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 last 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 are 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 “monoxygenase” in the NCBI (National Center for Biotechnology Information, www.ncbi.nlm.nih.gov) website will recover more than 90,000 monoxygenase 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, YbE 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 YbE 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 L⁻¹ OPBE were completely converted toethyl (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 Epoxyd Hydroxylases (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 dinitro

**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


Copyright: © 2012 Luo XJ, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
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 clopigdogrel, 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 clopigdogrel, 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 nitritase LaN mined from seven virtual selected candidates was reported to produce (R)-o-chloromandelic acid, also a key chiral synthon for (+)-CHBE, 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 knowledge-based libraries by preselecting mutation sites and limiting 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

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 10³-10⁴ 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 of industrial potential are rapidly developing. Apart from biocatalyst discovery and protein engineering.

References
