Identification of novel enzymes from soil metagenome

A review of the underlying procedure of enzyme discovery from metagenomic big data
Culture-dependent studies

Most of current knowledge on microorganisms based on pure cultivation of a few culturable microbes in artificial and naturally unreal media.

This method has resulted in a noticeable gap in our understanding since it eliminates the pivotal role of ecological relationships; hence, a great proportion of microorganisms are not culturable.
Culture-dependent vs. Culture-independent

Microbial communities are diverse, complex and powerful.

Only one gram of soil contains up to 4000 different species, however, current estimates indicate that less than 1% of these organisms are readily culturable with known methods.

This means that by classical studies, we are probably missing 99% of diversity of these communities.
Metagenomics

Also known as environmental and community genomics, is the genomic analysis of microbes by direct extraction and cloning of DNA from a microbial community.

1. **Whole metagenomics**
   a. Whole genome shotgun sequencing (WGS)

2. **Targeted metagenomics**
   a. Sequence based approach
   b. Function-based analysis
Targeted metagenomics

Isolating DNA from environmental sample
Cloning the DNA into a suitable vector
Transforming the clones into a host bacteria
Screening the resulting transforms

1. Sequence - driven analysis
2. Function – driven analysis
Construction of complex libraries derived from soil

Is dependent on **high-quality of the extracted soil DNA**, as the enzymatic modifications required during the cloning steps are sensitive to contamination.

The DNA isolation methods for soils are based either on the:

1. **Recovery of bacterial cells and subsequent lysis**
2. **Direct lysis of cells in the sample followed by DNA purification**.
Sample
Natural soils

Metagenomic libraries
Small-insert libraries (plasmids)

Screening
Activity-based

Sequencing and characterization

Enrichments
Large-insert libraries (cosmids, BACs)

Sequence-based

Current Opinion in Biotechnology
Function-based analysis

Its primary purpose is screening for enzymatic activities

1. isolating DNA from microbial communities
2. Cloning DNA fragments
3. Expressing genes in a surrogate host
4. Screening for enzymatic activities
5. Sequencing
6. Data analysis
Function Based mining for a Biomolecule

In this approach we screen cloned surrogate hosts based on a specific expression activity profile.

One drawback is its too much dependency on host’s expression of cloned foreign gene.

By this method we can search for presence of a certain metabolite in the media.

Thus far, many enzymes and other classes of proteins have been successfully discovered by this pipeline.
Sequence-based analysis

To retrieve a specific gene from a pool of DNA by PCR or hybridization.

Primers are designed specifically against an identified target gene.
Computational models have approved that sequential similarity between enzymes correlates with functional similarities.

This makes the amino acid sequence of proteins a reasonable target for finding a specific protein coding gene in the metagenomic data.
The availability of metagenomic big data, gives us the ability to explore these sequences for our product of interest.

Moreover, due to time and resource limitations and the fact that these data are being produced with an increased pace, makes it prohibitively impossible to screen probably thousands of candidate genes during wet lab experiments.
BLAST as a screening tool

BLAST is a popular algorithm which explores among entries or gene/protein sequences for the most sequential similarity. Sometimes, BLAST can be a fair estimation of the general attributes of a sequences. For example, general functionality, EC number,…

*But is that enough?*
What gives us a more precise prediction?

Although primary protein sequence can be a good indicator for some of its features, numerous attributes originate in more complex protein structures.

For some predictions, we might need a plethora of sequence based features such as:

AA, 2AA, 3AA compositions, Correlation, Hydrophobicity, Transition, Distribution, Evolutionary data, ...
This is where machine learning comes to the rescue!

Machine learning is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead.

It is seen as a subset of artificial intelligence.

*We can make sense of big, noisy, non-interpretable (by human!) data using machine learning techniques.*

**These methods have been used successfully to:**

**Predict** Enzymes’ optimum temperature & pH, EC,

to **find** Anti microbial peptides, Biofilm inhibitors,

To **Detect** special functionalities...
**Case Study:** Mining Thermo-active Xylanase in metagenome

**ML algorithms were used to classify xylanase enzymes based in their optimum temperature.**

Various metagenome data including soil, camel rumen, ... have been fed to the model and thermo and hyperthermo-active xylanase sequences were predicted with accuracy of 94%.

Some of candidates have been cloned and experimentally tested. experimental results approved model’s predictions.
Case Study: Metagenome-Wide Association Study and Machine Learning Prediction of Bulk Soil Microbiome and Crop Productivity

Bulk soil samples collected from a high and a low productivity area. Extracted DNA from these bulk soil samples were shotgun sequenced. A major proportion of the taxa variance in the bulk soil microbiome. Metagenome-wide association studies (MWAS) identified more *Bradyrhizodium* and *Gammaproteobacteria* in higher productivity areas and more *Actinobacteria, Ascomycota, Planctomycetales*, and *Streptophyta* in lower productivity areas. Successfully predicted productivity based on the microbiome composition with the best accuracy of 0.79 at the order level. Our study showed that crop productivity differences were associated with bulk soil microbiome composition and highlighted several nitrogen utility-related taxa.
Case Study: Metagenome-Wide Association Study and Machine Learning Prediction of Bulk Soil Microbiome and Crop Productivity