Cowan, D.A., et al. (2004). Metagenomics, gene discovery and the ideal biocatalyst. BIOCHEMISTRY SOCIETY TRANSACTIONS, 32(2): 298-302

Metagenomics, gene discovery and the ideal biocatalyst

D.A. Cowan[1], A. Arslanoglu, S.G. Burton, G.C. Baker, R.A. Cameron, J.J. Smith and Q. Meyer

Institute for Microbial Biotechnology and Metagenomics, Department of Biotechnology, University of the Western Cape, Bellville 7535, Cape Town, South Africa

Abstract

With the rapid development of powerful protein evolution and enzyme-screening technologies, there is a growing belief that optimum conditions for biotransformation processes can be established without the constraints of the properties of the biocatalyst. These technologies can then be applied to find the 'ideal biocatalyst' for the process. In identifying the ideal biocatalyst, the processes of gene discovery and enzyme evolution play major roles. However, in order to expand the pool genes for *in vitro* evolution, new technologies, which circumvent the limitations of microbial culturability, must be applied. These technologies, which currently include metagenomic library screening, gene-specific amplification methods and even full metagenomic sequencing, provide access to a volume of 'sequence space' that is not addressed by traditional screening.

The ideal biocatalyst

The concept of the 'ideal biocatalyst' has grown out of a perceived shift in the paradigm for the selection of biocata-lysts for industrial application [1]. The traditional approach to the implementation of a functional enzyme within the context of an industrial or commercial biocatalytic process has been the selection of an enzyme whose properties were less than ideal for the optimum process. Selection was typically via *de novo* screening or from the existing catalogue of commercially available enzymes. The catalyst would then be further modified using any of a battery of genetic and molecular tools in order to adapt the catalyst to better suit the process requirements. The new paradigm, which suggests that the process could be designed optimally without being constrained by consideration of the properties of the biocatalyst, has evolved from the recent development of new and powerful methods for enzyme redesign and enzyme discovery. The methods allow us to move more effectively through, and to probe more comprehensively, the variations in protein sequence that dictate enzyme functional properties.

Sequence space

Every undergraduate student in biochemistry learns that the hypothetical range of proteins which can be built from the 20 natural amino acids is staggeringly large; 20^{100} for a protein of only 100 amino acids in length. This concept has come to be known as 'sequence space' [2], a multidimensional volume representing all possible polypeptide combinations for a given sequence length. The occupancy of 'sequence space' is dictated by many factors. Most important are thermodynamic requirements for building stable and functional protein structures.

The protein sequences currently known to science from protein sequencing, from the translation of directly cloned genes and from sequenced genomes, total between 10^4 and 10^5. Clearly, these few

sequences occupy a relatively small proportion of the sequence space offered by a fully random selection of amino acids (10^{201} for 200-amino-acid proteins). There are a number of possible reasons why such a small proportion of total sequence space is occupied (Table 1). It is immediately and instinctively obvious that certain regions of protein sequence space are unlikely to be occupied (e.g. a protein of the sequence His200 is unlikely to be found in natural biological systems). Conversely, accumulated protein sequence data indicate that some regions are heavily occupied (e.g. highly conserved proteins such as members of the chaperone family will occupy quite localized regions of sequence space). To the authors' knowledge, no attempts have been made to develop tools for analysing and/or representing the occupied and unoccupied regions of sequence space. Such analyses may be particularly instructive in further developing an understanding of the thermodynamic basis of protein structure and function.

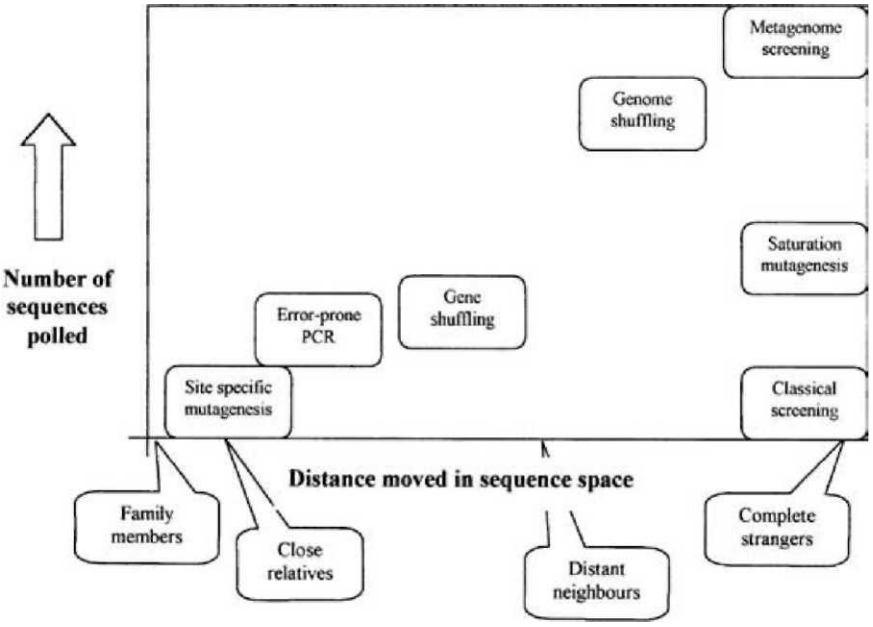
Moving around in sequence space

Our understanding of the occupancy of protein sequence space is derived from the sequences available in protein (and nucleic acid) databases. These sequences are, in turn, dependent on the range of known and available genomes. Developments in protein engineering and evolution have provided methods for moving short distances in sequence space. The 'classical' techniques of random and site-specific mutagenesis, which are used to replace a single or limited number of amino acids, only provide access to near neighbours (Figure 1). More recently developed methods, including error-prone PCR [3], saturation mutagenesis [4], domain shuffling [5] etc., give access to more distant neighbours (Figure 1). With the exception of saturation mutagenesis, the scope of these techniques is nevertheless limited by the range of genes and genomes that can be used as templates or sources of sequence data.

Table 1 Factors dictating occupancy of sequence space

Possible cause	Basis for occupancy/absence
Thermodynamic limitations	Some amino acid sequences could never fulfil the thermodynamic requirements of a functional pprotein
Functional limitations	The functional requirements of certain protein types (e.g. membrane-spanning proteins) reduce the range of amino acids that can be selected
Evolutionary limitations	Evolution from a relatively small population of progenitor sequences may not have had time to fully explore sequence-space options
Evolutionary pressure	The requirements of biological systems may be well satisfied with existing sequences, reducing the evolutionary pressure to explore new areas of sequence space
Database inadequacy	The occupancy of protein sequence space is much wider than currently perceived due to the limited protein sequence information available

Figure 1 Options for moving around in sequence



The extent of this limitation has only become evident in the past decade with the development of small subunit rRNA gene analysis as a tool for the study of molecular ecology, particularly of prokaryotic populations [6]. Molecular phylo-genetic analysis of microbial diversity has clearly demonstrated that a very high proportion of microbial 'species' have never been cultured. Such studies have led to the development of the concept of 'unculturables' [7], currently thought to represent between 90 and 99.9% of extant microbial species, depending on the biotope. Current estimates of prokaryotic and lower eukaryotic diversity may be in excess of 10^7 distinct species [8,9]. Extremophilic, particularly thermo-philic, biotopes may have a higher proportionofunculturable microbial species than many less 'extreme' biotopes.

The presence of such substantial proportions of 'uncultur-able' species in any environmental sample highlights the limitations of any gene discovery protocol that is dependent on culturing, no matter how sophisticated the amplification or selection process. While there is clearly an on-going need for new and innovative culturing technologies, there is also a requirement for alternative gene-/genome-discovery strategies that are culture-independent. These novel methods are now incorporated in the newly evolved field of 'metagenomics'; the culture-independent assessment of microbial ecology.

Metagenomic gene discovery as an option for wider access to sequence space

The application of metagenomics to (culture-independent) gene discovery was initiated with the simultaneous development in several laboratories of direct cloning and expression of multigenomic DNA extracts [8,10-16]. Depending on the choice of vector and host, either single genes and pri¬mary gene products, or secondary metabolites from the expression of complete operons, could be targeted (Table 2).

With the 'maturation' of this technology, efforts have been made to circumvent some of the limitations inherent in the preparation and screening of metagenomic libraries. Meta-genomic PCR amplification methods, which have been used successfully to identify families of homologous genes ([18] and R.A. Cameron and D.A. Cowan, unpublished work), suffer from the limitation that the primary PCR generates only partial gene sequences. Full-length sequences are subsequently obtained by hybridization screening of a complete metagenomics library or by genome walking [19].

It is generally assumed that PCR-based technologies, where known sequences are required for primer design, probe near-neighbour regions of sequence space (Figure 1). Inter¬estingly, our early results do not support this view. In amplifying metagenomic DNA extracts using primer sets specific to bacterial lipase and nitrile hydratase a-subunit sequences, we cloned and sequenced sets of putative gene fragments (Figure 2). While the nitrile hydratase genes show very high homology to each other and known full-length genes (i.e. we did not move far in sequence space), the putative lipase gene amplicons showed very low levels of homology, suggesting that we were probing a much larger volume of sequence space.

Table 2 Vector/host combinations for metagenomic gene targeting
NA, not applicable; BAC, bacterial artificial chromosome.

Vector	*Host*	Average insert size (kb)	Target gene/product	Reference
Lambda phage	*Escherichia coli*	1.8-4.2	Chitinase	[13]
NA	*Streptomyces lividans*	NA	Terragine antibiotics	[14]
BAC	*Escherichia coli*	10-20 and 40-50	DNase, lipase, amylase	[12]
BAC	*Escherichia coli*	10-20 and 40-50	N-Acyk-tyrosine antibiotics	[15]
Cosmid/lambda phage	*Escherichia coli*	>30	Biotin biosynthesis partway	[17]
pBluescript plasmid	*Escherichia coli*	5-8	Lipase	[10]
pCR-TOPO© plasmid	*Escherichia coli*	10-18	Tetrapyrrole biosynthesis genes	[16]

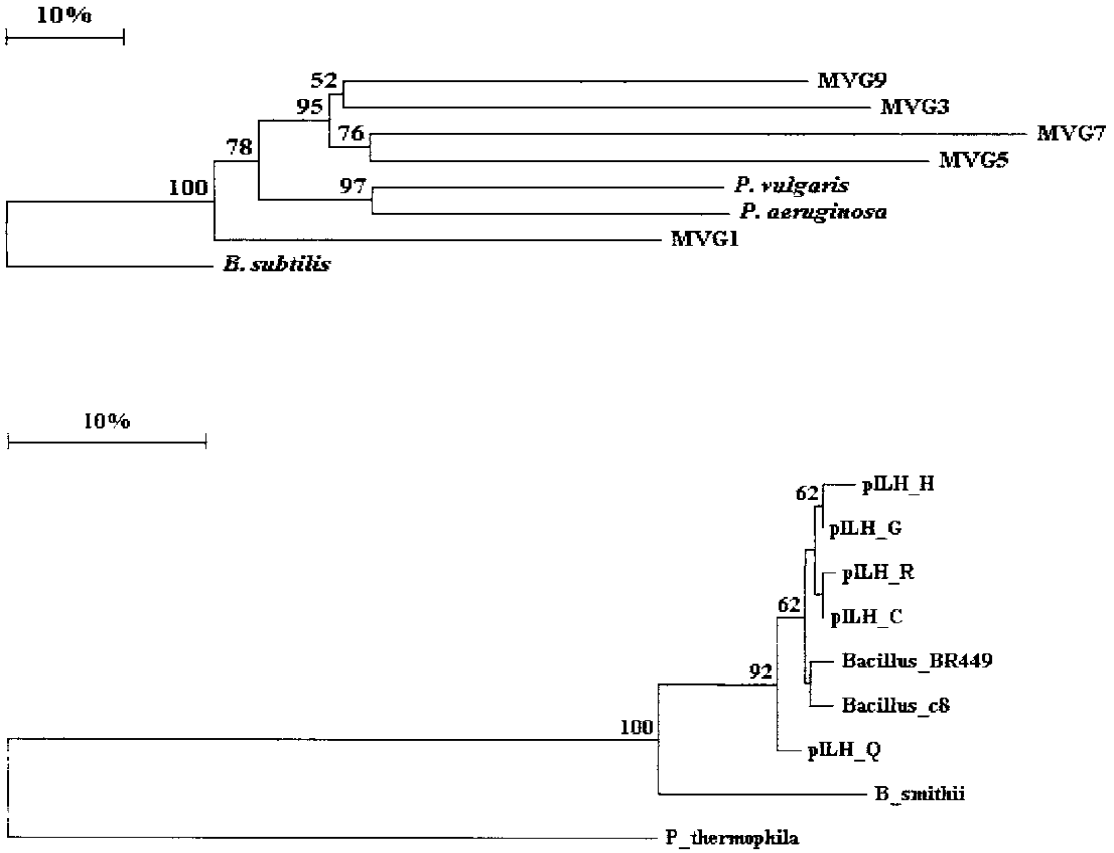
A number of PCR-based approaches designed to recover the flanking regions of a DNA fragment once its sequence is known have been reported (see review in [20]). Although suitable for use at a single-genome level these methods are technically more difficult to apply at the metagenomic level due to the increased complexity of a multigenomic DNA sample. A desire to simplify this process led us to look at the development of other novel approaches.

One potentially powerful approach is based on *in vitro* hybridization of a genomic DNA sample with the target gene fragment acting as a probe (Figure 3). Genomic DNA is fragmented and priming sites are introduced by ligation of adapters. The gene-specific PCR product is then used as a driver to selectively hybridize to full-length gene fragments in the DNA sample. These partially double-stranded full-length gene fragments can then be selectively separated from the single-stranded background (genomic DNA). To remove any residual background the adapters are removed; because the full-length gene fragments are only partially double stranded the priming sites will remain intact as the restriction enzyme can only act on double-stranded DNA within the priming site. The full-length gene can then be amplified. This method is particularly powerful for multigenomic cloning as the use of degenerate gene-specific primers on a metagenomic sample typically yields a population of target gene fragments (R.A. Cameron, J.J. Smith and

D.A. Cowan, unpublished work). Each member of the pool of target fragments can be used to probe the metagenomic sample, potentially yielding a number of novel full-length genes simultaneously.

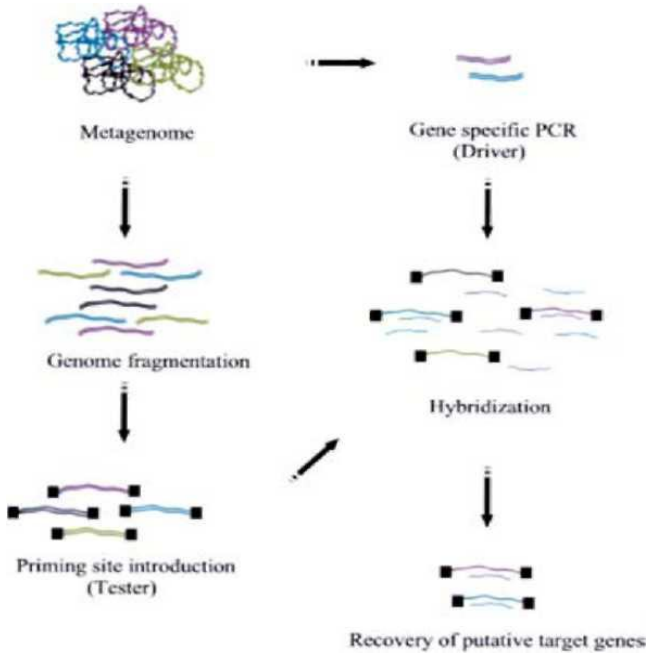
Figure 2 Neighbourhood joining tree representations of partial gene sequences derived from PCR a amplification of metagenomic extracts

(a) Putative lipase sequences derived from Antarctic Dry Valley soil DNA extracts. **(b)** Putative nitrile hydratase



a-subunit sequences derived from geothermal sediment DNA extracts.

Figure 3 Accessing full-length genes using gene-specific PCR products



.**Sequencing the metagenome**

With the rapidly increasing capacity for high-throughput screening, and the associated reduction in sequencing costs, the possibility exists of circumventing all current approaches to metagenomic gene discovery by merely sequencing the entire prokaryotic metagenome. While this is no trivial task, it is worth considering the scale and the likely cost effectiveness of the process.

An 'average' environmental sample will contain hundreds or even thousands of different prokaryotic and eukaryotic species [21-23]. A temperate eutrophic soil DNA extract (containing, conservatively, 500 bacterial 5 Mbp genomes and 100 fungal 20 Mbp genomes) will therefore constitute around 5 Gbp of DNA sequence (rather larger than a human genome). While the sequencing cost would be projected at a few hundreds of millions of dollars (for comparison, the U.S. Human Genome Project budget was predicted to exceed $US450 million), such a metagenome sequencing project would be expected to yield 2-3 million open reading frames, at a projected cost of around $200 per identified open reading frame. While the cost per open reading frame is surprisingly low, this is clearly not an efficient method for mining specific genes. For gene types which occur in limited numbers in each genome (such as protease genes) or only in some genomes (such as lipase genes), the costs could escalate to $10 000100 000 per open reading frame. Even if these costs do not seem unreasonable, the costs associated with the real commercial exploitation of a new enzyme (which includes all aspects of process development and optimization) are very much higher.

Metagenome sequencing becomes more technically and financially feasible when biotopes harbouring relatively low genomic diversity, preferably without eukaryotic genomes, are considered. Hyperthermophilic biotopes would seem to be ideal targets, where species diversity is generally lower than eutrophic neutrophilic environments. In addition, hyperthermophilic biotopes are dominated by small archaeal genomes! We predict that for some biotopes expression library and PCR-dependent

metagenomic gene-discovery methods may be rapidly outpaced by the speed, automation and efficiency of the sequencers of the future.

We gratefully acknowledge funding support from the BBSRC, the South African National Research Foundation, the Claude Harris Leon Foundation, the University of Waikato and Antarctica New Zealand.

References

1 Burton, S., Cowan, D.A. and Woodley, J.M. (2002) **30,** 35-46

2 Babajide, A., Farber, R., Hofacker, I.L., Inman, J., Lapedes, A.S. and Stadler, P.F. (2001*) J. Theor. Biol.* **212,** 35-46

3 Kuipers, O.P. (1996) *Methods Mol. Biol.* **57,** 351-356

4 Reidhaar-Olson, J.F. and Sauer, R.T. (1988) *Science* **241,** 53-57

5 Hiraga, K. and Arnold, F.H. (2003) *J. Mol. Biol.* **330,** 287-296

6 Pace, N.R., Stahl, D.A., Olsen, G.J. and Lane, D.J. (1985) *Am. Soc. Microbiol. News* **51**,4-12

7 Amann, R.I., Ludwig, W. and Schleifer, K. (1995) *Microbiol. Rev.* **59,** 143-169

8 Short, J.M. (1997) *Nat. Biotechnol.* **15,** 1322-1323

9 Bull, A.T., Goodfellow, M. and Slater, J.H. (1992) *Ann. Rev. Microbiol.* **46,** 219-252

10 Henne, A., Daniel, R., Schmitz, R.A. and Gottschalk, G. (1999*) Appl. Environ. Microbiol.* **65,** 3901-3907

11 Rondon, M.R., Raffel, S.J., Goodman, R.M. and Handelsman, J. (1999) *Proc. Natl. Acad. Sci.* U.S.A. **96,** 6451-6455

12 Rondon, M.R., August, P.R., Bettermann, A.D., Brady, S.F., Grossman, T.H., Liles, M.R., Loiacono, K.A., Lynch, B.A., MacNeil, I.A., Minor, C. et al. (2000) *Appl. Env. Microbiol.* **66,** 2541-2547

13 Cottrell, M.T., Moore, J.A. and Kirchman, D.L. (1999) *Appl. Env. Microbiol.***65,** 2553-2557

14 Wang, G., Graziani, E., Waters, B., Pan, W., Li, X., McDermott, J., Meurer, G., Saxena, G., Andersen, R.J. and Davies, J. (2000) *Organic Lett.* **2,** 2401-2404

15 Brady, S.F. and Clardy, J. (2000*) J. Am. Chem. Soc.* **122,** 12903-12904

16 Wilkinson, D.E., Jeanicke, T. and Cowan, D.A. (2002) *Biotechnol. Lett.* **24,** 155-161

17 Entcheva, P., Liebl, W., Johann, A., Hartsch, T. and Streit, W.R. (2001) *Appl. Env. Microbiol.* **67,** 89-99

18 Bell, P.J.L., Sunna, A., Gibbs, M.D., Curach, N.C., Nevalainen, H. and Bergquist, P.L. (2002) *Microbiology* **148,** 2283-2291

19 Morris, D.D., Reeves, R.A., Gibbs, M.D., Saul, D.J. and Bergquist, P.L. (1995) *Appl. Env. Microbiol.* **61,** 2262-2269

20 Ochman, H., Jose Ayala, F. and Hartl, D.L. (1993) *Methods Enzymol.* **218,** 309-321

21 Rondon, M.R., Goodman, R.M. and Handelsman, J. (1999) *Trends Biotechnol.* **17,** 403-409

22 Hill, G.T., Mitkowski, N.A., Aldrich-Wolfe, L., Emele, L.R., Jurkonie, D.D., Ficke, A., Maldonado-Ramirez, S., Lynch, S.T. and Nelson, E.B. (2000) *Appl. Soil Ecol.* **15,** 25-36

23 Ward, B.B. (2002) *Proc. Natl. Acad. Sci.* U.S.A. **99,** 10234-10236