Title:
Enhancing the SEED Framework for Curation and Analysis of Genomic Data and Genome-scale metabolic Models

Christopher Henry*[1] *(chenry@mcs.anl.gov),* ***Ross Overbeek***[2]*, Robert Olson*[1]*, Terrence Disz*[1]*, Bruce Parello*[2]*, Rob Edwards*[3]*, Matt DeJongh*[4]*, Aaron Best*[4]*, Fangfang Xia*[1]*, Scott Devoid*[1]*, and* ***Rick Stevens***[1]

[1]*University of Chicago, IL;* [2]*Fellowship for Interpretation of Genomes, IL;* [3]*San Diego State University, CA;* [4]*Hope College, MI*

Project Goals (1000 characters):

The objective of the SEED Knowledgebase project is encompassed by four goals: (1) enhance the computational infrastructure behind the SEED framework to improve extensibility, accessibility, and scalability; (2) integrate extensions into the SEED database to accommodate new computational and experimental data types including regulatory networks, biochemistry and models, eukaryotic genome data, and growth phenotype data; (3) develop web services providing access to all SEED data and algorithms including genome annotation, model reconstruction, flux balance analysis, and data query and access; and (4) apply the SEED framework to annotate and model organisms with applications to bioenergy, carbon cycle, and bioremediation. This project will provide users with programmatic access to SEED data and algorithms, it will produce new models of bioenergy organisms, and it will enable integration of expression data and regulation with annotations and models.

Abstract (limit 2 pages)

The SEED environment for the integration, annotation, and comparison of genomic data now includes thousands of microbial genomes and many eukaryotic genomes, all linked with a constantly updated set of curated annotations embodied in a large and growing collection of encoded subsystems and derived protein families. Additionally, the Model SEED resource has been developed to translate SEED annotations into functioning genome-scale metabolic models. In the SEED Knowledgebase project, we have endeavored to enhance the SEED environment in three areas: (1) development of a web API offering programmatic access to SEED data and algorithms (including trees and expression data); (2) extension of the SEED interface to enable curation of public genome annotations by registered users; and (3) reconstruction and analysis of metabolic models for all available prokaryotic genome sequences.

1. Web API for Programmatic Access to Data and Algorithms

We have developed a set of web-services for SEED that offer programmatic access to data and tools included within the SEED environment (find documentation at http://blog.theseed.org/servers/). The services include the ability to remotely submit genomes to RAST and Model SEED for annotation and modeling; enabling users to

query the SEED database for genome features, functional annotations, gene orthologs, and sequence similarities; and enabling users to apply flux balance analysis with genome-scale metabolic models to simulate cell growth in a variety of media conditions and with a variety of mutations. We highlight the powerful features of the SEED web services by demonstrating how multiple functions may be combined together to answer important questions in biology. Specifically, we apply the web services to: (1) identify new genome annotations based on model gapfilling; (2) identify commonly clustered and co-expressed sets of functional roles across all known genomes; and (3) to study redundancy of essential metabolic functions across all known genomes.

2. Extension of the SEED to Enable Curation of Public Genomes

We recently launched a new version of the SEED website called the Public SEED (pubseed.theseed.org). This new site now contains over 3000 annotated prokaryotic genomes, and it is continuously updated to include every complete prokaryotic genome sequence that is available in GenBank. In addition to offering a greatly enhanced database of genomes, the Public SEED also offers improved access to genomes, through a powerful new search feature. This feature enables users to rapidly search Public SEED content for organism names, gene names, locus IDs, and many other queries and returns results sorted by object type. The Public SEED also offers a unique ability enabling registered users to alter any genome annotation in the Public SEED, providing a unique resource for the community annotation of genomic data.

3. Reconstruction and Analysis of Metabolic Models for all Prokaryotic Genomes

Over the past decade, genome-scale metabolic models have emerged as a valuable resource for generating predictions of global organism behavior based on the sequence of nucleotides in the genome. These models can accurately predict essential genes, organism phenotypes, organism response to mutation, and metabolic engineering strategies. We have applied the Model SEED framework to produce draft metabolic models for over 3000 microbial genomes, representing nearly all complete microbial genomes currently available in GenBank. New algorithms were developed for the gap-filling of these models to enable the activation of every possible reaction in the models and improve the automated generation of biomass composition reactions. These algorithms were applied to assess the quality of annotations in the SEED framework, and to identify high-priority gaps to filled in these annotations. Finally, we applied SEED tools to identify gene candidates that may be associated with the gap-filled reactions. This work reveals insights into the diversity of microbial genomes, the completeness of our knowledge of these genomes, and the areas of our knowledge where more gaps presently exist.