Rational Strategies for Directed Evolution of Biocatalysts – Application to *Candida antarctica* lipase B (CALB)

Matthieu Chodorge,[a] Laurent Fourage,[a] Christophe Ullmann,[a] Vincent Duvivier,[a] Jean-Michel Masson,[b] Fabrice Lefèvre[a,*]

[a] Protéus S. A., 70 allée Graham Bell, Parc Georges Besse, 30000 Nîmes, France
Fax: (+33)-4-66-70-64-60, e-mail: flefevre@proteus.fr
[b] Institut National des Sciences Appliquées, Complexe scientifique de Rangueil, 31077 Toulouse Cedex, France
Fax: (+33)-5-61-17-59-94, e-mail: jean-michel.masson@insa-toulouse.fr

Received: January 26, 2005; Revised: March 29, 2005; Accepted: April 12, 2005

Abstract: Provided that the industrial constraints have been properly defined, the directed evolution technologies available today enable one to design tailor-made enzymes and biocatalytic routes for chemical processes. Family shuffling has proved to be a successful strategy using traditional recombination protocols. However, when starting from a single gene, the first step is to create the appropriate population of parental genes to ensure an efficient recombination during gene shuffling. Our recent work focused on the determination of rational directed evolution strategies that can be applied for the creation of an improved biocatalyst within the requested industrial timelines. For this purpose, we have developed a rational approach that first explores the "protein plasticity" (ability of the protein to accept mutations with a limited loss of activity) of the enzyme, which enables us to estimate (i) the "optimal mutation load" (number of mutations introduced per gene that gives the highest frequency of improved variants) and (ii) the "*ad minima* size sample" (minimal number of clones to be screened) that can be used to rapidly improve this enzyme. We have then applied this approach to create in a few weeks variants of the well known lipase B with a seven fold improvement in thermostability.

Keywords: biocatalysis; CALB; directed evolution; lipase B; strategies; thermostable

Introduction

Until recently, the use of biocatalysis in the field of fine chemistry has been pretty successful but somehow limited. The reason is that chemical process developers have access only to a limited range of commercially available biocatalysts. These "off-the-shelf" biocatalysts cannot fulfill the needs of the industry for the wide range of products required by the pharmaceutical industry. If chirality and reaction rate are two key parameters in fine chemistry production, substrate solubility is frequently a limiting factor. Thermoactive and thermostable enzymes are then usually required to enable the production process in organic solvents at a high temperature, thus increasing the substrate solubility for higher productivity, and decreasing the release of industrial wastes. For this purpose, we decided to run an optimization program from the natural backbone of the lipase B from *Candida antarctica*, which already catalyzes a great number of chemical synthesis reactions, but has a poor thermostability in solution.

Some work has already been done in the field of directed evolution of biocatalysts on the basis of random mutations: a small number of mutations (an average number of one mutation per gene) was introduced randomly to generate new enzymes. Although this strategy was successful in some cases,[1] other studies show that introducing a large number of mutations also lead to improved variants.[2] Up to now, few studies have been performed to rationalize the directed evolution strategies in order to run enzyme optimization programs in timelines that are compatibles with industrial requirements.

There was thus a need to define an appropriate methodology that would enable the design of the right strategy for each case of enzyme evolution. Using the *green fluorescent protein* (GFP) as a reference model, we first developed a protocol to rapidly determine the protein plasticity; i.e., the ability of a protein to accept mutations with a limited loss of activity. From these experimental data, we then built a mathematical model that determines the optimal number of mutations that must be introduced in a gene to reach the highest probability of generating improved variants. This model also determines the minimum number of clones that must be screened to detect these improved variants. We then applied this model to lipase B to rationally generate thermostable variants within a few weeks.

  DOI: 10.1002/adsc.200505055

Results and Discussion

Modeling the Mutation Load Effects on Protein Evolution

To create our statistical evolution model, we made the assumption that mutations can be classified in three types according to their effects on the protein: beneficial, neutral or deleterious. For a defined gene having a number (X) of mutations, the probability P to obtain an improved variant is a function of the number of beneficial mutations (B) and neutral mutations (N), among the number (E) of all the possible single mutations.

Thus for single mutations leading to an improved variant this probability is expressed as follows in Equation (1):

$$P_1^{\,X} = X.(B/E).(N/E)^{(X-1)} \quad (1)$$

When analyzing a statistically consistent population of variants, the evolution rate ER[X], defined as the fraction of improved variants in a mutated library, is a measure of P_1^X. The fraction of active variants AF[X], defined as the number of variants that are actives among the global population of variants, is a measure of (N/E)(X). Therefore Equation (1) becomes:

$$ER[X] = X.(B/E).AF[X-1] \quad (2)$$

As the fraction of active variants AF[X−1] is an exponentially decreasing function (when more mutations are introduced in a gene, less active variants are detected) of the average number of mutations per gene (also called mutation load),[3,4] Equation (2) can be solved as:

$$ER[X] = X.(B/E).e^{-\alpha.(X-1)} \quad (3)$$

The experimental measurement of the number of active variants AF[X−1] in a library bearing an average number of (X−1) mutations per gene allows the determination of the α factor. Equation (3) then enables us to rapidly estimate the value of frequency of improved variants in this random library.

We chose to evaluate the evolution of the *green fluorescent protein* (GFP) as a reference model to validate Equation (3). GFP is a simple reporter protein that emits a green fluorescence when excited at UV wavelength. When the *gfp* gene acquires a single specific mutation (T196C[5]), it turns into a *blue fluorescent protein* (BFP), which can easily be observed on plates by visual screening.

Eight randomly mutated libraries of the *gfp* gene, named library A to library H were generated with increasing error-prone PCR mutagenesis conditions. Fractions of active fluorescent GFP variants, AF[GFP], were observed from 82% to 3% for an average number of mutations per gene ranging from 1.6 to 14.8 respectively (Table 1).

As expected, the fraction of active fluorescent GFP variants was an exponentially decreasing function of the mutation load and the calculated α factor was 0.205.

Using Equation (3), the evolution rate of the GFP protein was calculated with the experimentally determined α factor, versus mutation load. Simulation of the GFP evolution rate is reported in Figure 1 (black line). To run this simulation, the B value was set at 1 (because a single specific mutation in the *gfp* gene leads to the BFP phenotype[5]) and E is equal to the number of total base pairs in the GFP gene (717 bp) × 3 (as the original base pair could be mutated into three other ones).

In parallel, the *gfp* libraries were screened for the appearance of blue fluorescent (BFP) variants and experimental results (experimental evolution rate) were overlaid with the simulation in Figure 1.

Table 1. Mutated *gfp* libraries analysis. AF[GFP] represents the fluorescent GFP clones fraction and ML(gfp) the mutation load per gfp gene.

Ep-PCR				
Library	$MgCl_2$ [mM]	$MnCl_2$ [mM]	AF[GFP] [%]	ML(*gfp*)
A	4	0	82	1.6
B	5	0	70	2.7
C	7	0	61	3.5
D	7	0.2	50	4.7
E	7	0.25	37	6.1
F	7	0.3	29	7.3
G	7	0.4	10	11.1
H	7	0.5	3	14.8

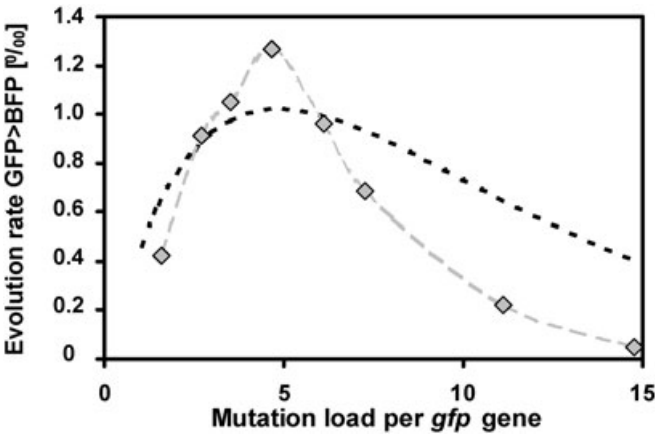

Figure 1. Evolution rate of GFP to BFP *versus* mutation load. Simulation is shown as dashed black line and experimental data are shown as dashed grey line and grey square.

Figure 1 shows that both simulation curve and experimental results reach the same asymptote, thus defining the optimal mutational load for this gene. A discrepancy is observed at high mutational loads between the simulation and the experimental curves. This is due to the cumulative effect of neutral mutations in the gene that have a lethal effect when combined. This phenomenon was intentionally not taken into account to simplify the mathematical model, as only the asymptote value is of interest in this work. The same optimal mutation load is reached for approximately 5 mutations corresponding to a frequency of BFP around 1.3‰. In the hypermutated library H (average number of 14.8 mutations), evolution rate is only 0.05‰, a value 25-fold lower than the one obtained at the optimum mutation load.

This simulation approach shows clearly that applying too low or too high mutation loads can hinder the identification of evolved variants. Our model thus enables us to create the optimized library in which improved variants could be rapidly identified and becomes a key tool for directed evolution studies.

Moreover, in addition to the optimal mutation load prediction, evolution rate simulations can be used to determine the *ad minima* screening sample size to ensure that all possible beneficial mutations have been analyzed. Assuming that all mutations occur at the same frequency, Equation (4) (as adapted from Moore[6]) describes the *ad minima* screening sample size S for a library with a given mutation load X, where c is the confidence limit and ER[X] is the simulated evolution rate.

$$(1 - ER[X].)^S < (1 - c) \quad (4)$$

Equation (4) can be used to rapidly define the number of variants that should be screened to be statistically consistent.

Directed Evolution to Increase Lipase CALB Thermostability

CALB is a highly versatile catalyst used successfully for the resolution and desymmetrization of numerous compounds in chemistry.[7,8] However, in aqueous solutions, the lipase denatures relatively quickly at temperatures as low as 40 °C.[9] We used our *in silico* simulation approach in order to determine the optimum mutation load for which an improved thermostable variant of CALB at 90 °C could be rapidly found. Six randomized *CalB* gene libraries, cloned in *E. coli*, were generated with increasing mutation load. 100 clones from each library were screened as described in the Experimental Section to determine active fractions AF[CALB] (see Figure 2).

Using AF[CALB] data, the α factor was then determined and introduced in Equation (3) in order to simulate the evolution rate ER[CALB] (see Figure 3).

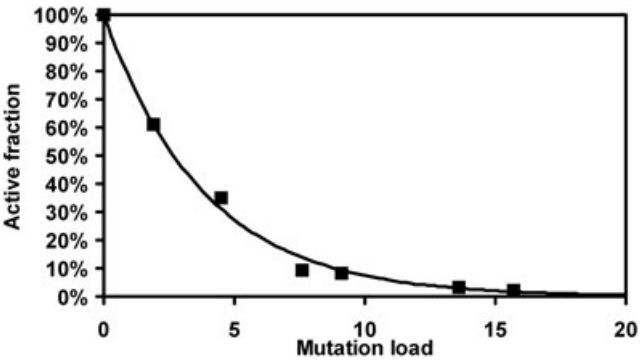

Figure 2. Active fraction of CALB *versus* mutation load.

The optimal mutation load was found to be 4.0 nucleic acid mutations per gene corresponding to a frequency of improved CALB of around 0.6‰. Using Equation (4), the screening of 10,000 variants was effective to be confident at 99.7% that all the single mutations will be screened. Based on this data, a screening of 10,000 clones from the corresponding "4 mutations per gene" library was done by selecting improved thermostable variants. The selection process consisted of incubating the mutant libraries for 1 h at 90 °C as described in the Experimental Section, and then assaying for the relative activity at 60 °C compared to the wild-type CALB (WT-CALB). Using this procedure, 24 mutants were selected. From these 24 mutants, one mutant (35E3) showed a 7.5-fold increase after 15 min at 90 °C compared with the WT-CALB as shown Figure 4.

In addition, the activity levels of both 35E3 mutant and WT-CALB, expressed at 37 °C using *E. coli* cells, were approximately the same and no significant differences in the expression level were observed (data not shown). The sequence of the 35E3 mutant was determined and one amino acid mutation (N317Y) was identified by comparison with the WT-CALB sequence. Analysis of the 3D structure[10] showed that this amino acid is located at the surface of the protein. The enhanced thermostability due to the replacement of the as-

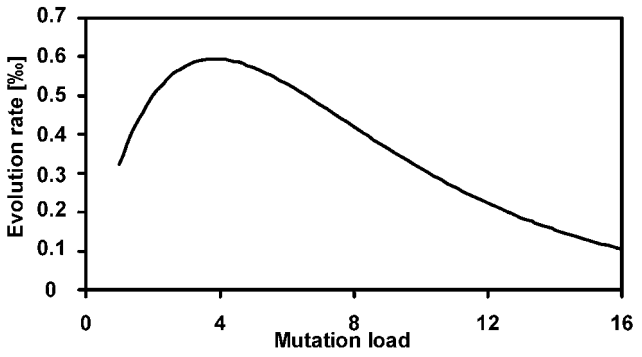

Figure 3. Simulation of the evolution rate of CALB variants *versus* mutation load.

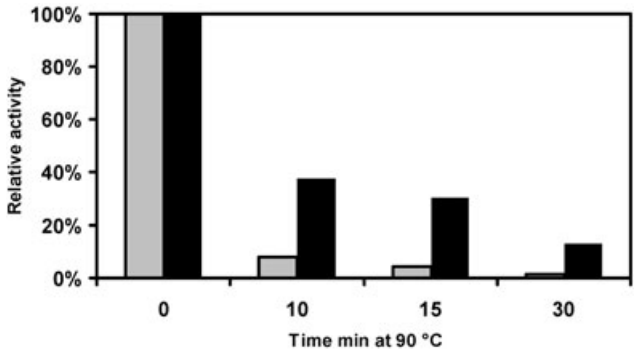

Figure 4. Thermal inactivation at 90 °C: temperature-dependency of WT-CALB (in grey) and the 35E3 mutant (in black).

paragine by a tyrosine residue at this position suggests that it is related to the limitation of potential deamidation processes.[11]

Conclusion

Molecules in fine chemistry are becoming more and more complex and an increasing number of projects involve biocatalytic steps to overcome the difficulties encountered. In such cases, there is a huge pressure on time for the chemical process R&D developers to find the appropriate biocatalyst as they need to accelerate the time to market. To fulfill this industrial need, we have developed an efficient model of enzymes evolution rate simulation that speeds up enzyme evolution.

Our model allows one to first determine the optimal number of random mutations that must be introduced in the considered gene to obtain libraries with a high proportion of improved variants. Optimal mutation loads for GFP and CALB evolutions were determined at around 5.0 and 4.0 mutations per gene, respectively. This optimal number of mutations reflects the level of protein plasticity. The better understanding of protein plasticity has major implications for recombination-based evolution experiments: genes that support high mutational loads can be more easily used to create initial diversified libraries of mutated variants, that can afterwards be recombined *in vitro*.

Secondly this model enables one to determine the *ad minima* screening sample size to be sure that all possible independent variants generated after random mutagenesis will be analyzed. One should be aware that this sample size varies greatly with the mutation bias, the gene length and the mutation load. This model can be applied for a wide range of enzyme evolution programs, provided that the appropriate screening conditions of the variants have been defined and calibrated to mimic the final industrial constraints.

Using this new methodology, we achieved a rational directed evolution of the lipase B from *Candida antarctica* to rapidly generate an improved resistance towards thermal inactivation at 90 °C for the industrial processing of low solubility raw materials. By screening only 10,000 clones, we obtained a set of enhanced variants of CALB with regard to thermostability (up to 7 times improvements for the best one as compared to the wild-type enzyme). These variants will now be used as a source of parental genes for a next step of directed evolution by gene L-Shuffling™.

Experimental Section

Gene Random Mutagenesis

For this study, the *Candida antarctica* strain was purchased from the CBS collection (CBS 214.83), and CalB gene has been cloned by PCR amplification. Random mutagenesis for the gfp and CalB genes were performed by error-prone PCR (ep-PCR) as described by Cadwell.[12] Reaction mixtures contained 10 mM Tris-HCl pH 9, 50 mM KCl, 0.1% Triton X-100, 0.2 mg/mL BSA, 2.5 U Taq polymerase (QBiogen), and 20 pmol of each primer:

pET5′ (5′-AGATCTCGATCCCGCGAAATTAATACG-3′) and,

pET3′ (5′-CAAAAAACCCCTCAAGACCCGTTTAG-3′).

Error-rate was controlled by concentrations in $MgCl_2$ ranging from 4 mM to 7 mM and $MnCl_2$ ranging from 0 mM to 0.5 mM. The dNTP concentrations were 0.2 mM dATP, 0.2 mM dGTP, 1 mM dTTP and 1 mM dCTP. 1 fmol of a pET vector containing the gfp or CalB gene was used as DNA template. A PCR program of 94 °C for 5 min; 91 °C for 30 s, 60 °C for 30 s, 72 °C for 1 min (30 times) followed by 10 min at 72 °C was used in an MJ Research PTC-200 thermocycler. PCR products were first gel-quantified by comparison with a known amount of DNA and purified using a QIAquick PCR purification column (Qiagen).

Cloning gfp and CalB Gene Libraries

PCR products were digested with *Nde* I and *Eco* RI (New England Biolabs), purified on a QIAquick PCR purification column and ligated in a *Nde* I-*Eco* RI digested pET26b + plasmid (Novozym). Resulting ligations were used to transform by electroporation *E. coli* MC1061(DE3) cells (hsdR2 hsdM + hsdS + araD139 · (ara-leu)7697 · (lac)X74 galE15 galK16 rpsL (Strr) mcrA mcrB1 DE3) as described by Maniatis.[13] For each library, transformed MC1061(DE3) cells were spread on LB agar plates containing 60 μg/mL of kanamycine in order to obtain around 25,000 independent clones per library with a density of 1,000 colonies per 12 cm × 12 cm plate. The number of recombinant clones was deduced from the total number of colonies and gene insertion yields after estimation by PCR on 96 randomly selected colonies.

Screening of Enzyme Variants (GFP and Lipase B)

For GFP, *E. coli* MC1061(DE3) colonies expressing a green or a blue fluorescent protein were plated and grown on agar plates

(with 60 μg/mL of kanamycine), and then visually detected on a UV bench with an excitation length at 355 nm.

For lipase B variants, *E. coli* MC1061(DE3) colonies expressing the CALB lipase mutants were grown in 96 wells microtiterplates at 37 °C in 150 μL of Luria-Bertani (LB) medium complemented with 60 μg/mL of kanamycine. After centrifugation (4 min at 4000 g) and resuspension in 50 μL of 200 mM PIPES buffer at pH 7.0, cells were incubated for 1 hour at 90 °C. After this incubation, residual activities were determined by incubating the cells at 60 °C with a synthetic C10 ester CLIPS-O™ as substrate (as described by Lagarde et al.[14]), and reading the corresponding absorbance at 414 nm. These values were compared to the value obtained with cells expressing the WT-CALB tested in the same conditions.

Characterization of CALB Variants

Growth conditions used for the characterization of the CALB variants were the same as the ones used during the screening. After centrifugation (4 min at 4000 g), E. coli MC1061(DE3) clones expressing the CALB improved lipase variants were resuspended in 50 μL of 200 mM PIPES buffer at pH 7.0 and incubated at different times: 5, 10, 15 and 30 min at 90 °C. Residual activities were determined at 60 °C using the synthetic C_{10} ester CLIPS-O™ as substrate (as described by Lagarde et al.[14]), and reading the corresponding absorbance at 414 nm.

Acknowledgements

We would like to thank Mr. Vincent Monziols for useful discussions and Protéus Directed Evolution Group for their contribution to this work.

References

[1] F. H. Arnold, *Acc. Chem. Res.* **1998**, *31*, 125–131.
[2] M. Zaccolo, E. Gherardi, *J. Mol. Biol.* **1999**, *285*, 775–783.
[3] P. S. Daugherty, G. Chen, B. L. Iverson, G. Georgiou, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 2029–2034.
[4] S. Shafikhani, R. A. Siegel, E. Ferrari, V. Schellenberger, *BioTechniques* **1997**, *23*, 304–310.
[5] R. Heim, D. C. Prasher, R. Y. Tsien. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 12501–12504.
[6] J. C. Moore, H. M. Jin, O. Kuchner, F. H. Arnold, *J. Mol. Biol.* **1997**, *272*, 336–347.
[7] K. E. Jaeger, M. T. Reetz, *Trends Biotechnol.* **1998**, *16*, 396–403.
[8] K. E. Jaeger, T. Eggert, *Curr. Opin. Biotechnol.* **2002**, *13*, 390–397.
[9] M. J. Homann, R. Vail, B. Morgan, V. Sabesan, C. Levy, D. R. Dodds, A. Zaks, *Adv. Synth. Catal.* **2001**, *343*, 744–749.
[10] J. Uppenberg, N. Ohrner, M. Norin, K. Hult, G. J. Kleywegt, S. Patkar, V. Waagen, T. Anthonsen, T. A. Jones, *Biochemistry* **1995**, *34*, 16838–16851.
[11] N. Declerck, M. Machius, G. Wiegand, R. Huber, C. Gaillardin, *J. Mol. Biol*. **2000**, *301*, 1041–1057.
[12] R. C. Cadwell, G. F. Joyce, *PCR Meth. Appl.* **1992**, *2*, 28–33.
[13] T. Maniatis, E. F. Fritsch, J. Sambrook, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, **1982**.
[14] D. Lagarde, H. K. Nguyen, G. Ravot, D. Wahler, J. L. Reymond, G. Hills, T. Veit, F. Lefevre, *Org. Proc. Res. Dev.* **2002**, *6*, 441–445.

