Genome Dynamics in Coevolved Genomes: Database Management System for Tracing Mutations

Sahith Sai Nareddy[*]
Computing and Software Systems
School of STEM
University of Washington Bothell
Bothell, WA 98011-8246, USA
sahithn@uw.edu

Erik Westover[†]
Computing and Software Systems
School of STEM
University of Washington Bothell
Bothell, WA 98011-8246, USA
erik.e.westover@comcast.net

Kristina Hillesland[‡]
Division of Biological Sciences
School of STEM
University of Washington Bothell
Bothell, WA 98011-8246, USA
hilleskl@uw.edu

Wooyoung Kim[§]
Computing and Software Systems
School of STEM
University of Washington Bothell
Bothell, WA 98011-8246, USA
kimw6@uwb.edu

ABSTRACT

We investigate coevolution of two microorganisms in terms of the adaptations and genome dynamics over a thousand of generations. A number of coevolved populations are generated in a lab, and preliminary results about the increased growth rate and density of evolved populations have proven that their interactions are mutually beneficial. In order to understand the underlying genotypic changes, it is important to examine the rates of new mutations in the original and evolved populations. We expect to identify beneficial mutations through computational analyses using the whole-genome DNA sequencing data which are already available in the lab. Initially, Breseq genomic sequencing tool is used to identify mutations. However, the simple results of 'predicted mutations' are not enough to identify real mutations, nor distinguish beneficial mutations from neural or deleterious mutations. Therefore, in this paper, we introduce a database management system (DBMS) that can rigorously trace the mutations over different generations by having the ability to update and retrieve data. The DBMS supports a query language and produces reports in a structural way. We also plan to build a web-based interface to increase the usability of the DBMS in near future.

[*]Corresponding author. Undergraduate Student
[†]Data Analyst
[‡]Co-advisor
[§]Corresponding author. Advisor

BCB'14, September 20–23, 2014, Newport Beach, CA, USA.
ACM 978-1-4503-2894-4/14/09.
http://dx.doi.org/10.1145/2649387.2660810.

Categories and Subject Descriptors

H.2.4 [**Information Systems**]: Systems—*Relational databases*

General Terms

Design

Keywords

DBMS; NGS; Mutation; Coevolution

1. INTRODUCTION

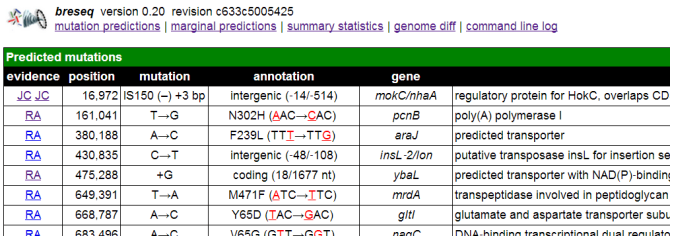
Coevolution is a phenomenon where two or more organisms interact with each other and results in unique adaptations, and gradually increasing rate of evolutionary change. These co-evolutionary changes have been mainly examined from antagonistic interactions, such as, those of prey and predators. However, mutually beneficial interactions have been also investigated as in the article [3], where the bacterium *Desulfovibrio vulgaris* and the archaeon *Methanococcus maripaludis* are paired into a number of different environments to examine an experimentally imposed mutualism. This mutualism is based on the exchange of hydrogen, which is produced by *D. vulgaris* and consumed by *M. maripaludis*. That is, *D. vulgaris* provides *M. maripaludis* with food, and *M. maripaludis* provides *D. vulgaris* with a permissive environment by keeping the toxic level of hydrogen low so that D. vulgaris can survive in return. Preliminary experimental results show that evolved cocultures grew up to 80% faster and were up to 30% more productive.

The experiment is extended to obtain more generations in more diverse environments. This is an ongoing collaboration project between the division of Biology and the division of Computing and Software Systems at the University of Washington Bothell. Currently more than 30 DNA genome data sequencing data are available and processed. Through this project, it is expected that the relationship between the two species will result in beneficial mutations. Therefore, the long term goal will be identifying what types of genetic changes accumulate in evolving populations over time that lead to beneficial mutualism.

To analyze the raw sequencing data, we need to align the sequences with known sequences, find base-pairs that have been altered, inserted, or deleted, and use computational and biological tests to distinguish true mutations from sequencing errors. In our initial computational analyses, we picked the Breseq genome sequencing analysis tool [2] to accomplish these tasks, as Breseq implemented a pipeline that is well customized to analyze the evolutionary changes, from read alignments to mutation predictions.

2. PRELIMINARY WORK

Breseq program generates a series of processed results in different folders, such as, BAM/SAM format files, which are the standard format for aligned sequences, gd files, for mutations, and web pages to show the predicted mutations. Figure 1 is an example output generated by Breseq, where main results are organized into 'predicted mutations' and 'marginal mutations' with alignment evidences and frequency of that mutation in population.



breseq version 0.20 revision c633c5005425

mutation predictions | marginal predictions | summary statistics | genome diff | command line log

Predicted mutations

evidence	position	mutation	annotation	gene	
JC JC	16,972	IS150 (–) +3 bp	intergenic (-14/-514)	*mokC/nhaA*	regulatory protein for HokC, overlaps CD
RA	161,041	T→G	N302H (AAC→CAC)	*pcnB*	poly(A) polymerase I
RA	380,188	A→C	F239L (TTT→TTG)	*araJ*	predicted transporter
RA	430,835	C→T	intergenic (-48/-108)	*insL-2/lon*	putative transposase insL for insertion se
RA	475,288	+G	coding (18/1677 nt)	*ybaL*	predicted transporter with NAD(P)-bindin
RA	649,391	T→A	M471F (ATC→TTC)	*mrdA*	transpeptidase involved in peptidoglycan
RA	668,787	A→C	Y65D (TAC→GAC)	*gltI*	glutamate and aspartate transporter subu
RA	683,496	A→C	V65G (GTT→GGT)	*nagC*	DNA-binding transcriptional dual regulato

Figure 1: A Breseq output example available in *http : //barricklab.org/*

In order to trace generic changes in evolved genomic sequences, however, the predicted mutations should be evaluated further. We need to distinguish real mutations from falsely predicted mutations, which might involves rigorous biological experiments through a number of time points. If the data involves many repeated regions, additional sequencing analysis tool, such as, BLAST [1] should be applied to selected regions. All these post-processes lead to the need of organizing the processed results in a structured way. Therefore, in this paper, we present a database management system (DBMS) that is specifically designed for structuring the mutations of evolved genomes using Breseq output files.

3. DBMS FOR COEVOLUTION

In this project, we will obtain more NGS data and will process computational analysis to predict beneficial mutations. Therefore, the ability to update and retrieve data is essential. With the irreplaceable raw sequencing data, and computationally expensive output data, we also need to recover data effectively. To meet all the requirements, we designed and implemented a database management system for this coevolution. We designed the database schema using *Microsoft Visio*, which we used to construct and visualize the Entity Relation Diagram, or ERD as shown in Figure 2. After creating the visualization, we created queries in Microsoft SQL to construct the tables which constituted the database. We created a script to parse mutation information from the 'annotated.gd' file, which was tab delimited for easy traversal, to actually put the data first into an excel file and into the database.

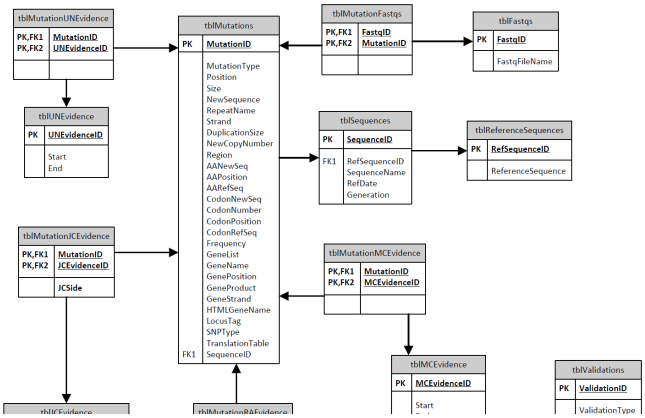


Figure 2: Partial view of relational database schema

4. FUTURE STUDY

Future plans include creating a web interface for the database using .asp net. The web interface will display commonly accessed information to the user and allow the user to view additional data through the interface. Another feature that could be added is the ability to download data in Microsoft Excel format. In the long term, various genome sequencing alignment tools are easily utilized and the results are easily manageable with the DBMS.

5. REFERENCES

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, Oct. 1990.

[2] D. Deatherage and J. Barrick. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. In L. Sun and W. Shou, editors, *Engineering and Analyzing Multicellular Systems*, volume 1151 of *Methods in Molecular Biology*, pages 165–188. Springer New York, 2014.

[3] K. L. Hillesland and D. A. Stahl. Rapid evolution of stability and productivity at the origin of a microbial mutualism. *Proceedings of the National Academy of Sciences*, 107(5):2124–2129, 2010.