Using synthetic biology to screen for functional diversity of GH1 enzymes

Sam Deutsch[1], Supratim Datta[2], Matthew Hamilton[1], Greg Friedland[2], Patrik D'Haeseleer[2], Jan-Fang Chen[1], Dylan Chivian[2], Rob Egan[1], Kenneth Sale[2], Blake Simmons[2], Eddy Rubin[1]

[1]DOE Joint Genome Institute, Walnut Creek, CA, USA [2] The Joint BioEnergy Institute, Emeryville, CA,USA

Introduction

Advances in next-generation sequencing technologies have enabled single genomes as well as complex environmental samples (metagenomes) to be comprehensively sequenced on a routine basis. Bioinformatics analysis of the resulting sequencing data reveals a continually expanding catalogue of predicted proteins (~ 14 million as of April 2011), 75% of which are associated with functional annotation (COG, Pfam, Enzyme, Kegg, etc). These predicted proteins cover the full spectrum of known pathways and functional activities, including many novel biocatalysts that are expected to significantly contribute to the development of 'clean technologies' including biomass degradation, lipid transformation for biodiesel generation, intermediates for polymer production, carbon capture, and bioremediation (**Figure 1**).

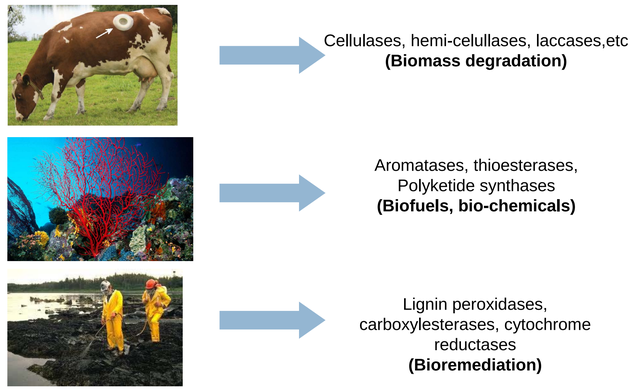

Figure 1: Metagenomics sequencing data is a rich source of novel biocatalysts

Translating sequencing data (digital information) into biochemical molecules that can be assayed for function remains challenging and represents a major bottleneck for fully exploiting the expanding gene catalogue.
Synthetic biology methods, including the *in-vitro* synthesis of genes in a template-independent manner, bridge the gap between digital information and biochemical characterization.

The aim of the project is to develop high-throughput and low-cost methods for gene synthesis and to apply this technology for the synthesis and characterization of genes involved in biomass degradation, starting with ~200 GH1 enzymes.
This project is part of the LBNL Carbon Cycle 2.0 initiative.

Specific Goals

- Development and validation of novel gene-synthesis pipeline.
- Discovery of novel GH1 enzymes.
- Biochemical characterization of enzymes (substrate specificity, thermostability, salt-resistance)
- Generation of a dataset for Structure-Activity Relationship (SAR)
- Optimization of GH1s for cocktail for different industrial settings.

.

Development of new synthesis technologies

The high cost of current DNA synthesis approaches, based on standard phosphoramidite oligonucleotide synthesis chemistry, has precluded the use of synthetic genes in large-scale functional characterization studies.
The recent development of 'programmable mask-less array' (Nimblegen) and '*in-situ* nucleotide inkjet printing' (Agilent, Figure 2) technologies, which direct the parallel synthesis of thousands of short oligonucleotides onto a solid support, enable substantial efficiency improvements and cost reduction.
We have developed a protocol for full-length gene synthesis based on the use of 'ink-jet' oligo arrays. In addition, we have developed and validated protocols for for performing clone screening for synthesis validation using the PacBio sequencing technology.

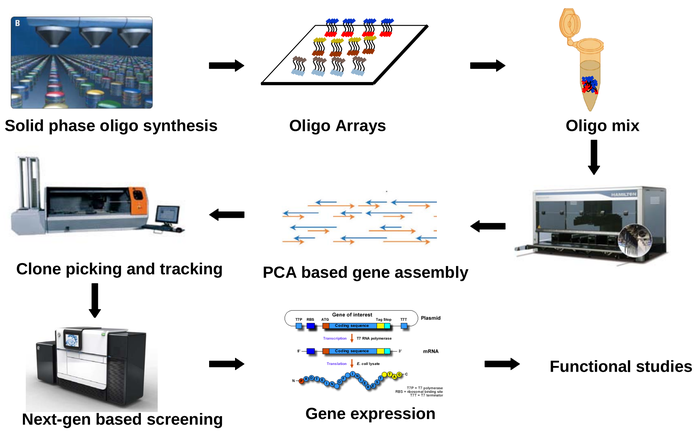

Figure 2: Development of gene-synthesis pipeline

Systematic survey of GH1 enzymes

GH1 enzymes participate in the last step of biomass breakdown whereby cellobiose (or cellotriose) is converted to glucose (Figure 3A).
To survey the diversity of GH1 enzymes in nature we searched the NR and CAZy databases for matches against the pfam domain 00232. We identified 2319 predicted GH1 proteins which could be grouped into ~1000 using a 90% homology filter.
We built a multiple sequence alignment and phylogenetic tree (Figure 3B) of all predicted proteins and applied the MaxPD algorithm to select 200 representatives that cover the maximum sequence space. This set was selected for synthesis and characterization.

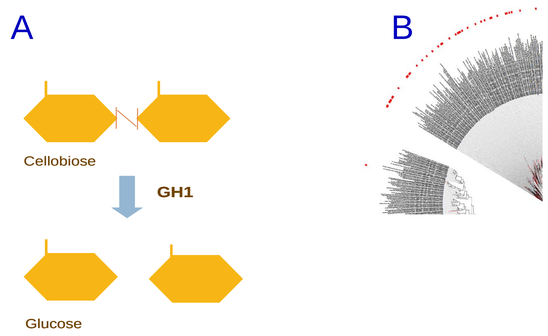

Figure 3: Selection of GH1 candidate genes. A. Shows main GH1 activity. B. Shows phylogenetic tree of all known GH1 enzymes. Selected proteins are marked by red outer dots.

Synthesis and Biochemical Characterization

Out of the 200 enzymes that were selected for characterization, 25 have been synthesized to date (Figure 4A). These genes were cloned into *E. coli* expression vectors, and protein expression was induced. Most predicted enzymes showed some level of expression in the soluble fraction (Figure 4B). Enzymatic activity was then measured over a range of temperatures (Figure 4C) and pH values (figure 4D). Many enzymes were active over a range of conditions.

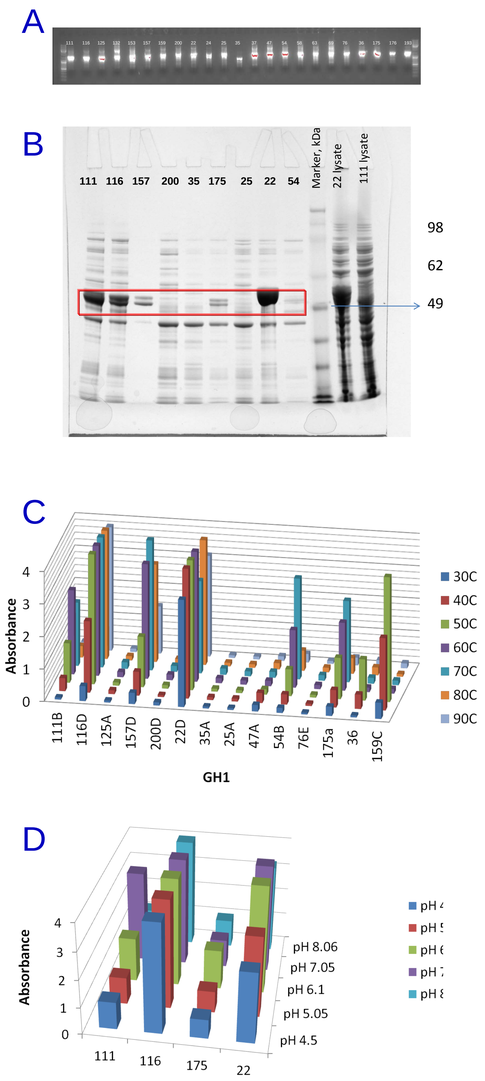

Figure 4: Gene synthesis and biochemical characterization. A. Gene-synthesis B. Protein expression C. Temperature profile D. pH profile

Conclusions

- We developed a pipeline for high-throughput gene- synthesis incorporating new technologies such as 'ink-jet' oligo printing and PacBio sequencing technology.
- We selected a set of highly diverse predicted GH1 enzymes for functional characterization.
- Initial data has shown good levels for protein expression, solubility and activity over a range of conditions.
- The fact that diverse temperature and pH optima have been observed validates the phylogenetics driven approach.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

DISCLAIMER