves at least two steps: (a) reversible unfolding of the native protein (N) and (b) irreversible alteration of the unfolded protein (U) to yield a final state (F) that is unable to fold back to the native structure. This two-step nature of irreversible denaturation is usually depicted by the Lumry and Eyring model: $N \Leftrightarrow U \rightarrow F$ [40]. Various mechanisms including autolysis, aggregation and chemical damage to certain amino acids may irreversibly inactivate subtilisins, however, at elevated temperatures autolysis is the major cause of inactivation [32,33, 41, 42]. Because of the broad specificity of subtilisins, conformational features rather than sequence characteristics of the subtilisin molecules are thought to determine the sites of autolytic attack. It has been shown that the rate of thermal inactivation is controlled by the rate of local unfolding processes that render the subtilisin molecule susceptible to autolysis [32,33, 41, 42], hence that stability to thermal unfolding results in increased resistance to inactivation under many conditions.

In accordance with the notion that unfolding is the rate-limiting step in the inactivation process, the kinetics of irreversible loss of activity of wild type subtilisin E at 60-65 °C, follows first-order kinetics and its half-life is independent of enzyme concentration (Chapter 2). The fact that the rates of inactivation of the evolved thermostable subtilisin E variants are still first-order and independent of enzyme

concentration (data are not shown) suggests that the beneficial mutations fixed during the directed evolution process do not change the rate-limiting step.

The kinetics of thermal inactivation of thermitase also follows this model. As shown in Table 4.3, the half-life of thermitase at 83 °C is virtually independent of the enzyme concentration. As another evidence, the kinetics of thermal inactivation also obeys first-order as exemplified in Fig. 4.3.

Sequence Comparisons

Subtilisin E belongs to the superfamily of subtilisin-like serine proteases, termed "subtilases", consisting of more than 200 members. Multiple sequence alignments indicate a high degree of sequence variability as the result of natural evolution. With the exception of the essential catalytic triad residues D32, H64, and S221 and a single glycine residue (G219), virtually every other residue can be replaced by one or more different residues. Large deletions and insertions are also frequently found. However, all subtilases have a similar overall core structure [43].

Many subtilisin homologs from thermophiles or hyperthermophiles have also been characterized, including thermitase from *Thermoactinomyces vulgaris* [35,36], aqualysin I from *Thermus aquaticus* YT-1 [44], aerolysin from *Pyrobaculum aerophilum* [45], and others. Based on the subtilisin sequences aligned by Siezen and coworkers [43], all the eleven new amino acid substitutions that occurred during the directed evolution process are found in the sequences of at least one other subtilisin. These amino acid substitutions therefore have been tried previously in nature and found to be acceptable. However, it is impossible to identify stabilizing mutations based on the sequence comparisons between subtilisins from mesophiles and (hyper-)thermophiles. For example, thermostabilizing mutations S161C and G166R are found only once in a mesophilic subtilase from *Drosophila melanogaster* and in one from *Bacillus subtilis*, respectively. Similarly, thermostabilizing mutations S9F and P14L are found exclusively in mesophilic subtilase,

but in more than four sequences. Furthermore, thermostabilizing mutation N218S is found in 67 out of 127 aligned sequences, including psychrophilic, mesophilic, thermophilic and hyperthermophilic subtilisins. On the other side, neutral mutation S37T is also found in thermophilic or even psychrophilic subtilisins.

To probe the origin of thermostability, several studies have been based on the sequence comparisons between thermophiles and mesophiles. It is suggested that certain amino acid residues or certain amino acid replacement pairs are used preferentially in thermophilic proteins [46-49]. For example, the top two replacement pairs are Lys→Arg and Ser→Ala (Argos' replacements) [49]. However, no such pairs were needed to convert mesophilic subtilisin E into its thermophilic counterpart. Nor were other proposed thermophilic transitions such as increase in alanine, isoleucine, tyrosine content [6]. Our results show that there is no clear correlation between the type of mutation and the effect on the protein thermostability. Thus, every single mutation should be evaluated in its specific structural context, as discussed below.

Structural Analysis

Two subtilisin structural models were generated to further examine the molecular basis of thermostability and activity for the stabilizing mutations. The first, shown in Fig. 4.6, maps these mutations onto a cartoon of the secondary structure topology drawn after Siezen and coworkers [43]. Most of the beneficial mutations are located in the loops connecting helices and strands; only two mutations S9F and P14L are located in helices. In particular, the substitution of proline with leucine (P14L) in helix-B can increase the helical propensity, thus increasing thermostability. The substitution of Ser194 with Pro rigidifies the flexible loop, reducing entropy and thus increasing thermostability. In addition, most of these mutations occur at the structurally variable regions (shown as dashed lines), while two mutation G166R and S9F occur at structurally conserved regions (shown as solid lines).

The second model, shown in Fig. 4.7, maps these mutations onto a 3-dimensional model generated on basis of the known crystal structure of wild type subtilisin E [38] using Biosym's InsightII program. The amino acid substitutions are distributed throughout the protein. All of them are located on the surface of the enzyme (G166R is partially exposed). To better visualize these mutations, Fig. 4.8a,b show the space-filling model. Most are also located far from the active site. Two residues, however, are directly involved in the substrate binding. Residue 166 is located at the bottom end of the distinct, large and elongated cleft known as S1 pocket, which determines the substrate specificity. Residue 218 involves the less tight binding of C-terminal or leaving portion P1'-P2' of the substrate. Thus, it is not surprising that mutation G166R decreases K_M by more than 2-fold and increases k_{cat} slightly, while N218S increases k_{cat} by two-fold and keeps K_M unchanged. Furthermore, the enhancement of thermostability caused by G166R may be attributed to the effect of reduced cavity volume since substitution of G166 by Arg significantly reduces the volume of the S1 pocket. N218S apparently increases thermostability by slight improvement of the hydrogen bonding scheme near a β-bulge, which is the thermostabilizing mechanism found in a thermostable subtilisin BPN' [11]. N76D is involved in the binding of calcium ions, the occurrence of a negative charge therefore increases the binding affinity, resulting in higher stability [13]. The molecular cause of thermostability brought by S9F, S161C, N181D and possibly N118S, are not clear.

The finding that all the stabilizing mutations are located on the protein surface is consistent with the model for thermal inactivation of proteases. This model predicts the stability of proteases is largely determined by the rate of local unfolding processes. Since the early steps of unfolding of a protein are thought to involve mainly surface-located structure elements [50,51], mutations stabilizing these elements will therefore have large effects on stability.

Only stabilizing mutations N218S and N181D are found in thermitase. The structural comparison between thermitase and two mesophilic subtilisins, subtilisin Carlsberg and subtilisin novo, has concluded that the unusual tight binding of calcium by thermitase was the most likey molecular cause for its increased thermostability [52]. However, our results show that eight stabilizing mutations are sufficient to account for the thermostability difference between thermitase and mesophilic subtilisins. The only stabilizing mutation N76D, which is supposed to increase the affinity for calcium is even not present in thermitase. These stabilizing mutations are distributed throughout the protein and structurally isolated. The stabilizing effect of each mutation is so small that it may be easily evade the detection by structural comparison. However, structural analysis did unveil the mechanism for the lower K_M of thermitase. As occurred in 5-3H5, G166 in thermitase is replaced by a residue with a long side chain (G$\rightarrow$N). The volume of S1 pocket of thermitase is further reduced by the deletion of a small loop (residue 160-163, numbering is based on subtilisin E) and substitution of P129T. Furthermore, the residue located at the bottom end of the S2 pocket is Thr in thermitase as compared to Ser in subtilisin E. The additional methylene group changing from Ser to Thr reduces the volume of S2 pocket. As a result, thermitase binds the substrate much more tightly than wild type subtilisin (x27), and to a less extent than 5-3H5 (x10).

CONCLUSIONS

A very limited number of amino acid substitutions are needed to convert a mesophilic enzyme, subtilisin E, into a variant which is more stable than its thermophilic counterpart, thermitase without sacrificing activity at lower temperatures. The mutations are scattered over the molecule and structurally isolated, which is presumably the physical basis for their additive or cumulative effects. Various mechanisms of thermostabilization

have been found, including better hydrogen bonding, enhanced secondary structure propensity, improved electrostatic interactions, which involve only minute local structural alterations. The present study strongly supports the notion that thermal stability is achieved by the cumulative effect of small improvements at many locations within the protein molecule [1]. As a consequence, the pursuit of a 'holy grail' for protein thermostabilization is deemed unsuccessful. However, directed evolution provides a general approach to increase protein thermostability. This approach requires no structural information or the principles that govern protein stability. Furthermore, by applying multiple selective pressures, novel thermophilic enzymes can be obtained, as exemplified by this study. Such thermostable and highly active enzymes will have many potential biomedical, chemical and industrial applications.

MATERIALS AND METHODS

Restriction enzymes were purchased from Boehringer Mannheim (Indianapolis, IN). Succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide (s-AAPF-*p*Na) was from Sigma (St. Louis, MO). Purified thermitase was a gift from Dr. Wolfgang Höhne at Humboldt University (Germany).

Subtilisin E working system. The construction of the *E. coli* - *B. subtilis* shuttle vector pBE3, preparation and transformation of competent *E. coli* HB101 cells and *B. subtilis* DB428 cells, and growth of cells were all described in Chapter 2.

Random Mutagenesis by Error-prone PCR. Random mutagenesis of subtilisin E genes was performed as described elsewhere [30]. Primers P5N (5'-CCGAG CGTTG CATAT GTGGA AG-3', underlined sequence is *NdeI* restriction site) and P3B (5'-

CGACT CTAGA GGATC CGATT C-3', underlined sequence is *BamHI* restriction site) were used to amplify ~1 kb fragments including a partial prosequence (a length of 15 residues), mature subtilisin E gene and 113 nt after the stop codon. The PCR reaction contained (100 µl final volume): 10 mM Tris (pH 8.3 at 25°C), 50 mM KCl, 7 mM $MgCl_2$, 0.01% (wt/vol) gelatin, 0.2 mM dGTP, 0.2 mM dATP, 1mM dCTP, 1 mM TTP, 0.15 mM $MnCl_2$, 0.3 uM of both primers, 5 ng of template and 5U *Taq* DNA polymerase (Promega). No mineral oil was overlaid since the lid of thin PCR tube was pre-heated. PCR was performed in a MJ Research (Watertown, MA) PTC-200 thermocycler for 13 cycles : 1 min 94 °C, 1 min 50°C and 1 min 72°C. The PCR products were purified using Wizards PCR Preps (Promega), followed by restriction digestion by *Nde* I and *Bam*HI. These digestion products were purified again using Wizards PCR Preps (Promega).

***In vitro* Recombination by Staggered Extension Process (StEP).** Recombination of the five first generation variants and the three third generation variants were all carried out as described elsewhere [30]. The same two primers P5N and P3B were used. StEP conditions (100 µl final volume): equal amounts of pBE3 plasmids containing the genes encoding thermostable variants were mixed and a total of 0.15 pmol plasmid DNAs were used as templates, 30 pmol of each primer, 1x *Taq* buffer, 0.2 mM of each dNTP, 1.5 mM $MgCl_2$ and 2.5 U *Taq* polymerase (Promega). Program: 5 min of 95°C, 80 cycles of 30 sec 94 °C, 5 sec 55 °C. StEP was performed in a MJ Research (Watertown, MA) PTC-200 thermocycler. A clear, discrete electrophoretic band of the correct size (~1 kb) among smears is typically obtained. The product of correct size (~1 kb) was cut from an 0.8% agarose gel after electrophoresis of the whole reaction mixture and purified using QIAEX II gel extraction kit (Qiagen, Chatsworth, CA). This purified product (~300 ng) was subjected to a standard restriction-digestion reaction by *Nde* I and *Bam*HI (a total of 20 µl of reaction volume), followed by electrophoresis. The DNAs of

correct size (~1kb) was again cut from the 0.8% agarose gel and purified using QIAEX II gel extraction kit (Qiagen, Chatsworth, CA).

Cloning, Expression, Thermostability Screening and DNA Sequencing. The purified restricted inserts from either error-prone PCR or StEP were ligated with vectors generated by *Bam*HI-*Nde*I digestion of the pBE3 *E. coli-B. subtilis* shuttle vectors. The ligation mixtures were used to transform *E. coli* HB101 by electroporation and selected by 100 ug/ml ampicillin. A typical ligation reaction (10 ul final volume) contained 1x T4 DNA ligase buffer, ~12 ng inserts, ~50 ng vector, and 0.5 U T4 DNA ligase. The reaction mixture was incubated at 15-16 °C for 16 h. The resulted library of plasmids were isolated and transformed into *B. subtilis* DB428 for enzyme expression. Colonies were picked by sterile toothpicks and grown in 96-well plates each well containing 200 ul of SG medium supplemented with 30 ug/ml kanamycin. After 20 h incubation at 37 °C, the 96-well plates were centrifuged at 3000 rpm for 10 min and 5 ul of supernatant from each well was transferred into the corresponding wells of two new fresh plates. 15 ul of SG medium was further added into each well of the plates in order to avoid a problem associated with evaporation during incubation at high temperatures. Initial activity was measured after 5 min of incubation at a specific temperature, and residual activity was measured after 20 min of incubation at the same temperature. The incubation was performed in an oven on an aluminum block machined to closely contact standard 96-well plates for uniform heating. The ratio between residual activity and initial activity was used as the thermostability index for each variant. The enzyme activity assay was carried out at 37 °C after addition of 100 ul of 0.2 mM s-AAPF-pNa in 10 mM Tris-HCl, 1 mM CaCl2, pH8.0 (prewarmed at 37 °C). The activity was determined by the absorbance change at 405 nm within 1 min monitored by a 96-well plate reader (Molecular Devices, Inc.) DNA sequencing of thermostable variants was carried out as described [39].

Enzyme Assays. All the subtilisin E variants including wild type were purified as described previously [39]. All the measurements were made using a thermostatted Milton-Roy Spectronic 3000 Array spectrophotometer. Purified subtilisins and thermitase were dialyzed in 10 mM Tris-HCl, 1 mM $CaCl_2$, pH 8.0 at 4°C overnight before characterization. The enzyme concentration was determined by absorbance at 280 nm (for subtilisin E, $\varepsilon = 35886\ M^{-1}cm^{-1}$, for thermitase, $\varepsilon = 59055\ M^{-1}cm^{-1}$). Purified subtilisins and thermitase were assayed by the initial rates of hydrolysis of s-AAPF-*p*Na substrate in 0.99 ml of 10 mM Tris-HCl, 1 mM $CaCl_2$, pH 8.0 at 37 °C as described previously [39]. Specific activities were determined as described [39].

$t_{1/2}$. A MJ Research (Watertown, MA) PTC-200 thermocycler was used as an incubator for precisely controlling the temperatures. Purified enzymes were incubated in 10 mM Tris-HCl (pH 8.0), 1 mM $CaCl_2$ at specific temperatures. Aliquots, taken at various time intervals, were removed and diluted into 1.0 ml of activity assay solution (0.2 mM s-AAPF-pNa, 10 mM Tris-HCl (pH 8.0), 1 mM $CaCl_2$) equilibrated at 37 °C. Typically, inactivation was followed until greater than 80% of the enzyme activity has been lost and plots of the logarithm of residual activity versus time were linear. The rate of thermal inactivation (k_{inact}) was the slope of the straight line fitted by linear regression algorithm and values of half-lives were calculated by the equation: $t_{1/2} = \ln 2 / k_{inact}$.

T_{50}. T_{50} of subtilisins were determined following the protocol described by Eijsink and coworkers [52]. Aliquots of purified subtilisins were incubated for 20 min at various temperatures. Subsequently, the residual activity was determined and expressed as percentage of the initial activity. T50 was the temperature of incubation at which 50% activity was preserved. Alternatively, for each subtilisin, $t_{1/2}$ was determined at various temperatures and T_{50} was the temperature of incubation at which $t_{1/2}$ was 20 min.

T_{opt}. The optimum temperature for activity was determined by incubating proteases (wild type subtilisin E, 5-3H5 and thermitase) with s-AAPF-*p*Na (2 mg/ml) in 10 mM Tris-HCl (pH8.0), 1 mM $CaCl_2$, at different temperatures for 10 minutes, after which the amount of released pNa was determined by absorbance at 410 nm. It was ensured that the substrate was in excess and the rate of digestion was proportional to the concentration of proteases.

k_{cat} and K_M. For each subtilisin, a series of initial rates were determined at eight different substrate concentrations over the range of 0.02 - 1.5 mM that bracketed K_M. For thermitase, the substrate concentration was ranged from 0.002 - 0.08 mM. Data from the reaction progress curve were fit to the Michaelis-Menten equation by a nonlinear regression algorithm and used to calculate the k_{cat} and K_M. Standard deviations in k_{cat} and K_M for all values reported are below ± 5%.

Generation	Subtilisin E variants	Temperature at which inactivation kinetics is measured (°C)	$t_{1/2}$ (min)	dT_{50} (°C)	$t_{1/2}$ at 65°C (variants) / $t_{1/2}$ at 65°C (wt)
0	wild type	65	4.9	0.0	1.0
1	4A5	65	44.2	8.2	9.0
	15C1		35.5	7.6	7.2
	32G11		9.8	2.3	2.0
	35F10		39.5	7.8	8.1
	36D10		15.2	3.1	3.1
2	47B5	65	~250	13.1	51.0
		75	8.7		
3	5H2	75	12.0	14.0	70.4
	16D11		12.8	14.1	75.1
	20E8		13.6	14.2	79.8
4	8B3	75	18.3	15.2	107.3
		80	5.3		
5	3H5	80	10.3	17.2	210.0
		83	3.53		

Table 4.1. The thermostability of evolved subtilisin E variants and wild type. Values of half-lives at each temperature are the average of triple experiments. Standard deviations in $t_{1/2}$ for all reported values are below ± 5%. dT50 is the difference of T50 between evolved thermostable variants and wild type subtilisin E. T50 of wild type subtilisin E is 59.2 °C.

Generation	Subtilisin E variants	Specific activity at 37 °C (U/mg)	k_{cat} (s^{-1})	K_M (mM)	k_{cat} / K_M ($s^{-1}mM^{-1}$)
0	wild type	20.0	25.4	0.385	66.0
1	4A5	40.0	52.4	0.388	135.1
	15C1	21.0	29.4	0.384	76.6
	32G11	44.2	30.6	0.158	193.7
	35F10	40.6	52.3	0.383	136.6
	36D10	19.8	25.6	0.385	66.5
2	45B7	69.9	55.6	0.152	365.8
3	5H2	71.6	58.8	0.155	379.4
	16D11	73.0	57.1	0.150	380.7
	20E8	72.2	56.2	0.151	372.2
4	8B3	72.5	56.5	0.153	369.3
5	3H5	71.0	55.8	0.151	373.3
	Thermitase	293.3	130.1	0.0146	8911.0

Table 4.2. The specific activity and kinetic constants of evolved subtilisin E variants and wild type measured in 0.2 mM s-AAPF-*p*Na, 10 mM Tris-HCl (pH 8.0), 1 mM $CaCl_2$ at 37 °C. Values of specific activies are the average of five measurements with standard deviations less than or equal to ± 5%.

concentration of thermitase (μM)	half-life (min) for thermal inactivation [a]
2.52	3.79
1.26	3.53
0.252	3.49
0.126	3.65

a. Inactivations were performed as described under Materials and Methods in 1 mM $CaCl_2$ and 10 mM Tris buffer (pH 8.0) at 83 °C.

Table 4.3. Concentration independence for the half-life of irreversible thermal inactivation of thermitase at elevated temperatures. Values of half-lives at different enzyme concentrations are the average of triple experiments. Standard deviations in $t_{1/2}$ for all reported values are below ± 5%.

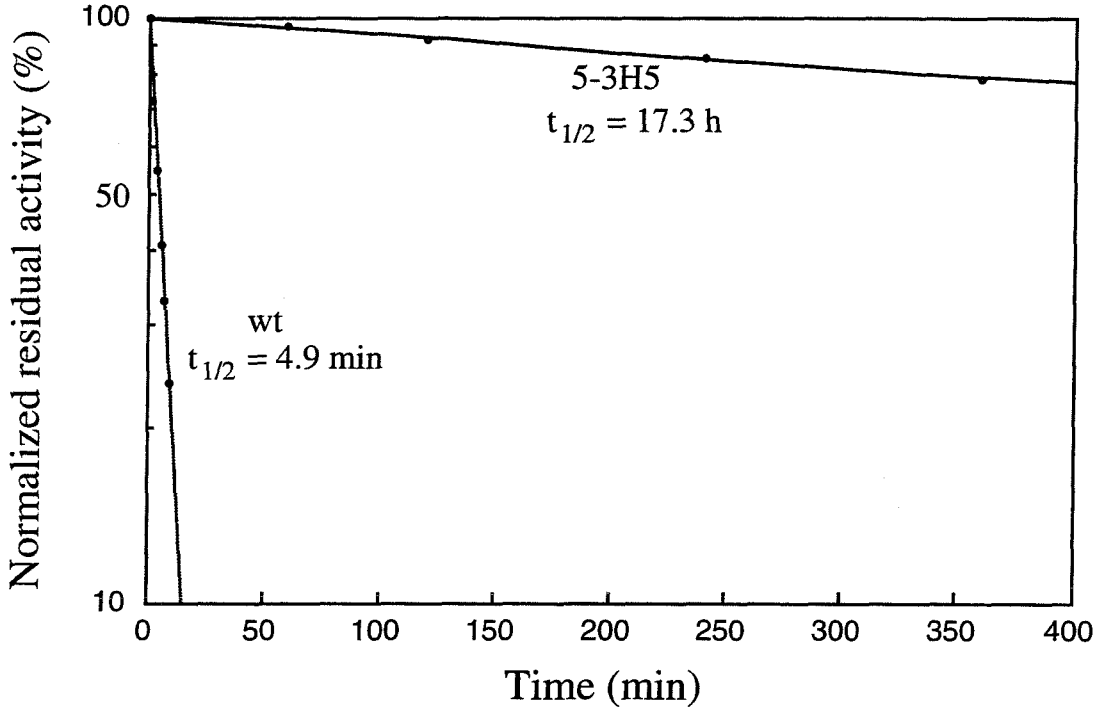


Fig. 4.1. Kinetics of thermal inactivation of 5-3H5 and wild type subtilisin E at 65 °C in 0.2 mM s-AAPF-*p*Na, 10 mM Tris-HCl (pH 8.0), 1 mM $CaCl_2$.

```
            1                                                    51
SubtE       AQSVPYGISQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDL--NVRGGASFV
                **  .*.**.  . . .**..*.*..*.*..*.****  .* ** .**
  THM   YTPNDPYFSSRQYGPQKIQAPQAWDIA-EGSGAKIAIVDTGVQSNHPDLAGKVVGGWDFV

        52                                                      110
SubtE   PSETNPYQDGSSHGTHVAGTIAAL-NNSIGVLGVSPSASLYAVKVLDSTGSGQYSWIING
         ....* *.*..**** **. **. ***.*. *..*.**. **.***..*** .. . **
  THM   DNDSTP-QNGNGHGTHCAGIAAAVTNNSTGIAGTAPKASILAVRVLDNSGSGTWTAVANG

        111                                                     170
SubtE   IEWAISNNMDVINMSLGGPTGSTALKTVVDKAVSSGIVVAAAAGNEGSSGSTSTVGYPAK
        *..* ... .**..****..*...*. .*. * ..* **.*****    ..... .***
  THM   ITYAADQGAKVISLSLGGTVGNSGLQQAVNYAWNKGSVVVAAAGN----AGNTAPNYPAY

        171                                                     230
SubtE   YPSTIAVGAVNSSNQRASFSSAGSELDVMAPGVSIQSTLPGGTYGAYNGTSMATPHVAGA
        *...***.... .....***. ** .** ***  * ** *..**.. .**********.
  THM   YSNAIAVASTDQNDNKSSFSTYGSVVDVAAPGSWIYSTYPTSTYASLSGTSMATPHVAGV

        231                                             275
SubtE   AALILSKHPTWTNAQVRDRLESTATYL-GNSFYYGKGLINVQAAAQ
        *.* *... .......*. .*.**. . *.. *..** .*.  *.*
  THM   AGL-LASQ-GRSASNIRAAIENTADKISGTGTYWAKGRVNAYKAVQY
```

Fig. 4.2. Amino acid sequence alignment of wild type subtilisin E and thermitase. Identical amino acids are marked by asterisks are identical (43.1% identity) and similar amino acids are marked by dots. Deletions are shown by dashes. SubtE: subtilisin E, THM: thermitase.

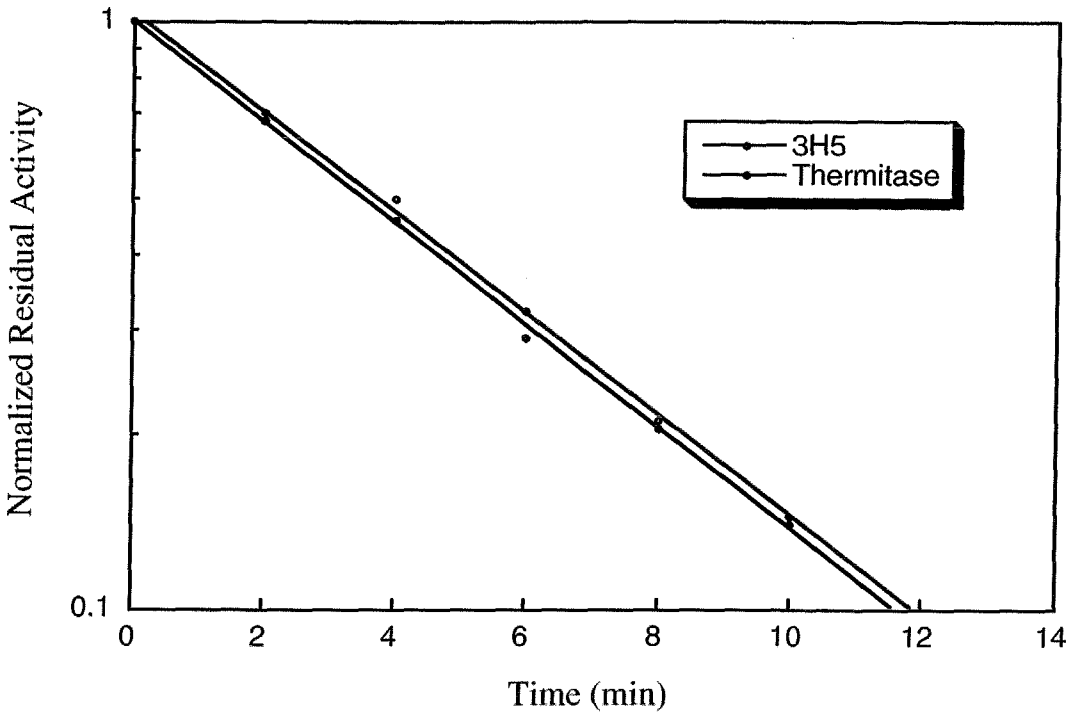


Fig. 4.3. Kinetics of thermal inactivation of 5-3H5 and thermitase at 83 °C. Shown are the first-order inactivation curves for themitase (red line) and 5-3H5 (blue line) in 10 mM Tris-HCl (pH 8.0), 1 mM $CaCl_2$ at 83 °C.

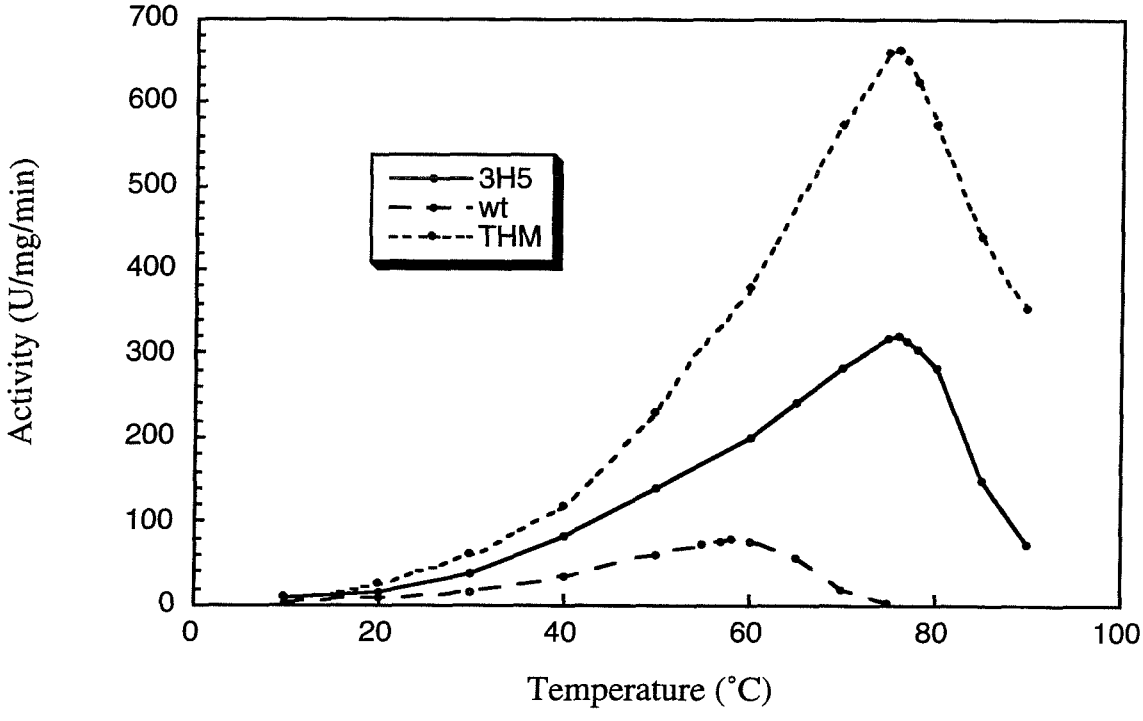


Fig. 4.4. Activity - temperature profiles of wild type subtilisin E (– –), 5-3H5 (——) and thermitase (----).

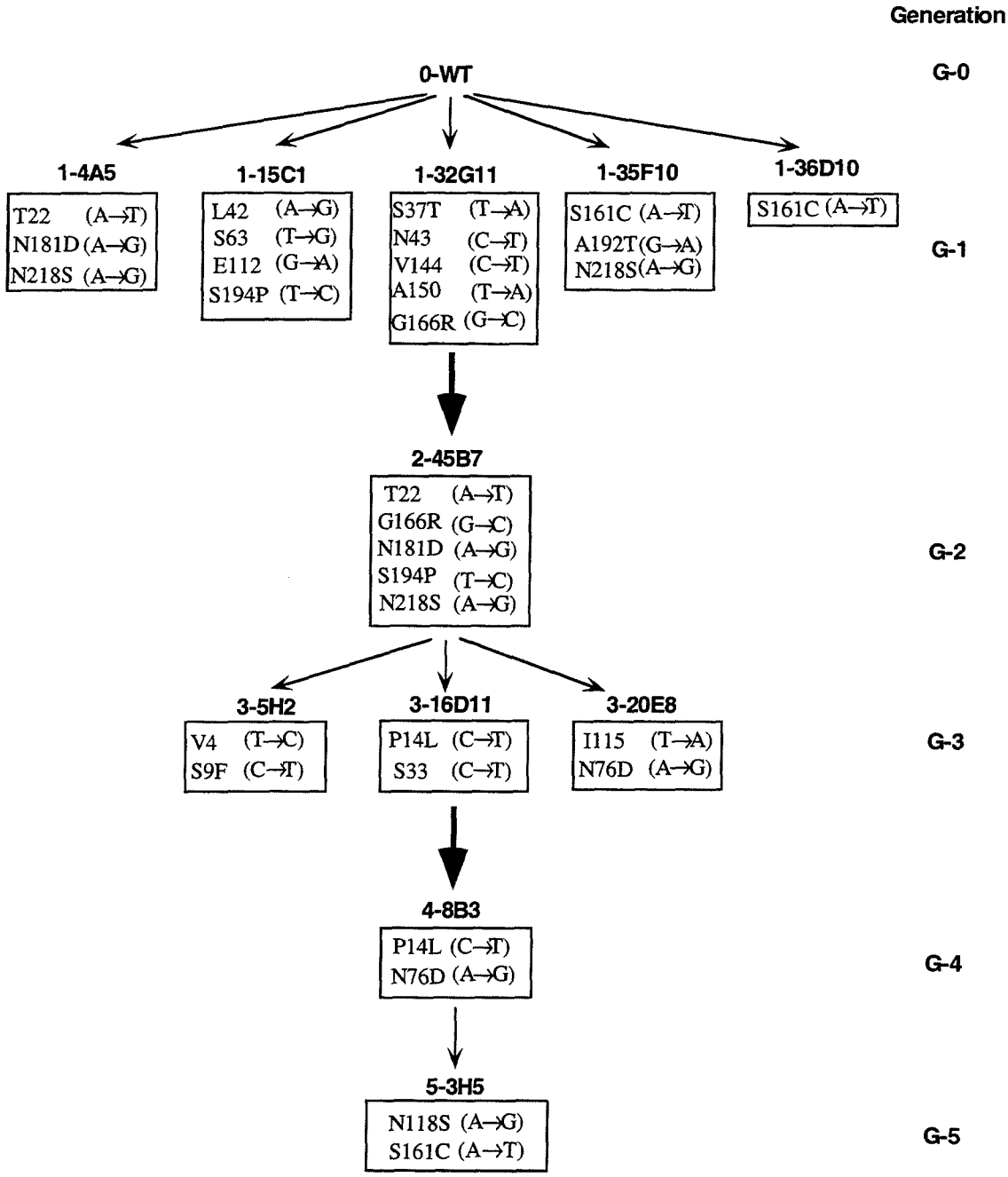


Fig. 4.5. Lineage and DNA sequencing results of the evolved thermostable subtilisin E variants.

Fig. 4.6. Schematic representation of the secondary structure topology of subtilisin E, with α-helices shown as cylinders and β-sheet strands as arrows. Solid lines indicate the conserved regions and dashed lines the variable regions. Approximate location is indicated of the main Ca^{2+}-binding sites (Ca1 and Ca2), catalytic triad residues D32, H64, S221 (by *) and substrate-binding region (between strands eI and eIII). (●), locations of the stabilizing amino acid substitutions found in the directed evolution process. The most thermostable variant 5-3H5 contains all these mutations except for S9F.

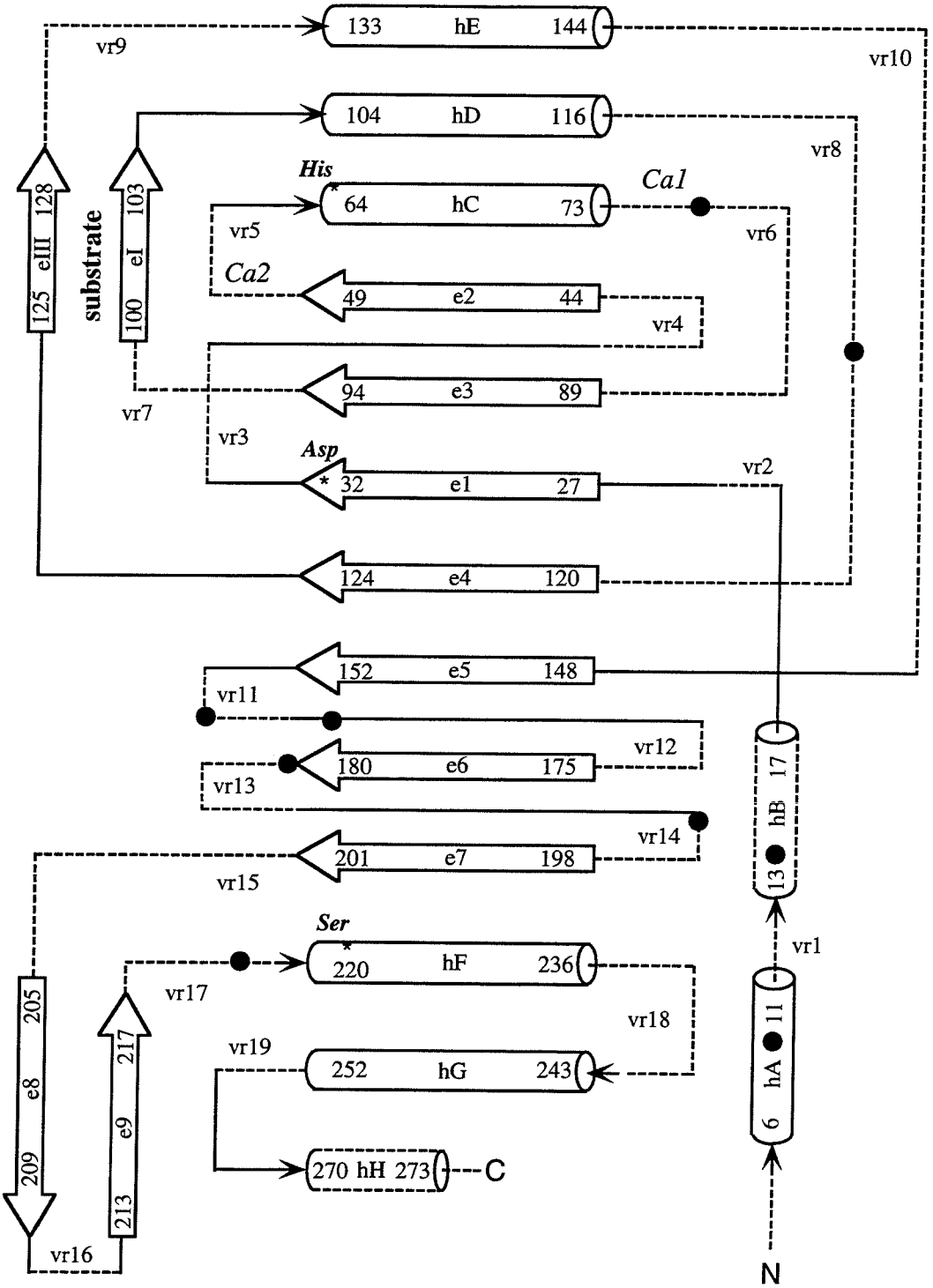

133 hE 144
vr9
vr10
104 hD 116
vr8
His
Ca1
64 hC 73
vr5
vr6
Ca2
49 e2 44
vr4
125 eIII 128
substrate
100 eI 103
94 e3 89
vr7
vr3
Asp
32 e1 27
vr2
124 e4 120
152 e5 148
vr11
vr12
180 e6 175
vr13
vr14
201 e7 198
vr15
Ser
220 hF 236
vr17
vr18
vr19
252 hG 243
270 hH 273
C
205 e8 209
217 e9 213
vr16
13 hB 17
vr1
6 hA 11
N

Fig. 4.7. Ribbon drawing of the three-dimensional model of the 5-3H5 subtilisin E variant. The eight stabilizing mutations are shown in cyan, and the catalytic triad is shown in red. The gray sphere indicates the binding calcium ion.

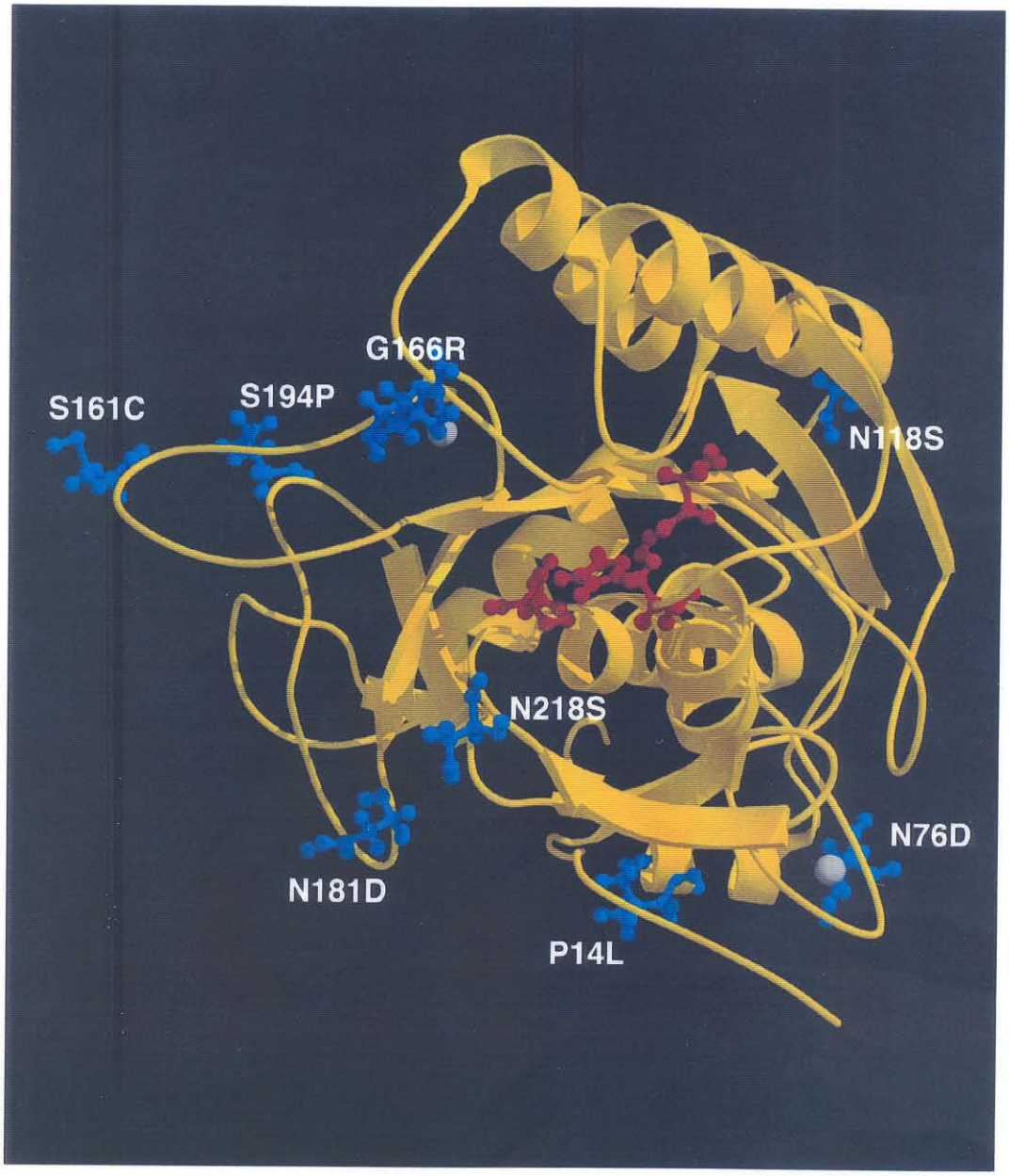
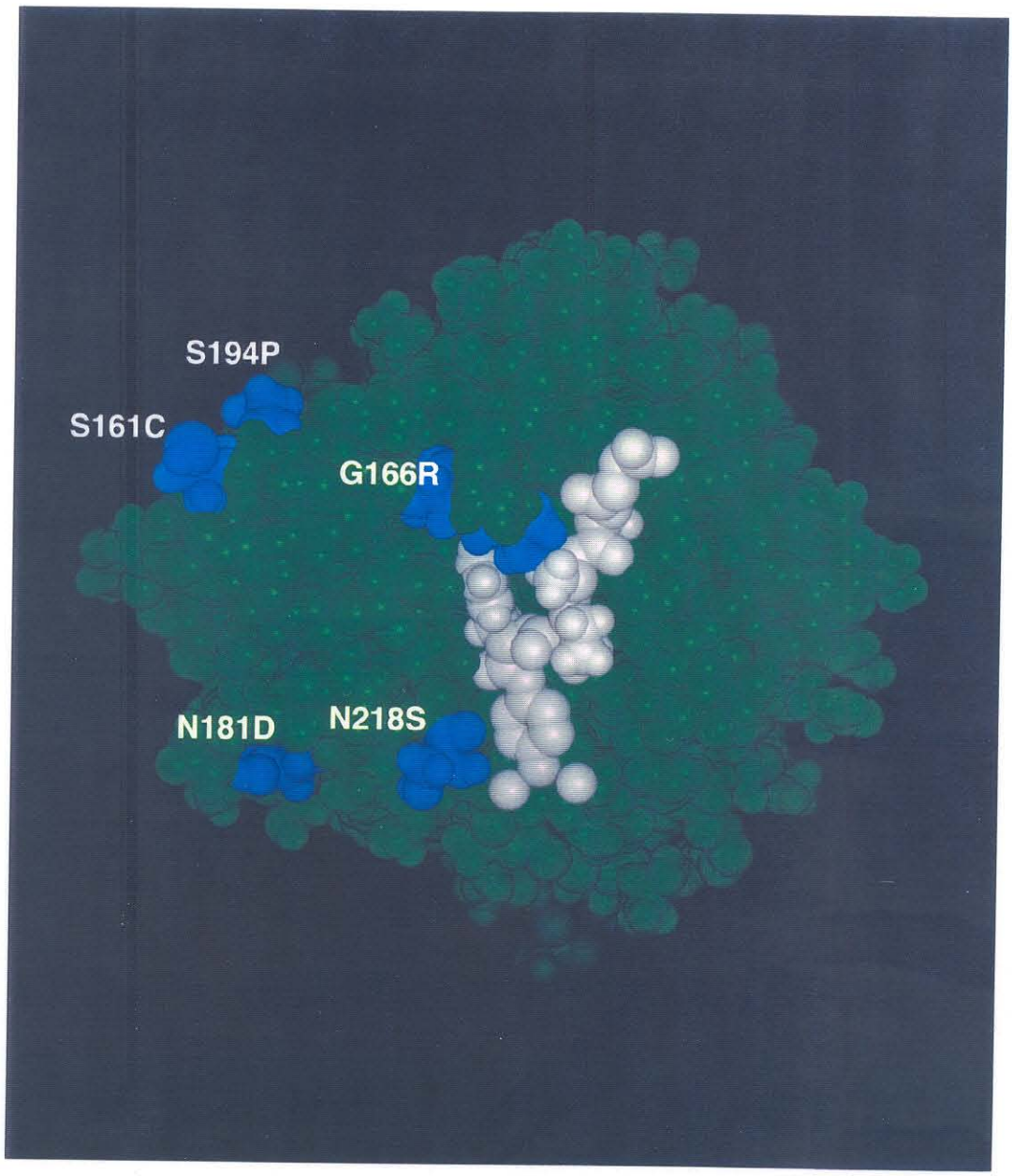
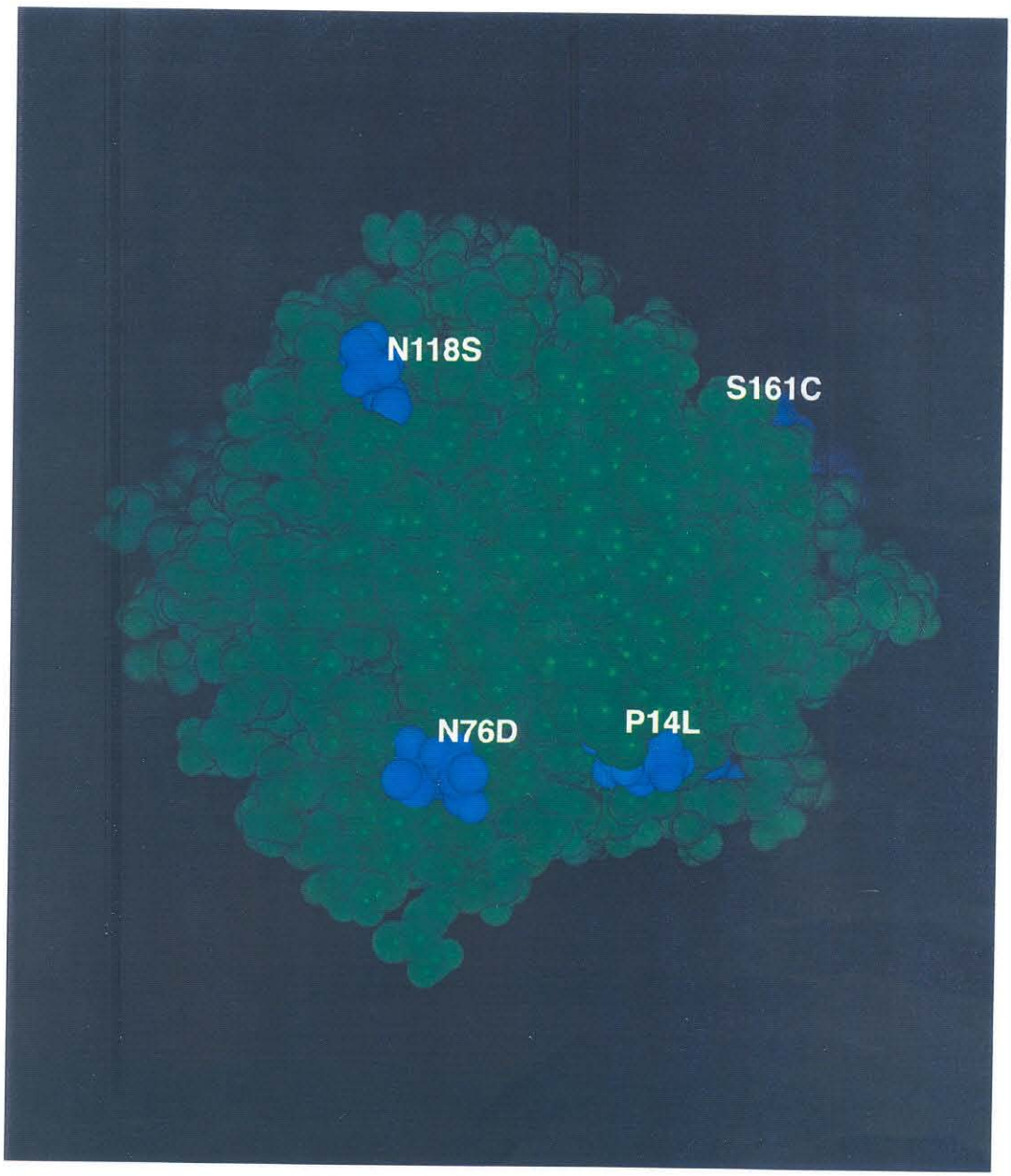

G166R
S161C
S194P
N118S
N218S
N76D
N181D
P14L

Figure 4.8. Space-filling model of the 5-3H5 subtilisin E variant. The eight stabilizing mutations are labeled and shown in cyan. The peptide substrate suc-AAPF-*p*Na is shown in gray. (b) is a view after turning (a) around 180°C.


S194P
S161C
G166R
N181D
N218S


N118S
S161C
N76D
P14L

REFERENCES

1. Jaenicke, H., Schurig, H., Beaucamp, N. and Ostendorp, R. (1996) Structure and stability of hyperstable proteins: glycolytic enzymes from hyperthermophilic bacterium *Thermotoga maritima. Advances in Protein Chemistry*, **48**, 181-269.

2. Querol, E., Perez-Pons, J. A. and Mozo-Villarias, A. (1996) Analysis of protein conformational characteristics related to thermostability. *Protein Engineering*, **9**, 265-271.

3. Vogt, G. and Argos, P. (1997) Protein thermal stability: hydrogen bonds or internal packing? *Folding and Design*, **2**, S40-S46.

4. Argos, P., Rossmann, M. G., Grau, U. M., Zuber, H., Frank, G. and Tratschin, J. D. (1979) *Biochemistry*, **18**, 5698-5703.

5. Menendez-Arias, L. and Argos, P. (1989) Engineering protein thermal stability - sequence statistics point to residue substitutions in alpha-helices. *J. Mol. Biol.*, **206**, 397-406.

6. Russell, R. J. M. and Taylor, G. L. (1995) Engineering thermostability: lessons from thermophilic proteins. *Curr. Opin. in Biotechnol.*, **6**, 370-374.

7. Fersht, A. R. and Serrano, L. (1993) Principles of protein stability derived from protein engineering experiments. *Curr. Opin. Struct. Biol.* **3**, 75-83.

8. Hecht, M. H., Sturtevant, J. M. and Sauer, R. T. (1984) Effects of single amino acid replacements on the thermal stability of the NH2-terminal domain of phage lambda-repressor. *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5685-5689.

9. Alber, T. and Wozniak, J. A. (1985) A genetic screen for mutations that increase the thermal stability phage-T4 lysozyme. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 747-750.

10. Liao, H., Mckenzie, T. and Hageman, R. (1986) Isolation of a thermostable enzyme variant by cloning and selection in thermophile. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 576-580.

11. Bryan, P.N., Rollence, M. L., Pantpliano, M. W., Wood, J., Finzel, B. C., Gilliland, G. L., Howard, A. J. and Poulos, T. L. (1986) Proteases of enhanced stability: Characterization of a thermostable variant of subtilisin, *Proteins Struct. Funct. Genet.* **1**, 326-334.

12. Cunningham, B. C. and Wells, J. A. (1987) Improvement in the alkaline stability of subtilisin using an efficient random mutagenesis and screening procedure. *Protein Engineering*, **1**, 319-325.

13. Pantoliano, M. W., Whitlow, M., Wood, J. F., Dodd, S. W., Hardman, K. D., Rollence, M. L.and Bryan, P. N. (1989) Large increases in general stability for subtilisin BPN' through incremental changes in the free energy of unfolding. *Biochemistry*, **28**, 7205-7213.

14. Zhang, X. J., Baase, W. A., Shoichet, B. K., Wilson, K. P., Matthews, B. W. (1995) Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive. *Protein Engineering* **8**, 1017-1022.

15. Shih, P. and Kirsch, J. F. (1995) Design and structural analysis of an engineered thermostable chicken lysozyme. *Protein Sci.* **4**, 2063-2067.

16. Arnold, F. H. (1998) Design by directed evolution. *Accounts of Chemical Research* **31**, 125-131.

17. Kuchner, O. and Arnold, F. H. (1997) Directed evolution of enzyme catalysts. *Trends in Biotechnology* **15**, 523-530.

18. Stahl, M.L., and Ferrari, E. (1984) Replacemnet of the *Bacillus subtilis* subtilisin structural gene with an *in vitro*-derived deletion mutation. *J. Bacteriol.* **158**, 411-418.

19. Wong, S. L. and Doi, R. H. (1986) Determination of the signal peptidase cleavage site in the preprosubtilisin of *Bacillus subtilis*. *J. Bio. Chem.* **261**, 10176-10181.

20. Ohta, Y. and Inouye, M. (1990) Pro-subtilisin E: purification and characterization of its autoprocessing to active subtilisin E *in vitro*. *Molecular Microbiology* **4**, 295-304.

21. Gallagher, T., Gilliland, G. and Bryan, P. (1995) The prosegment-subtilisin BPN' complex- crystal structure of a specific foldase. *Structure* **3**, 907-914.

22. Kuttner, G. A. S., Burger, E., Pfuller, B. and Frommel, C. (1987) Investigation on the long-term stability of enzymes in solution. *Biomed. Biochim. Acta* **46**, 39-52.

23. Mozhaev, V. V., Berezin, I. V. and Martinek, K. (1988) Structure-stability relationship in proteins - fundamental tasks and strategy for the development of stabilized enzyme catalysts for biotechnology. *CRC Crit. Rev. Biochem.* **23**, 235-281.

24. Varley, P. G. and Pain, R. H. (1991) Relationship between stability, dynamics and enzyme-activity in 3-phosphoglycerate kinases from yeast and *thermus thermophilus*. *J. Mol. Biol.* **220**, 531-538.

25. Shiochet, B. K., Baase, W. A., Kuroki, R. and Matthews, B. W. (1995) A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 452-456.

26. Takagi, H., Morinaga, Y., Ikemura, H. and Inouye, M. (1988) Mutant subtilisin E with enhanced protease activity obtained by site-directed mutagenesis. *J. Biol. Chem.* **263**, 19592-19596.

27. Takagi, H., Morinaga, Y., Ikemura, H. and Inouye, M. (1989) The role of Pro-239 in the catalysis and heat stability of subtilisin E. *J. Biochem.* **105**, 953-956.

28. Moore, J. C., Jin, H. M., Kuchner, O. and Arnold, F. H. (1997) Strategies for the *in vitro* evolution of protein function: enzyme evolution by random recombination of improved sequences. *J. Mol. Biol.* **272**, 336-347.

29. Zhao, H. and Arnold, F. H. (1997) Combinatorial protein design: strategies for screening protein libraries. *Curr. Opin. Struct. Biol.* **7**, 480-485.

30. Zhao, H., Giver, L., Shao, Z., Affholter, J. A. and Arnold, F. H. (1998) Molecular evolution by staggered extension process (StEP). *Nature Biotechnology* **16**, 258-262.

31. Mitchinson, C. and Wells, J. A. (1989) Protein engineering of disulfide bonds in subtilisin BPN'. *Biochemistry* **28**, 4807-4815.

32. Wells, J. A. and Powers, D. B. (1986) *In vivo* formation and stability of engineered disulfide bonds in subtilisin. *J. Biol. chem.* **261**, 6564-6570.

33. Vriend, G. and Eijsink, V. (1993) Prediction and analysis of structure, stability and unfolding of thermolysin-like proteases. *J. Computer-Aided Mole. Design* **7**, 367-396.

34. Eijsink, V., Burg, B., Vriend, G., Berendsen, H. and Venema, G. (1991) Thermostability of Bacillus subtilis neutral protease. *Biochemistry International* **24**, 517-525.

35. Frömmel, C., Hausdorf, G., Wöhne, W. E., Behnke, U. and Ruttloff, H. (1978) *Acta Biol. Med. Ger.* **37**, 1193-1204.

36. Meloun, B., Baudys, M., Kostka, V., Hausdorf, G., Frömmel, C. and Höhne, W. E. (1985) Complete primary structure of thermitase from *thermoactinomyces vulgaris* and its structural features related to the subtilisin-type proteinases. *FEBS Letters* **183**, 195-200.

37. Schreier, E., Fittkau, S., Höhne, W.E. (1984) Influence of synthetic peptide inhibitors on the thermal stability of thermitase, a serine proteinase from *Thermoactinomyces vulgaris. Int. J. Peptide Protein Res.* **23**, 134-141.

38. Chu, N. M., Chao, Y. and Bi, R. C. (1995) The 2Å crystal structure of subtilisin E with PMSF inhibitor. *Protein Eng.* **8**, 211-215.

39. Zhao, H. and Arnold, F. H. (1997) Functional and non-functional mutations distinguished by random recombination of homologous genes. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 7997-8000.

40. Lumry, R. and Eyring, H. (1954) Conformation changes of proteins. *J. Phys. Chem.* **58**, 110-120.

41. Braxton, S. and Wells, J. A. (1992) Incorporation of a stabilizing Ca2+-binding loop into subtilisin BPN'. *Biochemistry* **31**, 7796-7801

42. Frommel, C. and Sanders, C. (1989) Thermitase, a thermostable subtilisin: Comparison of predicted and experimental structures and the molecular cause of thermostability. *Proteins*, **5**, 22-37.

43. Siezen, R. and Leunissen, J. A. M. (1997) subtiliases: the superfamily of subtilisin-like serine proteases. *Protein Eng.* **6**, 501-523.

44. Matsuzawa, H., Tokugawa, K., Hamaoki, M., Mizoguchi, M., Taguchi, H., Terada, I., Kwon, S. T. and Ohta, T. (1988) Purification and characterization of aqualysin I (a thermophilic alkaline serine protease) produced by *Thermus aquaticus* YT-1. *Eur. J. Biochem.* **171**, 441-447.

45. Volkl, P., Markiewicz, P., Stetter, K. O. and Miller, J. H. (1994) The sequence of a subtilisin-type protease (aerolysin) from the hyperthermophilic archaeum *Pyrobaculum aerophilum* reveals sites important to thermostability. *Protein Science* **3**, 1329-1340.

46. Singleton, R., Middaugh, C.R. and McElroy, R.P. (1977) Comparison of proteins from thermophilic and nonthermophilic sources in terms of stratural parameters inferred from amino acid composition. *Int. J. Peptide Protein Res.* **10**, 39-50.

47. Ikai, A. (1980) Thermostability and aliphatic index in globular proteins. *J. Biochem.* (Tokyo) **88**, 1895-1898.

48. Ponnuswamy, P.K., Muthusamy, R. and Manavalan, P. (1982) Amino acid composition and protein thermostability. *Int. J. Biol. Macromol.* **4**, 186-191.

49. Menendez-Arias, L. and Argos, P. (1989) Engineering protein thermal stability - sequence statistics point to residue substitutions in alpha-helices. *J. Mol. Biol.* **206**, 397-406.

50. Matouschek, A., Kellis, J. T., Serrano, L. and Fersht, A. R. (1989) Mapping the transition-state and pathway of protein folding by protein engineering. *Nature*, **340**, 122-126.

51. Jackson, S. E. and Fersht, A. R. (1991) Folding of chymotrypsin inhibitor-2: 2. Influence of proline isomerization on the folding kinetics and thermodynamic characterization of the transition-state of folding. *Biochemistry* **30**, 10436-10443.

52. Eijsink, V.G., Vriend, G., van der Vinne, B., Hazes, B., van den Burg, B. and Venema, G. (1992) Effects of changing the interaction between subdomains on the thermostability of Bacillus neutral proteases. *Proteins: Structure, Function, and Genetics* **14**, 224-236.

Chapter 5

Functional and Non-functional Mutations Distinguished by Random Recombination of Homologous Genes

Huimin Zhao and Frances, H. Arnold

(appeared in *Proc. Natl. Acad. Sci. U.S.A.*, 1997, **94:**7997-8000)

Preface

According to neutral theory of evolution, the great majority of mutant substitutions are caused by random fixation through sampling drift of selectively neutral mutants. As a consequence, the sequences of evolutionarily-related proteins usually have diverged significantly, as we have seen for the subtilisin-like proteases. Thus, identification of those adaptive mutations (i.e., those affecting the growth and survival of the organism), neutral mutations and deleterious mutations or even identification of the important determinants in a specific case becomes an overwhelming task. This problem exists for enzymes evolved *in vitro* as well. While *in vitro* evolution can lead to the development of useful new protein functions, the responsible mutations almost always occur in a background of mutations which are neutral or even deleterious to the behavior(s) of interest.

To address these problems, I developed a convenient method based on what we have learned from *in vitro* evolution. As shown in this chapter, this approach has been used to identify two thermostable mutations out of ten mutations in a laboratory-evolved thermostable subtilisin E variant. This method involves the random recombination of homologous sequences followed by screening for the altered behavior. A similar approach, coupled with selection rather than screening, could be used to distinguish adaptive from neutral mutations.

In the end, I would like to point out that though the major technique used in this approach is an optimized high-fidelity DNA shuffling, other *in vitro* recombination techniques such as random priming recombination (RPR), staggered extension process (StEP) recombination may also be used, provided their associated mutagenesis rates are very low.

Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 7997–8000, July 1997
Evolution

Functional and nonfunctional mutations distinguished by random recombination of homologous genes

(adaptive evolution/neutral mutations/DNA shuffling/*in vitro* evolution)

HUIMIN ZHAO AND FRANCES H. ARNOLD*

Division of Chemistry and Chemical Engineering 210-41, California Institute of Technology, Pasadena, CA 91125

Communicated by Peter B. Dervan, California Institute of Technology, Pasadena, CA, May 16, 1997 (received for review March 3, 1997)

ABSTRACT **We describe a convenient method for distinguishing functional from nonfunctional or deleterious mutations in homologous genes. High fidelity *in vitro* gene recombination ("DNA shuffling") coupled with sequence analysis of a small sampling of the shuffled library exhibiting the evolved behavior allows identification of those mutations responsible for the behavior in a background of neutral and deleterious mutations. Functional mutations are expected to occur in 100% of the sequenced screened sample; neutral mutations are found in 50% on average, and deleterious mutations do not appear at all. When used to analyze 10 mutations in a laboratory-evolved gene encoding a thermostable subtilisin E, this method rapidly identified the two responsible for the observed protease thermostability; the remaining eight were neutral with respect to thermostability, within the precision of the screening assay. A similar approach, coupled with selection for growth and survival of the host organism, could be used to distinguish adaptive from neutral mutations.**

A fundamental problem in the study of evolutionarily related genes is to distinguish those mutations responsible for phenotypic differences from a background of neutral mutations that have little or no effect on function. In nature, adaptive changes may represent only a small fraction of all evolutionary events (1, 2). Some fraction of these may lead to functional differences measured *in vitro*. It is difficult to identify with certainty specific adaptive mutations; even mutations responsible for specific functional differences among proteins can evade identification when multiple nonfunctional mutations are present (3). Thus, the use of sequence comparisons is of limited utility in identifying the molecular mechanisms underlying differences in properties such as thermostability exhibited by proteins encoded by related genes.

This problem exists for sequences evolved *in vitro* as well. Although *in vitro* evolution can lead to the development of useful new protein functions, the responsible mutations almost always occur in a background of mutations that are neutral or even deleterious to the behavior(s) of interest. To derive key information on structure–activity relationships from these and nature's own experiments in molecular evolution, a convenient method for distinguishing functional from nonfunctional and/or deleterious mutations is needed. Site-directed mutagenesis requires the construction of multiple variants with different combinations of mutations and is far too laborious when many mutations are present. Here we demonstrate how a single experiment involving the random recombination of homologous sequences followed by screening for the altered behavior can be used to identify functional mutations. The experiment is based on the *in vitro* "DNA shuffling" method developed by Stemmer (4), with modifications to dramatically reduce the associated point mutagenesis rate (5). DNA shuffling of homologous genes creates a library of genes containing all possible combinations of mutations. As shown here, functional mutations are identified upon sequencing a set of the recombined genes that exhibit the evolved property. This approach, coupled with selection rather than screening, could be used to distinguish adaptive (i.e., those affecting the growth and survival of the organism) from neutral mutations.

MATERIALS AND METHODS

Restriction enzymes were purchased from Boehringer Mannheim. Succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide was from Sigma. *Bacillus subtilis* strain DB428 and *Bacillus* cloning vector pKWZ containing the subtilisin E gene were kindly provided by Dr. R. Doi of University of California, Davis, CA.

DNA Shuffling and Screening for Thermostability. High fidelity DNA shuffling of wild-type subtilisin E gene and 1E2A is described elsewhere (5). The PCR-amplified, reassembled product was purified by Wizard PCR prep kit (Promega), digested with *Bam*HI and *Nde*I, and electrophoresed in a 0.8% agarose gel. The 986-bp product was cut from the gel and purified by QIAEX II gel extraction kit (Qiagen, Chatsworth, CA). Products were ligated with vector generated by *Bam*HI–*Nde*I digestion of the pBE3 *Escherichia coli–B. subtilis* shuttle vector (Fig. 1). This gene library was amplified in *E. coli* HB101 and transferred into *B. subtilis* DB428-competent cells as described (6); 768 clones were picked with sterile toothpicks and grown in SG medium supplemented with 50 μg/ml kanamycin at 37°C for 24 h in eight 96-well plates. The cells were spun down, and samples of the supernatants were examined in the thermostability assay. Three replica 96-well assay plates were duplicated for each growth plate, with each well containing 5 μl of supernatant. Enzyme activity was measured as described (6) in 96-well plates using a Thermomax microplate reader (Molecular Devices). Activity measured at room temperature was used to calculate the fraction of active clones. Clones with activity <10% of that of wild type were scored as inactive. Initial activity (A_i) was measured on one assay plate after incubation at 65°C for 10 min by adding 100 μl of prewarmed (37°C) assay solution [0.2 mM succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide (suc-AAPF-pNA)/100 mM Tris·HCl, pH 8.0/10 mM $CaCl_2$] into each well. Residual activity (A_r) was measured after 40 min of incubation.

Sequence Analysis. Genes were individually purified from *B. subtilis* DB428 using a QIAprep spin plasmid miniprep kit (Qiagen) with the modifications that 2 mg/ml lysozyme was added to P1 buffer, and the cells were incubated for 5 min at 37°C, retransformed into competent *E. coli* HB101, and then purified again using QIAprep spin plasmid miniprep kit to obtain sequencing quality DNA. Sequencing was done on an

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*To whom reprint requests should be addressed. e-mail: frances@cheme.caltech.edu.

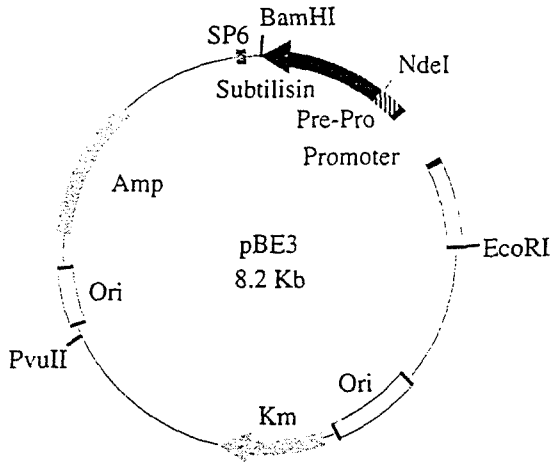


FIG. 1. *E. coli/B. subtilis* shuttle vector pBE3 containing the β-lactamase (Amp^r) gene and replicon from pGEM3 for growth in *E. coli* and the kanamycin nucleotidyl transferase (Km^r) gene and replicon from pUB110 for growth in *B. subtilis*. The subtilisin E gene including its natural promoter was subcloned from plasmid pKWZ by *Bam*HI and *Eco*RI restriction sites.

Applied Biosystems 373 DNA Sequencing System using the Dye Terminator Cycle Sequencing kit (Perkin–Elmer).

Construction of Subtilisin E Variants by Site-Directed Mutagenesis. Site-directed mutagenesis was carried out using an Altered Sites II *in vitro* mutagenesis kit (Promega). The wild-type gene was subcloned into pAlter-*Ex*1 vector with *Bam*HI and *Eco*RI. For each substitution (V93I, N109S, N181D, N218S), an oligonucleotide containing the desired mutation was used as the mutagenic primer. Mutants were screened by direct partial sequencing. Desired clones were subcloned back to the pBE3 shuttle vector digested with *Nde*I and *Bam*HI. All mutations were confirmed by DNA sequencing.

Enzyme Purification and Thermal Inactivation Assay. All of the subtilisin E variants were purified as described (6). Purified variants (≈10 μM) were dialyzed in 10 mM Tris·HCl/1 mM $CaCl_2$ (pH 8.0) at 4°C overnight. The samples were incubated at 65°C on a thermocycler. At various time intervals, 10-μl aliquots were added to 0.99 ml of activity assay solution (0.2 mM suc-AAPF-pNA/100 mM Tris·HCl/10 mM $CaCl_2$, pH 8.0, 37°C). Thermal inactivation at 65°C of all of the subtilisin E variants appears to obey first-order kinetics; $t_{1/2}$ is the half-life of enzyme activity at 65°C.

Specific Activity. Specific activity (unit/mg) was determined as the ratio of enzyme activity to protein concentration, under the same conditions as in thermal inactivation assay (0.2 mM suc-AAPF-pNA/100 mM Tris·HCl/10 mM $CaCl_2$, pH 8.0, 37°C). One unit will hydrolyze suc-AAPF-pNA to produce the color equivalent to 1.0 μmol of *p*-nitroanilide per min at pH 8.0, 37°C.

Differential Scanning Calorimetry. Melting temperatures (T_m) were determined by differential scanning calorimetry on a MicroCal MS1 calorimeter (Microcal, Amherst, MA). Experiments were carried out in 10 mM Tris·HCl/1 mM $CaCl_2$/1 mM phenylmethylsulfonylfluoride (an inhibitor of subtilisin E), pH 8.0. The temperature was increased at a rate of 1°C/min from 20°C to 90°C. The protein concentration (≈30 μM) was determined by absorbance at 280 nm (ε = 35886 $M^{-1}\cdot cm^{-1}$), and the sample size was 1.769 ml.

RESULTS AND DISCUSSION

Thermostable subtilisin E variant 1E2A was obtained by *in vitro* evolution of the wild-type enzyme (5, 6). At 65°C, its rate of thermoinactivation is about 8-fold slower than that of the wild-type protease. The mature gene encoding 1E2A differs from the wild-type sequence at 10 base positions; six mutations are synonymous with respect to the amino acid sequence, and four lead to amino acid changes (Table 1). All four nonsynonymous mutations exist in other naturally occurring subtilisins. To determine which mutations are responsible for the increased thermostability of the enzyme, we randomly recombined the 1E2A and wild-type subtilisin E genes to create a library of sequences containing all possible combinations of the mutations. This library was then expressed and screened to identify thermostable enzymes. In theory, the mutations responsible for the increased thermostability of the evolved enzyme should be present in 100% of the sequences coding for enzymes whose thermostability equals that of 1E2A subtilisin E. If deleterious mutations were present in the original evolved sequence, as can often happen when multiple mutations are accumulated in a single generation of directed evolution (7), then they would be removed in the recombined population that passed the screen for the evolved function. Thus, deleterious mutations would be present at 0% frequency in the sequenced genes. Mutations that have no effect on function should be present in roughly 50% of the screened sequences, if equimolar amounts of the parental genes are shuffled.

High Fidelity DNA Shuffling. As described (5), an equimolar mixture of the ≈1-kb wild-type and 1E2A genes was fragmented with DNase I, and the column-purified 20- to 50-bp fragments were reassembled (initially without primer) to a single PCR product of the correct size. The results of gene fragmentation, reassembly, and amplification are shown in Fig. 2. The overall rate of point mutagenesis associated with the high fidelity DNA shuffling of these genes is only 0.05% (5). After DNA shuffling, the gene library was subcloned and amplified in *E. coli* and then transferred into *B. subtilis*-competent cells for expression and screening. This was facilitated by an *E. coli/B. subtilis* shuttle vector, pBE3 (Fig. 1). This vector contains two sets of antibiotic resistance genes and replicons; one is functional in *E. coli* and the other in *B. subtilis* (although the kanamycin resistance gene is also expressed at low levels in *E. coli*). The transformation efficiency by ligated vectors is 3×10^5 cfu/μg for Ca^{2+}-prepared *E. coli* competent cells, and the transformation efficiency by plasmid DNA is 2×10^4 cfu/μg for *B. subtilis* DB428-competent cells. The direct cloning of recombined DNA into *B. subtilis* competent cells occurs at very low efficiency (less than a few hundred transformants per μg of DNA), and use of this shuttle vector greatly enhances the size of the recombined and/or randomly mutated gene libraries that can be created in *B. subtilis*.

Sequence Analysis. As a control, 10 unscreened clones from the recombined gene library were selected at random and sequenced (5). The results are summarized in Fig. 3*a*. All except clone 7 result from different recombination events. (Clone 7 is the intact 1E2A parent sequence.) The frequency

Table 1. DNA and amino acid substitutions in thermostable 1E2A subtilisin E (mature gene)

Base	Base substitution	Position in codon	Amino acid	Amino acid substitution
484	A → G	3	10	synonymous
520	A → T	3	22	synonymous
731	G → A	1	93	Val → Ile
745	T → C	3	97	synonymous
780	A → G	2	109	Asn → Ser
995	A → G	1	181	Asn → Asp
1107	A → G	2	218	Asn → Ser
1141	A → T	3	229	synonymous
1153	A → G	3	233	synonymous
1189	A → G	3	245	synonymous

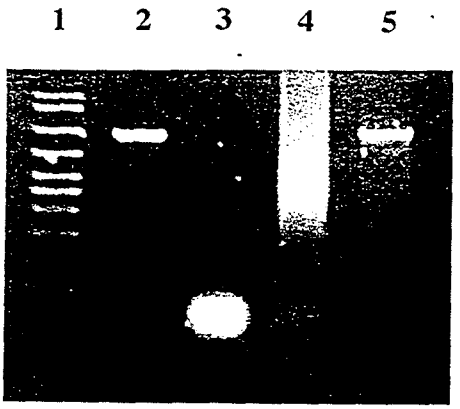


FIG. 2. DNA agarose gel (2%) showing process of high fidelity DNA shuffling of 1E2A and wild-type (wt) subtilisin E genes (5). Lanes: 1, AmpliSize DNA Size standards (Bio-Rad), from top to bottom: 2000, 1500, 1000, 700, 500, 400, 300, 200, 100, and 50 bp; 2, 1:1 mixture of 986-bp fragments from wt and 1E2A before DNase I digestion; 3, 1:1 DNA mixture of wt and 1E2A after 2 min of DNase I digestion in the presence of Mn^{2+}; 4, fragment reassembly with *Pfu* polymerase; and 5, PCR amplification of the reassembly product with *Taq*/*Pfu* (1:1). Primers 5'-CCGAG CGTTG CATAT GTGGA AG-3' (underlined sequence is *Nde*I restriction site) and 5'-CGACT CTAGA GGATC CGATT C-3' (underlined sequence is *Bam*HI restriction site) were used.

of occurrence of a particular point mutation from parent 1E2A in the shuffled genes ranged from 30 to 70%, fluctuating around the expected value of 50%. All 10 mutations can be recombined, even those that are only 12 bp apart. There is clear linkage, however, between such closely spaced mutations.

We then assayed the rates of subtilisin thermoinactivation at 65°C for 768 clones picked from SG agar plates with 50 μg/ml kanamycin and grown in eight 96-well plates. Initial activity (A_i) was measured on the culture supernatants in each well after incubation at 65°C for 10 min. Residual activity (A_r) was measured after 40 min of incubation. The normalized residual activity (A_r/A_i) was used as an index of thermostability. Thermostabilities measured on a typical 96-well plate are shown in Fig. 4, plotted in descending order. Approximately 23% of the clones exhibited thermostability comparable to 1E2A, which indicates immediately that only two mutations [$(1/2)^2$ = 25%] are responsible for the increased thermostability.† Twenty clones exhibiting the highest thermostability were selected, and their kinetics of thermoinactivation were verified in a second assay on the culture supernatants (assay used for purified enzymes as described in *Materials and Methods*). No false-positives were found. Genes from the 10 most thermostable were sequenced (Fig. 3*b*). Only two new point mutations were found among these 10 recombined, screened genes, as compared with five in the unscreened population (Fig. 3*a*). Most point mutations are deleterious to thermostability (data not shown). The lower rate of point mutations found in the screened population reflects the "cleansing" effect of the screen (8).

The two nonsynonymous mutations leading to amino acid substitutions N181D and N218S were found in all 10 of the recombined clones exhibiting high thermostability. The remaining eight mutations occurred at frequencies of 20–80%, very similar to the frequencies observed in the unscreened control sample (Fig. 4*a*). Thus, we can conclude that N181D and N218S are functional mutations and that the rest are neutral with respect to thermostability. In addition, we can conclude that none of the 10 mutations is deleterious, at least within the precision of the assay method, which is sensitive to changes on the order of 15% in deactivation rate.

Stability and Activity of Specific Subtilisin E Variants. To verify this result, all four nonsynonymous single mutants and the N181D+ N218S double mutant were constructed by site-directed

†The probability that any given mutation will appear in the randomly recombined population is 1/2. Thus, the probability that N-specific (functional) mutations will appear together in a sequence is $(1/2)^N$.

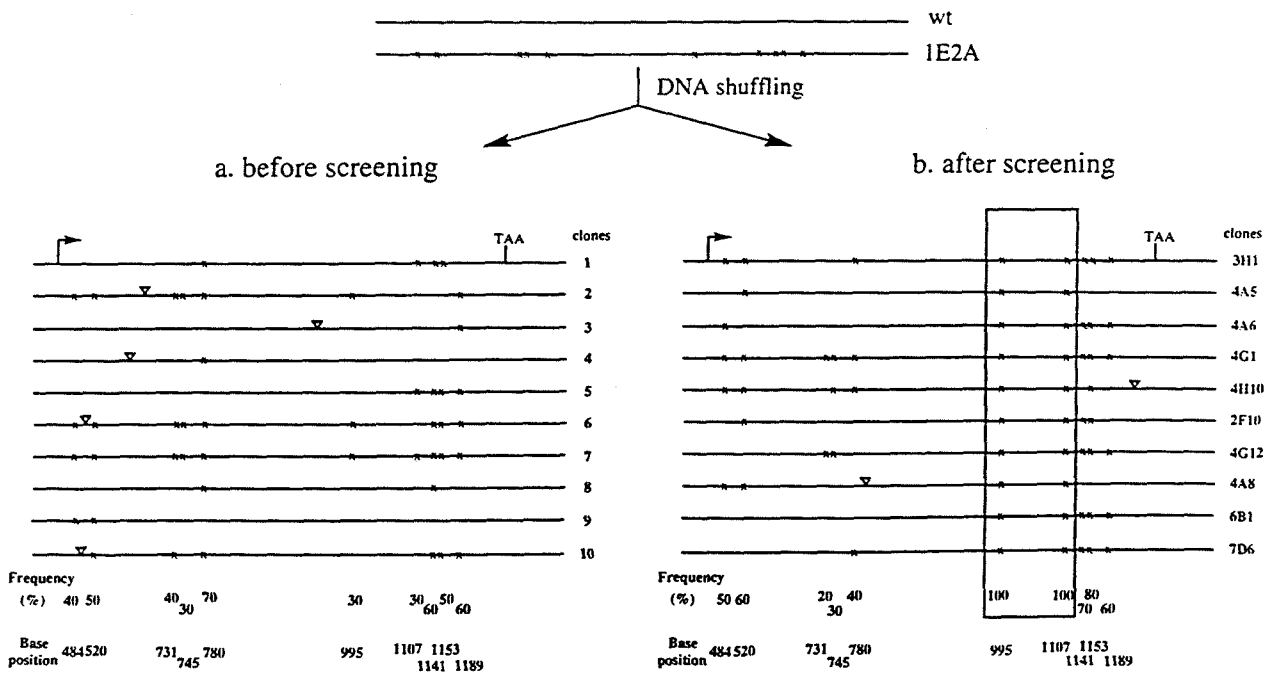


FIG. 3. Sequence analysis of randomly recombined gene libraries before (*a*) and after (*b*) screening for protease thermostability. Lines represent 986 bp of subtilisin E gene including 45 nt of its prosequence, the entire mature sequence, and 113 nt after the stop codon. Crosses indicate positions of mutations from 1E2A, and triangles indicate positions of new point mutations introduced during the DNA shuffling procedure.

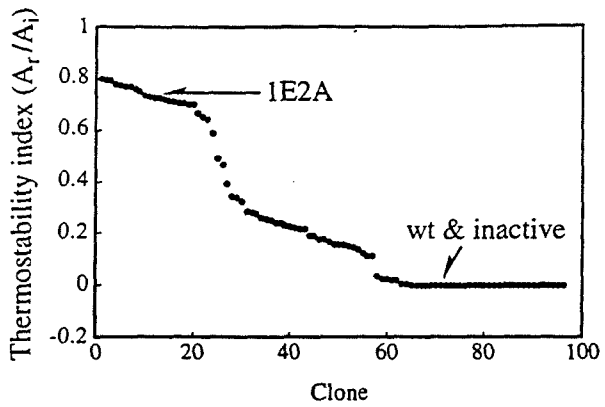


FIG. 4. Results of screening a typical 96-well plate for residual activity after incubation at 65°C for 10 (A_i) and 40 min (A_r). A_r/A_i was used as an index of the enzyme's thermostability. Data are sorted and plotted in descending order.

mutagenesis and characterized with respect to their activities and thermostabilities (Table 2). 1E2A and double mutant N181D+N218S have similar half-lives of thermoinactivation at 65°C and similar melting temperatures, T_m. The half-lives at 65°C of single mutants N181D and N218S are ≈3- and 2-fold greater than that of wild-type subtilisin E, respectively, and their T_ms are 3.7 and 3.2°C higher. Thus, both functional mutations identified by random recombination also confer thermostability in the wild-type background. Mutation N218S was found previously in a thermostable variant of subtilisin BPN′ (9). N181D and N218S also are found in thermostable subtilisin E homologs thermitase (10), proteinase K (10), aerolysin (11), and Ak1 proteinase (12). In contrast, the half-lives at 65°C of the single mutants V93I and N109S are very similar to that of wild-type, as are their T_m values. None of the four amino acid substitutions decreased the specific activity of the enzyme; N218S increased both specific activity and stability (Table 2).

Although thermophilic organisms in nature have evolved extremely stable enzymes, high stability has often come at the cost of specific activity at the lower temperature. This trade-off, however, may reflect evolutionary drift or even certain requirements for function within an adaptive biological network. Thus, it may not necessarily hold true for *in vitro* evolution, where the enzyme is decoupled from its natural function. Mutations that increase thermostability without decreasing specific activity are not extremely rare. Furthermore, it should be possible to combine evolved properties of thermostability and enhanced activity.

The method outlined above can be used to identify functional mutations in far more complicated systems than the laboratory-evolved thermostable subtilisin E used here for demonstration. For example, thermostable and nonthermostable members of a protein family often differ not at four but at dozens of amino acids. To identify those amino acid substitutions conferring thermostability using site-directed mutagenesis would require construction and characterization of an impractically large number of variants. For all practical purposes, functional mutations can be identified by screening a randomly recombined gene library when the number of functional mutations is on the order of 10 or less in a background of any number of neutral (or deleterious) mutations (provided there is sufficient homology between the genes for the *in vitro* recombination to succeed). With 10 functional mutations, the frequency of the thermostable phenotype would be $(1/2)^{10} \approx 0.1\%$. Thus, ≈10 thermostable clones could be identified by screening 10,000 clones, which is quite feasible using a simple 96-well plate assay. It is likely that an even larger number of functional mutations could be distinguished by coupling the evolved phenotype to a functional selection rather than a screen.

Table 2. Stabilities and activities of purified subtilisin E variants

Variant	$t_{1/2}$ at 65°C*, min	T_m†, °C	Specific activity, unit/mg‡
Wild type	5.1 ± 0.2	68.1	17.2 ± 0.1
1E2A	42.8 ± 0.1	74.4	33.7 ± 0.7
V93I	5.0 ± 0.1	68.1	21.6 ± 0.2
N109S	5.2 ± 0.1	68.2	16.1 ± 0.4
N181D	16.5 ± 0.6	71.8	18.0 ± 0.6
N218S	10.9 ± 0.1	71.3	38.6 ± 0.6
N218S + N181D	49.9 ± 0.8	74.6	38.4 ± 0.1

*Half-life of thermoinactivation, pH 8.0, 10 mM $CaCl_2$.
†Melting temperature, as measured by differential scanning calorimetry, pH 8.0, 1 mM $CaCl_2$.
‡Specific activity toward suc-AAPF-pNA at pH 8.0.

The experiment we have described differs in several important aspects from simple back-crossing with an excess of a parental sequence (or wild type) by DNA shuffling (4), which will effectively flush out both neutral and deleterious mutations because of the statistical preference for incorporating the parental sequence in the recombined genes (J. C. Moore, H. M. Jin, O. Kuchner, and F.H.A., unpublished work). Using an excess of parental sequence dramatically decreases the frequency of clones exhibiting the evolved property and therefore greatly increases the screening requirement. For example, in a library created by back-crossing a gene containing five functional mutations with a 10-fold excess of the wild-type sequence, the frequency of the evolved phenotype would be only $(1/11)^5$, or 6.2×10^{-6}. Furthermore, the relatively high rate of mutagenesis in the original Stemmer protocol, 0.7% (corresponding to ≈7 new mutations per gene) would mask the relationship between evolved phenotype and functional mutations that forms the basis of this experiment. Finally, this experiment is capable of distinguishing the deleterious mutations (0% expected frequency) from those that are neutral (50% frequency). Such a distinction would not be possible if the Stemmer method were used to back-cross with excess wild-type sequence.

We thank Prof. Steven Benner (University of Florida) for helpful discussions. This work was supported by the U. S. Department of Energy's program in Biological and Chemical Technologies Research within the Office of Industrial Technologies, Energy Efficiency and Renewables.

1. Stewart, C. B., Schilling, J. W. & Wilson, A. C. (1987) *Nature (London)* **330**, 401–404.
2. Perutz, M. F. (1983) *Mol. Biol. Evol.* **1**, 1–28.
3. Benner, S. A. (1989) *Chem. Rev.* **89**, 789–806.
4. Stemmer, W. P. C. (1994) *Nature (London)* **370**, 389–391.
5. Zhao, H. & Arnold, F. H. (1997) *Nucleic Acids Res.* **25**, 1307–1308.
6. Chen, K. & Arnold, F. H. (1991) *Bio/Technology* **9**, 1073–1077.
7. Moore, J. C. & Arnold, F. H. (1996) *Nat. Biotechnol.* **14**, 458–467.
8. Suzuki, M., Christians, F. C., Kim, B., Skandalis, A., Black, M. E. & Loeb, L. A. (1996) *Mol. Diversity* **2**, 111–118.
9. Bryan, P. N., Rollence, M. L., Pantoliano, M. W., Wood, J., Finzel, B. C., Gilliland, G. L., Howard, A. J. & Poulos, T. L. (1986) *Proteins Struct. Funct. Genet.* **1**, 326–334.
10. Siezen, R. J., Devos, W. M., Leunissen, J. A. M. & Dijkstra, B. W. (1991) *Protein Eng.* **4**, 719–737.
11. Volkl, P., Markiewicz, P., Stetter, K. O. & Miller, J. H. (1994) *Protein Sci.* **3**, 1329–1340.
12. Maciver, B., Mchale, R. H., Saul, D. J. & Bergquist, P. L. (1994) *Appl. Environ. Microbiol.* **60**, 3981–3988.

Appendices

Appendix A

Methods for Optimizing Industrial Enzymes by Directed Evolution

(Huimin Zhao, Jeffrey C. Moore and Frances H. Arnold)

(*ASM Manual of Industrial Microbiology and Biotechnology*, in press)

METHODS FOR OPTIMIZING INDUSTRIAL ENZYMES BY DIRECTED EVOLUTION

Huimin Zhao[1], Jeffrey C. Moore[2] and Frances H. Arnold[1]

[1] California Institute of Technology

MC 210-41

Pasadena CA 91125

Fax: (818) 568-8743

(818) 395-4162

[2] Current address: Merck & Co., Inc.

RY80Y-110

P. O. Box 2000

Rahway, NJ 07065

Fax: (908) 594-4400

(908) 594-4836

Table of Contents

Introduction

Enzymes exhibit exquisite catalytic power unmatched by conventional catalysts. Their many applications range from serving as catalysts for chemical synthesis to use in diagnostic testing, foods and pharmaceuticals. However, naturally occurring enzymes are often not well suited for industrial applications. Problems include enzyme instability and low catalytic activity on nonnatural substrates and in nonnatural environments. Because the extensive structural and mechanistic information required to guide rational approaches to engineering improved enzymes is available for only a tiny fraction of known sequences, alternative approaches are needed. Directed evolution has proven very effective for modifying enzymes in the absence of such knowledge. By directed evolution we have been able to tailor enzyme functions never required in the natural environment. Properties that can be improved include stability, catalytic activity, activity towards new substrates, expression level in a heterologous host, and others (1).

This laboratory has focused on developing both the methods and strategies for design by directed evolution and demonstrating them by engineering novel, industrially-useful enzymes (2,3). As outlined in Fig. 1, the first step in directed evolution is to create molecular diversity starting from a target gene or a family of related genes. The diversity can be created by introducing mutations and/or by recombination. The gene products are sorted by screening or selection, and those genes encoding improved products can be returned for further generations of evolution. This evolutionary process can be repeated until the goal is achieved (or until there is no further improvement).

We will illustrate the process of directed evolution and the methods presented here with two enzymes: the serine protease subtilisin E and p-nitrobenzyl esterase. P-nitrobenzyl esterase can be used during the synthesis of certain ß-lactam antibiotics to deprotect a p-nitrobenzyl ester intermediate (4). We will focus on the three major components of directed evolution efforts: 1) creating diversity by random point mutagenesis and/or *in vitro* recombination; 2) sorting the resulting enzyme libraries, and

3) data analysis, required for the accurate selection of positives and to improve further experiments.

Creating enzyme diversity

In contrast to natural evolution, directed evolution has a defined goal, and the key processes--mutation, recombination and selection or screening--must be carefully controlled by the experimenter. Due to the strict limitations on library sorting imposed by screening and selection, the mutation rate must be tuned to the power of the sorting method (1,2). Useful, reasonably-sized libraries can be created by introducing multiple mutations in a particular region or across a limited number of positions (e.g., by combinatorial mutagenesis using oligonucleotide cassettes (5)). However, such an approach will exclude the many useful solutions found in unexpected places (3). Protein engineers are becoming increasingly aware that many protein functions are not confined to a small number of amino acids, but are affected by residues far from active sites. We therefore usually try to evolve the entire gene, rather than to target particular positions. Because we can search only 10^4 -10^5 enzyme variants even with a good screening method, single- amino-acid-substitution libraries are preferred. Larger numbers of variants can of course be searched by good selections (10^8-10^9), allowing one to examine double- or even triple-amino-acid-substitution libraries, depending on the sequence length (1).

Most random mutagenesis methods create mutations in single bases. Due to the degeneracy of the genetic code, this provides access to only about 6 amino acid substitutions instead of 19, thus reducing the potential diversity significantly. In addition, many mutagenesis methods are not really random, further limiting the number of amino acid substitutions actually accessible in a given experiment. For example, the mutagenesis method we use, error-prone PCR (6), shows a strong bias for transitions over transversions. *In vitro* recombination methods include DNA shuffling (7), random-

priming recombination (8) and the staggered extension process (StEP) (9). All these methods have a (controllable) level of associated point mutagenesis. With recombination, directed evolution can begin from multiple, closely-related starting points rather than a single sequence (10). Random recombination of existing functional sequences (i.e., homologous enzymes) will create another level of diversity that point mutagenesis cannot generate. A library of recombined genes may provide an excellent starting point for creating novel functions.

Random Point Mutagenesis

In random point mutagenesis, the two most important factors to consider are mutation frequency and mutation bias. Mutation frequency is the average number of mutations per gene and is usually reported as a percentage. The target mutation frequency can be calculated from the length of the DNA coding sequence and an estimate of the number of mutations per sequence desired. For instance, a desirable mutation level for directed evolution is ~2-3 base substitutions per gene (2,11). Thus for a 1 kb sequence, the mutation frequency should be ~0.2-0.3%. Although numerous methods for making random DNA mutations exist, we choose to use error-prone PCR because the procedure is simple, rapid, robust, and, most importantly, the mutation frequency can be precisely controlled.

Error-prone PCR does not create truly random DNA substitutions (for instance, a common bias of error-prone PCR is a high occurrence of A to G substitutions). Some bias can be tolerated in directed evolution experiments, while large variations in error frequency usually cannot. Bias affects the **location** of mutations (i.e., mutations at AT base pairs occur much more frequently than mutations at GC base pairs) as well as their **type** (i.e., A is more frequently substituted with G). To a first approximation, however, the A's are well distributed throughout the gene, and mutations are still occurring throughout the protein. Additionally, error-prone PCR mutagenizes the A's on both

DNA strands during synthesis, effectively doubling the location of mutations. PCR modifications reduce bias, but do not eliminate it (12,13).

The error-prone PCR method we routinely use was originally outlined by Leung and coworkers (10) and further examined by Cadwell and Joyce (12) and Shafikhani *et al.* (13). A series of mutation frequencies ranging from 0.11% to 2% have been obtained under different reaction conditions. In particular, the mutation frequency can be controlled (from 0.11% to 0.49%) simply by adjusting the concentration of manganese in the reaction mixture (13). The following protocol is used in the directed evolution of subtilisin E. The overall mutation frequency is ~0.2%, or two base changes per gene, on average.

Error-Prone PCR protocol:

1. Prepare purified plasmid DNA.
2. Prepare a 10x mutagenic buffer containing 70 mM $MgCl_2$, 500 mM KCl, 100 mM Tris-HCl (pH 8.3 at 25 °C) and 0.1% (wt/vol) gelatin.
3. Prepare a 10x dNTP mix containing 2 mM dGTP, 2mM dATP, 10 mM dCTP, and 10 mM TTP.
4. Prepare a solution of 5 mM $MnCl_2$. (Do not combine with the 10x PCR buffer, which would result in precipitation.)
5. Combine 10 µl of 10x mutagenic PCR buffer, 10 µl of 10x dNTP mix, 30-50 pmols of each primer, 2 fmoles of template DNA (~10 ng for a 8 kb-plasmid), and an amount of distilled H_2O that brings the volume up to 96 µl.
6. Add 3 µl of 5 mM $MnCl_2$. Mix well.
7. Add 1µl of *Taq* polymerase (5U/ul, Promega). Mix gently. No mineral oil is needed if the lids are also heated.

8. Run a PCR program: 14 cycles of 94 °C for 30s, 50 °C for 30s and 72 °C for 30s. (For a gene of more than 1kb, increase the extension time at 72 °C accordingly.)

9. Purify the reaction products by Promega PCR DNA purification kit.

10. Run a small portion of the purified products on an agarose gel to estimate the yield of full-length gene (typically a yield of 0.5-1.0 μg per reaction is obtained).

11. Digest with appropriate restriction enzymes and clone into expression vector.

The overall mutation frequency in error-prone PCR is the product of three parameters: the error rate (fidelity) of the polymerase the reaction conditions, the length of the mutagenized gene and the number of effective doubling cycles. Varying the concentration of $MnCl_2$ will only affect the error rate of the polymerase. In the above protocol, the overall mutation frequency is ~0.1% at 0 mM $MnCl_2$ and ~0.5% at 0.5 mM $MnCl_2$. The number of effective doublings is another easily adjustable parameter. This can be altered by adjusting either the PCR cycle numbers or the ratio between template and primers.

Because polymerase fidelity also depends on the nature of target sequences, different mutation rates will be observed on different sequences, even when the exact same reaction conditions are used. A straightforward way to assess the overall mutation frequency and the nature of mutations is to sequence a few random clones from the amplified population. However, sequencing is time-consuming and expensive. A simple and efficient alternative is to estimate the mutation frequency from the fraction of active clones from the amplified population (13,14). The activity profiles of small samplings from libraries of subtilisin E variants produced by error-prone PCR with different $MnCl_2$ concentrations are plotted in Fig. 2. Clones exhibiting at least 10% of wild type subtilisin activity have been scored as active. The fraction of active clones varied from 90% at 0 mM $MnCl_2$ to 30% at 0.5 mM. Even a concentration difference as small as 0.05 mM $MnCl_2$ makes a clear difference in the activity profile. For subtilisin E, 65% active

clones corresponds to a mutation frequency of two base changes per gene, or 0.2%. Even though different enzymes will show different activity profiles for different mutation rates, the fraction of active clones is a convenient index of mutation frequency and can be used as a diagnostic check for the successful creation of the desired randomly-mutated library.

In Vitro Recombination

A convenient method for performing random recombination of DNA sequences by PCR has been described by Stemmer (7). This method, known as 'DNA shuffling', involves enzymatic digestion of the parental DNA into short fragments followed by reassembly of the fragments into full length genes. Since the fragments are free to associate with complementary fragments from other, similar genes, mutations from one parent can be combined with mutations from other parent(s) to generate novel combinations. The technique can also remove mutations which are deleterious or do not contribute to the desired property (neutral). The original method has been modified to simplify it and to yield better control over the associated mutagenic rate (14,15). We have recently developed two alternative random recombination methods that do not involve digestion of the parent genes (8,9). The recombination effected by these methods is similar to that obtained by DNA shuffling.

When recombining genes from multiple improved variants, it is important to first consider the recombined library size (16). The various *in vitro* recombination methods can recombine any number of parent genes. However, the resulting libraries may be impossibly large to find further improvements in function. If each mutation sorts independently from the others, the probability of generating a sequence containing all mutations M present in N separate sequences is $1/N^M$ (16). When four improved pNB esterases are DNA-shuffled (each parent contained one mutation responsible for increased activity), the probability of producing the sequence containing all four mutations is 1 in 256, assuming all mutations are sorted randomly. This probability, and

therefore the screening requirements, decreases rapidly as the number of mutations and sequences increase. For example, consider the case of shuffling the four most improved pNB esterases versus that of shuffling a pool containing all 15 sequences deemed more active than the parent for that generation. The latter pool of 15 improved sequences contained the 4 most active variants as well as 11 more variants that are less active, but still better than the parent sequence. While recombining all 15 sequences should eventually lead to a more 'fit' enzyme, the screening requirements may be beyond the screening capability. The four-variant pool will in fact provide the most fit recombined variants if screening is limited to a few hundred or thousand clones.

The *in vitro* recombination protocol included here is a modified high-fidelity DNA shuffling protocol (14) with a mutagenic rate of only 0.05%.

DNA shuffling protocol

1. Preparation of genes to be shuffled by restriction enzyme digestion of plasmid DNA. The DNA fragments were purified from an agarose gel after electrophoresis and dissolved in 10 mM Tris-HCl (pH 7.4). DNA concentrations were estimated by electrophoresis or spectrophotometrically, and the fragments were mixed 1:1 for a total of ~2 μg. (More DNA is preferred but it should not exceed 5 μg. Otherwise the following digestion reaction will not be complete.)

2. DNase I digestion in the presence of Mn^{2+}. The mixture was diluted to 45 μl in 10 mM Tris-HCl (pH 7.4), and 5 μl of 10x digestion buffer (500 mM Tris-HCl, pH 7.4, 100 mM $MnCl_2$) was added. This mixture was equilibrated at 15 °C for 5 min on a thermocycler before 0.30 U of DNase I was added (Boehringer Mannheim, 10 U/ul). The digestion was done at 15 °C and terminated after 2 min by heating at 90 °C for 10 min.

That the fragments were less than 50 bp was confirmed on a 2% agarose gel before purification on a Centri-Sep column (Princeton Separations, Inc., Adelphia, NJ). (In Stemmer's published protocol (7), this step requires Mg^{2+}, which usually gives smear such that the progress of the digestion must be closely monitored, and small fragments need to be further purified from the agarose gel. In the presence of Mn^{2+}, quite uniform fragments (20-50bp) are usually produced which can be purified through a column, thus greatly simplifying the procedure (14,15)).

3. Fragment reassembly. 10 µl of spin-column purified fragments were added to 10 µl of 2x PCR premix (five-fold diluted cloned *Pfu* buffer, 0.4 mM each dNTP, 0.06 U/µl cloned *Pfu* polymerase (Stratagene, La Jolla, CA)). The reaction mixture was overlaid with 30 µl of mineral oil (no oil is needed if lids are also heated). PCR program: 3 min 96 °C followed by 40 cycles of 1 min 94 °C, 1 min 55 °C, 1 min + 5 sec/cycle 72 °C, followed by 7 min at 72 °C.

4. PCR amplification of reassembled products. 1 µl of this reaction was used as template in a 25-cycle PCR reaction. PCR conditions (100µl final volume): 30 µmol of each primer, 1x *Taq* buffer, 1.5 mM $MgCl_2$, 0.2mM of each dNTP and 2.5 U of *Taq/Pfu* (1:1) mixture. PCR program: 2 min 96 °C, 10 cycles of 30 sec 94 °C, 30 sec 55 °C, 45 sec 72 °C, followed by another 14 cycles of 30 sec 94 °C, 30 sec 55 °C, 45 sec + 20 sec/cycle 72 °C, and finally 7 min 72 °C. This program gives a single band at the correct size.

5. Purify the reaction products by Promega PCR DNA purification kit.

6. Run a small portion of the purified products on an agarose gel to estimate the yield of full-length gene (typically a yield of 0.5-1.0 µg per reaction is obtained).

7. Digest with appropriate restriction enzymes and clone into expression vector.

As demonstrated elsewhere (14), the mutagenic frequency can be controlled over a wide range, from 0.05% to 0.7%, by the inclusion of Mn^{2+} or Mg^{2+}, by the choice of DNA polymerase and/or by using restriction enzyme digestion to prepare the starting DNA. If high fidelity is not required, it may be more convenient to prepare sufficient starting DNA by conventional PCR amplification (step 1). *Taq* or other polymerases can be used in the reassembly and final PCR amplification steps.

The finite error frequency associated with DNA shuffling has been used by Stemmer and coworkers to supply point mutations for recombination and evolution. When starting with a single sequence, we prefer to use error-prone PCR under controlled conditions to generate libraries of variants. When more than one improved sequence is identified during screening, they can be recombined.

Library sorting

Given a thoughtful approach to generating the mutant library, the development of an efficient method to search for the desired properties is probably the single most important element determining the success of a directed evolution experiment. Libraries can be sorted by selection or screening. Various *in vivo* and *in vitro* selection methods have been established (17). The prerequisite to biological selection is the generation of a function which confers a growth or survival advantage to the host organism. Selection can be a very efficient search mechanism, allowing an exhaustive search of libraries of 10^6 and more variants. The disadvantage of a biological selection is that the property or protein of interest cannot be decoupled from the biological function. Thus it can be difficult or even impossible to explore novel functions such as activity in a highly nonnatural environment (18).

The sorting method chosen should reflect the desired features as much as possible. Solutions can arrive through unanticipated mechanisms--an all too-common experience when selections are used! It is also often the case that improperly designed screens allow uninteresting mutants to pass the sorting criteria. We have found, for example, that bacteria grown in a controlled laboratory medium exhibit substantial increases in activity as a result of directed evolution efforts. When the bacteria are transferred to the industrial growth medium and conditions, however, the increased activity may no longer be apparent. Thus it is important to include in the sorting as many of the tasks the enzyme (and the organism) is expected to accomplish as possible. This can be accomplished in a single, comprehensive screen, or by using a tiered screening protocol, in which positives identified during a rapid assay are verified in a more rigorous series of tests.

Selection

In a biological selection, the enzyme is linked to the host organism survival such that only those organisms possessing the desired trait can grow. When the environmental conditions are set properly, the wild type (or parent) and mutants performing more poorly than wild type do not survive; all surviving colonies are positives. Large libraries (whose sizes are limited only by transformation efficiency) can be sorted in this way. Selections can be useful for evolving drug resistance enzymes, enzymes responsible for providing nutrients to supplement an auxotrophy, and thermostability when an essential enzyme in a thermophile can be replaced by the enzyme of interest, to name a few common examples.

Unfortunately, designing a biological selection around many enzymes activities is difficult, if not impossible. Selections often takes a large amount of time to implement, with no guarantee of success. Additionally, many desired features are nonnatural and by definition cannot be coupled to the growth and survival of the host organism. Where a selection is possible, care must be taken to ensure that positives are a result of mutations

in the targeted enzyme, rather than changes somewhere else in the organism. Other *in vitro* selection approaches for novel catalysts that include binding to a transition state or substrate analog attached to a column or catalysis of a single reaction followed by trapping are also problematic because they do not probe catalytic efficiency directly.

Screening

Screening is the most flexible sorting method for directed evolution (18). Experimental conditions can be tailored to meet the desired criteria. Additionally, the screens are often done in much the same way the enzyme is traditionally assayed (e.g., spectrophotometrically), so that one can be implemented relatively quickly. The screening conditions should ensure that the expected small enhancements brought about by single amino acid substitutions can be measured. This is usually accomplished by adjusting the assay conditions to the point where the wild type (or parent) enzyme is near (but not below) the lower detection limit. Improvements over the wild type enzyme will then be clearly discernible. (If wild type is below the detection limit, small improvements may not be measurable.) If a replica plate of the clones is prepared for storage, the clones can be subjected to an array of lethal procedures and the original clones can be recovered. For example, many intracellular enzymes require cell lysis before they can be assayed. This procedure could be performed in a screen, and when positives are found, they can be regrown from the replica plate. Furthermore, multiple measurements can be made on each sample to account for variability in protein expression levels or to check other key enzyme properties. The biggest drawback with screening is that the size of the library that can be screened is limited, usually to ~10^4, although the use of robotic hardware can increase this number by 10-100-fold.

Although the importance of screening under conditions close to what the enzyme would encounter in an actual process cannot be overstated, dealing with large numbers of variants requires approximations. (Screening 10^4 variants in high density, 10,000L

fermentations is clearly not feasible.) When approximations are made, the correlation between the approximate and actual conditions should be checked, if at all possible. In the directed evolution of the pNB esterase (11), two important approximations were made in the screening. First, the desired loracarbef p-nitrobenzyl ester hydrolysis reaction is usually assayed by HPLC, which is unsuitable for rapid screening. We therefore devised a screening assay based on the loracarbef p-nitrophenyl ester, which provides a colorimetric signal. To validate the screening method, we compared the activities of a set of mutants towards the p-nitrobenzyl and p-nitrophenyl substrates, as shown in Fig. 3. If the screening reaction perfectly mimicked the desired reaction, all the points would lie on the 45° line. Although there will be some false positives and false negatives using this screening reaction, the rapid screen provides sufficient information to make a rough cut of positive clones. Thus we needed to re-test only a small number of clones by HPLC to verify improved pNB esterase activity and to select which of the clones should parent the next generation. The second approximation concerned the concentration of organic solvent in the assay. The DMF concentration desired for the ultimate application was 25-30%, which was too high to generate any measurable activity at the beginning of the evolution experiments. Initial screening experiments were therefore performed in 10-15% DMF. The activities towards LCN-pNP in 5 and 25% DMF were found to be strongly correlated. Thus it was appropriate to screen in lower DMF concentrations, at least at the beginning.

Primary screening can often be performed in petri dishes. This involves either placing the substrate directly in the petri dishes, applying substrate to the already grown colonies, or transferring colonies to a second petri dish containing substrate. The assay conditions are then set such that wild type enzyme generates a weak positive signal, allowing observation of small improvements. When environmental conditions of interest preclude growth and the enzyme is secreted (e.g., when assaying proteases in organic solvents), the organisms can be grown on one or more filters. A "master" filter

containing the colonies can be maintained, and assay filter(s) capturing secreted enzymes can be placed on plates in any environment. One problem with this approach is the inability to keep a known wild type colony on each plate as a control for plate-to-plate variability. Because transformation mixes are usually plated directly onto the plates, the colony density depends on the ligation and transformation efficiency. Wild type is then placed on separate plates and under different colony densities. This difference in colony density on a plate can dramatically alter the intensity of a wild type signal and makes a control of this type less useful than expected. Positives found this way should be regrown on the same plate with wild type and reassayed. (For an example of this approach, see (19)).

Multi-well plates are very useful for screening large libraries, especially now that robotic equipment designed to handle these plates is becoming more widely available. Although one can pick, resuspend, and assay colonies in 96-well plates, growing liquid cultures in the plates is generally less tedious. Liquid cultures allow the inclusion of wild type controls under identical conditions with clones from the library to be screened and also tend to yield more uniform and reliable data. Growth for reasonably long durations (36 hours) can be achieved in regular incubators, provided there is some means to humidify the shaker. We routinely use beakers filled with water and paper towels in the shaker for this purpose.

When we set up an assay in 96-well plates, we always evaluate several plates that contain only wild type clones. This enables us to determine the reproducibility of the assay and whether there is any positional variability in the plate. This allows us to identify systematic errors due to pipetting, heat transfer, etc., and to determine the values of the screening parameter that should be associated with true positives. For example, in screening thermostable subtilisin E variants, each well of a 96-well plate contained an identical subtilisin E clone growing in 200 µl of SG medium. This plate was incubated for 20 h and cells were spun down. 5 µl of supernatant from each well was transferred

into two plates, into which we added 15 µl of SG medium (to avoid a problem with evaporation during incubation at high temperature). One plate was used to measure initial activity (after 5 min of incubation at 65 °C), and the other was used to measure the residual activity (after 20 min of incubation at 65 °C). The ratio between residual activity and initial activity was taken as index of thermostability (14). The 65 °C incubation was performed in an oven on an aluminum block machined to closely contact standard multi-well plates for uniform heating. Data from a control plate containing only wild type clones are shown in Fig. 4. A small amount of systematic variability is observed, possibly due to uneven heating of the wells. In general, however, the activity values do not vary widely outside the range 30-40% residual activity. This provides confidence that variants with more than 40% residual activity are likely to be more thermostable variants of subtilisin E.

Data Analysis

Screening hundreds of 96 well plates can generate a tremendous amount of data. Often the screen involves multiple measurements which have to be manipulated mathematically in order to determine whether a particular clone is more fit than the parent enzyme. For example, thermostability assays such as the subtilisin E assay described above use two measurements of activity, initial and residual activity. Screening pNB esterase for activity in DMF required three measurements: an absorbance reading of the cell suspension for the purpose of estimating cell density, an activity assay in a low concentration of DMF, and an activity assay in high DMF. The cell density measurement allowed us to determine which clones had increased total activity (increased specific enzyme activity and/or increased expression), while the ratio of activity measurements gave values which were independent of enzyme concentration and which reflected specific changes in the enzyme's ability to function in higher DMF concentrations.

(Similar sorts of estimates can be made visually with petri dish screens with less precision.)

These types of measurements can be easily performed using a 96-well plate reader interfaced to a computer. By downloading the data directly into a spreadsheet program, the calculations described above and the relationships between different properties can be quantified (as in Fig. 3 and 4). Additionally, as already shown in the above activity profile, spreadsheets permit sorting the data, so that the screening results can be better evaluated. For example, data from one or more 96-well plates can be sorted from best to worst and then plotted, as in Fig. 5, to give 'fitness profiles' from which additional interesting characteristics of the library and the property can be deduced. Apart from the application as a diagnostic check of mutation frequency, these profiles also allow us to assess the evolvability of a particular property of the enzyme. Fig. 5(a) and (b) show two different fitness profiles for activity and thermostability, respectively. Based on the fact that wild type is located near the top of the activity profile in Fig. 5(a), while quite a few clones have thermostability indices larger than wild type in Fig. 5(b)), it appears that there are more ways to improve thermostability than activity.

Acknowledgments

The authors wish to thank Dr. Pim Stemmer for his advice and assistance with the DNA shuffling method. This work was supported by the U.S. Office of Naval Research and the U.S. Department of Energy's program in Biological and Chemical Technologies Research within the Office of Industrial Technologies, Energy Efficiency and Renewables.

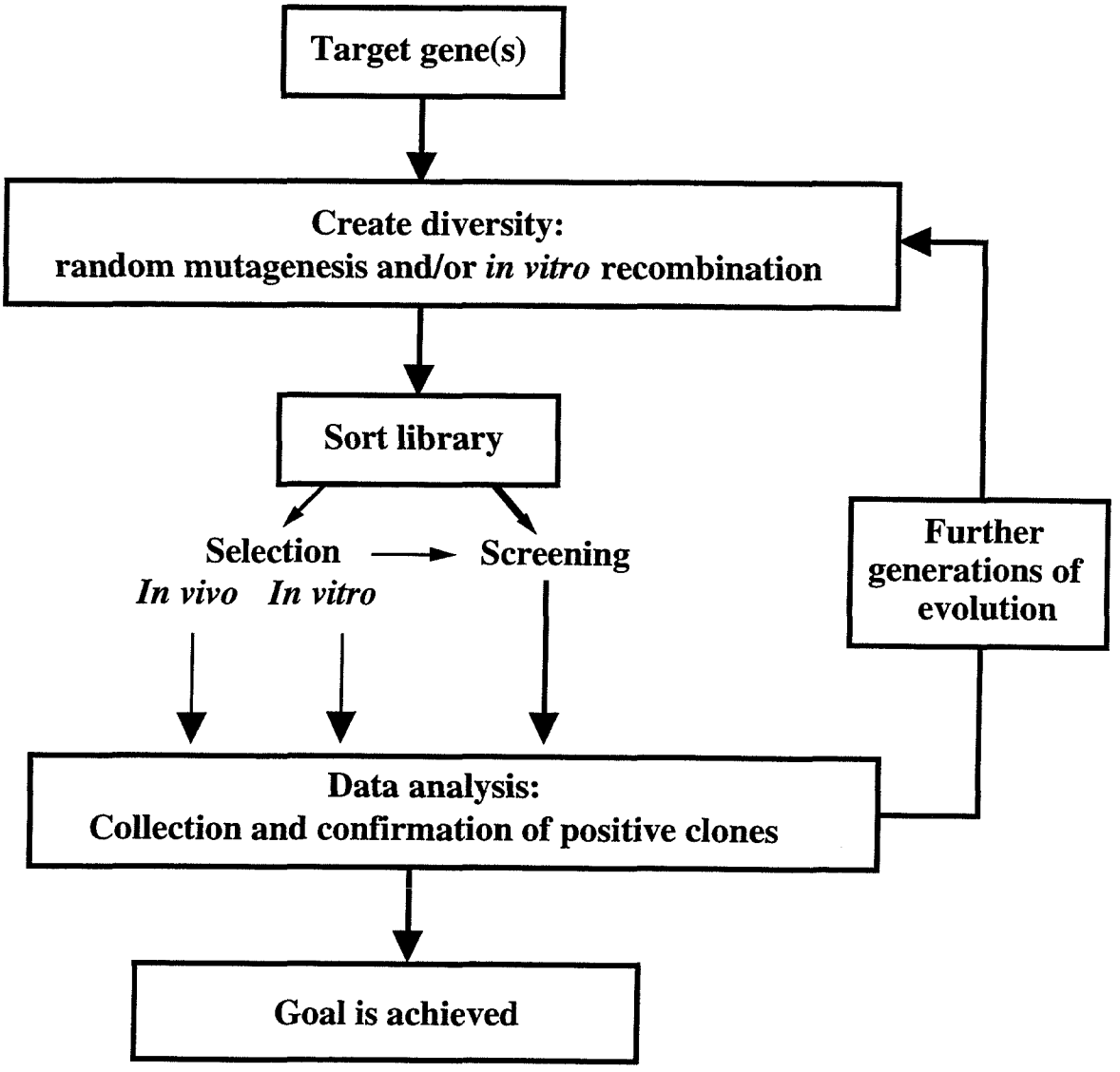


Fig. 1. Flowchart for directed enzyme evolution.

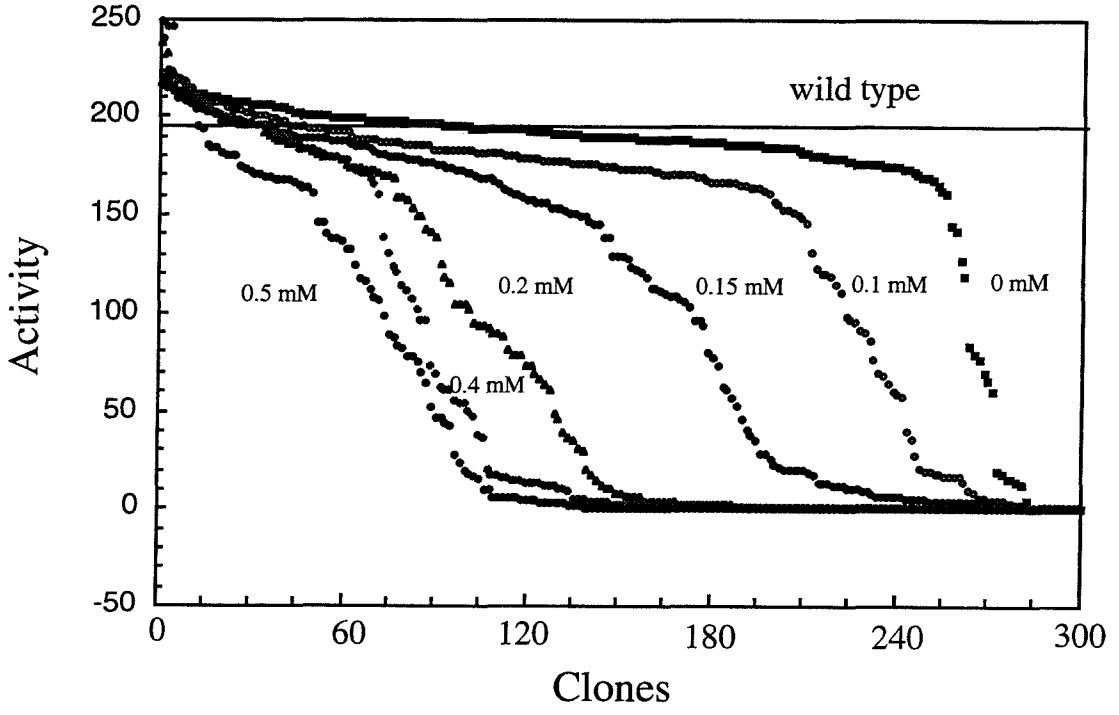


Fig. 2. The activity profiles of each variant library generated at varying concentrations of $MnCl_2$ (0 - 0.5 mM). About 300 variants from each library were taken for initial activity measurements. Data were sorted and plotted in descending order. The activity of wild type was shown by a horizontal line.

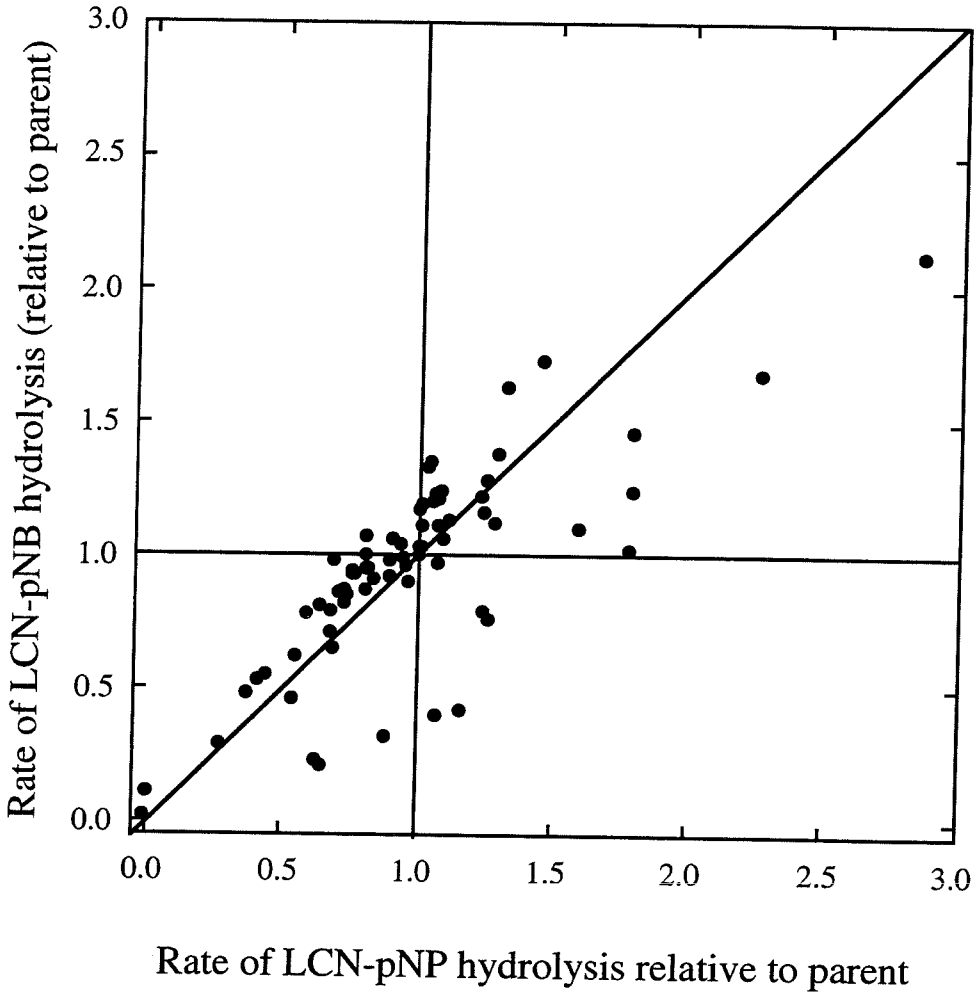


Fig. 3. Comparison of activities of pNB esterase variants towards desired LCN-pNB substrate and LCN-pNP substrate used in screening.

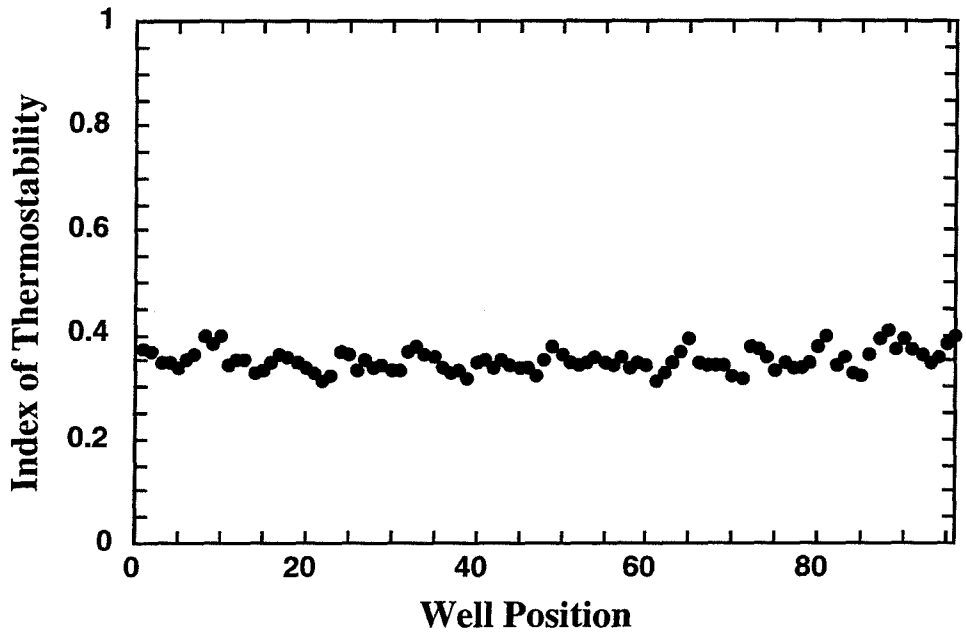


Fig. 4. Variations in the thermostability index of wild type subtilisin E clones in a 96-well plate assay.

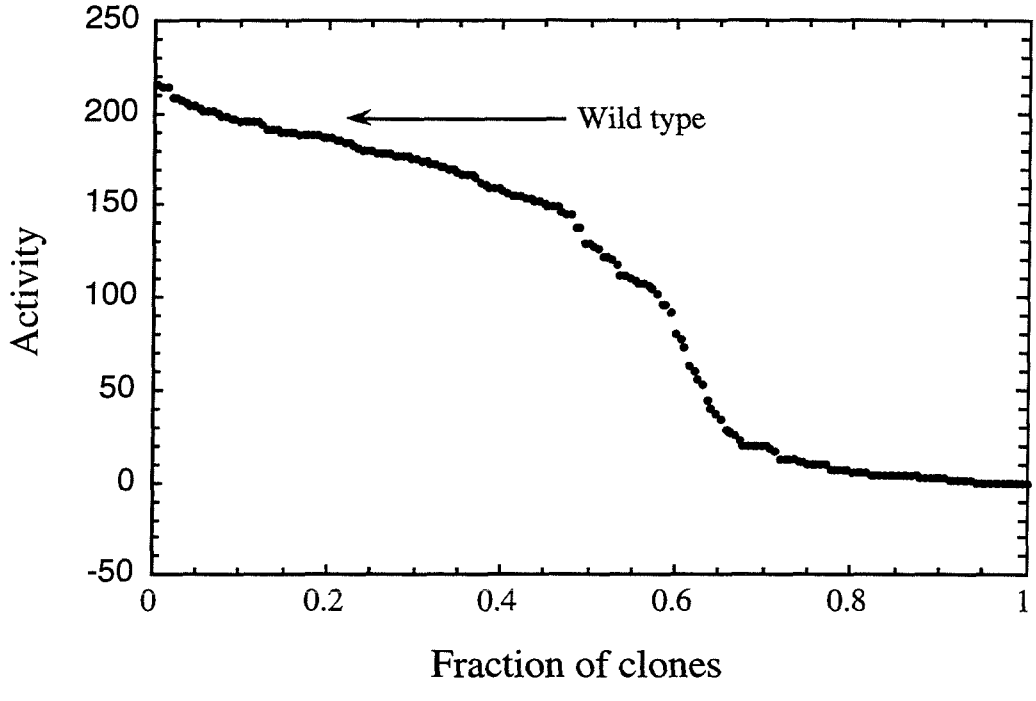


Fig. 5a. Typical profile of the activity of enzyme variants in a random mutant library.

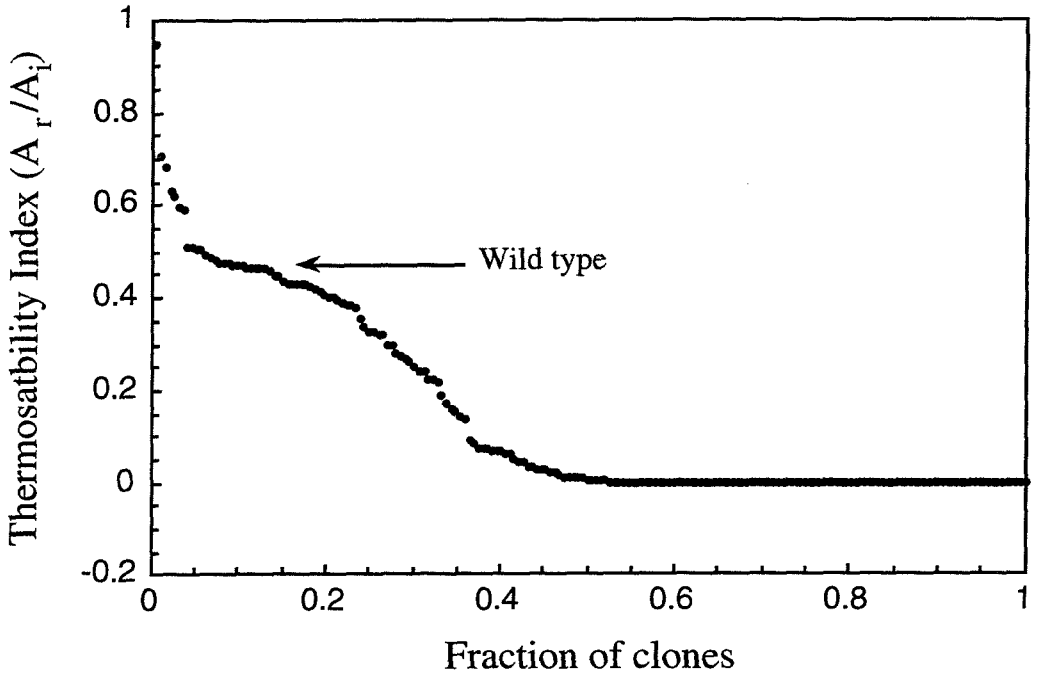


Fig. 5b. Stability of enzyme variants from the same mutant library.

References

1. **Kuchner, O. and Arnold, F. H.** 1997. Directed evolution of enzyme catalysts, Trends in Biotechnology, **15**:523-530.

2. **Arnold, F. H.** 1996. Directed evolution: creating biocatalysts for the future, Chemical Engineering Science, **51**:5091-5102.

3. **Arnold, F. H.** 1998. Design by directed evolution, Accounts of Chemical Research, **31**:125-131.

4. **Zock, J., Cantwell, C., Swartling, J., Hodges, R., Pohl, T., Sutton, K., Rosteck Jr., P., McGilvray, D., and Queener, S.** 1994. The *Bacillus subtilis* pnbA gene encoding p-nitrobenzyl esterase - cloning, sequence and high-level expression in *Escherichia coli*. Gene **151**:37-43.

5. **Reidhaar-Olson, J. F. and Sauer, R. T.** 1988. Combinatorial cassette mutagenesis as a probe of the information content of proteins. Science **241**:53-57

6. **Leung, D. W., Chen, E. and Goeddel, D. V.** 1989. A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. Technique **1**:11-15.

7. **Stemmer, W. P. C.** 1994. DNA shuffling by random fragmentation and reassembly - in-vitro recombination for molecular evolution. Proc. Natl. Acad. Sci. **91**:10747-10751.

8. **Shao, Z., Zhao, H., Giver, L., Arnold, F. H.** 1998. Random-priming *in vitro* Recombination: an Effective Tool for Directed Evolution. Nucleic Acids Research, **26**:681-683

9. **Zhao, H., Giver, L., Shao, Z., Affholter, J. A., Arnold, F. H.** 1998. Molecular evolution by staggered extension process (StEP). Nature Biotechnol., **16**:258-262..

10. **Crameri, A., Raillard, S.-A., Stemmer, W. P. C.** 1998. DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature, **391**:288-291.

11. **Moore, J. C. and Arnold, F. H.** 1996. Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents, Nature Biotechnology, **14**:458-467.

12. **Cadwell, R. C. and Joyce, G. F.** 1992. Randomization of genes by PCR mutagenesis. PCR Methods and Appl. **2**:28-33.

13. **Shafikhani, S., Siegel, R. A., Ferrari, E. and Schellenberger, V.** 1997. Generation of large libraries of random mutants in *Bacillus subtilis* by PCR-based plasmid multimerization. Biotechniques, **23**:301-310.

14. **Zhao, H. and Arnold, F. H.** 1997. Optimiziation of DNA shuffling for high-fidelity recombination. Nucleic Acids Research, **25**:1307-1308.

15. **Lorimer, I. A. J. and Pastan, I.** 1995. Random recombination of antibody single-chain fv sequences after fragmentation with DNase I in the presence of Mn^{2+}. Nucl. Acids Res. **23**:3067-3068.

16. **Moore, J. C., Jin, H. M., Kuchner, O. and Arnold, F. H.** 1997. Strategies for the in vitro evolution of protein function: enzyme evolution by random recombination of improved sequences. J. Mol. Biol. **272**:336-347.

17. **Kast, P. and Hilvert, D.** 1997. Three-dimensional structural information as a guide to protein engineering by genetic selection. Curr. Opin. Struct. Biol. **7**:470-479.

18. **Zhao, H. and Arnold, F. H.** 1997. Combinatorial protein design: strategies for screening protein libraries. Curr. Opin. Struct. Biol. **7**:480-485.

19. **Chen, K. Q. and Arnold, F. H.** 1993. Tuning the activity of an enzyme for unusual environments - sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. Proc. Natl. Acad. Sci. **90**:5618-5622.

Appendix B

Combinatorial Protein Design: Strategies for Screening Protein Libraries

(Huimin Zhao and Frances H. Arnold)

(appeared in *Current Opinion in Structural Biology*, 1997, **7**, 480-485)

Combinatorial protein design: strategies for screening protein libraries

Huimin Zhao and Frances H Arnold*

Powerful strategies for screening protein libraries further strengthen the arguments for applying 'irrational' approaches to understanding and designing new proteins. Developments during the past year include the application of functional complementation and automation to reduce screening loads, as well as the use of computerized data acquisition to characterize whole protein libraries rather than just selected individuals.

Addresses
Division of Chemistry and Chemical Engineering 210-41 California Institute of Technology, Pasadena, CA 91125, USA
*e-mail: frances@cheme.caltech

Current Opinion in Structural Biology 1997, 7:480–485

http://biomednet.com/elecref/0959440X00700480

Abbreviations

*Bs*LDH	*Bacillus stearothermophilus* L-lactate dehydrogenase
CSDL	cell surface display libraries
ELISA	enzyme-linked immunoadsorbent assay
FACS	fluorescence activated cell sorting
GFP	green fluorescent protein
NAD	nicotinamide adenine dinucleotide
PCR	polymerase chain reaction
SGPB	*Streptomyces griseus* protease B
SPA	scintillation proximity assay

Introduction

Protein engineers are becoming increasingly impatient with rational design approaches and the extensive structural and mechanistic information required to guide such efforts. Identifying the amino acids responsible for existing protein functions and those that might give rise to new functions remains an often overwhelming challenge. This, together with the growing appreciation that many protein functions are not confined to a small number of amino acids but are affected by residues far from active sites, has prompted a growing number of research groups to turn to 'irrational' design approaches, such as random mutagenesis and *in vitro* evolution, to engineer novel proteins [1] as well as to better understand existing ones [2,3]. These irrational approaches involve the generation and selection or screening of molecular repertoires with sufficient diversity for the altered function to be represented. Such approaches have been used to create novel functional nucleic acids [4], peptides and other small molecules [4], antibodies [4], as well as enzymes and other proteins [1–6,7•]. Given an intelligent approach to generating the mutant library (i.e. one that considers the enormous size of protein sequence space), the development of efficient methods to search protein libraries for desired properties is probably the single most important element determining the success of these experiments.

The prerequisite to selection is the generation of a function that confers a growth or survival advantage to the host organism. Selection can be a very efficient search mechanism, allowing an exhaustive search of libraries of 10^6 or more protein variants. The disadvantage of a biological selection is that the property or protein of interest cannot be decoupled from a biological function. Thus, exploring novel functions such as stability or activity in non-natural environments (e.g. organic solvents) or activity on non-natural substrates can be difficult or even impossible [5]. The creative researcher who is tempted to apply a 'synthetic' selection approach should also know that these can be extremely tedious to develop and are often ineffectual because cells find alternative ways to solve the problem posed. *In vitro* selections have been developed in an attempt to mimic the power of natural selection, including a variety of selections based on column binding or 'panning' [4,6] and a recently reported chemical selection for catalysis [8]. Thus member(s) with desired functions can be selected directly from the rest of the library by either preferential binding or covalent interaction due to the accomplishment of a chemical feat. When choosing a search strategy, it is useful to remember the principle, "You get what you select (screen) for" [9•]. Many selections do not yield the desired result! Selection approaches are discussed in a separate review by Kast and Hilvert (pp 470–479) in this section on engineering and design.

For many problems, and especially those of practical interest, libraries of variants must be screened rather than selected, which means that the library members must be examined separately (often one at a time). A typical screening strategy thus involves the construction of an arrayed protein library and the application of a rapid assay that is sufficiently sensitive and specific to identify positives. The screen can be more or less sensitive, depending on the willingness of the researcher to accept false positives (and to apply additional tests). In this review, we focus on recent developments in screening protein libraries for specific functions.

Semi-quantitative visual screens

Simple visual screens are widely used when the function of interest can generate a visible signal. Green fluorescent protein (GFP) variants are easily identified in visual screens based on the intrinsic fluorescence of GFP under UV illumination [10–12]. GFP libraries generated by error-prone polymerase chain reaction (PCR) [11,12] and

DNA shuffling [10] have been expressed in *Escherichia coli* and arrayed on agar plates. Visual screening of 10^4–10^5 clones yielded variants with increased intensity of the fluorescence signal [10,11] and suppressed thermosensitivity [12].

Simple colorimetric assays are also widely used in screening [13,14,15•,16–18]. Brunet *et al.* [13] have assessed the role of turns in stabilizing the structure of an α-helical protein by completely randomizing a tripeptide turn of cytochrome *b-562* using cassette replacement mutagenesis. 45 clones have been screened on the basis of the observation that cells expressing mutants that fold correctly will bind heme and yield bright red periplasmic extracts, whereas cells expressing mutants that fold incorrectly yield colorless extracts. More recently, a similar approach has been applied to studying a β turn in a β-barrel protein, plastocyanin, which is blue in its native, folded state and colorless when denatured [14]. Because of the need for concentration and other pretreatments, however, this approach is not yet suitable for screening large libraries. Hawrani *et al.* [15•] have completely randomized two amino acids in a solvent-exposed surface loop of *Bacillus stearothermophilus* L-lactate dehydrogenase (*Bs*LDH) with the goal of engineering new substrate specificities. The expressed *Bs*LDH library is immobilized spatially on nitrocellulose membranes after cell lysis and several washing steps and tested for each variant's ability to reduce NAD^+ in high and low concentrations of a target substrate (to detect substrate inhibition), and in the presence and absence of an allosteric activator (to find variants that do not require the activator). Enzyme activity is detected by coupling the hydrogen transfer system of NADH/NAD^+ to further redox reactions involving phenazine methosulphate and nitroblue tetrazolium, which yields a blue insoluble dye on reduction.

Two *in vivo* color screening strategies have been described recently [19•,20]. Bacterial membrane permeability increases after synthesis of the poliovirus protein 3AB. Lama and Carrasco [19•] have developed a screen in which permeabilizing wild-type 3AB and nonpermeabilizing 3AB mutants are differentiated because of the different rate of entry of a chromogenic β-galactosidase substrate, X-Gal. *E. coli* clones with wild-type 3AB are stained dark blue, whereas the mutants lacking pore-forming activity are stained light blue. To isolate α-factor pheromone receptor mutation(s) that constitutively signal in the absence of α-factor, Konopka *et al.* [20] have constructed a yeast strain which contains the pheromone-responsive *FUS1-lacZ* reporter gene. Constitutive receptor mutants are detected as blue colonies on an agar plate containing X-Gal, as a result of induced expression of *lacZ*-encoded β-galactosidase. Approximately 600,000 colonies have been screened.

Clones secreting active proteases produce a zone of clearing or 'halo', the size of which is proportional to the hydrolytic activity when grown on agar plates containing casein or skim milk proteins. Utilizing this well-known visual screen, You and Arnold [9•] have applied sequential rounds of random mutagenesis and screening to significantly enhance the expression level of subtilisin E and further increase its specific activity in aqueous organic solvent. Sidhu *et al.* [21] have varied seven residues involved in conferring primary specificity of *Streptomyces griseus* protease B (SGPB) by DNA cassette mutagenesis. An *E. coli* expression library containing 29,952 possible SGPB mutants has been screened for secretion of active protease by halo formation on agar plates containing skim milk. Sidhu *et al.* [21] have found that the substrate specificity of recombinant SGPB is constrained by the sequence at the promature junction, and active protease production is dependent on the efficiency of self-processing by pro-mature junction cleavage. Easily observed colony phenotypes are also exploited in rapid screening methods [22,23].

Screening in 96-well plate formats suitable for automation

Although visual screening on the basis of color or halo formation is rapid and efficient, its limitations are also obvious: visual screens are nonquantitative and often insensitive to small changes in properties, and they are not generally applicable. Designing a visual screen for protein functions such as catalysis of a specific reaction or of a specific substrate may be difficult or impossible. Digital imaging spectroscopy has been developed to increase the sensitivity and throughput of filter and agar plate based screens [24]; however, the 96-well microtitre plate remains the standard format for automated, high-throughput screening [7•,25•,26]. Screening automation and quantification of the results are highly desirable; 96-well plates appear to be the format most compatible with currently available robotic arms, liquid handling systems and plate readers.

Data collection and analysis are greatly facilitated by computerized data acquisition. Assay automation and computerized data acquisition are being exploited extensively by researchers at Recombinant Biocatalysis (La Jolla, CA), who are using high-throughput 96-well plate assays to identify enzyme catalysts from genomic libraries, to characterize their substrate specificities, and to further evolve specific properties [26]. Useful information about the protein library can be retrieved from the large amount of data generated during screening. We have used library screening to distinguish functional from nonfunctional and deleterious mutations in a laboratory-evolved thermostable subtilisin E, 1E2A [27•]. The evolved gene has been randomly recombined with the wild-type gene to create a library of all possible mutation combinations. The resulting library has been screened for thermostability using normalized residual activity after incubation at high temperature in 96-well plates (Figure 1). The fact that roughly 25% of the clones exhibit thermostability comparable with that of 1E2A immediately indicates that only

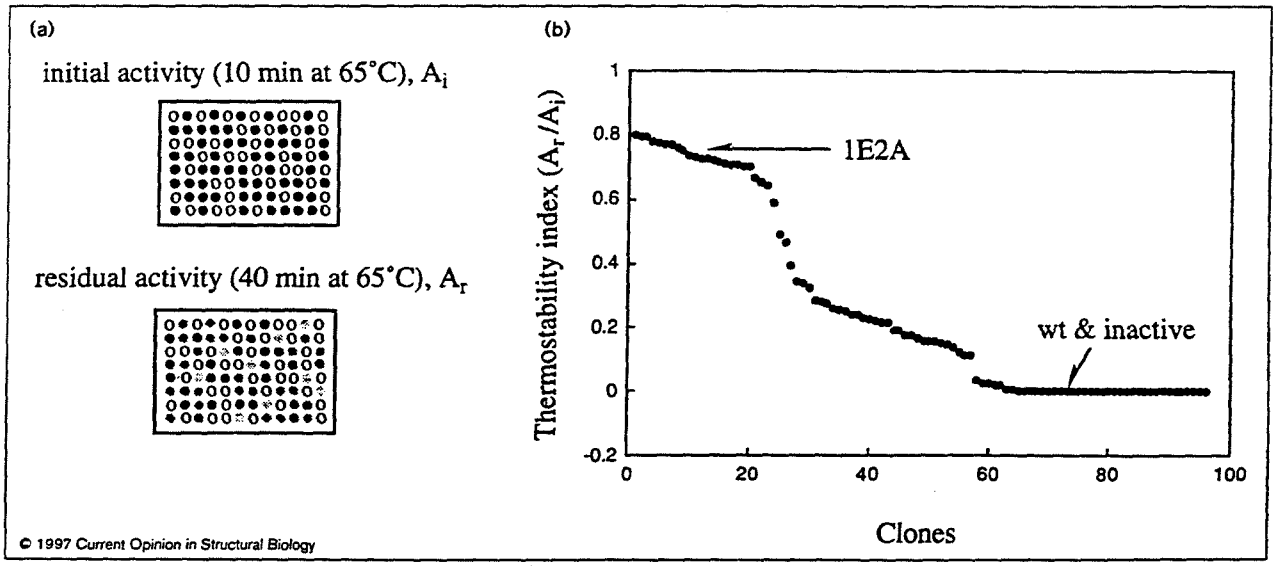
two of the ten DNA mutations in 1E2A are responsible for the increased protein thermostability, a result that has been confirmed by further biochemical analysis. Similar sorted library profiles are very useful for estimating the mutagenic rates associated with error-prone PCR or other random mutagenesis approaches [28,29]. Computerized data acquisition will certainly facilitate exploitation of the enormous amounts of information available in protein libraries as opposed to single individuals.

Many conventional assays such as enzyme-linked immunoadsorbent assays (ELISA), radioimmunoassays, and enzyme–substrate assays can be readily converted into automated formats. Two new assays that require no steps to separate free from bound tracer have been described: scintillation proximity assay (SPA) [7•] and fluorescence polarization equilibrium binding assay [30]. Several novel screening approaches for catalytic activity [31–34] and enantioselectivity of catalysis [35•] have also been developed for the 96-well plate format. Tawfik *et al.* [31,32] have devised a sensitive method, catELISA, on the basis of immobilized substrates, and immunodetection of the end product of the catalyzed reaction. This strategy was used by MacBeath *et al.* [33] to screen antibody libraries for catalysis of a bimolecular Diels–Alder reaction. The main limitation of this approach is that it requires a specialized antibody that can strongly discriminate between substrate and product. Fenniri *et al.* [34] have described an encoded reaction cassette for the detection of chemical bond breakage and formation for biocatalyst screening. The reaction cassette is immobilized on a matrix and encoded by a polynucleotide containing two primers. Thus, when the functionalized matrix is exposed to a library of biocatalysts that are able to selectively cleave the reaction cassette at the substrate juncture, the polynucleotide is released and can be amplified by PCR. For large library screening, the PCR reaction can be carried out in 96-well plates, and the PCR products can be further detected by the addition of a fluorescent probe.

Janes *et al.* [35•] have developed a rapid spectrophotometric method to measure the enantioselectivity of hydrolases, based on the addition of a chromogenic reference compound that introduces competitive binding for both enantiomers. The relative hydrolysis rates of the two enantiomers competing with the reference compound are measured separately; the ratio of the two relative rates gives the enantioselectivity. Compared with an endpoint method, this 'quick E' method has advantages in terms of speed, the amounts of enzyme required, and the higher enantioselectivities that can be measured.

A number of other screens based on 96-well plates have been described during this past year. Moore and Arnold [36•] have utilized a chromogenic substrate (an antibiotic *p*-nitrophenyl ester, pNP) that serves as a surrogate for

Figure 1



A rapid screen for enzyme thermostability based on residual activity after incubation. **(a)** Catalytic activity is measured before (A_i) and after (A_r) incubation at high temperature. **(b)** Results from a typical 96-well plate. *Bacillus* transformants of a randomly recombined subtilisin E library are picked and grown in 96-well plates. Initial subtilisin activity (A_i) and residual activity (A_r) after incubation at 65°C of supernatants are measured (towards succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide) using a microtitre plate reader. A_r/A_i values are sorted and plotted in descending order. Residual activities of wild-type (wt) and 1E2A, a thermostable variant of subtilisin E, are indicated by arrows. The probability that any given mutation will appear in the randomly recombined population of two equally mixed genes is 1/2. Thus the probability that *N* specific (functional) mutations will appear together in a sequence is $(1/2)^N$. In this example, two mutations are responsible for the enhanced thermostability of 1E2A [27•].

the desired substrate of an esterase (the *p*-nitrobenzyl ester, pNB) in a rapid screening assay coupled with sequential random mutagenesis and recombination. While the reaction with the pNB ester must be analyzed by HPLC, the cell growth and pNP screening reactions can all be carried out in 96-well plates and analyzed using a microtitre plate reader. The surrogate assay has been validated by comparing the activities of random mutants on the two substrates, which shows a relatively high correlation. Thus, the positives identified during the rapid screening assay can be verified during a second level screen using HPLC. Panchal *et al.* [37] have described the screening of libraries generated by combinatorial cassette mutagenesis to isolate mutants of the pore-forming toxin α-hemolysin that are rapidly and preferentially activated by a tumor protease, cathepsin B. Mutants exhibiting the desired activation after preliminary screening with clostripain have been reassayed more carefully for their hemolytic activity after treatment with human liver cathepsin B or clostripain. Cells are grown and screened in 96-well plates.

Functional complementation coupled with screening

An attractive approach to screening large libraries is to couple functional complementation with screening [2,3,24,38•,39,40,41•,42,43]. By requiring that at least some of the biological function of the protein is retained, functional complementation can greatly reduce the subsequent screening requirements. A major limitation, however, is finding or constructing an appropriate complementation system. Furthermore, the retention of biological function may preclude acquisition of other functions. Loeb and coworkers [38•] have applied this strategy effectively in several studies whose goals have been to identify residues responsible for protein functions as well as to engineer novel properties. The proteins studied recently include herpes simplex virus type 1 thymidine kinase (HSV-1 TK), human immunodeficiency virus reverse transcriptase (HIV RT) [3] and human DNA alkyltransferase [39]. Six residues adjacent to the putative nucleotide-binding site of HSV-1 TK have been completely randomized in an effort to increase the enzyme's specificity for the phosphorylating nucleoside analogs gancyclovir (GCV) and/or acyclovir (ACV) [38•]. 426 active mutants selected from more than one million mutants by functional complementation have been screened for enhanced sensitivity to GCV and/or ACV. One drug-sensitive mutant produced stable mammalian cell transfectants that are 43-fold more sensitive to GCV and 20-fold more sensitive to ACV.

Warren *et al.* [40] have altered each polar residue within 6 Å of the catalytic center of glycinamide ribonucleotide transformylase using saturation site-directed mutagenesis. Approximately 50 transformants from each mutagenesis have been picked, sequenced and screened for their activity using functional complementation of auxotrophic cells. None of the polar residues close to the catalytic center of the enzyme has been found to be irreplaceable. Further study of the three key polar active-site residues has revealed that none is irreplaceable, although any change leads to substantially decreased catalytic activity and no more than one mutation can be tolerated [2].

Axe *et al.* [41•] have developed a sensitive biological screen on the basis of the extreme autotoxicity of barnase when produced in *E. coli* in the absence of its natural inhibitor, barstar. Two amber stop codons have been introduced to replace serine codons at positions 28 and 57 of the synthetic barnase gene (*synbar*), which prevents lethal production of barnase in a nonsuppressing (sup^-) *E. coli* strain. When plasmid DNA prepared from the sup^- strain is used to transform a suppressing (*supD*) strain, these two amber stop codons are read as serine codons, producing wild-type barnase. Only *synbar* mutations yielding variants with dramatically reduced activity allow the *E. coli supD* strain to grow. The activity of these mutants is then qualitatively estimated. A barnase mutant library in which 12 hydrophobic core residues have been randomly replaced by hydrophobic alternatives has been screened. A significant fraction (23%) of these mutants retain activity, which implies that hydrophobicity is a nearly sufficient criterion for the construction of a functional core. A similar screening method has been developed to detect low activity barnase mutants, in which a barnase–barstar plasmid inversion system has been constructed [42].

Sorting single cells or proteins

In order to move toward automated, high-throughput screening, it will be important to increase the level of miniaturization. Silicon wafers with many thousands of compartments housing liquid volumes in the nanoliter and picoliter range have been developed [44]. Liquid handling systems have also been developed to accurately dispense biological samples of ~5 picoliters to a few nanoliters and speeds up to 10,000 drops per second [45]. Innovative approaches that may be applicable to large-scale screening of protein libraries have been described in recent years, including fluorescence correlation spectroscopy coupled with single molecule trapping devices [46–48] and cell surface display libraries (CSDL) coupled with fluorescence activated cell sorting (FACS) [49•]. Fluorescence correlation spectroscopy coupled with devices for trapping single molecules in an electric field may be used in the future to sort single cells (or parts of their surfaces), organelles, viruses, single genes, proteins, or even small molecular entities such as peptide hormones or other oligomeric compounds [46–48]. This approach is capable of monitoring very low concentrations ($< 10^{-15}$ M) without the need for amplification; however, actual applications to library screening have not yet been reported. For protein library screening, phage displayed [6] or bacterial/yeast CSDLs [49•] could be used with the advantage that an isolated sample can be amplified since the genotypic information is linked with the phenotypic information. Because of the relatively large size of cells, CSDLs can be

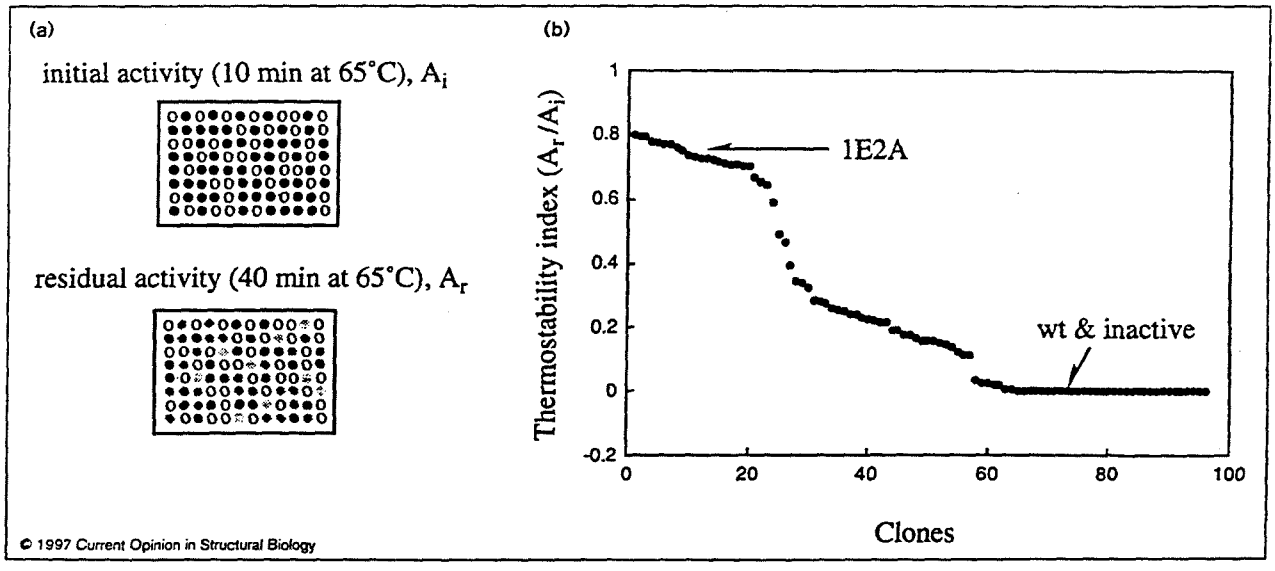
two of the ten DNA mutations in 1E2A are responsible for the increased protein thermostability, a result that has been confirmed by further biochemical analysis. Similar sorted library profiles are very useful for estimating the mutagenic rates associated with error-prone PCR or other random mutagenesis approaches [28,29]. Computerized data acquisition will certainly facilitate exploitation of the enormous amounts of information available in protein libraries as opposed to single individuals.

Many conventional assays such as enzyme-linked immunoadsorbent assays (ELISA), radioimmunoassays, and enzyme–substrate assays can be readily converted into automated formats. Two new assays that require no steps to separate free from bound tracer have been described: scintillation proximity assay (SPA) [7•] and fluorescence polarization equilibrium binding assay [30]. Several novel screening approaches for catalytic activity [31–34] and enantioselectivity of catalysis [35•] have also been developed for the 96-well plate format. Tawfik *et al.* [31,32] have devised a sensitive method, catELISA, on the basis of immobilized substrates, and immunodetection of the end product of the catalyzed reaction. This strategy was used by MacBeath *et al.* [33] to screen antibody libraries for catalysis of a bimolecular Diels–Alder reaction. The main limitation of this approach is that it requires a specialized antibody that can strongly discriminate between substrate and product. Fenniri *et al.* [34] have described an encoded reaction cassette for the detection of chemical bond breakage and formation for biocatalyst screening. The reaction cassette is immobilized on a matrix and encoded by a polynucleotide containing two primers. Thus, when the functionalized matrix is exposed to a library of biocatalysts that are able to selectively cleave the reaction cassette at the substrate juncture, the polynucleotide is released and can be amplified by PCR. For large library screening, the PCR reaction can be carried out in 96-well plates, and the PCR products can be further detected by the addition of a fluorescent probe.

Janes *et al.* [35•] have developed a rapid spectrophotometric method to measure the enantioselectivity of hydrolases, based on the addition of a chromogenic reference compound that introduces competitive binding for both enantiomers. The relative hydrolysis rates of the two enantiomers competing with the reference compound are measured separately; the ratio of the two relative rates gives the enantioselectivity. Compared with an endpoint method, this 'quick E' method has advantages in terms of speed, the amounts of enzyme required, and the higher enantioselectivities that can be measured.

A number of other screens based on 96-well plates have been described during this past year. Moore and Arnold [36•] have utilized a chromogenic substrate (an antibiotic *p*-nitrophenyl ester, pNP) that serves as a surrogate for

Figure 1



A rapid screen for enzyme thermostability based on residual activity after incubation. **(a)** Catalytic activity is measured before (A_i) and after (A_r) incubation at high temperature. **(b)** Results from a typical 96-well plate. *Bacillus* transformants of a randomly recombined subtilisin E library are picked and grown in 96-well plates. Initial subtilisin activity (A_i) and residual activity (A_r) after incubation at 65°C of supernatants are measured (towards succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide) using a microtitre plate reader. A_r/A_i values are sorted and plotted in descending order. Residual activities of wild-type (wt) and 1E2A, a thermostable variant of subtilisin E, are indicated by arrows. The probability that any given mutation will appear in the randomly recombined population of two equally mixed genes is 1/2. Thus the probability that *N* specific (functional) mutations will appear together in a sequence is $(1/2)^N$. In this example, two mutations are responsible for the enhanced thermostability of 1E2A [27•].

coupled with FACS for high throughput screening. As with phage display, however, there are difficulties in using these systems for screening for properties other than binding.

Conclusions

Most current screening methods are labor intensive and can be used to search small libraries of only a few thousand members. Automation and miniaturization, which reduce labor and materials requirements and increase reproducibility, can extend screening capabilities by a factor of 10–100. Methods for screening single cells or even single molecules, however, may dramatically increase the numbers of variants that can be screened and will significantly enhance the power of *in vitro* evolution. By screening much larger protein libraries, one can hope to identify rare events such as the acquisition of a novel catalytic activity.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Shao ZX, Arnold FH: **Engineering new functions and altering existing functions.** *Curr Opin Struct Biol* 1996, 6:513-518.

2. Warren MS, Benkovic SJ: **Combinatorial manipulation of three key active site residues in glycinamide ribonucleotide transformylase.** *Protein Eng* 1997, 10:63-68.

3. Kim B, Hathaway TR, Loeb LA: **Human immunodeficiency virus reverse transcriptase–functional mutants obtained by random mutagenesis coupled with genetic selection in *Escherichia coli*.** *J Biol Chem* 1996, 271:4872-4878.

4. Abelson JN (Ed): **Combinatorial chemistry.** In *Methods in Enzymology*, vol 267. San Diego: Academic Press; 1996.

5. Arnold FH: **Directed evolution creating biocatalysts for the future.** *Chem Eng Sci* 1996, 51:5091-5102.

6. Wang C-I, Yang Q, Craik CS: **Phage display of proteases and macromolecular inhibitors.** *Methods Enzymol* 1996, 267:52-68.

7. • Cole JL: **Approaches to high-volume screening assays of viral polymerases and related proteins.** *Methods Enzymol* 1996, 275:310-328.

This review describes several techniques for developing high-volume *in vitro* screening assays for viral polymerases and related proteins. Several examples are illustrated.

8. Janda KD, Lo LC, Lo CH, Sim MM, Wang R, Wong CH, Lerner RA: **Chemical selection for catalysis in combinatorial antibody libraries.** *Science* 1997, 275:945-948.

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

This paper discusses the first law of random mutagenesis: you get what you screen for.

10. Crameri A, Whitehorn EA, Tate E, Stemmer WPC: **Improved green fluorescent protein by molecular evolution using DNA shuffling.** *Nat Biotechnol* 1996, 14:315-319.

11. Heim R, Tsien RY: **Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer.** *Curr Biol* 1996, 6:178-182.

12. Siemering KR, Golbik R, Sever R, Haseloff J: **Mutations that suppress the thermosensitivity of green fluorescent protein.** *Curr Biol* 1996, 6:1653-1663.

13. Brunet AP, Huang ES, Huffine ME, Loeb JE, Weltman RJ, Hecht MH: **The role of turns in the structure of an α-helical protein.** *Nature* 1993, 364:355-358.

14. Ybe JA, Hecht MH: **Sequence replacements in the central β-turn of plastocyanin.** *Protein Sci* 1996, 5:814-824.

15. • Hawrani AS, Sessions RB, Moreton KM, Holbrook JJ: **Guided evolution of enzymes with new substrate specificities.** *J Mol Biol* 1996, 264:97-110.

The paper describes a systematic approach to screening variants of an NAD-dependent dehydrogenase for several features simultaneously: thermal stability and catalytic activity on various substrates, without substrate inhibition and in the absence of an allosteric activator.

16. Shinkai A, Hirano A, Aisaka K: **Substitutions of Ser for Asn-163 and Pro for Leu-264 are important for stabilization of lipase from *Pseudomonas aeruginosa*.** *J Biochem* 1996 120:915-921.

17. Sattler A, Kanka S, Maurer KH, Riesner D: **Thermostable variants of subtilisin selected by temperature-gradient gel electrophoresis.** *Electrophoresis* 1996, 17:784-792.

18. Zhang JH, Dawes G, Stemmer WPC: **Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening.** *Proc Natl Acad Sci USA* 1997, 94:4504-4509.

19. • Lama J, Carrasco L: **Screening for membrane-permeabilizing mutants of the poliovirus protein 3AB.** *J Gen Virol* 1996, 77:2109-2119.

An *in vivo* screen, which may be applicable to other eukaryotic proteins known to form pores, is developed to differentiate permeabilizing wild-type 3AB and nonpermeabilizing 3AB mutants.

20. Konopka JB, Margarit SM, Dube P: **Mutation of Pro-258 in transmembrane domain 6 constitutively activates the G protein-coupled α-factor receptor.** *Proc Natl Acad Sci USA* 1996, 93:6764-6769.

21. Sidhu SS, Borgford TJ: **Selection of *Streptomyces griseus* protease B mutants with desired alternations in primary specificity using a library screening strategy.** *J Mol Biol* 1996, 257:233-245.

22. Kim JY, Caterina MJ, Milne JLS, Lin KC, Borleis JA, Devreotes PN: **Random mutagenesis of the cAMP chemoattractant receptor, cAR1, of *Dictyostelium*.** *J Biol Chem* 1997, 272:2060-2068.

23. Parent CA, Devreotes PN: **Isolation of inactive and G protein-resistant adenylyl cyclase mutants using random mutagenesis.** *J Biol Chem* 1996, 270:22693-22696.

24. Youvan DC, Goldman E, Delagrave S, Yang MM: **Digital imaging spectroscopy for massively parallel screening of mutants.** *Methods Enzymol* 1995, 246:732-749.

25. • Broach JR, Thorner J: **High-throughput screening for drug discovery.** *Nature* 1996, 384:14-16.

This review describes the implementation, limitations and future directions of high-throughput screening for drug discovery. Many of the discussions are applicable to screening protein libraries.

26. Robertson DE, Mathur EJ, Swanson RV, Marrs BL, Short JM: **The discovery of new biocatalysts from microbial diversity.** *SIMS News* 1996, 46:3-7.

27. • Zhao H, Arnold FH: **Functional and non-functional mutations distinguished by random recombination of homologous genes.** *Proc Natl Acad Sci* 1997, 94:7997-8000.

Screening of a randomly recombined gene library allows identification of functional mutations in a laboratory-evolved gene encoding a thermostable subtilisin.

28. Moore JC, Jin H, Kuchner O, Arnold FH: **Strategies for the *in vitro* evolution of protein function: enzyme evolution by random recombination of improved sequences.** *J Mol Biol* 1997, in press.

29. Shafikhani S, Siegel RA, Ferrari E, Schellenberger V: **Generation of large libraries of random mutants in *Bacillus subtilis* by PCR-based plasmid multimerization.** *Bio-Techniques* 1997, in press.

30. Checovich WJ, Bolger RE, Burke T: **Fluorescence polarization a new tool for cell and molecular biology.** *Nature* 1995, 375:254-256.

31. Tawfik DS, Green BS, Chap R, Sela M, Eshhar Z: **catELISA: a facile general route to catalytic antibodies.** *Proc Natl Acad Sci USA* 1993, 90:373-377.

32. Tawfik DS, Lindner AB, Chap R, Kim SH, Green BS, Eshhar Z: **catELISA: ELISA-based detection of catalytic antibodies and enzymes.** In *Immunology Methods Manual*, chapter 8.8. San Diego: Academic Press; 1997:553-560.

33. Macbeath G, Hilvert D: **Monitoring catalytic activity by immunoassay: implications for screening.** *J Am Chem Soc* 1994, 116:6101-6106.

34. Fenniri H, Janda KD, Lerner RA: **Encoded reaction cassette for the highly sensitive detection of the making and breaking of chemical bonds.** *Proc Natl Acad Sci USA* 1995, 92:2278-2282.

35. • Janes LE, Kazlauskas RJ: **Quick E. A rapid spectrophotometric method to measure the enantioselectivity of hydrolases.** *J Org Chem* 1997, in press.
A rapid spectrophotometric method is developed to measure the enantioselectivity of hydrolases on the basis of the addition of a chromogenic reference compound to introduce competitive binding for both enantiomers.

36. • Moore JC, Arnold FH: **Directed evolution of a para-nitrobenzyl esterase for aqueous–organic solvents.** *Nat Biotechnol* 1996, 14:458-467.
This paper describes a rapid screening assay using a chromogenic substrate (an antibiotic *p*-nitrophenyl ester) which serves as a surrogate for the desired substrate of an esterase (the *p*-nitrobenzyl ester). The screen is validated against a set of random mutants.

37. Panchal RG, Cusack E, Cheley S, Bayley H: **Tumor protease-activated, pore-forming toxins from a combinatorial library.** *Nat Biotechnol* 1996, 14:852-855.

38. • Black ME, Newcomb TG, Wilson HMP, Loeb LA: **Creation of drug-specific herpes simplex virus type 1 thymidine kinase mutants for gene therapy.** *Proc Natl Acad Sci USA* 1996, 93:3525-3529.
Functional complementation is used to select 426 active mutants from a library of more than 10^6 HSV-1 TK mutants. Active mutants are then screened for enhanced sensitivity to gancyclovir and/or acyclovir.

39. Christians FC, Loeb LA: **Novel human DNA alkyltransferases obtained by random substitution and genetic selection in bacteria.** *Proc Natl Acad Sci USA* 1996, 93:6124-6128.

40. Warren MS, Marolewski AE, Benkovic SJ: **A rapid screen of active site mutants in glycinamide ribonucleotide transformylase.** *Biochemistry* 1996, 35:8855-8862.

41. • Axe DD, Foster NW, Fersht AR: **Active barnase variants with completely random hydrophobic cores.** *Proc Natl Acad Sci USA* 1996, 93:5590-5594.
The autotoxity of barnase is used to identify functional barnases from a mutant library in which 12 hydrophobic core residues are randomly replaced by hydrophobic alternatives. A strikingly high proportion of the mutants are found to retain enzymatic activity *in vivo*.

42. Jucovic M, Hartley RW: ***In vivo* system for the detection of low level activity barnase mutants.** *Protein Eng* 1995, 8:497-499.

43. Venekei I, Hedstrom L, Rutter WJ: **A rapid and effective procedure for screening protease mutants.** *Protein Eng* 1996, 9:85-93.

44. Köhler JM, Schober A, Schwienhorst A: **Micromechanical elements for microchemical systems.** *Exp Tech Phys* 1994, 40:35-56.

45. Schober A, Günther R, Schwienhorst A, Döring M, Lindemann BF: **Accurate high-speed liquid handling of very small biological samples.** *Bio-Techniques* 1993, 15:324-329.

46. Eigen M, Rigler R: **Sorting single molecules: application to diagnostics and evolutionary biotechnology.** *Proc Natl Acad Sci USA* 1994, 91:5740-5747.

47. Rigler R: **Fluorescence correlations, single-molecule detection and large number screening applications in biotechnology.** *J Biotechnol* 1995, 41:177-186.

48. Eigen M: **The origin of genetic information: viruses as models.** *Gene* 1993, 135:37-47.

49. • Georgiou G, Stathopoulos C, Daugherty PS, Nayak AR, Iverson BL, Curtiss R III: **Display of heterologous proteins on the surface of microorganisms: from the screening of combinatorial libraries to live recombinant vaccines.** *Nat Biotechnol* 1997, 15:29-34.
This review describes recent progress in the expression and display of heterologous proteins on the surfaces of microorganisms. This approach can be used with cell-sorting techniques for screening large protein libraries.