Evolutionary Molecular Engineering Based on the Structure of Fitness Landscape

1999

Tomoaki Matsuura

Evolutionary Molecular Engineering Based on the Structure of Fitness Landscape

1999

Tomoaki Matsuura

Evolutionary Molecular Engineering Based on the Structure of Fitness Landscape

(適応度地形に基づく進化分子工学的手法)

Tomoaki Matsuura

1999

Contents

General Introduction ---- 3

Chapter 1

Nonadditivity of Mutational Effects on the Properties of Catalase I and its Application to Efficient Directed Evolution. ---- 8

1.1 Introduction

1.2 Material and methods

1.3 Results and discussion

1.4 Summary

Chapter 2

Evolutionary Molecular Engineering by Random Elongation Mutagenesis. ---- 36

2.1 Introduction

2.2 Material and methods

2.3 Results and discussion

2.4 Summary

Conclusion ---- 58

Appendix ---- 62

Reference ---- 64

List of Publications ---- 68

Acknowledgment ---- 69

Introduction

Natural enzymes have long been utilized in industries as well as in agriculture. Hence, the importance of enzymes as part of the progress in the life of human being is of no question. However, utilizing these enzymes *in vitro* as obtained from nature face several drawbacks when it comes to the issues like stability, effectiveness, and level of the property. This may due to the fact that enzymes when placed in a foreign environment become unfit for exerting its function. Recently, the problem involving the utilization of enzymes are now being resolved through the advancement in technology, for instance, DNA manipulation.

Enzyme, with or without subunits is a linear polymer composed of 20 different kinds of amino acids. Polymerization of amino acids makes up the primary structure of the proteins that folds into the proper tertiary structure for its function. Though gathering the information on the relationships among primary structure, tertiary structure and the functions of proteins have aided in understanding features contributing to protein functions, still the know-how to manipulate and design an enzyme as desired is not achieved (Arnold, 1993; Arnold, 1996). Nevertheless, these knowledge on hand provided us ideas as to how proteins are folded for its function. In consideration of the know-how, possibility exist that the missing link may be conquered by imitating the way nature has done to

evolve proteins we know at present.

Darwinian principle states that natural proteins have evolved and developed by alteration of amino acid residues and by natural selection. This principle becomes the basis of the advancement in evolutionary molecular engineering which first involves the creation of mutated gene library. Gene products were screened and those with improved property are selected and subjected to further cycles of mutation and selection in order to accumulate beneficial mutations. The experimental mutation and selection, hence, allows the observation of protein evolution in much a shorter span when compared to that of natural evolution.

Polymers abundant in nature are widely diversified. The constituent of polymers can simply be comprised of the 4 different bases as in RNA and DNA or the 20 essential amino acids as in enzymes. The varieties of possible sequences in each polymers are prone to mutation during evolution which way lead to accumulation of beneficial mutations. The turn of events in mutation and selection can be the cause of improvement in functions of the polymers or enzymes. Therefore, evolutionary molecular engineering can be applicable for designing the desired polymers like the RNA (Ekland *et al.*, 1995; Ellington & Szostak, 1990) and DNA (Cuenoud & Szostak, 1995) molecules.

The concept of fitness landscape was first introduced by Wright (Wright, 1932). This concept was then the basis of Smith (Smith, 1970) for

the "protein sequence space" and adaptive walk through the fitness landscape. The sequence space consists of all 20^N proteins, length N, arranged such that each protein is a vertex next to all 19N single mutant variants obtained by replacing one amino acid at one position by one of the 19 remaining possible amino acids. Each protein is assigned some "fitness" with respect to its ability to perform a specific function, such as catalyzing a given reaction, binding a given ligand, or a stability in a given condition. The distribution of these fitness values across the space constitutes the fitness landscape. Sequential cycles of random mutation and selection corresponds to the adaptive walk which is a successive movement to a fitter variants toward the attainable local or global optima.

Sequence space and fitness landscape have been widely used for theoretical works to study the evolution of polymers and enzyme molecules (Aita & Husimi, 1996; Huynen *et al.*, 1996; Kauffman & Levin, 1987; Kauffman & Weinberger, 1989; Schuster, 1995). However, whether or not will the concept brought to the experimental level is an issue to tackle. Previously, Trakulnaleamsai, S. *et al* (1995) had shown that indeed the fitness landscape can be visualized through the analysis of amino acid sequence and the property of an enzyme. Local sequence space around the wild-type protein of catalase I from *Bacillus stearothermophilus*, with respect to its catalatic activity, peroxidatic activity and thermostability showed that the landscape of thermostability was of a single hill with the

wild-type enzyme located near the top of the hill, and that of the catalatic and peroxidatic activities are rugged with hills and valleys and the wild-type enzyme was located on the side of the hill. Furthermore, the fitness landscape of catalase I indicated that as compared to other two activities improvement on thermostability might face slim chances if additional cycles of mutation and selection will be imposed. Therefore, it was shown that fitness landscape can provide a guide map for adaptive walk.

Through the analysis of the fitness landscape, certain features essential for the adaptive walk were brought to attention. In this thesis, we thereby dealt with these features for an effective application of the fitness landscape and the introduction of a new method for evolutionary molecular engineering as based from the analysis of fitness landscape.

In the first chapter, we have elucidated parameters that is necessary to estimate the efficient mutation rate and appropriate population size that will lead to the improvement of the enzyme function during sequential cycles of mutation and selection. The average and the variance of the effects of all the one-point mutations on a property of an enzyme, $E(z_O)$ and $V(z_O)$,respectively, and the degree of nonadditivity of mutational effect on the property of an enzyme had been estimated through the statistical analysis of the experimental data obtained from the mutant libraries of catalase I from *Bacillus stearothermophilus*. The application of these parameters for the prediction of efficient mutation rate and population size to prompt the

effective adaptive walk is discussed.

In the second chapter, we developed a new method of mutagenesis to conquer limitations of point mutagenesis and DNA shuffling in cases where an enzyme's property is supposedly optimized as deduced from the fitness landscape. The method involved the expansion of the sequence space through the addition of a peptide tail to the C- terminus of an enzyme which gives rise to new landscape. This new landscape on the expanded sequence space renders additional allowance for improving enzyme function. Here, the effectively of peptide tail addition in the study of evolutionary molecular engineering are discussed.

Chapter 1

Nonadditivity of mutational effects on the properties of catalase I and its application to efficient directed evolution

1.1. Introduction

Directed evolution of enzymes by sequential cycles of random mutagenesis and screening have proved to be useful for obtaining new or improved properties (Arnold, 1993; Arnold, 1996; Moore & Arnold, 1996). This strategy is thought to be convenient and even primitive as no structural information is required. Theoretical works on fitness landscapes (Aita & Husimi, 1996; Aita & Husimi, 1998) is also underway with the same aspects and motivations. Therefore, it may be best to integrate both experimental and theoretical aspects so as to provide a more efficient strategy for the adaptive walk in directed evolution.

Catalase I from *Bacillus stearothermophilus* is a member of the bacterial catalases with broad-spectrum peroxidatic activity (Trakulnaleamsai *et al.*, 1992) having both the catalatic and peroxidatic activity. Kinetic studies revealed that the reaction of this enzyme is 95% specific to catalase and 5% to peroxidase of the total activity(Yomo *et al.*, 1997). In our previous work, we prepared a mutant library of the wild-type

enzyme of catalase I by random mutagenesis of the gene (Trakulnaleamsai *et al.*, 1995). A mutant having a higher peroxidatic activity and a lower catalatic activity than that of the wild-type enzyme was selected from the library (Trakulnaleamsai *et al.*, 1995) and named as D130N based on the amino acid substitution at position 130 from Asp to Asn. D130N mutant was then used in the present work. There are many examples of mutant enzymes with different substrate specificity, but few, if any, with different reaction specificity (Lewis *et al.*, 1997). Therefore, it will be interesting to walk towards different reaction specificity, i.e., catalase to peroxidase. The gene encoding the D130N mutant was subjected to random mutagenesis and one mutant was selected and characterized from the second mutant library. The reaction specificity of this mutant turn out to be 58% specific to peroxidase.

The above results demonstrated the usefulness of directed evolution. However, to construct an efficient strategy of adaptive walk in directed evolution, it is essential to know the following basic parameters characterizing a random mutant population: average and variance of the effects of one-point mutations and degree of nonadditivity of the mutational effects.

When the sum of the free energy changes of a functional property caused by single mutations is equal to that by multiple mutations, then the mutational effects on the property are defined to be additive. A review of compiled data of site-directed mutagenesis of proteins shows that, in most

cases, combination of mutations that affect substrate or transition-state binding, protein-protein interactions, DNA-protein recognition, or protein stability is simply additive (Wells, 1990). In addition, some deviations from simple additivity have also been observed when the sites of mutations interact with one another (Wells, 1990). To date, these data are still limited to a relatively small number of cases.

In this work, we have developed a new method for estimating the degree of nonadditivity of the mutational effects using two sample populations derived from the first and second random mutagenesis. The average degree of nonadditivity was estimated to be 0.13 irrespective of the properties of the enzyme. Taking into consideration of the degree of nonadditivity, we have estimated the average and variance of the effects of one-point mutations from the data of the first mutant population. The reliability of these basic parameters in the prediction of the distribution pattern of a property of the second mutant population was demonstrated. Based on the accuracy of the prediction, appropriate mutation rate and effective sample size for further mutation and selection can be determined. In short, the strategy of an efficient adaptive walk in directed evolution involves the consecutive processes of sequential cycles of random mutagenesis, analysis of the mutant population, selection of a mutant with desired properties, and determination of the conditions for the next mutagenesis.

1.2. Material and Methods

1.2.1. Bacterial strains and plasmids

Escherichia coli UM228, a catalase HPI-deficient mutant (Triggs-Raine & Loewen, 1987), was kindly provided by Dr. P. C. Loewen, University of Manitoba, Winnipeg, Canada. Plasmid pOD64, a derivative of pUC19 containing the gene of catalase I from *B. stearothermophilus*, was prepared previously (Loprasert *et al.*, 1990).

1.2.2. Preparation of random mutant population

The random mutant population used in this study was prepared by exactly the same way as that in our previous work (Trakulnaleamsai *et al.*, 1995). The gene encoding the D130N mutant of catalase I obtained previously (Trakulnaleamsai *et al.*, 1995) was subjected to random mutagenesis with sodium nitrite (Myers *et al.*, 1985). Plasmids harboring mutant genes were introduced into *E. coli* UM228 cells, and 2,624 transformants were obtained. The catalatic and peroxidatic activities of the transformants were examined by colony assay as described previously (Trakulnaleamsai *et al.*, 1995). The transformants were classified into H group when both activities were of similar level to those of D130N, M group when both or either one of the activities was weaker than those of

D130N, and B group when neither of the activities was detectable. There were 1356, 654, and 614 transformants that belong to H, M, and B groups, respectively. A sample population was drawn out from the mutant library as based on the productivity of the mutant enzymes. The transformants with too low productivity of the enzymes in accordance to the detectable level of sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) protein measurement were excluded from the sample population. In addition, the transformants in the H group were not included to exclude the probability of incorporating the starting mutant material (D130N) into the sample population. From the 599 M-group transformants producing detectable amount of the enzyme protein, 141 transformants were arbitrarily chosen and used as the sample population. The sample population was prepared and chosen by exactly the same methodology as that in the previous work (Trakulnaleamsai *et al.*, 1995) for direct comparison on the effect of mutagenesis.

1.2.3. Characterization of the mutant population

For enzyme characterization, cell lysates of the *E. coli* before and after heat treatment were prepared. *E. coli* UM228 cells harboring plasmid bearing a mutant gene of catalase I were grown at 37℃ for 22 h in 2xTY medium (Maniatis *et al.*, 1998) containing 50 μg/ml ampicillin. The cells obtained from a 40-ml culture were suspended in 2 ml 0.1 M potassium

phosphate (pH 7.0), disrupted by sonication, and centrifuged. The supernatant obtained after centrifugation was the cell lysate before heat treatment. The supernatant obtained after centrifugation of the lysate heated at 70℃ for 10 min was the heat-treated lysate.

Both types of lysates were analyzed by SDS-PAGE (Laemmli, 1970), and the catalase protein concentration in the lysates was determined from the intensity of the protein bands visualized by Coomassie brilliant blue as described previously (Trakulnaleamsai *et al.*, 1995). Thermostability (S) of catalase was calculated as the ratio of the catalase protein concentration measured after heat treatment (70℃,10 min) of the cell lysate to that before heat treatment. This is based on the fact that heat denaturation causes insolubilization and precipitation of this enzyme.

Catalatic and peroxidatic activities of the lysate before heat treatment were measured spectrophotometrically at 30℃ as described previously (Trakulnaleamsai *et al.*, 1992). Reaction mixtures contained 10 μl cell lysate and 3 ml substrate solution. For the catalatic activity, the substrate solution contained 20 mM H_2O_2 in 0.1 M potassium phosphate (pH 7.0). Enzyme reactions were recorded as the decrease in absorbance at 240 nm (H_2O_2 concentration), and the reaction rate was calculated from the maximum slope using the molar absorption coefficient of H_2O_2 of 43.6 $M^{-1}cm^{-1}$ (Hildebraunt & Roots, 1975). The value of the catalatic activity was

expressed as half of the consumption rate of H_2O_2, because two molecules of H_2O_2 are consumed per catalytic cycle. For the peroxidatic activity, the substrate solution contained 20 mM H_2O_2, 4.1 mM 2,4-dichlorophenol, and 0.67 mM 4-aminoantipyrine in 0.1 M potassium phosphate (pH 7.0). Reactions were recorded as the increase in absorbance at 500 nm due to the formation of a red dye. The reaction rate was calculated from the maximum slope using the absorption coefficient of the red color of 1.63 $x10^4 M^{-1}cm^{-1}$ (Trakulnaleamsai *et al.*, 1992). Both activity values were divided by the subunit protein concentration obtained from SDS-PAGE, and the specific activity was expressed as the turnover number of catalytic cycle per second. The background catalatic activity detected from the host cell was subtracted with the assumption that this activity is proportional to the protein concentration of the host cell (Trakulnaleamsai *et al.*, 1995). There is no background peroxidatic activity detected from the host cell.

1.2.4. Reaction specificity of purified enzymes

The reaction specificity was evaluated as the percentage of conversion of added H_2O_2 into the dye product (A500) in the presence of excess amount of substrates for the peroxidatic reaction. Reactions (30℃) were initiated by adding 20 μl enzyme solution into 3 ml substrate solution containing 20 μM H_2O_2, 4.1 mM 2,4-dichlorophenol, and 0.67 mM 4-aminoantipyrine in 0.1 M potassium phosphate (pH 7.0). The increase in

absorbance at 500 nm was monitored until completion of the reaction, and the total amount of dye produced by the reaction was calculated using the absorption coefficient of 1.63 x 10^4 $M^{-1}cm^{-1}$ (Trakulnaleamsai *et al.*, 1992). Horseradish peroxidase (Toyobo Co. Ltd., Osaka) was used as a typical peroxidase with no catalatic activity. It was confirmed that the values of the conversion obtained for the purified wild-type and mutant enzymes were independent of the H_2O_2 concentration (3-30 μM) used for the assay.

1.2.5. Equations for the estimation of the degree of nonadditivity of the mutational effects and for the prediction of the statistical properties of the next random mutant population

Let us consider the first mutant population prepared by random mutagenesis on the gene encoding a wild-type model enzyme. The difference (z) in a property between the wild-type and a mutant enzyme is defined by

$$z = x - x_w \qquad (1)$$

where x_W and x are the values of the property of the wild-type and the mutant enzyme, respectively; these values are expressed on the free energy level. If the number of mutations in the mutant sequence is m, and if these mutations occur one by one in an arbitrary order, the difference, z, is

expressed as:

$$z = \sum_{i=1}^{m} z_i + \frac{1}{2!} \sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} z_{ij} \qquad (2)$$

where z_i is the change in the property caused by the single mutation at the i'th site, and z_{ij} reflects the extent of the nonadditivity of the effects of double mutations caused by the interaction between the two mutation sites. Here, for simplicity, the terms reflecting the nonadditivity due to the interactions among three or more mutation sites are neglected. In addition, the extent of the nonadditivity is considered to be negatively proportional to the sum of the effects of two single mutations; i.e.,

$$z_{ij} = -b(z_i + z_j) \qquad (3)$$

where b is a constant showing the degree of nonadditivity. Take note that Equations 2 and 3 are adapted for a simple model protein where b is constant. Therefore, simplicity is applied for expressing the essential effects of nonadditivity on statistical average and variance of the mutational effects. From Equations 2 and 3, the average of z of the first population, $E(z_1)$, is expressed as:

$$
\begin{aligned}
\mathrm{E}(z_{\mathrm{I}}) &= \mathrm{E}(\sum_{i=1}^{m} z_i) - \frac{b}{2!}\mathrm{E}(\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m}(z_i + z_j)) \\
&= \mathrm{E}(m)\mathrm{E}(z_{\mathrm{O}}) - \frac{b}{2}\times 2\,\mathrm{E}(z_{\mathrm{O}})\mathrm{E}(m(m-1)) \\
&= \mathrm{E}(m)\mathrm{E}(z_{\mathrm{O}}) - b\mathrm{E}(z_{\mathrm{O}})\mathrm{E}(m^2 - m) \\
&= \mathrm{E}(m)\mathrm{E}(z_{\mathrm{O}}) - b\mathrm{E}(z_{\mathrm{O}})(\mathrm{V}(m) + \mathrm{E}(m)^2 - \mathrm{E}(m))
\end{aligned} \tag{4}
$$

where $\mathrm{E}(m)$ and $\mathrm{V}(m)$ are the average and variance of the number of mutations in a sequence, respectively, and $\mathrm{E}(z_{\mathrm{O}})$ is the average of the effects of one-point mutations. On the other hand, the variance of z of the first population, $\mathrm{V}(z_{\mathrm{I}})$, is expressed as:

$$
\begin{aligned}
\mathrm{V}(z_{\mathrm{I}}) &= \mathrm{E}[\{z_{\mathrm{I}} - \mathrm{E}(z_{\mathrm{I}})\}^2] \\
&= \mathrm{E}[\{\sum_{i=1}^{m} z_i - \frac{b}{2!}\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m}(z_i + z_j) - \mathrm{E}(z_{\mathrm{I}})\}^2]
\end{aligned} \tag{5}
$$

Assuming a Poisson distribution for the number of mutation, Equation 5 is converted to

$$
\begin{aligned}
\mathrm{V}(z_{\mathrm{I}}) &= \mathrm{E}(m)[\mathrm{E}(z_{\mathrm{O}})^2\{2b\mathrm{E}(m) - 1\}^2 + \mathrm{V}(z_{\mathrm{O}})\{b\mathrm{E}(m) - 1\}^2 \\
&\quad + b^2\mathrm{E}(m)\{\mathrm{V}(z_{\mathrm{O}}) - \mathrm{E}(z_{\mathrm{O}})^2\}]
\end{aligned} \tag{6}
$$

where $\mathrm{V}(z_{\mathrm{O}})$ is the variance of the effects of one-point mutation. Derivation

of Equation 6 is shown in the appendix.

In Equation 4, the values of $E(z_I)$, $E(m)$, and $V(m)$ can be obtained from experimental analysis of the first mutant population. However, to obtain the values of b and $E(z_O)$, a second mutant population is required. Suppose that the second mutant population is prepared by the same random mutagenesis on a mutant, M, in the first population, and that the occurring mutation sites are different from those exist in mutant M, the average of z of the second population, $E(z_{II})$, is written similarly to Equation 4 as:

$$\begin{aligned} E(z_{II}) &= E(\sum_{k=1}^{n} z_k) - \frac{b}{2}E[\sum_{k=1}^{n}\sum_{j=1,j\neq k}^{n}(z_k + z_j)] \\ &\quad + E(\sum_{i=n+1}^{n+m} z_i) - \frac{b}{2}E[\sum_{i=n+1}^{n+m}\sum_{j=n+1,j\neq i}^{n+m}(z_i + z_j)] \\ &\quad - bE[\sum_{k=1}^{n}\sum_{j=n+1}^{n+m}(z_k + z_j)] \\ &= z_M + [E(m) - b\{V(m) + E(m)^2 - E(m)\}]E(z_O) \\ &\quad - bE(m)[nE(z_O) + \sum_{k=1}^{n} z_k] \end{aligned} \tag{7}$$

here z_M is the difference in the property between the wild-type and mutant M, and n is the number of mutations in the sequence of mutant M. As $E(m)$ and $V(m)$ of the second population is the same as those of the first one, Equation 7 is written using Equation 4 as:

$$E(z_{II}) = z_M + E(z_I) - bE(m)[nE(z_O) + \sum_{k=1}^{n} z_k] \quad (8)$$

Assuming a Poisson distribution for the number of mutations, the variance of z of the second population, $V(z_{II})$, is expressed as:

$$\begin{aligned} V(z_{II}) = E(m)[&\{E(z_O)(2bE(m) + bn - 1) + b\sum_{k=1}^{n} z_k\}^2 \\ &+ V(z_O)\{b(n+E(m)) - 1\}^2 \\ &+ b^2E(m)\{V(z_O) - E(z_O)^2\}] \end{aligned} \quad (9)$$

Derivation of Equation 9 is shown in the appendix.

Rearranging Equation 8 gives

$$E(z_{II}) - [E(z_I) + z_M] = -b\,E(m)[\sum_{k=1}^{n} z_k + nE(z_O)] \quad (10)$$

The left-hand side of Equation 10 shows the observed value of nonadditivity, NA_{ob}, i.e.,

$$NA_{ob} = E(z_{II}) - [E(z_I) + z_M] \quad (11)$$

On the other hand, the right-hand side shows the theoretical value, NA_{th}, and is transformed by using Equation 4 to

$$NA_{th} = -b\,E(m)\sum_{k=1}^{n}[z_k + E(z_I)/[E(m) - b\{V(m) + E(m)^2 - E(m)\}]] \quad (12)$$

In these equations, $E(m)$, $V(m)$, and n are experimentally obtainable constants. Though NA_{ob}, $\sum_{k=1}^{n} z_k$ and $E(z_I)$ are also experimentally obtainable, the values vary depending on the observed properties. Assuming that the b value is constant independent of the observed properties, the best estimate of b value is a value that minimizes the sum of $(NA_{ob} - NA_{th})^2$ of the observed properties. As shown in our results described below, the estimated b value, b, can commonly be used as the average degree of nonadditivity for thermostability, catalatic activity, and peroxidatic activity.

The values of $E(z_{II})$ and $V(z_{II})$ of any second mutant population can be estimated from the data of the first population. The $E(z_O)$ and $V(z_O)$ values are obtained from Equations 4 and 6, respectively. Using these values, $E(z_{II})$ and $V(z_{II})$ can be estimated from Equations 7 and 9, respectively. Using the estimated values, appropriate conditions for the second mutagenesis can be predicted as described under Results and discussion. The consecutive steps of mutagenesis, analysis, and prediction can continue as long as required, and will be an efficient strategy of directed evolution.

1.3. Result and Discussion

1.3.1. Second random mutagenesis

Catalase I from *B. stearothermophilus* has high catalatic and low peroxidatic activities (Trakulnaleamsai *et al.*, 1992; Yomo *et al.*, 1997). That is, the reaction specificity of the wild-type enzyme is more of catalase than peroxidase. The first random mutagenesis of the gene encoding the wild-type enzyme gave rise to a mutant enzyme having a lower catalatic activity than that of the wild-type enzyme and a higher peroxidatic activity (Trakulnaleamsai *et al.*, 1995). The nucleotide sequence of the gene encoding the mutant enzyme was determined using a Toyobo Sequence Kit (Toyobo Co. Ltd., Osaka). The deduced amino acid sequence revealed that Asp-130 of the wild-type enzyme is altered to Asn, from which the mutant is named D130N. To change the reaction specificity further to peroxidase side, the gene encoding D130N was subjected to random mutagenesis, and the second mutant population comprising of 141 mutants was obtained as described under Materials and methods. It should be noted that the second mutant population was prepared by exactly the same way as that of the first mutant population.

Thermostability (S), specific catalatic activity (CA), and specific peroxidatic activity (PA) of the 141 mutant enzymes were measured as described under Materials and methods. Fig. 1-1 shows the distribution of

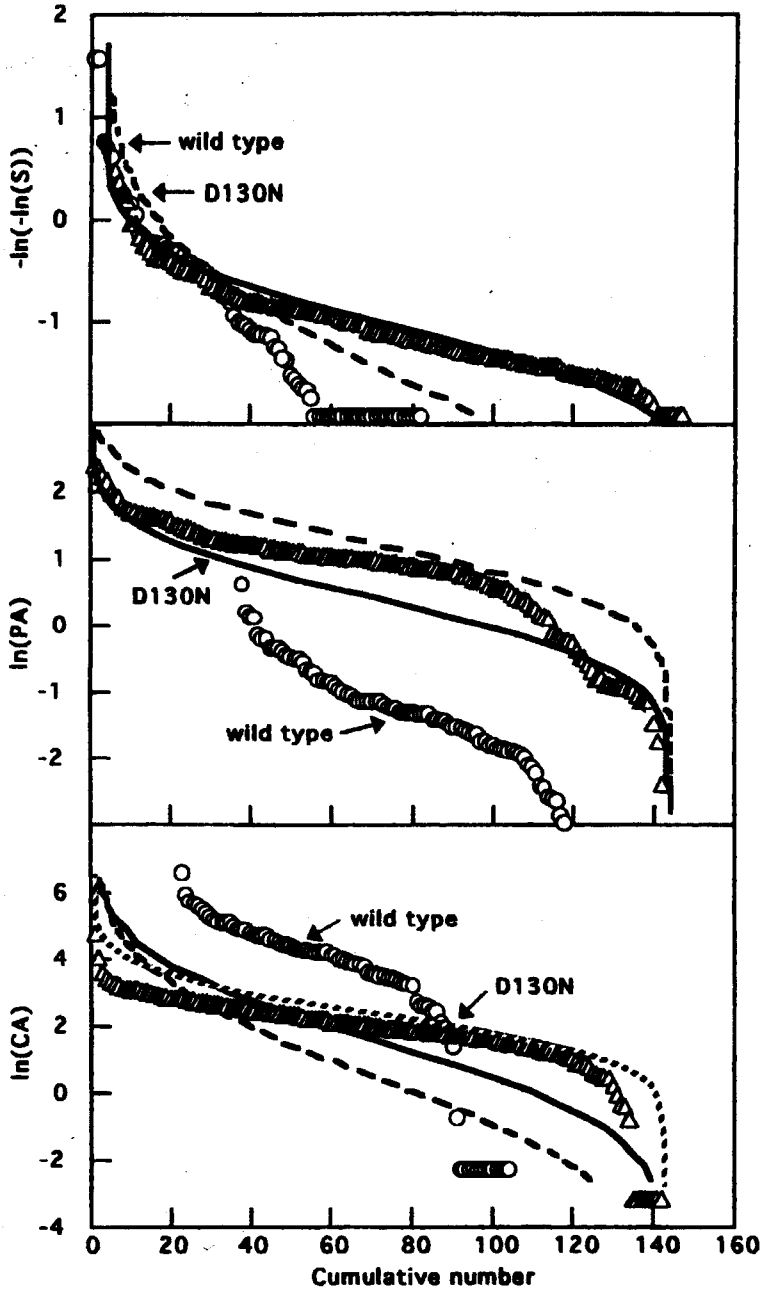

Fig. 1-1. Distribution pattern of the three properties of the first (○) and second (△) random mutant population. Thermostability (S), specific peroxidatic activity (PA), and specific catalatic activity (CA) of the mutant enzymes in the second population were measured as described under Materials and methods, and the values were arranged in the order of the highest to the lowest values; the positions of the mutants with regard to the cumulative number are 6-147, 1-142, and 1-142 for the thermostability, peroxidatic activity, and catalatic activity, respectively, including the position of the D130N mutant. For the thermostability, S values of less than 0.001 were regarded as 0.001, and for the catalatic activity, CA values of less than 0.04 s^{-1} were regarded as 0.04 s^{-1}. The data of the first mutant population obtained in our previous work (Trakulnaleamsai *et al.*, 1995) are also shown; the positions of the mutants including the wild-type with regard to the cumulative number are 1-82, 37-118, and 23-104, for the thermostability, peroxidatic activity, and catalatic activity, respectively. The positions of the wild-type enzyme (●) and the D130N mutant (▲) are indicated by the filled marks. The curved lines represent the expected distribution patterns of the second mutant population prepared by the random mutagenesis of the D130N mutant with $E(m) = V(m) = 2.5$. The lines were drawn assuming a normal distribution with $E(z_n)$ and $V(z_n)$ values calculated using Equations 7 and 9, respectively, where $n = 1$ and the z_M values are those given in Table 1-1. The $E(z_O)$ and $V(z_O)$ values in Equations 7 and 9 were calculated from Equations 4 and 6, respectively, under the following conditions: for the solid lines, the whole data of the first population and a b value of 0.13 were used (i.e., values in Table 1-1); the dashed lines calculated with $b = 0$; the broken line calculated after omitting the samples from no. 92 to 104 of the first population with $b = 0.13$ (i.e., values in parentheses in Table 1-1).

the first and second mutant populations as to the values of the above mentioned properties. As -ln(S), CA, and PA correspond to the rate constants, the position of the three properties in the distribution is expressed with the activation free energy level (Δ G*/RT) by using -ln[-ln(S)], ln(CA), and ln(PA) in Fig. 1-1.

Fig. 1-1 shows results as predicted from the landscape around catalase I in case of introducing further mutations (Trakulnaleamsai *et al.*, 1995). Almost all the mutants in the second population have lower thermostability than the D130N mutant as the thermostability of the wild-type enzyme is near the local optimum level and the probability of obtaining a mutant with much higher stability is very slim. For peroxidatic activity, many mutants at higher level than the highest mutant in the first population are found in the second population. This indicates that the peroxidatic activity of the D130N mutant, though is the highest in the first population, is still at a much lower level than the optimum and can be increased further by successive adaptive walks. It should be pointed out, however, that while the peroxidatic activity of the wild-type enzyme is almost the same as the average value obtained for the mutants of the first population, that of the D130N mutant is higher than the average of the second population. This indicates that the fraction of mutants having higher peroxidatic activity will decrease with the increase in the number of steps of the adaptive walk. The same tendency is observed for the catalatic activity such that the level of the

D130N mutant is lower than that of the wild-type and the fraction of the mutant with higher activity is increased in the second population. This tendency comes from the presence of nonadditivity of the mutational effects as described below.

1.3.2. Change in the reaction specificity

From the second mutant population, one mutant was selected as based on the shift of reaction specificity. The gene of the mutant enzyme was sequenced, and the amino acid sequence deduced. The mutant enzyme was found to have two alterations in the amino acid sequence of the D130N mutant, namely, Ile-108 to Thr and Ile-222 to Thr. Hence, the mutant is named as triple mutant (I108T/D130N/I222T).

To trace the shift of reaction specificity during the two-step adaptive walk, the wild-type, the D130N mutant, and the triple mutant were purified according to the procedures described previously (Trakulnaleamsai *et al.*, 1992). The reaction specificity of these enzymes was calculated as the percent conversion of the added H_2O_2 into the product of peroxidatic reaction, a red dye. The values obtained for the wild-type, the D130N mutant, and the triple mutant were 2%, 20%, and 58%, respectively. These results evidently show a shift in the reaction specificity of catalase I. Catalase I that was originally 98% specific to catalatic activity and only 2% to peroxidatic activity was converted to 58% specific to peroxidatic activity

after two-step walks. The results so far obtained confirm the effectiveness of directed evolution by adaptive walk for improving the properties of an enzyme.

1.3.3. Nonadditivity of the effects of mutation

From the data of the two random mutant populations prepared for the adaptive walk described above, the degree of nonadditivity of the mutational effects was estimated using the equations derived for the model protein as presented in Materials and methods. The observed values of nonadditivity, NA_{ob}, defined by Equation 11 were calculated for the three properties using the observed values of $E(z_I)$, $E(z_{II})$, and z_M listed in Table 1-1, where $E(z_I)$ and $E(z_{II})$ are the average of the differences in a property between the wild-type and the mutant enzymes in the first and the second mutant populations, respectively, and z_M is the difference in the property between the wild-type and the mutant M, the D130N mutant, used for the preparation of the second population. The average $E(m)$ and variance $V(m)$ of the number of mutations determined previously to be 2.5 and 1.5, respectively (Trakulnaleamsai *et al.*, 1995), are constant for the first and second mutant populations, as the same methods were used for the preparation of both mutant populations. The D130N mutant being one-point mutant of the wild-type, $n = 1$, and hence, $\sum_{k=1}^{n} z_k = z_M$. Accordingly, the theoretical value of nonadditivity, NA_{th}, defined by Equation 12, is

expressed as:

$$NA_{th} = -2.5\{E(z_I)/(2.5 - 5.25b) + z_M\} \quad (13)$$

where b is the degree of nonadditivity defined by Equation 3. Assuming that the b value is constant independent of the observed properties, the sum of $(NA_{ob} - NA_{th})^2$, i.e., the sum for the three properties was taken at the minimum level to estimate the common b value. The estimated b value, $\overline{b}$, is 0.13. For simplicity, a Poisson distribution for the number of mutations was assumed, i.e., $V(m) = E(m) = 2.5$, which indeed resulted to the same $\overline{b}$ value of 0.13. Accordingly, the Poisson distribution is assumed for the number of mutation in further analysis. Using this $\overline{b}$ value, the average and variance of the effects of one-point mutations, $E(z_O)$ and $V(z_O)$, can be calculated from Equations 4 and 6, respectively, and the values are listed in Table 1-1. The high correlation coefficient (R^2=0.999, p=0.01) with a regression line having a slope of -0.13 (Fig. 1-2) demonstrates that with the estimated $\overline{b}$ value, the observed nonadditivity for the three properties can be explained by the following relationship:

$$NA_{ob} = -\overline{b}E(m)[z_M + E(z_O)] \quad (14)$$

This means that Equation 10 derived for a simple model protein is applicable to catalase I with a common $\overline{b}$ value of 0.13 for the three

properties. It should be pointed out that when the values of b and $E(z_O)$ are calculated independently for each property from Equations 4 and 8, nearly the same value were obtained as shown in Fig. 1-2, hence, an identical regression line confirms that the average degree of nonadditivity of 0.13 is common for the three properties. This relationship will be applicable not only to the case of the D130N mutant but also to other mutations in the sequence of catalase I and to those of other enzymes. In fact, the $\overline{b}$ values calculated using the data of various proteins listed by Wells (Wells, 1990) are 0.10 and 0.07 for the transition-state stabilization and unfolding, respectively. Therefore, the average degree of nonadditivity of the mutational effects can be concluded to be in the range around 0.1 for globular proteins. In addition, as shown in Fig. 1-2, the observed value of nonadditivity, NA_{ob}, negatively correlates with the mutational effect, $E(m)[z_M + E(z_O)]$. Accordingly, from Equation 11, the diminishing effect of nonadditivity on the average of z of the second mutant population, $E(z_{II})$, will increase with the increase in the mutational effect.

Although the degree of nonadditivity is not large, it was confirmed by t test with unequal variances that the values of $E(z_{II})$ and $[E(z_I) + z_M]$ are significantly different ($p < 0.001$) for the three properties. In addition, as discussed below, nonadditivity is an essential factor in predicting the distribution of the next mutant population on the successive adaptive walks.

Table 1-1. The values of $E(z_I)$ and $V(z_I)$ were calculated from the data of the first mutant population (82 mutants) prepared previously (Trakulnaleamsai *et al.*, 1995). The z_M value is for the D130N mutant. The values of $E(z_{II})$ and $V(z_{II})$ were calculated from the data of the second mutant population (141 mutants) prepared in this work. The values of $E(z_O)$ and $V(z_O)$ were calculated from Equations 4 and 6, respectively, using a common b value of 0.13 and $E(m) = V(m) = 2.5$.

a :The values in parentheses were calculated after omitting the samples from no. 92 to 104 in the first population.

Properties	$E(z_I)$	$V(z_I)$	z_M	$E(z_{II})$	$V(z_{II})$	$E(z_O)$	$V(z_O)$
Thermo-stability	-1.75	0.81	-0.52	-1.85	0.23	-1.04	0.48
Peroxidatic activity	0.05	0.61	2.51	2.00	0.81	0.03	0.49
Catalatic activity	-1.35 (-0.35)[a]	6.62 (1.45)[a]	-2.63	-2.67	2.51	-0.80 (-0.21)[a]	5.21 (1.16)[a]

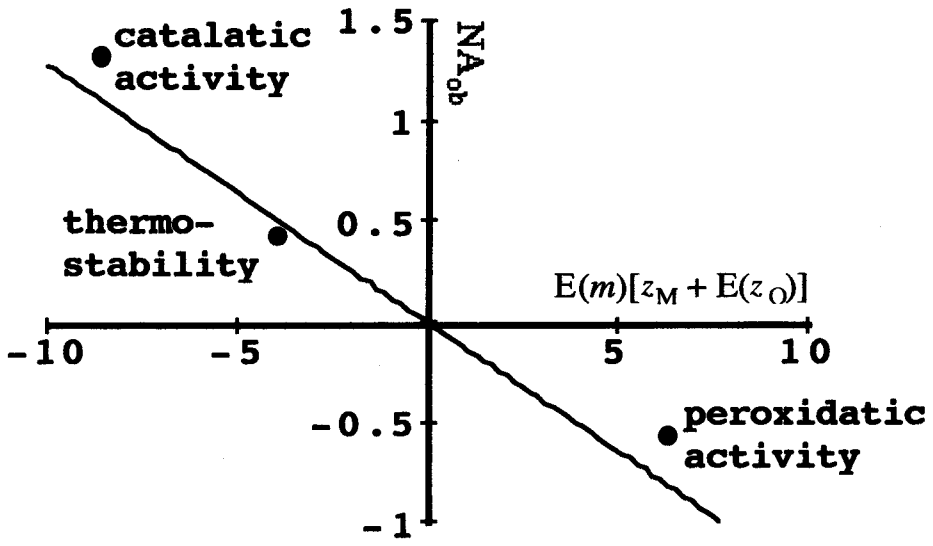

Fig. 1-2. Plot of the observed nonadditivity (NA_{ob}) versus $E(m)[z_M + E(z_O)]$ for the three properties of catalase I. The values of NA_{ob} defined by Equation 11 and $E(m)[z_M + E(z_O)]$ were calculated using the data listed in Table1-1 and an $E(m)$ value of 2.5.

1.3.4. Strategy of adaptive walk

The first step in the directed evolution of an enzyme involves the selection of the best mutant from the first mutant population. The gene encoding the selected mutant is further subjected to random mutagenesis, from which a mutant with better properties is again selected. The process of the consecutive steps of selection and mutation continues until a mutant enzyme having the desired properties is obtained. To have an effective directed evolution, it is essential to know the appropriate mutation rate and effective sample size in the process of mutation and selection.

From the analysis of the first mutant population, the values of $E(z_I)$, $V(z_I)$, and $E(m)$ are obtained. Assuming a Poisson distribution for the number of mutations, $V(m) = E(m)$. Using the common b value of 0.13 obtained above, the values of $E(z_O)$ and $V(z_O)$ are calculated from Equations 4 and 6, respectively. Supposing that a mutant M is selected from the first population and its gene is subjected to second random mutagenesis, the values of z_M, n, and $\sum_{k=1}^{n} z_k$ can then be obtained. Consequently, the values of $E(z_{II})$ and $V(z_{II})$ for the second mutant population can be estimated from Equations 7 and 9, respectively. Accordingly, the distribution of the properties of the mutants in the second population can be predicted assuming a normal distribution.

Taken as an example, the distribution pattern of the second mutant

population prepared by introducing random mutagenesis to the gene encoding the D130N mutant with an E(m) value of 2.5 was predicted through the data of the first mutant population derived from the wild-type catalase I, and compared with the observed data. Fig. 1-1 shows that the expected distribution pattern (solid lines) fits well to the observed data except that of catalatic activity. The observed deviation with the catalatic activity was analyzed and was attributed to the gap in the values of the activities between sample number 90 and 92 of the first population (Fig. 1-1) which causes overestimation of the $V(z_{II})$ value. In fact, omission of the samples from no. 92 to 104 for the calculation resulted to a well fitted expected pattern of the catalatic activity (Fig. 1-1, dotted line). These results show that unless the distribution pattern of the first population has a large gap, the above procedures can be applied directly for predicting the distribution pattern of the second population.

Fig. 1-1 also includes the results when the mutational effects are assumed to be additive, i.e., $b = 0$ (broken lines). Comparison of the broken lines with the solid ones clearly indicates that the nonadditivity of the mutational effects must be taken into account for the prediction of the distribution pattern of the second population.

As stated above, the distribution pattern of the second mutant population prepared from a mutant M by random mutagenesis with an average number of mutation of E(m) can be predicted assuming a normal

distribution for the second population. Accordingly, supposing that a mutant with a level of a property higher than z exists with a probability of P in the second population and a sample size of N, z can be expressed as a function of E(m). This relationship can be used as a guide for determining the conditions of the second mutagenesis and selection. That is, when we desired for a mutant with a level of thermostability higher than z at a probability of 0.5, we can determine the appropriate population size for analysis that is to be prepared at a mutation rate of E(m); or when we fixed the values of N and E(m), we can expect with a probability of 0.5 to obtain a mutant with thermostability higher than z (Fig. 1-3A).

When $E(z_O)$ value is positive, the chance to have a better mutant increases with the increase in E(m). However, as the property of the enzyme improves, $E(z_O)$ will decrease and become negative due to the increase in the chance of deleterious mutation and the decrease in that of beneficial one (Aita & Husimi, 1996). Interestingly, even $E(z_O)$ is negative, still a better mutant can be included in the population (Fig. 1-3A). That is, there will be an optimum E(m) value when $E(z_O)$ is negative. The presence of an optimum E(m) value comes from the balance of the effects of E(m) on the average [$E(z_{II})$] and standard deviation [$V(z_{II})^{1/2}$] of the second mutant population. When E(m) is small, the increase in the variety of the population exceeds the decrease in $E(z_{II})$ and increase the chance to take a better mutant, while with a larger E(m), the decrease in $E(z_{II})$ exceeds the increase in the

variety. In other words, when E(m) is too large, the probability of occurrence of beneficial mutations in one mutant sequence is much smaller than that of deleterious mutations, and hence, the chance to take a better mutant is very slim. In addition, the nonadditivity of the mutational effects suppresses the increase in the variety of the population.

In our previous paper, a parameter "smoothness" was introduced, which is defined as $| E(z_O) | /(V(z_O)^{1/2})$ (Trakulnaleamsai *et al.*, 1995). Above consideration indicates that when $E(z_O)$ is negative, the chance to obtain a better mutant increases with the decrease in the smoothness of the landscape in protein sequence space (Fig. 1-3B). Hence, b, $E(z_O)$, and $V(z_O)$, being the basic parameters for characterizing a random mutant population, are essential to the construction of an efficient strategy for adaptive walk in protein sequence space.

Fig. 1-3A also shows that although the z value increases with the increase in the sample size (N), the effect will not come up to the different numerical order of increase in the sample size, indicating that the stepwise adaptive walks with a small sample size has more advantage over the one-step selection from a large sample population. Therefore, an efficient adaptive walk will be achieved by determining appropriate values of E(m) and N at each step of mutagenesis based on the analysis shown in Fig. 1-3 with the knowledge of the basic parameters of the random mutant

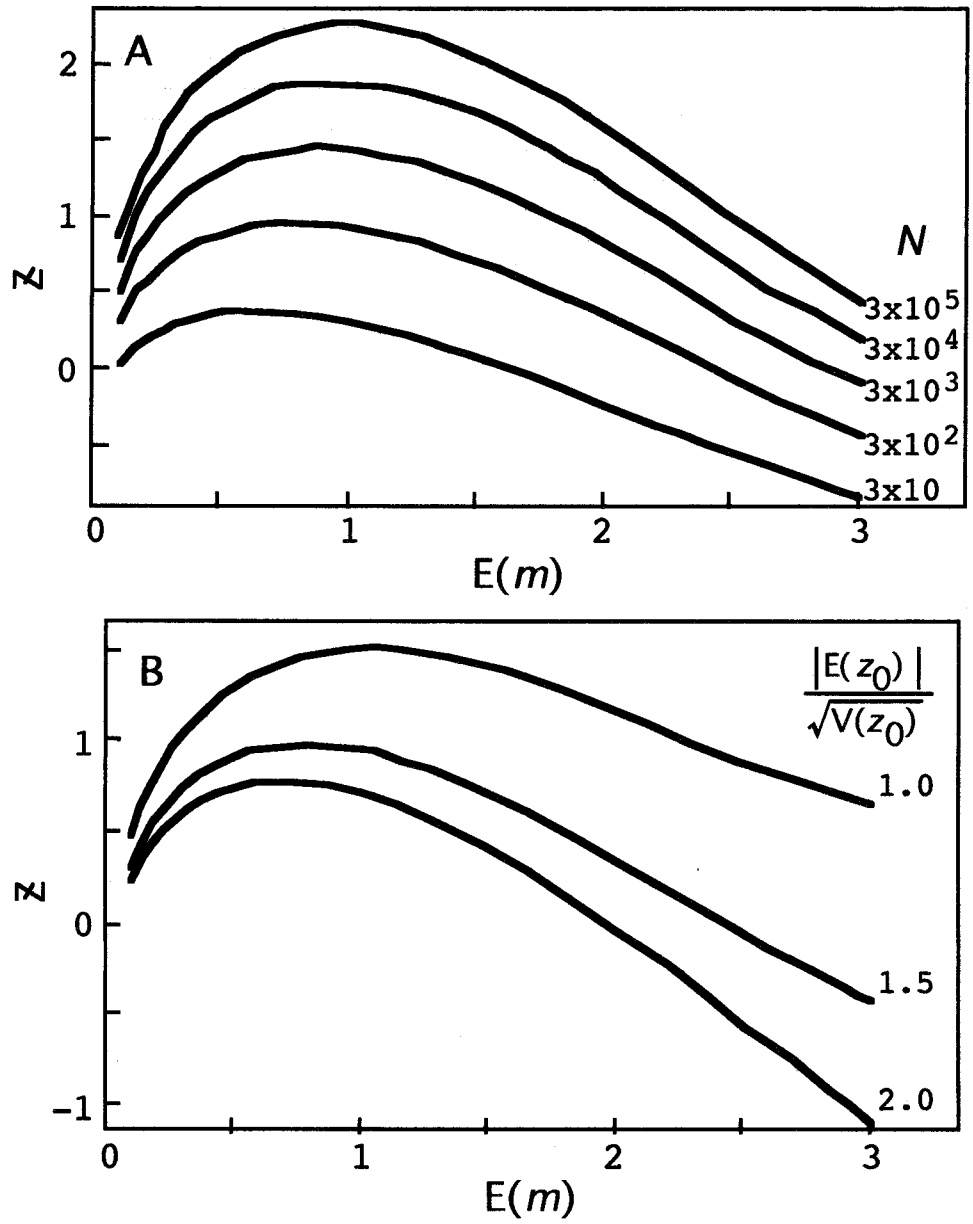

Fig. 1-3. Correlation between z and E(m) at different values of sample size, N, (A) and smoothness, $|E(z_O)|/(V(z_O)^{1/2}$, (B). For (A), calculation is based on the following situation: the second mutant population is prepared by introducing random mutagenesis to the gene encoding the D130N mutant with an average number of mutation of E(m); a Poisson distribution is assumed for the number of mutations, i.e., $V(m) = E(m)$; a normal distribution is assumed for the second population with $E(z_{II})$ and $V(z_{II})$ values calculated from Equations 7 and 9, respectively, where $n = 1$, $b = 0.13$, and the values of z_M, $E(z_O)$, and $V(z_O)$ are those given in Table 1-1 for thermostability; a mutant having a z value defined by Equation 1 exists with a probability of 0.5 in the second population with a sample size of 3×10, 3×10^2, 3×10^3, 3×10^4 or 3×10^5. For (B), calculation is based on the same situation except that the sample size is fixed at 300 and the value of $V(z_O)$ is changed so as to make the value of smoothness becomes 0.5, 1.5, or 4.0.

population.

1.4. Summary

Catalase I from *Bacillus stearothermophilus* has high catalatic and low peroxidatic actives. The mutant from the first random mutant population, the D130N mutant, which has higher peroxidatic and lower catalatic activities than those exhibited by the wild-type enzyme, was subjected to second random mutagenesis in observance of the change in reaction specificity. From the second mutant population, triple mutant (I108T/D130N/I222T) was selected and examined. The reaction specificity of the purified enzyme revealed that catalase I that was originally 98% specific to catalatic activity and only 2% to peroxidatic activity was brought to 58% specific to peroxidatic activity after two-step walks. From the statistical analysis of the two random mutant populations, the average degree of nonadditivity of the mutational effect was estimated to be 0.13 irrespective of the properties of the enzyme. As b value was estimated to be 0.13 irrespective of the three properties of catalase I, the value obtained for the average [$E(z_I)$] and variance [$V(z_I)$] of z of the first population, and the average of the number of mutations in a sequence [$E(m)$] were used to estimate $E(z_O)$ and $V(z_O)$. It was demonstrated that the distribution pattern of a property of the second mutant population can be predicted well from

the data of the first mutant population which is $E(z_I)$ and $V(z_I)$ by taking into consideration the degree of nonadditivity. Accordingly, with the average $[E(z_{II})]$ and variance $[V(z_{II})]$ of z of the second population as a function of $E(m)$, if a mutant has a level of a property higher than z with a probability of P in the second population and a sample size of N, z can be expressed as a function of $E(m)$. Thereafter, appropriate mutation rate and efficient sample size for further mutation and selection was shown to be predictable. Hence, we conclude that $E(z_O)$, $V(z_O)$ and b value are essential parameters for efficient adaptive walk.

Chapter 2

Evolutionary molecular engineering by random elongation mutagenesis

2.1. Introduction

Mutation and selection have always been a part in the evolution of enzymes and have been adapted in the evolutionary molecular engineering for obtaining enzymes with new or improved properties (Matsuura *et al.*, 1998; Moore & Arnold, 1996; You & Arnold, 1996; Zhang *et al.*, 1997). As the directed evolution is considered to be an optimization process through adaptive walk on the fitness landscape (Aita & Husimi, 1996) , it is important to understand the characteristic feature of the landscape around an enzyme (Trakulnaleamsai *et al.*, 1995) and to construct an efficient strategy for the adaptive walk (Matsuura *et al.*, 1998).

Studies on the landscapes around catalase I from *Bacillus stearothermophilus* revealed that both catalatic and peroxidatic activities can easily be improved by random mutagenesis (Trakulnaleamsai *et al.*, 1995). However, the landscape on thermostability showed that the probability of obtaining a mutant with a much higher stability is very slim as the thermostability of the enzyme is near the local optimum level

(Matsuura *et al.*, 1998; Trakulnaleamsai *et al.*, 1995). In this regard, we have developed a method of mutagenesis to overcome such cut-off crippling *in vitro* molecular evolution. Instead of the random substitution on the main gene, we added peptides with random sequences to the C-terminus of the enzyme to increase the dimension of the sequence space in order to create a new landscape. The random elongation mutagenesis indeed generated mutant library containing mutants with higher thermostability. The method is beneficial for the study of *in vitro* molecular evolution since it relieves us from the limitations set in by substitution mutagenesis such as further improvement of a supposedly optimized property of an enzyme. By expanding the sequence space, random elongation mutagenesis provides a new landscape where further diversification of a property is acceptable.

2.2. Material and Methods

2.2.1. Bacterial strains and plasmids

The bacterial strains and plasmids were as in Chapter 1.

2.2.2. Preparation of mutant population by random elongation

The schematic diagram for random elongation mutagenesis is illustrated in Fig. 2-1. A synthetic linker DNA containing stop codons for all three frames between *Pst* I and *Kpn*I sites was ligated to pOD64EX that

had been digested with *Bst*EII. The resulting plasmid was named pOD64EXI. The mixture of 70-mer single-stranded synthetic oligonucleotide (random DNA) containing two randomized portions each composed of 15-mer random oligonucleotides was made double-stranded by Klenow fragment before digestion with *Pst* I. The *Pst* I fragments were ligated with pOD64EXI that had been digested with *Pst* I. To prevent multiple insertion of random DNA, the ligated DNA was digested with *Bgl*II and religated before introducing to *E.coli* UM228. The hybrid plasmids were isolated from the transformants, and the gene structure of the added random DNA was examined. Out of the hybrid plasmid examined, 72 contained one unit random DNA as judged from the length of *Bam*HI-*Kpn*I fragment by PAGE. The 58 elongation mutants with one-unit random DNA were used as the sample mutant population, and the nucleotide sequence of the region containing random DNA were identified.

2.2.3. Characterization of the mutant population

Catalatic activity, peroxidatic activity and thermostability of mutant enzymes were characterized as described in Chapter 1.

2.2.4. Preparation and heat inactivation of catalase I mutants DI and DI-Tail

E. coli UM228 harboring the plasmid bearing the gene for DI

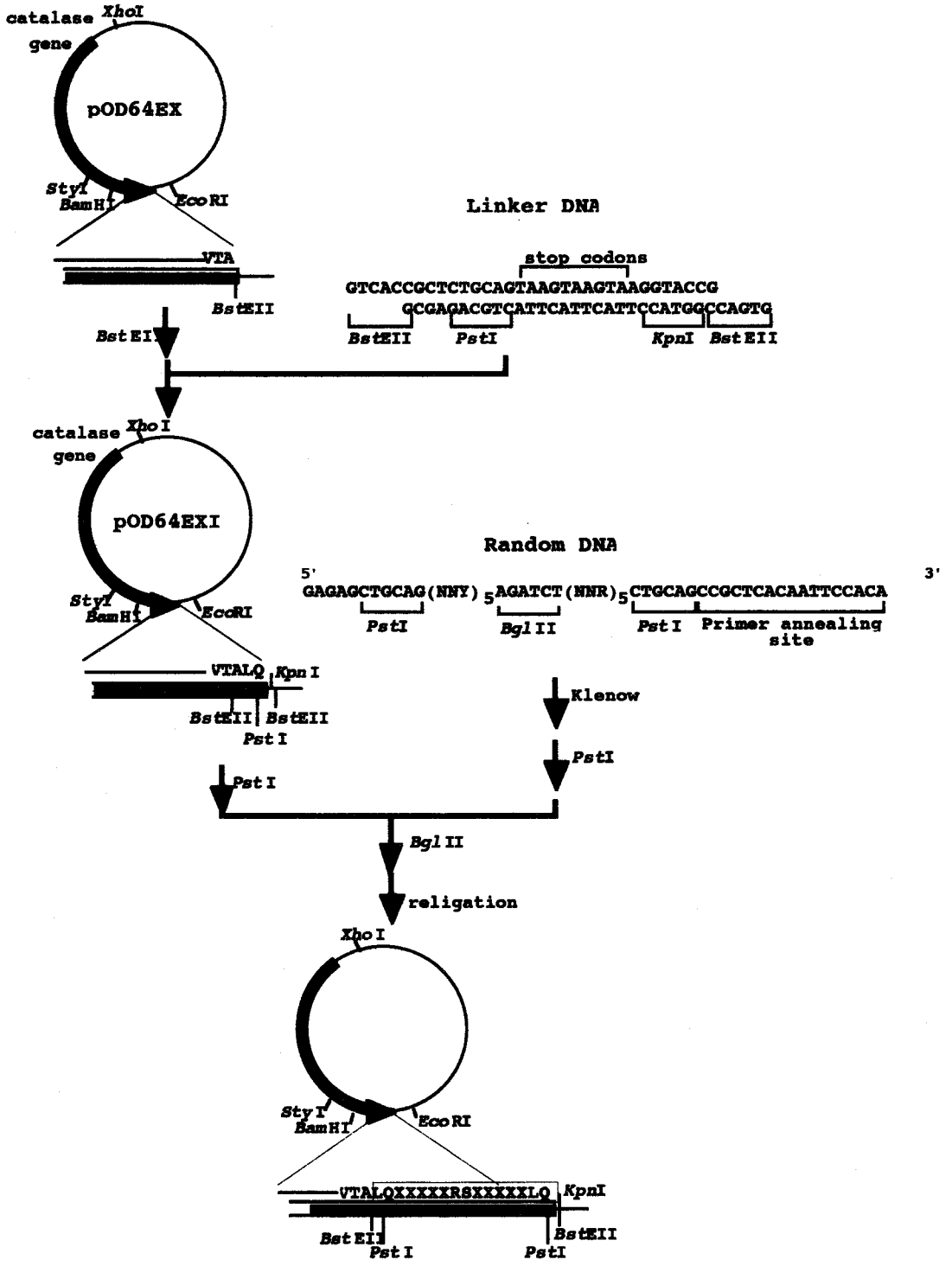

Fig. 2-1. Construction of plasmids for random elongation mutagenesis. pOD64 contains the catalase gene in the *Pst* I-*Eco*RI fragment. The unique *Pst* I site was converted to *Xho* I site by site-directed mutagenesis and the plasmid obtained was named pOD64EX.

mutant (D87N/I311T) was obtained from the mutant library prepared previously by random point mutagenesis of catalase I gene (Trakulnaleamsai *et al.*, 1995). The gene for DI-Tail was prepared by replacing the *Sty* I- *Eco* RI fragment of the plasmid bearing the gene for DI mutant with that contained in the plasmid of the elongation mutant indicated by the open arrow in Fig. 2-2 (top panel). The mutant enzymes, DI mutant and DI-Tail, were purified according to the procedure described previously (Trakulnaleamsai *et al.*, 1992) , except that the heat treatment of the crude enzyme solution was done at 55℃ and 60℃ for DI mutant and DI-Tail, respectively. The purified enzymes in 50 mM potassium phosphate (pH 7.0) were heated at 70℃ for various time intervals. The residual catalatic activity were measured as described above. The amino acid sequence of the fused peptide tail was deduced to be LQYRSLDRSGDVENLQ.

2.3. Result and Discussion

2.3.1. Random elongation mutagenesis of wild-type catalase I

A mutant population comprising of 58 elongation mutants were prepared as described under Material and methods. Though the elongation mutants were designed to fuse artificial peptide tails of 16 amino acid residues containing 10 random ones at the C-terminus of catalase I, sequencing of the elongated region revealed some deviations on the size of

the peptide tails (Table 2-1). These deviations from the originally intended length can arise during the synthesis of the random DNA.

The catalatic and peroxidatic activities of the mutants expressed in terms of activation free energy level (Δ G*/RT) vary from -1.7 to 6.4 and from -3.8 to -0.24, respectively (Fig. 2-2). Hence, the distributions of both activities in the elongation mutant population are as wide as those obtained previously by random point mutagenesis, where the values for the catalatic and peroxidatic activities ranged between -2.3 and 6.6 and between -3.0 and 1.2, respectively (Trakulnaleamsai *et al.*, 1995). It is to be noted that the wild-type, the starting material for mutagenesis, has the catalatic and peroxidatic values of 4.4 and -1.3, respectively. These results show that addition of small peptide at the C-terminus can bring about diversity in the properties of catalase I. The random elongation mutagenesis, therefore, is as effective as the random point mutagenesis in producing a library of mutants with functional diversity.

For thermostability, the distribution of the values of the elongation mutants arising from the wild-type (0.76) is wider (-1.7 to 2.6) than that of point mutagenesis (-2.0 to 1.6) (Trakulnaleamsai *et al.*, 1995) (Fig. 2-2). In addition, assuming normal distribution with the observed mean and standard deviation of the values, the expected frequency of elongation mutants having higher thermostability than the wild-type were calculated to be about 27% of the total population, a ten times higher than the 2.6%

obtained after point mutagenesis (Trakulnaleamsai *et al.*, 1995). This indicates that the new landscape in the sequence space expanded by increasing the dimension includes mutants with much higher level of thermostability.

For the relationship among the understudied properties of the elongation mutant population, the two catalytic activities are positively correlated where mutants with high catalatic activity do have high peroxidatic activity (Fig. 2-3A). The positive correlation means that the elongation mutagenesis causes little effect on the reaction specificity. Similar correlation was observed for the previous point mutant population (Trakulnaleamsai *et al.*, 1995). This may be attributed to the fact that catalatic and peroxidatic reactions are catalyzed at the same active site and share a common intermediate of compound I (Yomo *et al.*, 1997). The same positive correlation is observed between the thermostability and the two enzyme activities (Figs. 2-3B and 2-3C), a characteristic of the elongation mutant population that was not observed with the point mutant population which showed no clear correlation (Trakulnaleamsai *et al.*, 1995).

2.3.2. Random elongation mutagenesis of mutant catalase I

As mentioned in Chapter 1, two sequential cycles of random point mutagenesis and selection produced the triple mutant (I108T/D130N/I222T) with reaction specificity of 42% catalase and 58%

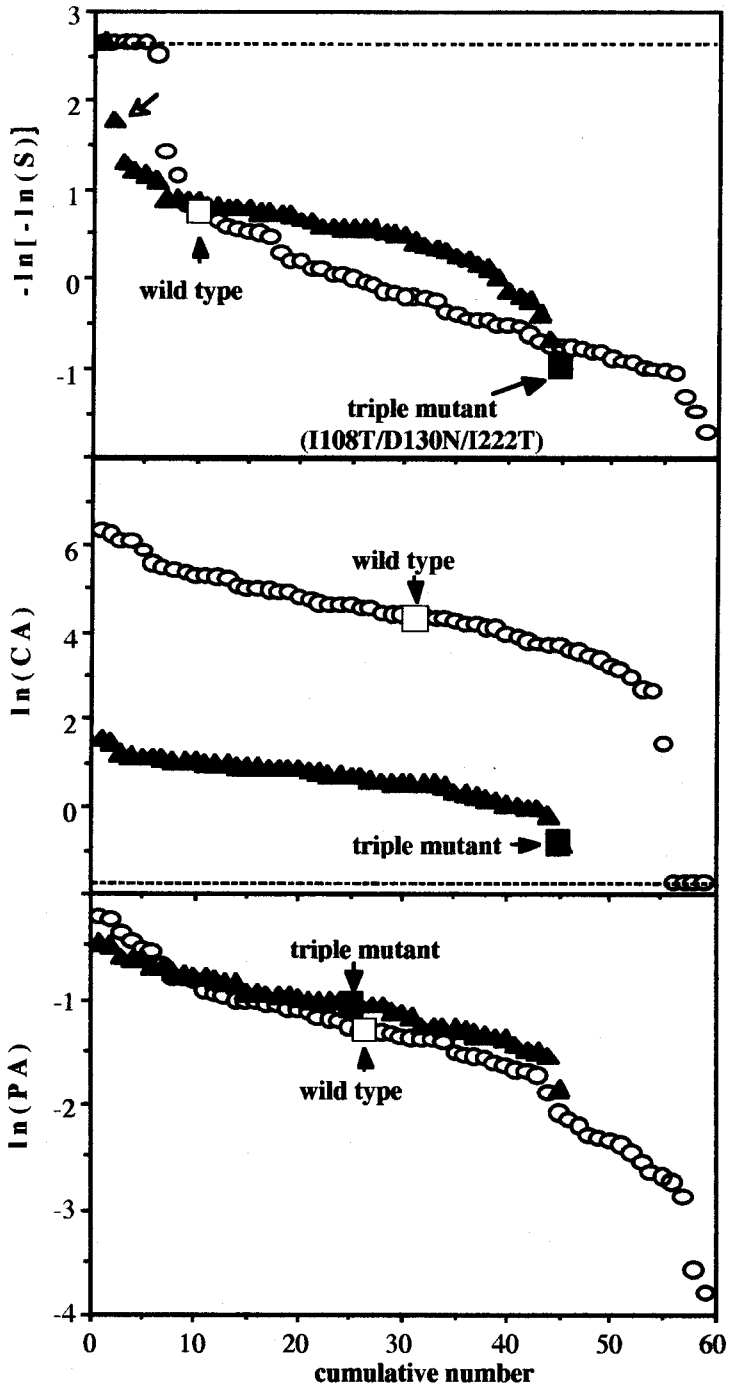

Fig. 2-2. Distribution pattern of the three enzymatic properties of the two elongation mutant populations prepared from the wild-type enzyme (○) and the triple mutant (I108T/D130N/I222T) (▲). Thermostability (S), specific catalatic activity (CA), and specific peroxidatic activity (PA) of the mutant enzymes were measured as described under Experimental protocol. As -ln(S), CA, and PA correspond to the rate constant, the three properties are expressed on the ordinate in terms of activation free energy level (Δ G*/RT) by using -ln[-ln(S)], ln(CA), and ln(PA), and the values were arranged in the order of the highest to the lowest values for each population. For the thermostability, S values of more than 0.93 were regarded as 0.93, and for the catalatic activity, CA values of less than 0.18 s^{-1} were regarded as 0.18 s^{-1}. The positions of the wild-type enzyme (□) and the triple mutant (■) are indicated by the filled arrow. The open arrow in the top panel indicates the elongation mutant of which peptide tail was added to the C terminus of DI mutant (D87N/I311T).

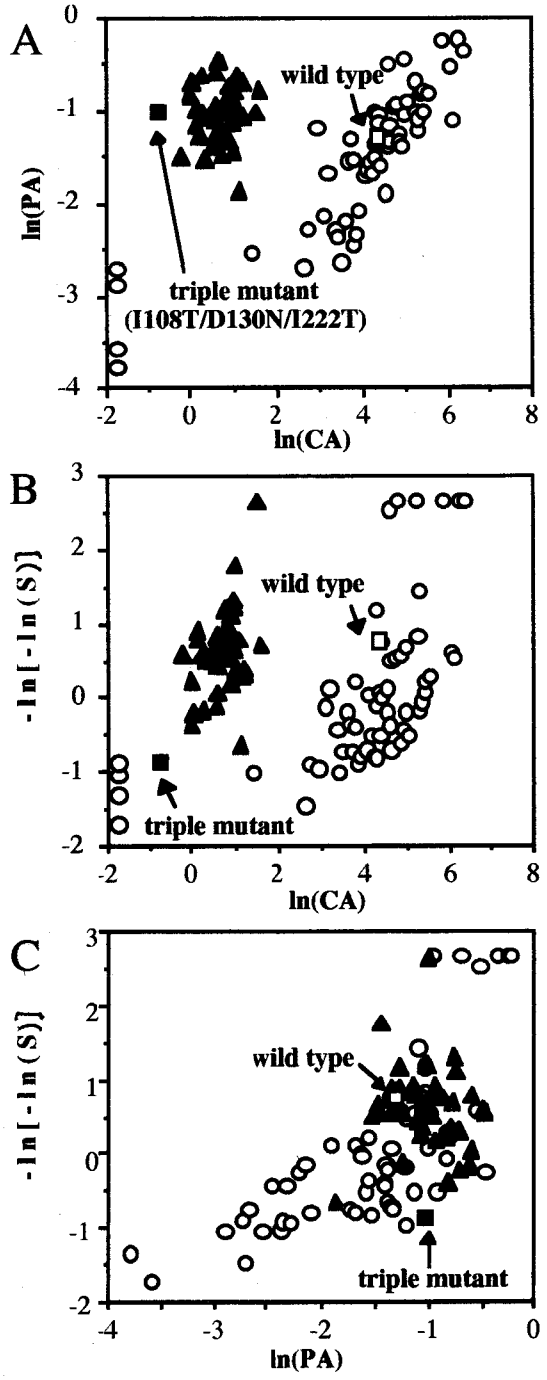

Fig. 2-3. Correlation between two properties of the elongation mutants in the populations prepared from the wild-type enzyme (○) and the triple mutant (▲). Each mutant is located at the position having the coordinates of the following properties: (A) catalatic activity vs. peroxidatic activity; (B) catalatic activity vs. thermostability; (C) peroxidatic activity vs. thermostability. The positions of the wild-type enzyme (□) and the triple mutant (■) are indicated by the filled arrow. The value of the upper or lower detection limit for each property are shown by the broken line (the values are the same as those in Fig. 2-1).

Table 2-1. Properties of elongation mutants prepared from the wild-type catalase I and its triple mutant (I108T/D130N/I222T). Thermostability (S), specific catalatic activity (CA), and specific peroxidatic activity (PA) of the enzymes were measured as described under Experimental protocol. For the thermostability, S values of more than 0.93 were regarded as 0.93, and for the catalatic activity, CA values of less than 0.18 s^{-1} were regarded as 0.18 s^{-1}. Amino acid sequences were deduced from the nucleotide sequences.

Amino acid sequence of peptide tail	-ln(S)/600 (s^{-1})	CA (s^{-1})	PA (s^{-1})
wild-type	7.8×10^{-4}	7.9x10	2.7×10^{-1}
LOHCCHLRSSAETVLO	1.2×10^{-4}	1.9×10^{2}	5.0×10^{-1}
LQTLILTPM	1.2×10^{-4}	1.2×10^{2}	3.9×10^{-1}
LQLLVFDRYEKGTAVSK	1.2×10^{-4}	5.3×10^{2}	7.9×10^{-1}
LQSFTINRSTTNRLQ	1.2×10^{-4}	3.6×10^{2}	7.8×10^{-1}
LQTSNRDRSNRDVTLQ	1.2×10^{-4}	5.7×10^{2}	6.9×10^{-1}
LQLVSFNRSSIINRLQ	1.3×10^{-4}	1.0×10^{2}	5.9×10^{-1}
LQGIASNRSINNKGLQ	4.0×10^{-4}	2.0×10^{2}	3.4×10^{-1}
LQCVLATDL	5.3×10^{-4}	7.6x10	3.6×10^{-1}
LQISRYPRSNTTIACSK	7.4×10^{-4}	1.9×10^{2}	3.6×10^{-1}
LQNNCDNRSDVVVTLQ	8.1×10^{-4}	8.1x10	3.5×10^{-1}
LQLSDESRSINEKTAVSK	8.7×10^{-4}	1.4×10^{2}	3.5×10^{-1}
LQPDNINRSTGTRKLQ	9.3×10^{-4}	4.4×10^{2}	5.8×10^{-1}
LQGSNYPRSRITKDLQ	9.7×10^{-4}	1.3×10^{2}	2.8×10^{-1}
LQTGDVDRSGESDIRK	9.9×10^{-4}	4.4×10^{2}	3.3×10^{-1}
LQYTHFDRSVDAAVSK	9.9×10^{-4}	1.1×10^{2}	3.8×10^{-1}
LQDHSNVRSEVSDMLQ	1.0×10^{-3}	1.1×10^{2}	3.0×10^{-1}
LQGNYFYRSGGIIRLQ	1.3×10^{-3}	2.7×10^{2}	4.4×10^{-1}
LQFSTPPRSIVNNGLQ	1.4×10^{-3}	2.3×10^{2}	4.5×10^{-1}
LQRFDRSMNTATLQ	1.4×10^{-3}	4.3x10	2.2×10^{-1}
LQSSRCFRSTDNAVLQ	1.5×10^{-3}	9.6x10	1.5×10^{-1}
LQCTTVSRSNEKAGLQ	1.5×10^{-3}	2.5x10	1.8×10^{-1}
LQFFLNRSKSIAVLQ	1.6×10^{-3}	2.3×10^{2}	3.6×10^{-1}
LQVSDGIRSMGTANLQ	1.6×10^{-3}	7.9x10	2.6×10^{-1}
LQNFGADRSMVEVALQ	1.7×10^{-3}	6.2x10	2.0×10^{-1}
LQCCHTPRSDMIGGLQ	1.7×10^{-3}	8.4x10	2.0×10^{-1}
LQDDCLGRSIIEIMLQ	1.8×10^{-3}	2.1×10^{2}	4.4×10^{-1}
LQGNGHNRSSIDEVLQ	1.9×10^{-3}	7.5x10	2.5×10^{-1}
LQDCSIIRSDVTNKLQ	2.0×10^{-3}	2.3x10	1.2×10^{-1}
LQSTFLRRSG	2.0×10^{-3}	2.0×10^{2}	2.9×10^{-1}
LQLDSAPDLARLMICSK	2.0×10^{-3}	9.8x10	2.5×10^{-1}
LQTNRINRSNNGTVLSK	2.1×10^{-3}	1.5×10^{2}	6.3×10^{-1}

LQTYCYPRSKSKENLQ	$2.1x10^{-3}$	3.6x10	$1.1x10^{-1}$
LQCPHARRSAVEAGLQ	$2.4x10^{-3}$	4.1x10	$2.1x10^{-1}$
LQFNSSIRSIITKALQ	$2.5x10^{-3}$	$1.0x10^{2}$	$2.5x10^{-1}$
LQPYSFIRSEGSAGLQ	$2.6x10^{-3}$	4.5x10	$8.5x10^{-2}$
LQHFCHPRSVNGIELQ	$2.6x10^{-3}$	3.0x10	$9.8x10^{-2}$
LQVSAATRSVGGRNLQ	$2.6x10^{-3}$	$1.4x10^{2}$	$2.5x10^{-1}$
LQCFCNFRSVGINTLQ	$2.8x10^{-3}$	8.2x10	$3.2x10^{-1}$
LQDTTGNR	$2.8x10^{-3}$	6.7x10	$2.1x10^{-1}$
LQSFNLRRSAGTTRLQ	$2.8x10^{-3}$	$1.6x10^{2}$	$4.0x10^{-1}$
LQALDNTDLRGGCTAVSK	$3.1x10^{-3}$	$1.3x10^{2}$	$2.6x10^{-1}$
LQYAYYSRSTTKNMLQ	$3.4x10^{-3}$	$1.1x10^{2}$	$2.6x10^{-1}$
LQGGNGVRSNIKGILQ	$3.5x10^{-3}$	6.0x10	$1.8x10^{-1}$
LQCFNDARSVRTGTLQ	$3.5x10^{-3}$	4.1x10	$2.7x10^{-1}$
LQFNISSRSVSSGNLQ	$3.6x10^{-3}$	3.4x10	$7.0x10^{-2}$
LQHVDYVRSGRVSALQ	$3.7x10^{-3}$	6.7x10	$1.9x10^{-1}$
LQIDDYNRSKDNINLQ	$3.7x10^{-3}$	5.0x10	$1.2x10^{-1}$
LQLSGIIDLTG	$3.8x10^{-3}$	7.3x10	$2.2x10^{-1}$
LQSYTTPFLDLRLVLNCSK	$4.1x10^{-3}$	$1.8x10^{-1}$	$6.5x10^{-2}$
LQNNSYRRSENDGDLQ	$4.2x10^{-3}$	4.9x10	$9.5x10^{-2}$
LQVTIFYRSAAGEELQ	$4.3x10^{-3}$	1.5x10	$1.0x10^{-1}$
LQVYPYTRSKGVGVLQ	$4.5x10^{-3}$	2.0x10	$3.0x10^{-1}$
LQINLLHRSERGNELQ	$4.6x10^{-3}$	3.2x10	$9.4x10^{-2}$
LQYFISIRSKIAG	$4.7x10^{-3}$	4.2	$7.8x10^{-2}$
LQTGALPDLLLIINCSK	$4.8x10^{-3}$	$1.8x10^{-1}$	$5.6x10^{-2}$
LQCIIFSRSDVAIALQ	$6.3x10^{-3}$	$1.8x10^{-1}$	$2.3x10^{-2}$
LQYFLNFY	$7.4x10^{-3}$	1.4x10	$6.7x10^{-2}$
LQLIFHIRSKMVTAVSK	$9.4x10^{-3}$	$1.8x10^{-1}$	$2.8x10^{-2}$
triple mutant	$4.0x10^{-3}$	$5.9x10^{-1}$	$3.6x10^{-1}$
LQCICTDRSATIIKLQ	$1.2x10^{-4}$	4.5	$3.6x10^{-1}$
LQYRSLDRSGDVENLQ	$2.8x10^{-4}$	2.7	$2.4x10^{-1}$
LQGCSRSRSDVIEGLQ	$4.5x10^{-4}$	2.5	$4.7x10^{-1}$
LQTNTPCRSGNNMNLQ	$5.0x10^{-4}$	2.6	$3.6x10^{-1}$
LQSVHSNRSDTARNLQ	$5.1x10^{-4}$	2.2	$2.9x10^{-1}$
LQYPAPFRSVSVGSLQ	$5.5x10^{-4}$	2.5	$4.8x10^{-1}$
LQYDSSDRSSAAIDLQ	$6.7x10^{-4}$	1.1	$3.2x10^{-1}$
LQTGDSDRSMKKTILQ	$6.7x10^{-4}$	2.1	$3.9x10^{-1}$
LQIHSGIRSMTGANLQ	$6.8x10^{-4}$	2.2	$2.8x10^{-1}$
LQGSIYNRSIKGEALQ	$6.9x10^{-4}$	2.4	$2.6x10^{-1}$
LQHTSPVDLS	$7.3x10^{-4}$	1.8	$2.6x10^{-1}$
LQDSHHLRSGAGTDLQ	$7.3x10^{-4}$	1.7	$2.8x10^{-1}$

LQLLPNTRSIGSDMLQ	$7.5x10^{-4}$	2.6	$3.2x10^{-1}$
LQSDLSFRSVGVDILQ	$7.5x10^{-4}$	2.9	$5.5x10^{-1}$
LQDCYSVRSNDANTLQ	$7.5x10^{-4}$	2.8	$3.8x10^{-1}$
LQFDTSTRSGSGENLQ	$7.7x10^{-4}$	1.1	$3.8x10^{-1}$
LQLNCHNRSMGSTTLQ	$7.8x10^{-4}$	2.6	$4.2x10^{-1}$
LQDIVLARSAAIASLQ	$8.0x10^{-4}$	1.7	$2.5x10^{-1}$
LQSRYRCRSNTVIGLQ	$8.2x10^{-4}$	5.0	$4.6x10^{-1}$
LQTFGDPRSVAEDVLQ	$8.6x10^{-4}$	2.1	$2.3x10^{-1}$
LQTPRYVRSRDGNELQ	$8.7x10^{-4}$	2.7	$3.5x10^{-1}$
LQIVTSTDLE	$9.2x10^{-4}$	1.4	$3.5x10^{-1}$
LQNIHGPRSGKKVNLQ	$9.3x10^{-4}$	2.0	$6.1x10^{-1}$
LQFYSFVRSVTEEKLQ	$9.4x10^{-4}$	$8.1x10^{-1}$	$2.2x10^{-1}$
LQYHSGSRSVITSGLQ	$9.4x10^{-4}$	1.6	$3.5x10^{-1}$
LQFRYNRRSGDVTVLQ	$9.5x10^{-4}$	2.3	$2.6x10^{-1}$
LQDYTNIRSGGVDTLQ	$9.5x10^{-4}$	1.2	$2.8x10^{-1}$
LQHFVAARSKTAEELQ	$9.6x10^{-4}$	1.9	$6.2x10^{-1}$
LQTIISYRSAGDVTLQ	$9.7x10^{-4}$	2.4	$3.6x10^{-1}$
LQTGFYHRSSTKVSLQ	$9.9x10^{-4}$	1.7	$3.8x10^{-1}$
LQDALSYRSNTAVRLQ	$1.0x10^{-3}$	1.4	$2.2x10^{-1}$
LQVLFNTRSGTVAKLQ	$1.1x10^{-3}$	1.8	$3.3x10^{-1}$
LQCSPISRSTVGRVCSK	$1.1x10^{-3}$	3.4	$3.5x10^{-1}$
LQILTGLRSVGAVVLQ	$1.2x10^{-3}$	2.9	$4.5x10^{-1}$
LQIIYDDRSEKTVDLQ	$1.2x10^{-3}$	3.3	$4.9x10^{-1}$
LQNIHGSRSKKVRVLQ	$1.3x10^{-3}$	3.1	$3.4x10^{-1}$
LQVNYITDLPLG	$1.3x10^{-3}$	1.0	$4.3x10^{-1}$
LQNSFVCRSVGTIDLQ	$1.4x10^{-3}$	2.5	$3.9x10^{-1}$
LQTFVSHRSKKESRLQ	$1.6x10^{-3}$	1.8	$5.5x10^{-1}$
LQGCAALRSKIGKRLQ	$1.9x10^{-3}$	1.7	$2.9x10^{-1}$
LQSSNRIRSVVGESLQ	$2.0x10^{-3}$	1.3	$5.4x10^{-1}$
LQYVRSHRSNGGRGLQ	$2.1x10^{-3}$	1.0	$5.0x10^{-1}$
LQGYLTVRSMTRSNLQ	$2.4x10^{-3}$	$9.8x10^{-1}$	$4.4x10^{-1}$
LQGCLSCRSIVIVGLQ	$3.2x10^{-3}$	3.0	$1.6x10^{-1}$

peroxidase. However, the mutant was found to have low thermostability. Furthermore, analysis of the point mutant population showed that thermostability tends to decrease with the increase in the number of mutations on the wild-type enzyme (Trakulnaleamsai *et al.*, 1995). This indicates that the probability to improve the thermostability of the triple mutant by another point mutagenesis is quite low. On such a particular situation, random elongation mutagenesis can be an alternative method for it was effective in improving the thermostability of the wild-type enzyme.

The same process of random elongation mutagenesis applied to the wild-type gene was carried out on the triple mutant gene. A mutant population comprising of 44 elongation mutants was obtained with the indicated size distribution of the peptide tails (Table 2-1). All mutant enzymes possess higher thermostability than the triple mutants to an extent that some are even higher than the wild-type (Fig. 2-2). The results further support that random elongation mutagenesis can be an efficient tool in improving the properties of enzymes.

2.3.4. Correlation between the amino acid sequence of the added peptide tails and the mutational effect

Each amino acid residue has its own physicochemical properties such as hydrophobicity (Kyte & Doolittle, 1982), electric charge, molecular weight, α-helix forming tendency (Chou & Fasman, 1978) and β-sheet

forming tendency (Chou & Fasman, 1978). Now, let w_i (i=1,2,....,20) be the value of a property of i'th amino acid residue, and h_{ij} be the number of the i'th residues in the added peptide tail of j'th mutant in a population. The value of the property (d_j) of the peptide tail of j'th mutant is then defined as:

$$d_j = \sum_{i=1}^{20} (h_{ij} w_i)$$

It follows that the value of a physicochemical property is determined by the amino acid composition of the peptide tail. Therefore, any correlation found between the physicochemical property and the enzymatic property being optimized can be basis for improving the design of the random DNA used for the elongation mutagenesis to increase the chance of obtaining better mutant population.

Of the two elongation mutant population prepared, correlation between each of the three enzymatic properties of the mutants and each of the stated physicochemical properties of the peptide tails was examined. As shown in Fig. 2-4., negative correlation was observed between the hydrophobicity of the peptides and the thermostability of the mutants; the slopes of the regression lines of both populations are almost the same (0.07). This means that elongation mutants with less hydrophobic peptides tend to be more thermostable than those with more hydrophobic peptides. It should be noted that this trend is observed within the detection limit of thermostability. Analysis of both elongation mutant populations showed that

aside from the combinations of hydrophobicity and thermostability, no general correlation's can be drawn between other combinations involving one of the enzymatic properties to one of the physicochemical properties mentioned above (data not shown). Hence, other physicochemical properties except hydrophobicity can be rated insignificant in designing the random DNA for elongation mutagenesis. On the other hand, it was also confirmed that the size distribution of the peptide tails does not affect the correlation discussed above.

The thermostability of the triple mutant was profoundly increased by elongation mutagenesis(Fig. 2-2). If the effect of elongation and amino acid substitution are simply additive, then the stabilizing effect of the peptide tail on the mutant will be similar when added to another mutant, i.e., the degree of stabilizing effect of a peptide tail will be independent of the amino acid sequence of the enzyme. To test the additivity, the peptide tail in the open arrow in Fig. 2-2. was added to the C-terminus of a mutant DI, and the resulting mutant was named DI-Tail. The thermostability of DI mutant (-ln[-ln(S)]=-0.47) is higher than the triple mutant (-ln[-ln(S)]=-0.88) but is much lower than the wild-type enzyme (-ln[-ln(S)]=0.76). Fig. 2-5 shows the course of the heat inactivation of the purified DI mutant and DI-Tail at 70℃ with rate constants of 5.4×10^{-4} s^{-1} and 4.1×10^{-4} s^{-1}, respectively. For the wild-type enzyme, the heat inactivation rate constant at 70℃ was

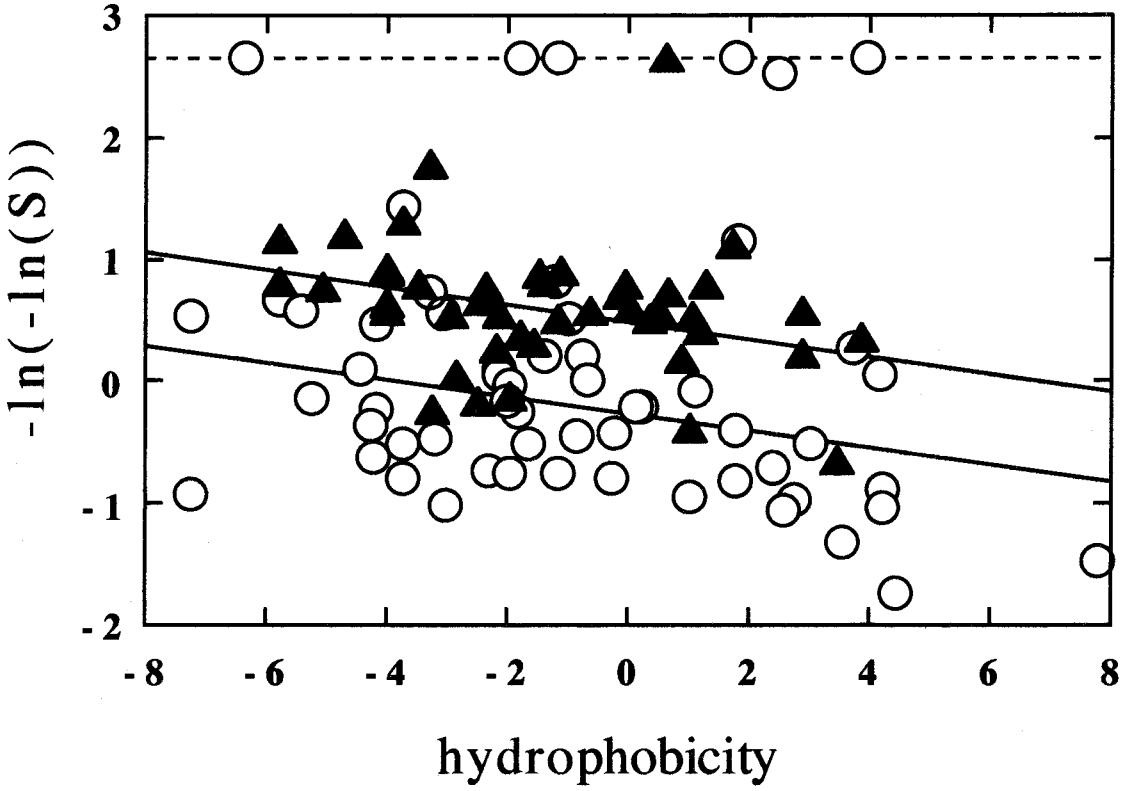

Fig. 2-4. Correlation between hydrophobicity of the peptide tail and thermostability of the elongation mutants in the population prepared from the wild-type enzyme (○) and the triple mutant (I108T/D130N/I222T) (▲).The regression lines for each population were obtained by the least square method, omitting the points for the mutants of which thermostabilities are above the upper detection limit shown by the broken line. Correlation coefficients (R^2) for the populations prepared from the wild-type and the triple mutant, respectively, are 0.09 (p=0.03) and 0.15 (p=0.01), whereas slopes are 7.08×10^{-2} and 7.11×10^{-2}.

calculated to be 5.6×10^{-5} s^{-1} based on the results reported in (Kobayashi *et al.*, 1997); the slope is shown by the broken line in Fig. 5. These results indicate that although the added peptide tail enhances the thermostability of the DI mutant, the stabilizing effect on the DI mutant is

not as great as it did to the triple mutant, of which derived mutant bearded thermostability higher than that of the wild-type. Therefore, the effect of elongation and substitution are not simply additive.

We have demonstrated that random elongation mutagenesis can generate mutant population with properties variation as wide as that of random point mutagenesis (Fig. 2-2). This means that the added peptide tail does impose some effects on the original enzyme molecule and modifies the original structure. Hence, there will be deviation from simple additivity of the two mutational effects of elongation and substitution, as pointed out by Wells for two point mutations (Wells, 1990). Accordingly, the mutational effect of a peptide tail is not constant but varies depending on the amino acid sequence of the enzyme, and a peptide tail exerting most beneficial effect to an enzyme may not do with another. Hence, there is no specific tail sequence that can be universal for improving a property of enzymes. However, as the resulting diversity of a property is not dependent on the enzyme sequence and is wide enough to include better mutants with fitted peptide tail, the issue on the universality of peptide tail sequence is not a concern for the random elongation mutagenesis.

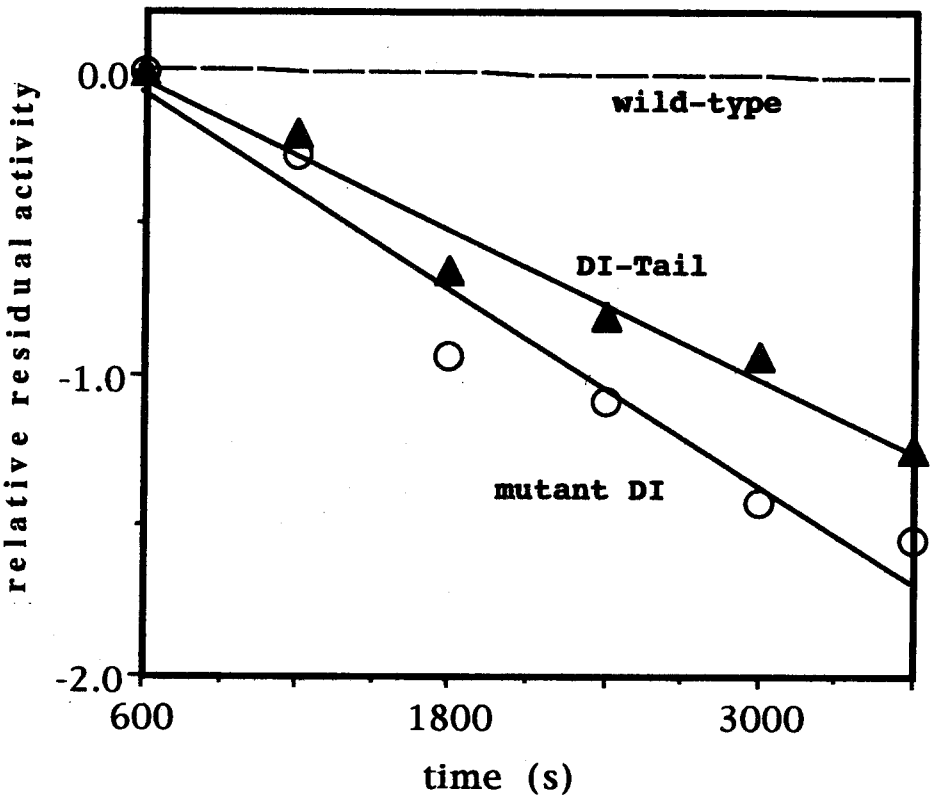

Fig. 2-5. Time course of the heat inactivation of catalase I mutants DI and DI-tail. *E.coli* UM228 harboring the plasmid bearing the gene for DI mutant was obtained from the mutant library prepared previously by random point mutagenesis of catalase I gene. The gene for DI-tail was prepared as described in Material and methods. DI mutant and DI-tail were purified according to the procedure described previously, except the heat treatment of the crude enzyme solution was done at 55℃ and 60℃ respectively. These purified enzymes dissolved in 50mM potassium phosphate (pH 7.0) were heated at 70℃ for various time intervals. The remaining catalatic activity measured as described under Material and method was expressed relative to the value after 600 s heating.

2.3.5. Random elongation mutagenesis as a tool for evolutionary molecular engineering

Improvement of an enzyme by molecular evolution, repeated cycles of sequential random mutagenesis and screening, is an optimization process on the fitness landscape around the enzyme. In this regard, random substitution and random elongation involve different concepts of random mutagenesis. For an enzyme with a length of " m " amino acid residues, a variety of 20^m sequences are located in the m -dimensional sequence space. Random substitution is a process that aids in the pursue of the best sequence "X" for a property in the sequence space. On the other hand, elongation mutagenesis opens a new n -dimensional space for each sequence in the m -dimensional space through the addition of a peptide tail with a length of n. In the newly added n-dimensional sequence space corresponding to the length of peptide tail n, the location of the unextended sequence expressing the optimized enzyme "X" in the original m-dimensional space will fall not on the peak but will generally be on the slope (Fig. 2-6), as the enzyme has not been optimized by any means, either natural or artificial, in the new n-dimensional space. Thereafter, random elongation mutagenesis will provide allowances for adaptive walk on the expanded (m+n)-dimensional sequence space to pave the way for a new peak. That is, after locating the highest peak in the m-dimensional landscape by random substitution, random elongation will provide new peaks on the newly drawn n-

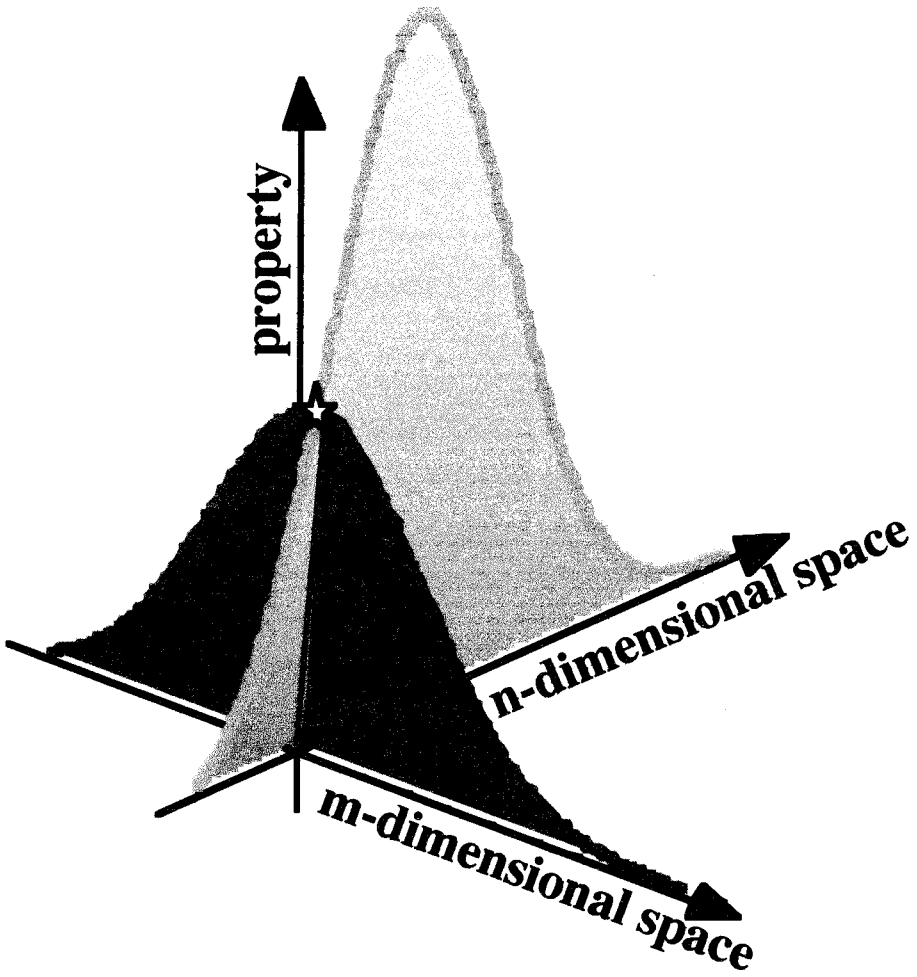

Fig. 2-6. Sketch of the landscape of a property of an enzyme in the ,m and n-dimensional sequence space. The landscape in the m-dimensional space is illustrated by plotting the values of the property of all the mutants with m amino acid residues over the assigned axis for m-dimensional space. The landscape in the n-dimensional space is for all the elongation mutants, prepared by adding a peptide tail with n amino acid residues to the C-terminus end of the enzyme, locating at the highest peak in the m-dimensional landscape. The m- and n-dimensional sequence space are degenerated to the corresponding axes.

dimensional landscape, and consequently, random substitution will do the work in searching for the new optimum peak in the (m+n)-dimensional landscape. Therefore, combining the two types of random mutagenesis, substitution and elongation, can be the basic strategy for evolutionary molecular engineering. The alternating events of the two random mutagenesis as required can be similar to those events occurring on the actual evolution of natural enzymes.

2.4. Summary

The new method of random mutagenesis employs the addition of peptide tails with random sequences to the C-terminus of enzyme molecules, and was termed random elongation mutagenesis. The mutant population prepared as such from catalase I of *Bacillus stearothermophilus* has a diversity in thermostability and enzyme activity as that prepared by random point mutagenesis. When a triple mutant of catalase I (I108T/D130N/I222T) of which thermostability is much lower than that of the wild-type was subjected to elongation mutagenesis, we generated a mutant population containing only mutants with higher thermostability than the triple mutant, and to the extent that some had even higher stability than the wild-type enzyme which possessed supposedly an optimized thermostability. These results indicate that the peptide addition allows the

protein sequence space to expand, which results to a new fitness landscape. The enzyme can then move along the routes of the new landscape until it reaches a new optimum. We therefore conclude that combination of random elongation mutagenesis with random point mutagenesis is no doubt a powerful tool for evolutionary molecular engineering.

Conclusion

Based on the fitness landscape of catalase I from *Bacillus stearothermophilus*, we elucidated the essential parameters determining effective mutation rate and population size to bring about improvement in a property of an enzyme. Statistical analysis of the understudied properties of the 1st and 2nd mutant libraries gave the *b* value, which is the degree of nonadditivity of mutational effect on the property of an enzyme. As *b* value was estimated to be 0.13 irrespective of the three properties of catalase I, the value obtained for the average [$E(z_I)$] and variance [$V(z_I)$] of z (differences in property from wild-type in a free energy level) of the first population, and the average of the number of mutations in a sequence [$E(m)$] were used to estimate the average [$E(z_O)$] and the variance [$V(z_O)$] of the effects of all the one-point mutations on a property of an enzyme. Thereafter, assuming normal distribution for the mutant population, with a mutant M of the first population, the values of the difference in the property between the wild-type and mutant M [z_M], the number of mutations in the sequence of mutant M [n], and the sum of property difference caused by each mutation on mutant M [$\sum_{k=1}^{n} z_k$] give the average [$E(z_{II})$] and the variance [$V(z_{II})$] for second mutant population. As mentioned above, the *b* value for catalase I was estimated to be 0.13. Therefore, though *b* value may differ

from one enzyme to another, the value may fall in the range of 0.1, as that calculated for various proteins which are 0.10 and 0.07 for the transition-state stabilization and unfolding, respectively (Wells, 1990).

Through the statistical analysis, the distribution pattern of the properties of second mutant population was predicted. Accordingly, with $E(z_{II})$ and $V(z_{II})$ as a function of $E(m)$, if a mutant has a level of a property higher than z with a probability of P in the second population and a sample size of N, z can be expressed as a function of $E(m)$.

$$P = 1 - \left(1 - \int_{z}^{+\infty} \frac{1}{\sqrt{2\pi V(z_{II})}} \exp\left(\left(\frac{(t - E(z_{II}))}{V(z_{II})}\right)^2\right) dt\right)^N$$

This relationship can be used as a guide for determining the conditions for the second mutagenesis and selection. That is, when a mutant with a higher level of thermostability than z is desired at a probability of 0.5, the appropriate population size required within a mutation rate of $E(m)$ can be determined. On the other hand, if values of N and $E(m)$ were fixed initially, a mutant with thermostability higher than z can be expected.

Random substitution mutagenesis including DNA shuffling has been adopted in improving the properties of an enzyme. However, the probability of obtaining an enzyme with a higher property will be very slim if the property is near the optimum level as deduced from the fitness landscape. In this regard, we have shown that random elongation

mutagenesis, a method of adding a random peptide tails to the C-terminus of an enzyme, is effective in overcoming such cut-off crippling *in vitro* molecular evolution. Indeed, the method proved to be useful in the further improvement of the thermostability of catalase I, a property that was supposedly optimized as deduced from the fitness landscape. By random elongation mutagenesis, the expected frequency of elongation mutants having higher thermostability than the wild-type were calculated to be about 27% of the total population, a ten times higher than the 2.6% obtained after point mutagenesis. Moreover, when a mutant catalase I, triple mutant (I108T/D130N/I222T), of which thermostability is much lower than that of the wild-type was subjected to random elongation mutagenesis, we generated a mutant population containing only mutants with higher stability than the triple mutant, to an extent that some had even higher stability than the wild-type enzyme.

This thesis thereby shows that by integrating the concept of sequence space and fitness landscape in engineering enzymes at molecular level, we may obtain desired enzyme or polymers for industrial, medical or agriculture use. Furthermore, we have shown that statistical analysis of the fitness landscape can be handy in establishing guideline for molecular engineering of enzyme. This thesis then provides new frontier in the field of enzyme optimization and for the studies of molecular evolutionary engineering.

As most existing proteins have a length of more than a hundred, the fitness landscape are mostly in a high dimensional space. Therefore, visualization of the landscape in these dimensional space is difficult with the knowledge we have now. At present, we have dig into some parameters relating to the feature of the fitness landscape which allow us to predict the outcome of an experimental molecular engineering of an enzyme. However, further elucidation on other essential parameters might be relevant to foresee more of the high dimensional space. Moreover, we might need to consider further relationship between the required population size for adaptive walk and the measurement limitation of the assay system of an enzyme property, particularly in improving novel property of an enzyme. That is, when the enzyme property increases minimally or no significant increase in each cycle of mutation and selection, minimizing the error in measurement will be the essential factor for further property improvement. With regard to this aspect, one may ask, "How far can an assay system hold true for the detection of property to be within the standard error in measurement?", "How much deviation can we allow to accommodate error in measurement?" and "Can increasing mainly the population size be the answer in obtaining the desired range of a property?". Answering some of the questions raised may give an possibility to bring a novel property of an enzyme to existence with evolutionary molecular engineering.

Appendix

I) Derivation of Equation 6

z_I is the difference between the property of a mutant enzyme and the wild-type. I indicates the first mutant library. The variance of z_I is expressed as:

$$\begin{aligned}
V(z_I) &= E[\{z_I - E(z_I)\}^2] \\
&= E[\{\sum_{i=1}^{m} z_i - \frac{b}{2!}\sum_{i=1}^{m}\sum_{j=1,i\neq j}^{m}(z_i + z_j) - E(z_I)\}^2] \\
&= E[\{(1-b(m-1))\sum_{i=1}^{m} z_i - E(z_I)\}^2] \\
&= E[m\,E(z_O^2)(1-b(m-1))^2 + m(m-1)(1-b(m-1))^2 E(z_O)^2 \\
&\quad - 2\,E(z_I)(1-b(m-1))\sum_{i=1}^{m} z_i + E(z_I)^2] \\
&= E[m(1-b(m-1))V(z_O)] + E[\{m(1-b(m-1))E(z_O) - E(z_I)\}^2]
\end{aligned} \tag{a}$$

Assuming the Poisson distribution, $E(m)$ is then expressed as:

$$E(m)=V(m)=E(m(m-1))=E(m(m-1)(m-2))=..... \tag{b}$$

Hence, from Equation 4 and b, Equation a is converted to

$$\begin{aligned}
V(z_I) &= E[m(1-b(m-1))V(z_O)] \\
&\quad + E[\{m(1-b(m-1))E(z_O) - E(m)E(z_O)(1-bE(m))\}^2] \\
&= E(m)[E(z_O)^2\{2bE(m)-1\}^2 + V(z_O)\{bE(m)-1\}^2 \\
&\quad + b^2 E(m)\{V(z_O) - E(z_O)^2\}]
\end{aligned} \tag{6}$$

II) Derivation of Equation 9

z_{II} is the difference between the property of a mutant enzyme and the wild-type. II indicates the second mutant library. The variance of z_{II} is expressed as:

$$
\begin{aligned}
V(z_{II}) &= E[\{z_{II} - E(z_{II})\}^2] \\
&= E[\{\sum_{k=1}^{n} z_k - \frac{b}{2}\sum_{k=1}^{n}\sum_{j=1,k\neq j}^{n}(z_k + z_j) \\
&+ \sum_{i=n+1}^{n+m} z_i - \frac{b}{2}\sum_{i=n+1}^{n+m}\sum_{j=n+1,j\neq i}^{n+m}(z_i + z_j) - b\sum_{k=1}^{n}\sum_{j=n+1}^{n+m}(z_k + z_j) \\
&- z_M - E(z_I) + bE(m)[nE(z_O) + \sum_{k=1}^{n} z_k]\}^2] \\
&= E[\{b\sum_{k=1}^{n}\{E(m)(E(z_O) + z_k) - \sum_{j=n+1}^{n+m}(z_k + z_j) \\
&+ \sum_{i=n+1}^{n+m}\{z_i - b(m-1)z_i\} + E(z_I)\}^2] \\
&= E[\{b\sum_{k=1}^{n}\{E(m)(E(z_O) + z_k) - \sum_{j=n+1}^{n+m}(z_k + z_j) \\
&+ \sum_{i=n+1}^{n+m}\{z_i - b(m-1)z_i\} + E(z_I)\}^2]
\end{aligned}
\qquad (c)
$$

Assuming Poisson distribution and from Equation 4, Equation c is converted to:

$$
\begin{aligned}
V(z_{II}) &= E[m(1-b(m+n-1)^2 V(z_O) \\
&+ \{m(1-b(m+n-1))E(z_O) + b(E(m)-m)\sum_{k=1}^{n} z_k + bE(m)nE(z_O) \\
&- \{E(m) - b(V(m) + E(m)^2 - E(m))\}E(z_O)\}^2 \\
&= E(m)[\{E(z_O)(2bE(m) + bn - 1) + b\sum_{k=1}^{n} z_k\}^2 \\
&+ V(z_O)\{b(n+E(m)) - 1\}^2 \\
&+ b^2 E(m)\{V(z_O) - E(z_O)^2\}]
\end{aligned}
\qquad (9)
$$

Reference

Aita, T. & Husimi, Y. (1996). Fitness spectrum among random mutants on Mt. Fuji-type fitness landscape. *J. Theor. Biol.*, **182**(4), 469-85.

Aita, T. & Husimi, Y. (1998). Adaptive Walks by the Fittest among Finite Random Mutants on a Mt. Fuji- type Fitness Landscape. *J. Theor. Biol.*, **193**(3), 383-405.

Arnold, F. H. (1993). Engineering proteins for nonnatural environments. *Faseb. J.*, **7**(9), 744-9.

Arnold, F. H. (1996). Directed evolution: creating biocatalysts for the future. *Chemical Engineering Science,* **51**(23), 5091-5102.

Chou, P. Y. & Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas. Mol. Biol.*, **47**, 45-148.

Cuenoud, B. & Szostak, J. W. (1995). A DNA metalloenzyme with DNA ligase activity. *Nature,* **375**(6532), 611-4.

Ekland, E. H., Szostak, J. W. & Bartel, D. P. (1995). Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science,* **269**(5222), 364-70.

Ellington, A. D. & Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature,* **346**(6287), 818-22.

Hildebraunt, A. G. & Roots, I. (1975). Reduced nicotinamide adenine dinucleotide phosphate (NADPH)-dependent formation and breakdown of hydrogen peroxide during mixed function oxidation reactions in liver microsomes. *Arch. Biochem. Biophys.*, **171**(2), 385-97.

Huynen, M. A., Stadler, P. F. & Fontana, W. (1996). Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. U S A*, **93**(1), 397-401.

Kauffman, S. & Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.*, **128**(1), 11-45.

Kauffman, S. A. & Weinberger, E. D. (1989). The NK model of rugged fitness landscapes and its application to maturation of the immune response. *J. Theor. Biol.*, **141**(2), 211-45.

Kobayashi, C., Suga, Y., Yamamoto, K., Yomo, T., Ogasahara, K., Yutani, K. & Urabe, I. (1997). Thermal conversion from low- to high-activity forms of catalase I from *Bacillus stearothermophilus. J. Biol. Chem.*, **272**(37), 23011-6.

Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**(1), 105-32.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature,* **227**(259), 680-5.

Lewis, E. R., Winterberg, K. M. & Fink, A. L. (1997). A point mutation leads to altered product specificity in beta-lactamase catalysis. *Proc. Natl. Acad. Sci. U S A*, **94**(2), 443-7.

Loprasert, S., Urabe, I. & Okada, H. (1990). Over production and single-step purification of *Bacillus stearothermophilus* peroxidase in *Escherichia coli. Appl. Microbiol. Biotechnol.*, **32**, 690-692.

Maniatis, T., Fritsch, E. F. & Sambrook, J. (1998). *Molecular cloning: A laboratory manual, Cold Spring Harbor Laboratory, Cold spring harbor, N Y.*

Matsuura, T., Yomo, T., Trakulnaleamsai, S., Ohashi, Y., Yamamoto, K. & Urabe, I. (1998). Nonadditivity of mutational effects on the properties of catalase I and its application to efficient directed evolution. *Protein Eng.*, **11**(9), 789-795.

Moore, J. C. & Arnold, F. H. (1996). Directed evolution of a *para*-nitrobenzyl esterase for aqueous-organic solvents [see comments]. *Nature Biotechnol.*, **14**(4), 458-67.

Myers, R. M., Lerman, L. S. & Maniatis, T. (1985). A general method for saturation mutagenesis of cloned DNA fragments. *Science,* **229**(4710), 242-7.

Schuster, P. (1995). How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *J. Biotechnol.*, **41**(2-3), 239-57.

Smith, J. M. (1970). Natural selection and the concept of a protein space. *Nature,* **225**(232), 563-4.

Trakulnaleamsai, S., Aihara, S., Miyai, K., Suga, Y., Sota, M., Yomo, T. & Urabe, I. (1992). Revised sequence and activity of *Bacillus stearothermophilus* catalase I (formerly peroxidase). *J. Ferment. Bioeng.*, **74**, 234-237.

Trakulnaleamsai, S., Yomo, T., Yoshikawa, M., Aihara, S. & Urabe, I. (1995). Experimental sketch of landscape in protein sequence space. *J. Ferment. Bioeng.*, **79**, 107-118.

Triggs-Raine, B. L. & Loewen, P. C. (1987). Physical characterization of katG, encoding catalase HPI of *Escherichia coli. Gene,* **52**(2-3), 121-8.

Wells, J. A. (1990). Additivity of mutational effects in proteins. *Biochemistry,* **29**(37), 8509-17.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings 6th international congress on genetics* **1**, 356.

Yomo, T., Yamano, T., Yamamoto, K. & Urabe, I. (1997). General equation of steady-state enzyme kinetics using net rate constants and its application to the kinetic analysis of catalase reaction. *J. Theor. Biol.*, **188**(3), 301-12.

You, L. & Arnold, F. H. (1996). Directed evolution of subtilisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylformamide [published erratum appears in *Protein Eng.*, 1996 Aug;9(8):719]. *Protein Eng.*, **9**(1), 77-83.

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

List of publications

Publications related to this thesis.

1. Matsuura, T., Yomo, T., Trakulnaleamsai, S., Ohashi, Y., Yamamoto, K., and Urabe, I. (1998). Nonadditivity of mutational effects on the properties of catalase I and its application to efficient directed evolution. *Protein Eng.* 11, 789-795.

2. Matsuura, T., Miyai, K., Trakulnaleamsai, S., Yomo, T., Shima, Y., Miki, S., Yamamoto, K., and Urabe, I. (1998). Evolutionary molecular engineering by random elongation mutagenesis. *Nature Biotechnol.*, in press.

Other publication.

3. Yomo, T., Habu, T., Soga, S., Matsuura, T., Shima, Y., and Urabe, I. (1996). In vitro self-replication system as a minimum set of life. 403-405. in *Artificial life V*. Langton, C. G. and Shimohara, K.(eds.). The MIT press, Cambridge.

Acknowledgment

I am greatly indebted to the staff of the Department of Biotechnology, Faculty of engineering, Osaka university, especially the staff of the Laboratory of enzyme engineering. I must mention long and frequent discussion with Prof. Itaru Urabe, Assoc. Prof. Tetsuya Yomo and Asst. Pro. Yasufumi Shima during the many difficult and crucial stages in the accomplishment of this thesis. Their kind advise and heartful support encouraged me to overcome many difficulty in my studies.

I would like to express my gratitude to Prof. Kanaya and Prof. Husimi who gave me a chance to further improve my knowledge. Gratitude is also extended to the members in our laboratory. I could learn so many thing from them. I am also thankful to Dr. Elizabeth Ko Mitamura for her encouragement to complete my thesis.

Lastly, I would like to thank my family, my father for a financial as well as mental support, my mother for her heartful kindness. Also would like to appreciate to my wife Kumiko for supporting me in everything!!