Functional Metagenome Mining: A sequence directed strategy for the retrieval of functional enzymes for biocatalysis*

Jack W. E. Jeffries, Natalie Dawson, Christine Orengo, Thomas S. Moody, Derek J. Quinn, Helen C. Hailes and John M. Ward*

Abstract: Biocatalytic reactions are increasingly being used as a sustainable strategy in organic synthesis and it is recognised that there is need for new enzyme discovery. To establish the utility and versatility of a metagenomics approach, metagenomic DNA extracted from the oral cavity was sequenced and used to create an in silico contig library. This enables individual open reading frames, operons or all the enzymes of a particular family to be identified and then retrieved from the original DNA by PCR. As proof of principle a lactate dehydrogenase, a malate dehydrogenase and transketolases were identified in silico, successfully cloned and assayed. This new enzyme retrieval sequence directed method gives constructive access to metagenomic diversity and importantly improves on the low hit rate experienced when using conventional metagenomic screens.

Over the last 30 years enzymes have increasingly been used in commercial chemical processes.[1] Enzyme catalysts have distinct advantages over chemical catalysts such as cost, sustainability, low toxicities and the use of moderate reaction conditions. The greatest advantage however, is the wide range of catalytic activities displayed by naturally occurring enzymes and the stereoselectivities that can be achieved. Drawbacks when using enzymes can include a low organic solvent tolerance, narrow pH working ranges, and low thermostability. However these properties along with enhanced stereoselectivities can be engineered using random or directed evolution techniques.[2]

One of the key requirements for this type of engineering is multiple amino acid sequences. Having many enzymes that catalyze the same reaction, with variation in their primary sequence is a much more effective starting point for the engineering process than any single example of an enzyme family. One successful method for identifying large numbers of novel enzymes has been to mine fully sequenced and annotated genomes of laboratory cultivable strains. Such sequence directed genome mining lends itself to the retrieval of multiple targets, however a limitation to this method is the number of annotated available species.[3] With only 0.1-5% of bacteria cultivable in the laboratory, much of the existing diversity of bacteria are inaccessible in this way.[4] At the same time, with the advances in high throughput sequencing, whole or partial genomes of uncultivable bacteria are being made available for study. The nascent field of metagenomics gives insight into the genomes of previously unstudied bacteria and by extension a potential wealth of new biocatalysts for the generation of small molecules and large bioactive molecules such as polyketides and glycopeptides.[5]

The use of metagenomics to obtain new enzymes and biocatalysts for industrial applications is steadily growing but still in its infancy.[6] Efforts to capture enzymes from metagenomic samples have relied upon the creation of physical metagenomics libraries in more tractable bacteria.[7] Poor transcription and translation, enzymes with activities that are difficult to screen for all lead to a low “hit rate”, the number of enzymes discovered compared with the number of clones screened, for physical functional libraries. This low “hit rate” is disappointing especially when considering the wealth of genomic diversity available through metagenomics.[3]

At the same time high throughput sequencing and in silico functional annotation can be used to study microbial communities as they exist in vivo. Studies have used this burgeoning amount of sequence data to aid enzyme retrieval. The large numbers of annotated proteins are used to build consensus sequences from which degenerate PCR primers are designed.

Here a strategy is described which uses the large number of individual protein annotations, generated through high throughput sequencing, to design specific primer pairs for the retrieval of enzymes in a way analogous to genome mining. By building an in silico metagenomic contiguous read (contig) library from high throughput sequencing data a resource is created that can be quickly and importantly repeatedly mined for enzymes. Together with conventional molecular biology techniques such as Polymerase Chain Reaction (PCR), DNA cloning and protein overexpression, this strategy will now allow constructive access to the wealth and diversity of enzymes that so far have been out of reach. A comparison of our approach with existing functional metagenomics approaches is summarized in Figure 1.

To test the strategy an in silico metagenomic contig library of the oral cavity was created and formatted into a BLAST database. As a proof of principle that specific sequences identified from the in silico library can be retrieved by PCR, two lactate dehydrogenase (LDH) sequences and a malate dehydrogenase (MDH) sequence were chosen to query the in silico metagenomic contig library. Furthermore five transketolases (TKs) were retrieved as exemplars of industrially relevant enzymes.[3]
The in silico metagenomic contig library was created by sequencing a mixed sample of DNA extracted from the oral cavity (tongue scrapings) of 9 volunteers. Roche 454 sequencing produced 1.1 million reads, which were combined and built into larger contigs using the MIRA assembler. In total 39,971 contigs were generated; the largest was 41.9 kbp with an N50 value of 2416 for the data set. 125,211 Open reading frames (ORFs) and 88,247 protein annotations were marked using metagene mark and the Pfam standatlon tool respectively. This contig library was formatted into a BLAST database.

Two LDH amino acid sequences were chosen from Uniprot with which to query the contig library; one each from Rothia mucalaginoso (Uniprot id: D2NTCO) and Streptomyces parasanguinis (Uniprot id: I1ZLP7), species known to have DNA present in the metagenomic sample. Also chosen was the sequence of a MDH from Streptomyces venezuelae (Uniprot id: F2RK50) an organism that does not reside in the mouth but has members of the same genus in the mouth and data set. These enzymes were chosen due to the likelihood of them existing in the dataset and straightforward activity assays.

A BLAST search of the library with the LDH sequences from both Rothia and Streptococcus identified the same 19 contigs. In addition to these 19 contigs, 2 contigs were exclusive to the search with Rothia and 2 found only when querying with Streptococcus. When querying with the MDH sequence from Streptomyces 4 further contigs were identified not matching any found before. In total 27 contigs where identified; 23 found when searching with an LDH sequence, 10 containing full length enzyme sequences. From the 4 contigs found searching with the MDH sequence, only 2 contained full length non redundant sequences.

From this subset of 12 full length enzyme sequences, three were selected as proof of principle to take forward for PCR from the original DNA used to generate the in silico library. Two enzymes were chosen with the highest similarity (>95% identity) to the LDH query sequences from Rothia and Streptococcus to examine the accuracy of the retrieval from BLAST sequence, through contig to retrieved protein. A third enzyme sequence with only 46% identity to the MDH from Streptomyces was also chosen. The query sequence from Streptomyces venezuelae and a contig containing a low similarity sequence were deliberately chosen to test the limits of the BLAST search approach, to see if low similarity sequences identified through BLAST searching retained the activity of the original query sequence.

All three enzymes were retrieved through PCR and were expressed in Escherichia coli: Rothia 2665 (LDH) and Streptomyces 3443 (MDH) were well expressed and could be successfully purified. Streptococcus 5169 (LDH) was less well expressed and perhaps for this reason could not be purified. Initial rates of reaction for the purified enzymes for varying substrate concentrations were plotted against the change in substrate concentration. Values for the $V_{\text{max}}$, $K_{\text{cat}}$ and $K_{\text{cat}}/K_{\text{m}}$ for both enzymes were calculated using Michaelis-Menten kinetics and cooperative and inhibitory equations for the LDH and MDH respectively (Table 1). The oligomeric state of the LDH and MDH were inferred as a homotetramer and homodimer and the kinetic data calculated based on these assumptions. The data was consistent with existing studies on enzymes from related bacteria and fit with the assumption of multimeric states for the enzymes. The DNA of the enzymes was sequenced and compared to the original query sequences. The enzymes Rothia 2665 and Streptococcus 5169 had >98% similarity to the sequences used for BLAST. Streptomyces 3443 showed 62% identity to the original query sequence, an increase in percentage identity due to the relatively poorer quality of the initial contig sequence. Full alignments of the DNA and amino acid sequences for all enzymes are given in supplementary information.

Table 1. Characterization of the LDH and MDH enzymes

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>$V_{\text{cat}}$ µkat.mg$^{-1}$</th>
<th>$K_{\text{cat}}$ mM</th>
<th>$K_{\text{cat}}$ s$^{-1}$</th>
<th>$K_{\text{cat}}/K_{\text{m}}$ M$^{-1}$ s$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDH 2665 Michae</td>
<td>42.8 (±3.7)</td>
<td>9.6 (±2.6)</td>
<td>1.2 x10$^6$</td>
<td>1.3 x10$^8$</td>
</tr>
<tr>
<td>LDH 2665 coop.</td>
<td>32.9 (±0.9)</td>
<td>7.4 (±0.4)</td>
<td>9.5 x10$^5$</td>
<td>1.3 x10$^8$</td>
</tr>
<tr>
<td>A. viscos LDH</td>
<td>35.0 (±0.7)</td>
<td>0.8</td>
<td>8.8 x10$^6$</td>
<td>1.0 x10$^9$</td>
</tr>
<tr>
<td>MDH 3443 Michae</td>
<td>16.2 (±0.7)</td>
<td>0.6 (±0.1)</td>
<td>2.2 x10$^7$</td>
<td>3.6 x10$^{7}$</td>
</tr>
</tbody>
</table>
An alternative method of enzyme identification was used when searching for transketolase (TK) enzymes. Contigs containing TKs were identified using the Pfam ids for TK C-terminal, N-terminal and pyrimidine binding domains. In total 73 contigs were identified with at least one of the Pfam ids; 12 of the contigs contained full length enzymes. Seven of the 12 were non redundant from which primers were designed for retrieval, 5 of which were successfully retrieved. All 5 of the amino acid sequences had >90% identity to sequences deposited in the NCBI database. None of the sequences in the database have been cloned and assayed having been identified via homology from sequencing data. One of the metagenomics TK, identified as having 90% similarity to a sequence from Atopobium parvulum, had an internal stop codon and appeared to be a split domain TK. Sequencing of the ORF product maintained this internal stop codon and SDS gel of the purified enzyme showed two distinct bands corresponding to the computed molecular weight of the two domains Figure 2.

Characterisation of purified enzymes at 25 °C following the decrease in absorbance at 340 nm as NADH is oxidised and substrate is reduced. [a] The reaction catalysed by LDH 2665 reducing pyruvate to lactate. [b] Graph showing initial velocities for LDH 2665 plotted against substrate concentration fit with the Michaelis Menton equation (dashed line), and the Hill equation for co-operativity (solid line). [c] The reaction catalysed by MDH 3443 reducing oxaloacetate to malate. [d] Graph showing initial velocities for MDH 3443 plotted against substrate concentration fit with the Michaelis Menton equation (dashed line) and an equation for substrate inhibition (solid line). [e] Biochemical characteristics for both enzymes calculated from the assay data. Values are also shown of enzymes from closely related species, errors for the values are given in parentheses where available.

Table 2. Purification and kinetic values for Transketolase

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>$V_{max}$ µkatol.mg$^{-1}$</th>
<th>$K_m$ mM</th>
<th>$K_{cat}$ s$^{-1}$</th>
<th>$K_{cat}/K_m$ M$^{-1}$.s$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TK1</td>
<td>0.13 ±(0.8x10$^{-1}$)</td>
<td>3.3 (±0.47)</td>
<td>213</td>
<td>6.5 x10$^4$</td>
</tr>
<tr>
<td>TK2</td>
<td>0.44 ±(±1.3x10$^{-1}$)</td>
<td>6.5 (±0.77)</td>
<td>1269</td>
<td>1.9 x10$^5$</td>
</tr>
<tr>
<td>E.coli</td>
<td>0.13</td>
<td>35</td>
<td>391</td>
<td>1.1 x10$^4$</td>
</tr>
</tbody>
</table>

Characterization of 2 purified TKs following the production of erythulose from the reaction of β-hydroxy pyruvate with increasing concentrations of glycoaldehyde. All kinetics calculated based on the formation of erythulose [a] The reaction catalysed by the TK enzyme where a two carbon ketol unit is transferred from β-hydroxy pyruvate irreversibly to glycoaldehyde to produce erythulose and CO$_2$. [b] Graph showing initial velocities for TK1 plotted against substrate concentration fitted with the Michaelis Menton equation. [c] Graph showing initial velocities for TK2 plotted against substrate concentration fitted with the Michaelis Menton equation. Biochemical characteristics for both enzymes calculated from the assay data compared to a purified E.coli TK. Errors are given in parenthesis.

Physical functional metagenomics libraries are a powerful tool for finding novel enzymatic activities however they can have a low “hit rate”, due in part to the failure of host systems to transcribe and translate heterologous DNA. In our strategy DNA is first sequenced and analysed in silico, then single open reading frames or larger segments of DNA are amplified and can be placed in the correct orientation in front of a strong promoter/transcription system. While E.coli is a tractable and well studied system, not all sequences retrieved from the metagenome will express well in this host. However because our method provides isolated DNA sequences they can be cloned into multiple expression vectors and expressed in different host systems much more easily than larger Fosmid or BAC libraries. Metagenomic clone libraries rely on visualizable outputs for enzyme identification; enzymes which are not readily visualized in some manner are unobtainable using physical strategies. Using our in silico approach all types of enzymes can be
identified before any physical manipulation of the DNA takes place. After PCR capture and cloning individual enzyme activity can be followed using the whole range of spectrophotometric and mass spectroscopy techniques as well as colorimetric assays. To date research in retrieving enzymes from metagenomes via PCR have relied upon degenerate primers from consensus sequences to retrieve enzymes. The research presented here uses primers designed for specific non redundant sequences from a specific sequenced metagenome in an effort to increase the number and variety of sequences retrieved. Furthermore using a sequence directed approach gives insight beyond any single open reading frame. The *in silico* approach may also aid in the characterization of an enzyme by providing contextual information about substrate specificity from the annotation of neighboring genes especially within operons.

The *in silico* library from which the enzymes where identified was created from a relatively small data set using technology that has now been surpassed in terms of base pair out put and read quality. Sequencing of other environmental samples, at larger sequence depth with current technologies, and the creation of multiple *in silico* libraries would allow fast, productive access to the huge enzymatic diversity of chemically/biologically interesting niches. Having shown it is possible to identify and retrieve targeted active enzymes from a metagenomic sample using sequence directed, specific primers, the next step is to retrieve larger numbers of commercially relevant enzymes from multiple environmental samples.

**Experimental Section**

LDH and MDH Enzymes were assayed at varying concentration of substrate either sodium pyruvate or oxaloacetic acid (Sigma) respectively. NADH concentration was kept constant at 1 mM. The reaction was followed by the decrease in absorbance at 340 nm as NADH was oxidised to NAD+ concomitantly with reduction of the substrate. 80 µl of a 1.25 mM NADH solution initiated the reaction of 10 µl of enzyme with 10 µl of substrate, total reaction volume 100 µl. NADH and substrates were prepared in 0.5 M Hepes buffer pH 7.5. Assays were run in 96 well plates in triplicate at 25 °C, pH 7.5 using a fluorostar Optima plate reader. Zero enzyme controls were set up as well as an NADH standard across a concentration range of 0.1-1 mM. Substrate range for LDH 1.0, 2.5, 5.0, 7.5, 10.0, 12.5, 15.0, 20.0, 25.0, 30.0, 35.0 mM of pyruvate. Substrate range for MDH 0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0, 3.0, 5.0, 10.0, 15.0, 20.0 mM of oxaloacetic acid. An LDH enzyme concentration of 0.126 mg/ml and MDH of 0.293 mg/ml was ascertained by absorbance at 280 nm. The molecular weight of 144.88 kDa and 78.3 kDa were calculated using the Bradford assay and the TK contribution calculated using densitometry of protein SDS gels. Purified transketolase enzymes were assayed on a smaller scale than the lysate. 10 µl of purified enzyme was incubated for 5 min with 10 µl of 48 mM ThDP and 180 mM MgCl2. After incubation for 5 min 90 µl of 110 mM β-hydroxy pyruvate and 90 µl of glycolaldehyde at a range of concentrations, 1 - 45 mM, was added to initiate the reaction. Reactions were run at 25 °C in triplicate for 5 min with 50 µl samples being taken and quenched in 250 µl of a 0.5% TFA solution. Production of erythrulose was followed by HPLC as described above. Enzyme concentrations of 3.9 and 2.5 mg/ml for TK1 and TK2 respectively were calculated by absorbance at 280 nm. Molecular weight and extinction coefficients for the TKs were calculated using the EXPASY protoparm tool, a combined molecular weight of 63.7 kDa for TK1 and 72.8 kDa for TK2 were computed.

**Keywords:** Functional Metagenomics • Polymerase Chain Reaction • Enzymes • Biocatalysis


A combination of high throughput sequencing and bioinformatic tools are used to generate an in silico metagenomic library. Such a library can grant fast, productive access to the wealth of enzyme diversity promised by metagenomics.