ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Biotechnology

# Metagenomic era for biocatalyst identification

Lucía Fernández-Arrojo, María-Eugenia Guazzaroni, Nieves López-Cortés, Ana Beloqui and Manuel Ferrer

Microbial enzymes have many known applications as biocatalysts. However, only a few of them are currently employed for biocatalysis even though an annotated collection of more than 190 billion bases is available in metagenome sequence databases from uncultured and highly diverse microbial populations. This review aims at providing conceptual and technical bases for the translation of metagenome data into both experimental and computational frameworks that facilitates a comprehensive analysis of the biocatalysts diversity space. We will also briefly present the status of the current capabilities that assess and predict catalytic potential of environmental sites and track its diversity and evolution in large-scale biocatalysis process resulting from studies applying metagenomics in association with gene fingerprinting, catabolic arrays and complementary '-omics'.

**Address**
Department of Applied Biocatalysis, Institute of Catalysis, CSIC, Madrid, Spain

Corresponding author: Ferrer, Manuel (mferrer@icp.csic.es)

**Current Opinion in Biotechnology** 2010, **21**:725–733

This review comes from a themed issue on
Chemical biotechnology
Edited by Phil Holliger and Karl Erich Jaeger

Available online 9th October 2010

0958-1669/$ – see front matter

**DOI** 10.1016/j.copbio.2010.09.006

## Introduction

Given the wide-spread occurrence of molecules and their potential used through chemical modification, it is long-held desire to design efficient biocatalysis processes via amenable microbial activities [1•]. However, despite their promising performance in the laboratory, the application of natural enzymes in industrial scale or near-industrial situations has mostly ended failure [2•]. This limitation is primarily due to the lack of availability of microbial enzymes that can perform the desired chemical reactions. This in turn is attributable to the fact that only a limited number of sequenced microbes are available since most microbes cannot be cultured [3••]. This yet uncultured majority of microbes, adapted to a wide range of physical–chemical parameters matching industrial requirements (pH, pressure and temperature), potentially conceal an enormous treasure of unknown biological functions locked in their hereunto unidentified genes, proteins and biochemical pathways, the elucidation of which can indisputably increase the chance to design new biotech processes [4].

There is also a technical problem associated to enzyme fitness itself: enzymes currently used or those to be identified may not be 'ideal' enzymes for a given bioprocess, and sometimes the industrial processes have to be designed to purposely fit these suboptimal enzymes [5]. After a series of publications dedicated to the beneficial insertion of random mutations we expected a shift of the functional diversity paradigm: it appears that the diversity of protein variants is not as great as it was thought few years ago and it is impractical, costly and often impossible to test all emerging variants [6]. Moreover, changing the substrate selectivity of an enzyme remains a hurdle for larger implementation of this approach and it is impossible to start a protein engineering project for every single protein actually deposited in databases [7].

The nature appears the veteran protein engineer since it began its bioengineering 'experiments' billions years ago and has already created and tested *in vivo* an intriguing functional diversity of biochemical pathways and their constituent enzymes that perform numerous transformations of molecules in diverse biological settings with great precision and specificity [8]. Therefore, it is conceivable that a biocatalyst we are looking for may already exist in nature, and thus an interesting strategy would be to augment our knowledge base by exploring the inherent diversity of natural environments. To this end, a particularly large number of culture-independent techniques have been developed, which now allow the determination of microbial diversity and activity, easy screenings for particular gene diversity, gene quantification and whole genome sequencing of prokaryotic and eukaryotic isolates and communities [9]. Such techniques have turned out to be especially amenable for detecting molecule-metabolizing organisms in a natural setting and the identification of active biocatalysts [10]. Only a detailed understanding of the functioning and diversity of enzymes in an environmental context will allow their rational utilization and manipulation for the purpose of optimizing biocatalysis efforts.

We will present here the status of the current capabilities that access and predict the catalytic potential of microbial communities, an objective that capitalize the utilization of biological activities for biocatalysis.

## Interpreting an incomplete scroll saw: functional miss-annotation

Genomic and (meta-) genomic studies allow a reconstruction of the enzyme space relevant for transformation of a wide range of molecules, providing a holistic (or system) view on the enzyme network of a particular organism or microbial community [8,11]. Quite importantly, among current bottlenecks in metagenome analysis the lack of knowledge and insufficient efforts on enzymology [9] and amplifying annotation mistakes in databases [12,13] are of great hindrance for functionally analyzing the enzymatic landscape.

Overall it is evident that a large proportion of the open reading frames of newly sequenced metagenomes have little sequence homology with known enzymes, so their potential activities remain hidden [14••]. Moreover, of the new data from metagenome sequencing projects, the majority of protein sequences in public databases have not been experimentally characterized [3]. Even though annotation strategies have become more sophisticated in recent years [15••], the most common approach in use today continues to be the assignment of molecular function from the inference of homology followed by annotation transfer [13].

The miss-annotation level for molecular functions in public protein sequence databases has been extensively debated. Recently, from a model set of 37 enzyme families from which extensive experimental information exist for a high number of members [13], it has been observed a surprisingly high level (~80%) of miss-annotation for some of the subfamilies studied. This was mainly due to the over-estimation of molecular functions. Accordingly, an exponential increase in miss-annotations from 1993 to 2005 was found, as consequence of the rapid increase of sequences deposited in databases [3••]. Thus, miss-annotation in enzyme superfamilies containing multiple families that catalyse different reactions is a large problem that has been recognized. As an example, gallate dioxygenases are typically annotated as protocatechuate 4,5-dioxygenases in at least 3% of the metagenome sequencing projects, and their computational screen for biocatalysis programmes remains hidden.

Whole metagenome sequences have enabled us to identify genes and to predict (and possibly confirm) their biological roles through homologous comparisons and to identify novel biocatalysts [16••]. In concert with specific proteomics and transcriptomic information new insights into the biocatalytic potential of novel gene encoding enzymes could be obtained [17–19,20•,21]. In this context, progress has now been made to unravel and understand full bacterial genomes and metagenomes enzyme networks under conditions of substrate feeding [11,22,23•]. For that, new high-throughput sequencing technologies such as the 454 GS FLX (Roche) or the Genome Analyzer (Illumina) make it possible to query the sequence space of an organism in an efficient unbiased manner [24,25]. This technology can be used as a tool for the detection of functional genes coding enzymes of interest [16••].

The Genomes Online Database (GOLD) compiles data related to sequencing projects of (meta-) genomes (last update February of 2010) and the resulting data are available through the IMG/M website (Integrated Microbial Genomes with Microbiome Samples http://img.jgi.doe.gov/cgi-bin/m/main.cgi; Carbohydrate-active Enzymes (CAZy) Database http://www.cazy.org) [26••]. From a total of 239 projects, 134 of these (56%) are associated to environmental communities. Sequencing of these environmental samples has provided valuable insight into the lifestyles and metabolic capabilities of uncultured organisms occupying various environmental niches, such as freshwater (29 projects, 12.1%), marine water (52 projects, 21.7%), soil (26 projects, 10.9%), rhizoplane (7 projects, 2.9%), sediment (12 projects, 5%) and bioremediation of polluted zones (5 projects, 2.1%). Among the 5 projects related to communities affected by recalcitrant compounds, the GOLD database collects information related to studies with terephthalate (1 project), tetrachloroethylene (1 project), benzene (1 project) and chloroethene (2 projects).

By searching through the list of ‘genes with Pfam’ (the protein family database) from every metagenome on the IMG/M website, each group may retrieve enzyme homologues. Recent reports revealed that the number of putative enzymes of potential use in biocatalysis found per million base pairs range from 1.4 to 19, depending on the metagenome samples [27••]. Further, full-length genes can be identified, synthesized and expressed resulting in active enzymes. A good example of the potential of bioinformatics based gene-centric metagenome analysis for mining the sequence space for (un) known enzymes is the analysis of glycosyl hydrolases with potential use as energy vector. About 4874 glycosyl hydrolase homologues from 46 completed metagenome databases have been identified [27••], with an increase abundance in samples that are characterized by a steady input and turnover of complex plant cell wall biomass: the metagenomes from microbial communities derived from rumen, termite, human and mouse guts displayed more putative GHase homologues (approximately 1.5% total gene count) than those from other samples such as soil and water sample (approximately 0.3% total gene count) [28,29•].

The above data confirmed the potential of sequence-based bioinformatics as screening system; however, the sequence information provide serious problem in relation to the quality and degree of completeness of the annotation of metagenomes [12,30•]. Most troublesome are the large numbers of open reading frames that have been