Introduction to Bioinformatics Project 2 - Identification of Xylanase Enzyme Genes from the Rumen Metagenome of Ruminants

University of Tehran, Faculty of Sciences, School of Computer Science
Asal Rabiee

Introduction

Xylan molecules, through their linkage with lignin and cellulose, play an important role in maintaining the integrity of plant cell fibers and cell walls, and are one of the main obstacles in lignocellulose degradation. Lignocellulose is an abundant and low-cost renewable carbon source, composed mainly of cellulose, hemicellulose, and lignin, forming a strong structure through covalent and non-covalent bonds.

Xylanase is a key auxiliary enzyme in lignocellulose hydrolysis. By hydrolyzing xylan and disrupting its fibers, xylanase creates pores that loosen the structure and facilitate lignocellulose degradation. This property makes xylanase valuable in industry. Its use can effectively reduce production costs of biofuels and improve the yield of reducing sugars, which can be used as bioenergy. Therefore, studying xylanase for lignocellulose degradation is of great importance, as understanding the mechanisms by which xylanase acts provides a foundation for biofuel production [1].

Beyond its role in pulp and paper industries, xylanase is also used as a feed additive in poultry nutrition and for improving the nutritional properties of silage. Thus, isolating and cloning xylanase genes are crucial steps in microbial engineering. Microbial xylanases have been applied in various industries such as animal feed, pharmaceuticals, paper, baking, and waste treatment [2].

Action of Xylanase on Xylan Fibers

Research on xylanase currently focuses on screening high-yield strains, improving enzyme properties, and optimizing culture conditions. However, obtaining high-yield, low-cost xylanase remains the bottleneck for industrial production.

The ruminant rumen hosts a diverse community of microorganisms including bacteria, fungi, archaea, and protozoa. Natural selection and long-term adaptation have established dynamic interactions between ruminants and their rumen microbiota. These microbes, by secreting various lignocellulolytic enzymes, act as natural bioreactors for lignocellulose degradation.

Compared with xylanases from other sources, rumen-derived xylanases show superior properties and greater development potential [1]. The gut microbiota is a diverse ecosystem that plays a crucial role in maintaining host physiology and immune function. Gut bacteria ferment food to produce beneficial metabolites, influencing host health through ecological interactions. Microbiota also depend on other community members to acquire nutrients such as carbohydrates and amino acids. For example, *Bacteroides ovatus* and *Lactobacillus paracasei* secrete enzymes that break down polysaccharides like inulin and xylan into monosaccharides, which are then used by *Faecalibacterium prausnitzii* and *Bacteroides vulgatus* to produce short-chain fatty acids.

Xylanase can be secreted by many bacteria, including *Bacillus*, *Streptomyces*, and *Ruminococcus*. By hydrolyzing xylan into oligosaccharides, it alters gut microbiota composition. However, the exact relationship between xylanase-driven microbial shifts and host health remains unclear [4].

Rumen-derived xylanases play a crucial role in lignocellulose degradation, highlighting the rumen microbiome as a valuable genetic reservoir for discovering effective xylanases [3]. Discovering novel xylanases from this rich source can significantly benefit industries reliant on biomass degradation. Moreover, characterizing new xylanases is attracting attention as it could help address global challenges in food and energy supply.

In this study, the main objective was to identify the most probable candidate sequences for xylanase enzymes from the ruminant rumen metagenome using bioinformatics tools. Results indicated similarity between selected contigs and thermostable xylanases.

Methods

Before starting, group members conducted a literature review to better understand xylanase activity and project goals. Findings were shared and a roadmap was drawn. Early challenges included choosing appropriate tools and preparing required data, which were resolved through group discussions.

Step 1 – Identification of Candidate Sequences

- Downloaded the assembled rumen metagenome contigs and 11 known thermostable xylanase sequences.
- Installed and ran **BLAST+** on an Ubuntu virtual machine.
- Compared contigs against known thermostable xylanases.
- Set similarity threshold using E-value ≤ 1e-5.
- Filtered results accordingly.
- Extracted contig IDs and sequences of filtered hits into FASTA files.
- Translated filtered nucleotide sequences into protein sequences.

Step 2 – Clustering and Representative Selection

- To reduce redundancy, used **CD-HIT** clustering with 97% similarity threshold.
- From each cluster, the longest sequence was chosen as representative.
- Generated two outputs: clustered contigs (.clstr, .fasta) and representative (non-redundant) sequences (.fasta).

Step 3 – Conserved Region Modeling and Filtering

- Performed multiple sequence alignment using **Clustal Omega**.
- Built an HMM model of conserved regions with **HMMER**.
- Used the HMM to filter representative sequences.
- Retained sequences with the strongest match as high-confidence xylanase candidates.

Commands:

Part 1 – Candidate Sequence Identification (BLAST+)

```
makeblastdb –in metagenome.fasta –dbtype nucl –out metagenme_db

tblastn –query xylanase_reference.fasta –db metagenome_db –out blast_results.txt –outfmt 6
```

```
awk '{if($11 < 1e-5 && $3 > 50) print $0}' tblastn_results.txt > filtered_results.txt

cut –f2 tblastn_results.txt > selected_hits.txt

seqtk subseq metagenome_contigs.fasta selected_hits.txt > potential_xylanase.fasta

transeq –sequence extracted_sequences.fasta –outseq translated_sequences.fasta
```

Part 2 – Clustering & Representative Selection (CD-HIT)

```
cd-hit –i potential_xylanase.fasta –o clustered_xylanase –c 0.97 –n 5

grep ">Cluster" –A 1 clustered_xylanase.clstr | grep –v ">Cluster" | awk –F'[>. ]' '{print $3}' >
longest_sequence_ids.txt

seqtk subseq potential_xylanase.fasta longest_sequence_ids.txt > longest_representatives.fasta
```

Part 3 – Conserved Region Modeling & Filtering (Clustal Omega + HMMER)

```
clustalo –i longest_representatives.fasta –o xylanase_aligned.fasta --auto

hmmbuild xylanase.hmm xylanase_aligned.fasta

hmmsearch --tblut filtered_xylanase.tsv xylanase.hmm translated_sequences.fasta
```

Results

Step 1 – BLAST+

- Five main output files were generated.
- Filtered contigs showed significant similarity with thermostable xylanases.
- Protein translations of these candidate sequences were obtained.

Step 2 – Clustering

- Generated .clstr files listing contig IDs per cluster.
- FASTA files containing clusters and representative sequences were produced.
- Representatives were typically the longest sequence in each cluster.

Step 3 – HMM Modeling

- Multiple sequence alignment performed with Clustal Omega.
- HMM profile generated from aligned sequences.
- Filtered .tsv output file containing candidate sequences that matched the model.

Discussion

The results highlight the power of bioinformatics tools in analyzing biological data. For decades, researchers have sought new enzymes like xylanases to better understand metabolism of various substrates

in digestive systems [5]. Studying xylanase genes and their conserved regions brings us closer to understanding their function and industrial applications.

With bioinformatics, the challenge of analyzing vast amounts of biological data can be managed effectively. The main difficulty now lies in identifying optimal tools and algorithms for minimizing error.

As attention grows on mammalian gut microbiomes, metagenomic studies are becoming increasingly valuable. Since recent research shows the microbiome's role in neural and immune system function [6], rumen metagenome studies may have applications not only in industry but also in medicine and psychology.

Conclusion

Xylanase, produced by the rumen microbiome of many ruminants, has gained increasing attention due to its wide applications in lignocellulose-related industries.

In this project:

- Candidate thermostable xylanase-like sequences were extracted from the rumen metagenome.
- These sequences were translated into proteins.
- They were clustered and representatives were selected.
- Conserved regions were modeled using an HMM, enabling identification of strong xylanase gene candidates.

This project demonstrates how bioinformatics tools can advance genetic engineering studies and support industrial applications. It also provided students with teamwork experience and insight into the intersection of computer science and life sciences.

Such approaches may ultimately lead to deeper understanding of genetic and biological complexity.

References

1. Ruyue Dong et al. Effectiveness of ruminal xylanase with an extra proline-rich C-terminus on lignocellulosic biomass degradation. *Bioresource Technology*, 2023.
2. Ariaeenejad, S. et al. Identification and characterization of a novel thermostable xylanase from camel rumen metagenome. *Int J Biol Macromol*, 2019.
3. Zhang, M. et al. Characterization of novel multifunctional xylanase from rumen metagenome. *Fermentation*, 2024.
4. Wang, T. et al. Xylanase enhances gut microbiota-derived butyrate to exert immune-protective effects. *Microbiome*, 2024.
5. Christel Mattéotti et al. Identification and characterization of a new xylanase from Gram-positive bacteria isolated from termite gut. *Protein Expression and Purification*, 2012.
6. Toader, C. et al. Mind, Mood and Microbiota—Gut–Brain Axis in Psychiatric Disorders. *Int. J. Mol. Sci.*, 2024.