Enzyme Engineering

Luo, Enzyme Engg 2012, 1:1
DOI: 10.4172/2329-6674.1000104

Review Article Open Access

Genomic Data Mining: An Efficient Way to Find New and Better Enzymes

Xiao-Jing Luo, Hui-Lei Yu and Jian-He Xu*

Laboratory of Biocatalysis and Synthetic Biotechnology, State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai 200237, China

Abstract

It is top priority nowadays for biocatalysis researchers to discover novel, potent enzymes and to redesign and engineer for tailor-made enzymes. Genomic data mining, which depends on the burgeoning computational algorithms and bioinformatics tools, accelerates the process by *in silico* screening and constructing focused/smart mutant libraries.

Keywords: Genomic data mining; Genome hunting; Protein engineering; Semi-rational design; High-throughput screening

Introduction

Biocatalysts have been used in manufacturing food or fabrics for thousands of years. In the context of global concern about reducing chemical pollution and energy consumption, biocatalysis and biotransformation are of surging interest for both scientific and industrial researchers in the past few decades. As an environmentally benign process with higher efficiency and selectivity, enzyme-based catalysis is rapidly becoming a fresh paradigm for pharmaceutical, petrochemical, flavor and food industries. In the current third wave of biocatalysis, there have been numerous commercial enzymes with remarkable capabilities available for chemists [1]. However, the pool of enzymes catering to industrial demands is still insufficient. Hence, it is still necessary to discover and engineer novel and better enzymes. Recently the trends in discovering and engineering enzymes for organic synthesis have been reviewed in depth [2].

Enrichment cultivation, a classical approach for enzyme screening, is an effective but time-consuming procedure, which includes sampling, microorganism cultivation, strain isolation, genomic DNA extraction (or protein purification), and gene identification. Typically these take one to two years for scientists to find a new enzyme. Fortunately, genomic data mining, combined with recombinant DNA technique, circumvents these excruciating steps, shortening the screening period from years to months, or even weeks. This mini-review focuses mainly on the basic strategies of data mining for enzyme discovering, and selected recent examples are given to illustrate the potency of these methods in efficiently finding new and better enzymes with excellent selectivity, versatility and stability. In addition, data mining approaches in protein engineering are also introduced and discussed.

Discovering New Enzymes

The advances in bioinformatics have spurred significant progress in biocatalysis since the new millennium. DNA sequencing of bacterial genome or metagenomic fragments is much easier, cheaper and faster, which generates a large amount of genomic information. The massive sequence information is deposited in integrated or cross-linked public databases. As an example of the abundance of the information, currently searching "monooxygenase" in the NCBI (National Center for Biotechnology Information, www.ncbi.nlm.nih.gov) website will recover more than 90,000 monooxygenase amino acid sequences. It is worthy of noting that most of the enzyme sequences are not functionally confirmed, which undoubtedly makes sequence databases as new treasures for biocatalysis researchers.

How to exploit the rapidly expanding information? Thanks to the development of predictive bioinformatics tools, genomic data mining for new enzymes becomes an established routine. Generally speaking, there are two approaches [3]. One practical approach, dubbed genome hunting, is to "hunt" for enzymes within a specified microorganism. Open reading frames are searched in the genome of a certain microorganism which was selected from either soil samples or culture collections. Sequences that are annotated as reviewed or putative enzymes are subjected to subsequent molecular cloning, over-expression and activity screening. Such strategy proved to be effective in searching reductases from *Bacillus* sp. ECU0013 without constructing the genomic library. Ketone reductase-producing *Bacillus* sp. strain ECU0013 was previously isolated from soil samples, exhibiting excellent stereoselectivity and substrate tolerance [4]. After *in silico* mining, eleven oxidoreductases from the strain were heterologously overexpressed in *E. coli* BL21 (DE3). Subsequent screening revealed three recombinant reductases, BYueD, YtbE and FabG, with good activity and high stereoselectivity towards various prochiral ketones. Among them, BYueD could reduce all the 14 tested β-ketoesters or aromatic ketones to corresponding chiral alcohols in almost enantiomerically pure forms [5] and YtbE exhibited high prochiral selectivity in the reduction of various carbonyl substrates (>99% *ee*) [6]. An impressive example is the asymmetric reduction of ethyl 2-oxo-4-phenylbutyrate (OPBE) using *E. coli* co expressing FabG and GDH, in which 620 $g \cdot L^{-1}$ OPBE were completely converted to ethyl (*S*)-2-hydroxy-4-phenylbutyrate [(*S*)-HPBE] in very high enantiomeric excess within 12 hours without external addition of expensive cofactor [7].

Although sequence annotation is predictive, some amino acid sequences can be unnamed, or erroneously annotated. To "rescue" these potential enzymes, mathematical and computational methods are helpful. Based on the homology search of conserved regions shared among α/β-hydrolase fold Epoxide Hydrolases (EHs), a recombinant BMEH cloned from *Bacillus megaterium* ECU1001 showed a very high activity in kinetic resolution of *rac*-glycidyl ethers. Interestingly, the enantioselectivity of BMEH was switched by different nitro

***Corresponding author**: Jian-He Xu, Laboratory of Biocatalysis and Synthetic Biotechnology, State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai 200237, China, Tel: +86-21-64252498; Fax: +86-21-64250840; E-mail: jianhexu@ecust.edu.cn

Received May 29, 2012; **Accepted** June 25, 2012; **Published** June 27, 2012

Citation: Luo XJ, Yu HL, Xu JH (2012) Genomic Data Mining: An Efficient Way to Find New and Better Enzymes. Enzyme Engg 1:104. doi: 10.4172/2329-6674.1000104

substitution sites of substrates. In both the two cases, *E*-values of the bioreaction were excellent (> 200), despite of the completely opposite enantiopreference [8]. Since BMEH shows only 24% sequence identity to the closest structurally known homologue, discovery of the novel enzyme would be more difficult without the prowess of local similarity search within distant homologues.

Another approach, dubbed data mining, is to search enzymes of interest by homology alignment among all the sequences deposited in the whole database. Reported enzymes with desired properties are preliminarily chosen as templates. A BLAST-search finds conserved regions between sequences, and yields homologous protein sequences. Usually the sequences with moderate identity are selected as candidates. This approach is simple and effective, which has inspired researchers to discover various promising biocatalysts. While screening ten oxidoreductases sharing 40-80% sequence identities with confirmed COBE reductases, a new reductase from *Streptomyces coelicolor*, designated as ScCR, was discovered with high activity and excellent stereoselectivity towards β-ketoesters. After simple optimization, ScCR was able to asymmetrically synthesize ethyl (*S*)-4-chloro-3-hydroxybutanoate [(*S*)-CHBE] in a biphasic system with a high total turnover number of 12,100 [9]. An exciting example of reductase reported recently is the identification of CgKR2 from *Candida glabrata* for the synthesis of ethyl (*R*)-2-hydroxy-4-phenylbutyrate [(*R*)-HPBE], an important building block of Angiotensin-Converting Enzyme (ACE) inhibitors. Coupled with a cofactor regeneration system, the recombinant *E. coli* could synthesize optically pure (*R*)-HPBE at a remarkable productivity of 700 g/L/d, which is 27-fold higher than the best record reported so far [10]. Newly mined enzymes in our laboratory have greatly improved the biocatalytic production of chiral intermediates of clopidogrel, a best-selling anticlotting drug. A yeast-origin carbonyl reductase CgKR1 was heterologously expressed in *E. coli* and exhibited best stereoselectivity among 6 homologues of the template reductase Gre2p. Space-time yield of methyl (*R*)-o-chloromandelate [(*R*)-CMM], an important intermediate for synthesis of clopidogrel, reached as high as 700 g/L/d by using crude enzymes of CgKR1 and BsGDH (*Bacillus megaterium* glucose dehydrogenase) without any addition of external NADP$^+$ [11]. Likewise, a new nitrilase LaN mined from seven virtual selected candidates was reported to produce (*R*)-o-chloromandelic acid, also a key chiral synthon for clopidogrel, with good enantioselectivity and high substrate tolerance [12].

Large scale application of biocatalysts is often hampered by the severe industrial conditions (e.g. high temperature) which cause irreversible inactivation of enzymes and lead to low productivity. Thus reactions driven by thermostable enzymes will be advantageous because of higher reaction rates and shorter equilibrium time. Normally, extremophiles are the major sources of thermally tolerant enzymes. Hunting in the genomes of (hyper-) thermophilic organisms and mining the homologues of thermostable templates, pave the way for finding novel, robust enzymes with industrial potential. Recently, many thermostable enzymes have been reported based on the genomic data mining approach, including endoglucanases [13], cytochrome P450 monooxygenase [14], xylanase [15], and laccase [16]. Significant progress has been made in searching new thermostable β-glucosidase lately. Dotsenko and co-workers discovered glycoside hydrolase Bxl5 from fungus *Chrysosporium lucknowense* C1 by genome hunting method. The half life of Bxl5 was up to 6 h at 65°C under acidic conditions [17]. On the other hand, β-glucosidase DtGH was mined from a series of homologous candidates of an apple seed glycosidase. DtGH was highly thermostable, with a half life of more than 500 h at 70°C, and therefore considered as a promising enzyme for faster synthesis of various glucosides at a higher productivity [18].

Engineering for Better Enzymes

Protein engineering provides versatile ways to generate tailored enzymes. Directed evolution, exemplified by error-prone PCR (epPCR) and DNA shuffling techniques, proves to be powerful in generating enzyme variants with desired properties. However, such strategies are not flawless. Directed evolution requires screening or selection of 10^3-10^6 variants in each round, thus a high-throughput screening system is indispensable. Simultaneous multi-sites mutation seems to be impractical for directed evolution because two substitutions would theoretically generate more than 30 million combinations within a 300-amino acid protein [19]. In addition, synergetic effects might be neglected. Compared with laborious and lengthy conventional methods, rational design and semi-rational approach are preferred, in which small but smart libraries are designed. Methods and examples related to the two approaches were extensively reviewed elsewhere [20-25]. In the post-genomic era, advances in bioinformatics and computational algorithms facilitate the creation of focused and knowledge-based libraries by preselecting mutation sites and limiting amino acids diversity. The *in silico* protein design strategies comprise sequence-guided, structure-guided or hybrid methods.

Sequence-guided design aligns the target protein with various homologous enzymes to identify the conserved or differing amino acids. It is based on the hypothesis that consensus amino acids play more important roles in sequence-property relationship than those nonconsensus ones. Substitution or saturation mutagenesis of these hot-spots is more convincing than random mutagenesis. Successful examples, such as altering cofactor specificity of *Bacillus stearothermophilus* lactate dehydrogenase [26], or improving thermostability of *Pseudomonas fluorescens* esterase (PFE) [27] underline the importance of the "consensus approach". Drastic changes in amino acids could lead to loss of stability. To ensure more variants are functional and correctly folded, 3DM database is exploited to improve activity and enantioselectivity of PFE. Frequently appeared amino acids were chosen from structural-based sequence alignment (3DM) of 1751 sequences of α/β-hydrolase fold enzymes. The limited amino acids diversity dramatically reduces the screening effort by more than 300-fold, while the improvement of activity and enantioselectivity is still significant [27]. In a recent example, multiple sequence alignments and phylogenetic tree were analyzed to design small libraries depending on predictive ancestral sequences. As much as 50-fold activity improvement could be achieved by screening only 300 variants [28].

In parallel with sequence-guided design, structure-based design sheds light on the structure-property relationship. RCSB PDB databank (www.rcsb.org) contains a large amount of experimentally-determined protein structures or protein-ligand complexes. With the structure information at hand, rational design can be applied. For instance, switching the bulky residues to smaller ones around substrate binding sites may change the substrate spectrum or reverse stereoselectivity. In addition, Combinatorial Active site Saturation Test (CAST) or crystallographic B-Factor Iterative Test (B-FIT), integrated with Iterative Saturation Mutagenesis (ISM), can rapidly redesign enantioselective or thermostable enzymes, respectively [29,30]. Unfortunately, enzyme structures are not always available. In this case, homology structure modeling enables us to view computational structures in no time. Many web servers or programs have been developed, including 3D-JIGSAW, ESyPred3D, HHpred, HMMSTR/Rosetta, SWISS-MODEL and Modeller. Numerous successful examples have demonstrated the aptitude of bioinformatics tools in protein structure prediction. Redesigning thermostable L-aminoacylase TliACY reported lately exemplified that homology model facilitates the rapid identification of "hotspot" sites when crystal structure is not available [31].

Conclusions

Approaches to discovering and engineering new and better enzymes of industrial potential are rapidly developing. Apart from laborious conventional screening methods, the *de novo* enzyme design is in its infancy due to limited structure-function understanding, while nascent method of incorporating Unnatural Amino Acids (UAAs) in directed evolution is still less efficient, unpredictable, and difficult to operate. As an alternative, genomic data mining is taking advantage of the enlightening and ever increasing sequence information and bioinformatics tools. Combined with established recombinant DNA techniques, these computational methods (say, genome hunting or data mining) limit the time scale of enzyme discovery to months or even weeks. Higher efficiency could be achieved by refining sequence-based or activity-based screening strategies. Höhne et al. [32] proposed an ingenious strategy to spot desired enzymes directly from the protein sequences inventory. Sequence-based alignment of key motifs featured by structural information revealed 17 amine transaminases with excellent (*R*)-selectivity. An impressive one-pot High-throughput in vitro Glycoside Hydrolase (HIGH) method [33] made possible expression and screening of 82 putative GHs in 3 hours! Furthermore, prompt *in silico* analysis of multiple sequence and structure information provided by data mining, web servers or programs brings about small and smart libraries, shifting the redesign focus from reliable high-throughput system to accurate prediction strategies. Together with advanced DNA technologies and improved evolution conceptions, genomic data mining offers practical and efficient routes to both biocatalyst discovery and protein engineering.

References

1. Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, et al. (2012) Engineering the third wave of biocatalysis. Nature 485: 185-194.
2. Behrens GA, Hummel A, Padhi SK, Schätzle S, Bornscheuer UT (2011) Discovery and protein engineering of biocatalysts for organic synthesis. Adv Synth Catal 353: 2191-2215.
3. Ni Y, Xu JH (2011) Biocatalytic ketone reduction: A green and efficient access to enantiopure alcohols. Biotechnol Adv.
4. Xie Y, Xu JH, Xu Y (2010) Isolation of a *Bacillus* strain producing ketone reductase with high substrate tolerance. Bioresour Technol 101: 1054-1059.
5. Ni Y, Li CX, Wang LJ, Zhang J, Xu JH (2011) Highly stereoselective reduction of prochiral ketones by a bacterial reductase coupled with cofactor regeneration. Org Biomol Chem 9: 5463-5468.
6. Ni Y, Li CX, Ma HM, Zhang J, Xu JH (2011) Biocatalytic properties of a recombinant aldo-keto reductase with broad substrate spectrum and excellent stereoselectivity. Appl Microbiol Biotechnol 89: 1111-1118.
7. Ni Y, Li CX, Zhang J, Shen ND, Bornscheuer UT, et al. (2011) Efficient reduction of ethyl 2-oxo-4-phenylbutyrate at 620 g·L^{-1} by a bacterial reductase with broad substrate spectrum. Adv Synth Catal 353: 1213-1217.
8. Zhao J, Chu YY, Li AT, Ju X, Kong XD, et al. (2011) An uusual (*R*)-selective epoxide hydrolase with high activity for facile preparation of enantiopure glycidyl ethers. Adv Synth Catal 353: 1510-1518.
9. Wang LJ, Li CX, Ni Y, Zhang J, Liu X, et al. (2011) Highly efficient synthesis of chiral alcohols with a novel NADH-dependent reductase from *Streptomyces coelicolor*. Bioresour Technol 102: 7023-7028.
10. Shen ND, Ni Y, Ma HM, Wang LJ, Li CX, et al. (2012) Efficient synthesis of a chiral precursor for angiotensin-converting enzyme (ACE) Inhibitors in high space-time yield by a new reductase without external cofactors. Org Lett 14: 1982-1985.
11. Ma HM, Yang LL, Ni Y, Zhang J, Li CX, et al. (2012) Stereospecific reduction of methyl *o*-chlorobenzoylformate at 300 g·L^{-1} without additional cofactor using a carbonyl reductase mined from *Candida glabrata*. Adv Synth Catal 354: 1765-1772.
12. Zhang CS, Zhang ZJ, Li CX, Yu HL, Zheng GW, et al. (2012) Efficient production of (*R*)-*o*-chloromandelic acid by deracemization of *o*-chloromandelonitrile with a new nitrilase mined from *Labrenzia aggregate*. Appl Microbiol Biotechnol 95: 91-99.
13. Qiu LH, Li CX, Sun J, Wang Z, Ye Q, et al. (2011) Thermostable bacterial endoglucanases mined from Swiss-Prot database. Appl Biochem Biotechnol 165: 1473-1484.
14. Schallmey A, den Besten G, Teune IG, Kembaren RF, Janssen DB (2011) Characterization of cytochrome P450 monooxygenase CYP154H1 from the thermophilic soil bacterium *Thermobifida fusca*. Appl Microbiol Biotechnol 89: 1475-1485.
15. Hung KS, Liu SM, Tzou WS, Lin FP, Pan CL, et al. (2011) Characterization of a novel GH10 thermostable, halophilic xylanase from the marine bacterium *Thermoanaerobacterium saccharolyticum* NTOU1. Process Biochem 46: 1257-1263.
16. Reiss R, Ihssen J, Thöny-Meyer L (2011) *Bacillus pumilus* laccase: a heat stable enzyme with a wide substrate spectrum. BMC Biotechnol 11: 9.
17. Dotsenko GS, Sinitsyna OA, Hinz SW, Wery J, Sinitsyn AP (2012) Characterization of a GH family 3 β-glycoside hydrolase from *Chrysosporium lucknowense* and its application to the hydrolysis of β-glucan and xylan. Bioresour Technol 112: 345-349.
18. Zou ZZ, Yu HL, Li CX, Zhou XW, Hayashi C, et al. (2012) A new thermostable β-glucosidase mined from *Dictyoglomus thermophilum*: properties and performance in octyl glucoside synthesis at high temperatures. Bioresour Technol 118C: 425-430.
19. Tracewell CA, Arnold FH (2009) Directed enzyme evolution: climbing fitness peaks one amino acid at a time. Curr Opin Chem Biol 13: 3-9.
20. Bornscheuer UT, Pohl M (2001) Improved biocatalysts by directed evolution and rational protein design. Curr Opin Chem Biol 5: 137-143.
21. Chica RA, Doucet N, Pelletier JN (2005) Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. Curr Opin Biotechnol 16: 378-384.
22. Lutz S (2010) Beyond directed evolution—semi-rational protein engineering and design. Curr Opin Biotechnol 21: 734-743.
23. Dalby PA (2011) Strategy and success for the directed evolution of enzymes. Curr Opin Struct Biol 21: 473-480.
24. Goldsmith M, Tawfik DS (2012) Directed enzyme evolution: beyond the low-hanging fruit. Curr Opin Struct Biol.
25. Illanes A, Cauerhff A, Wilson L, Castro GR (2012) Recent trends in biocatalysis engineering. Bioresour Technol 115:48-57.
26. Flores H, Ellington AD (2005) A modified consensus approach to mutagenesis inverts the cofactor specificity of *Bacillus stearothermophilus* lactate dehydrogenase. Protein Eng Des Sel 18: 369-377.
27. Jochens H, Bornscheuer UT (2010) Natural diversity to guide focused directed evolution. Chembiochem 11: 1861-1866.
28. Alcolombri U, Elias M, Tawfik DS (2011) Directed evolution of sulfotransferases and paraoxonases by ancestral libraries. J Mol Biol 411: 837-853.
29. Reetz MT, Wang LW, Bocola M (2006) Directed evolution of enantioselective enzymes: iterative cycles of CASTing for probing protein-sequence space. Angew Chem Int Ed Engl 45: 1236-1241.
30. Reetz MT, Carballeira JD, Vogel A (2006) Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. Angew Chem Int Ed Engl 45: 7745-7751.
31. Jochens H, Aerts D, Bornscheuer UT (2010) Thermostabilization of an esterase by alignment-guided focussed directed evolution. Protein Eng Des Sel 23: 903-909.

32. Höhne M, Schätzle S, Jochens H, Robins K, Bornscheuer UT (2010) Rational assignment of key motifs for function guides *in silico* enzyme identification. Nat Chem Biol 6: 807-813.

33. Kim TW, Chokhawala HA, Hess M, Dana CM, Baer Z, et al. (2011) High-throughput in vitro glycoside hydrolase (HIGH) screening for enzyme discovery. Angew Chem Int Ed Engl 50: 11215-11218.