A Peer-to-Peer Environment for Annotation of Genomes: The SEED

An open source tool is being designed, developed, disseminated, and used to help scientists manage and understand the growing body of genomic data

Ross Overbeek, Terry Disz, Rick Stevens

A genome may be thought of as a set of genes that encode protein sequences. The function of each gene is determined by the activity of the protein it encodes. Genome annotation is the process of assigning functions to genes. Functions are assigned by any of several methods. The most direct form of function assignment involves determining the function of a gene by experiment. Since vastly more gene sequences are available in genome databases than the number with directly determined functions, most genes are assigned a function by indirect methods. These methods include assigning a function to a gene based on sequence similarity to genes with known function, assigning a function to a gene based on its position in a conserved gene cluster through comparative analysis of many genomes, and inferring function via other techniques for detection of functional coupling. Genome annotation is an iterative process that can exploit a variety of domain knowledge sources (see Figure 1). For genes that code for enzymes involved in core metabolism, much is known about the biochemical reaction networks in which the enzymes participate. The existence of a known reaction pathway (such as those available in biochemistry databases) can provide valuable information that supports inference of function through processes of systematic elimination. We believe this approach will be highly valuable, especially in comparison to simple similarity methods (which are often unreliable, particularly in the case of paralogous genes, genes that have common ancestry but have evolved divergent functions).

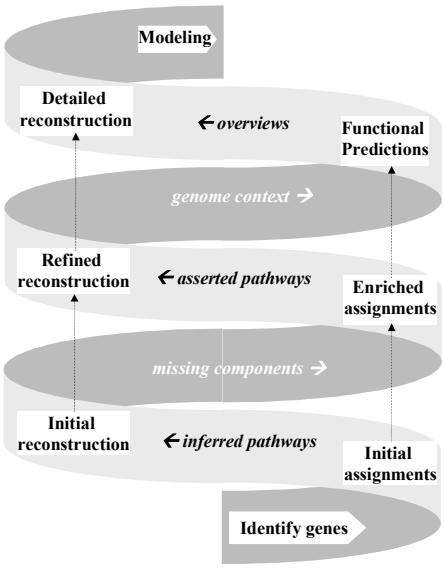

Figure 1. The iterative process of genome annotation.

Good access to and straightforward navigation of annotated genomes is critical for scientists who wish to use genomic data in biomedical research. Scientists studying the structure and function of genes require access to a broad integration of verified, publicly available data (including newly sequenced genomes) in a computing environment that supports comparison of genomes. Additionally, scientists often wish to include analysis of genomes that have not yet been added to the public integrations, but few resources today permit integration of data from public sources with data not yet in public resources. Environments for genome comparison and curation typically either include only final public data or are not generally available for use by the scientific community because they are proprietary tools created by commercial firms.

The individual researcher today simply does not have access to an effective framework for comparative analysis of genomes that can incorporate publicly available genomes, non public data to which the researcher has access (but under restrictions or license terms), and the researcher's own private data. We have created and are distributing a suite of tools, called the SEED, that meets this need and makes the creation and maintenance of new integrations of genomic data feasible by taking advantage of the collective capabilities of many distributed researchers.

The objective of the SEED project is to provide to the biology community a suite of open source tools that will enable distributed teams of researchers to rapidly annotate new genomes, particularly microbial genomes. The initial goal for the SEED is to support rapid annotation of the first 1,000 genomes that are sequenced. In particular, The SEED supports genome annotation and enables researchers to create, collect, and maintain sets of gene annotations organized by groups of related biological and biochemical functions across many organisms. These groups of related functions are called *subsystems*, and each subsystem is a set of biological functions that together implement a specific process. For any organism that is believed to contain a variant of the process, the subsystem includes a designation of precisely which genes (if any) implement each of the functions. This organizational approach enables biologists using the SEED to rapidly project function assignments to sets of genes in newly sequenced organisms. This subsystem approach is unique among genome annotation systems. Indeed, most existing annotation systems support annotation of one genome at a time (achieving improvements in speed by accelerating the rate at which each genome is annotated), while the SEED supports annotation of a single subsystem over hundreds of organisms simultaneously (i.e., annotating one subsystem at a time).

The SEED, still being developed and expanded, will allow users to easily examine the way a given gene relates to other genes, in order to expose the clues relevant to determination of function. It will enable:

- rapid location and visualization of clusters of genes that are relevant to the analysis of a specific subsystem,
- identification of similar genes from other organisms,

- visualization of a neighborhood on the chromosome,
- comparison of one neighborhood of genes with other neighborhoods around corresponding genes in other genomes,
- examination of genes that implement closely related metabolic functions,
- addition of and updating of function assignments and annotations,
- detection of inconsistent representation of function and
- development of packages of assignments and annotations corresponding to a single subsystem.

The ability to study each gene as it relates to all of these critical sources of clues to function is what distinguishes integrations from simple collections of genomic data. As a result, the SEED should dramatically reduce the effort and cost required to construct genome integrations.

The SEED supports distributed workflow, thus enabling teams to share draft subsystem annotations and their corresponding gene function assignments in a peer-to-peer fashion. thus supporting a natural non-hierarchical work style. This non- hierarchical work style is reflected in the two sources of the name SEED. One is FIG SEED, a play on the initials of the Fellowship for the Interpretation of Genomes, the company that developed the initial SEED. The other source is a passage in Neil Stephenson's novel The Diamond Age, which refers to the non hierarchical nature of self-reproducing nanotechnology (in other words, biological systems) that do not require a centralized infrastructure to function.

The SEED's organization is designed to support the various work styles of biologists who may collaborate in collections of loosely coordinated distributed teams or in tightly coupled teams located at a single institution. Each SEED instance is a self-contained genome annotation system that permits multiple users to access, update, and extend the annotation database via a Web-based user interface (local developers also have access to a rich API and programming shell in PERL and Python). The design of the SEED ensures that each user has at hand on a local machine all the data and tools required to do annotations. To support distributed teams, we have developed a basic peer-to-peer synchronization facility that supports sharing information between ad hoc collections of SEED installations. In particular the SEED supports the ability to:

- install and share new genomes between servers,
- share subsystem definitions and associated gene function assignments,
- share gene function assignments (these can be selectively installed),
- share gene annotations (notes, pathway diagrams, etc.), and
- share naming rules (sets of function translations)

SEED development includes three major software projects: development of scalable software infrastructure to support the construction of large-scale integrations of biological data (with genomic data being the primary initial target), the construction of a peer-to-peer software infrastructure to enable the distributed curation of integrated databases and synchronization of the integrations, and the development of software technology to

enable extensions of the integrated database environment to allow for broad adoption and adaptation to new biological applications. The SEED is an open source system, and it is being cooperatively developed by researchers from a number of institutions worldwide. The SEED architecture supports the distribution of:

- SEED source code, including rapid exchange of updates and new features,
- gene function assignments and translation rules, and
- complete subsystem annotation packages.

There are currently 30 SEED installations. These SEED distributions contain the RefSeq data distributed by National Center for Biotechnology Information, along with numerous genomes that have not yet been deposited in the public archives. The version in use now contains 295 complete or near-complete genomes and almost 700 partial genomes. The SEED offers access to the following basic object types: (1) genome – one or more contigs associated with an organism; (2) contig – contiguous DNA sequence; (3) feature – region of DNA sequence that has some properties associated (CDS, RNA, etc.); (4) protein sequence – amino acid sequence; (5) functional role – text description of the role of a protein or DNA sequence; and (6) subsystem – collections of functional roles that are related in some way. This information is organized by genome but is searchable by many different indices, such as annotation, pathway, homology, and subsystem context. It also contains subsystem maps with links to the SEED or external databases. Since different researchers sometimes use different terminology to indicate the same functionality, the SEED also maintains a set of translation rules to convert between various notations. The SEED data exchange model is outlined in Figure 2.

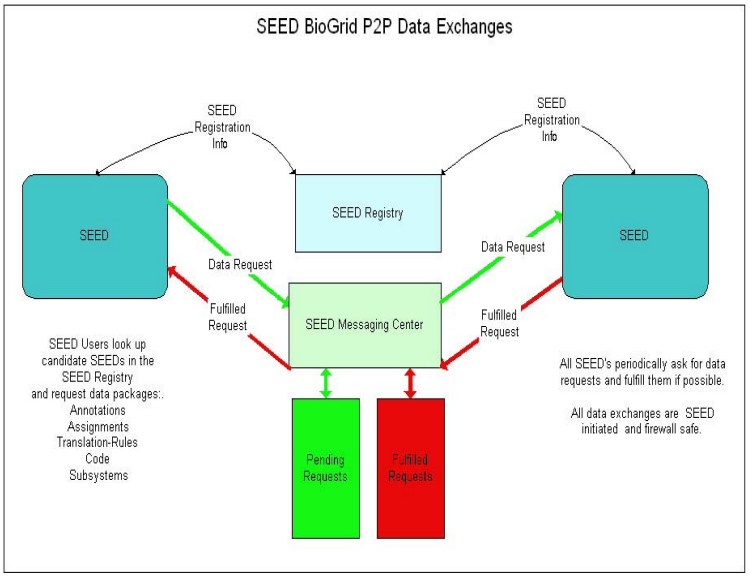

Figure 2. The SEED data exchange model

Our commitment to a peer-to-peer protocol as part of the SEED design has many positive implications. Since each researcher has a complete copy of the code and database and works privately, he is free to include any data, from any source, restricted or otherwise,

for personal use. The common notion of peer-to-peer communications is one of anonymous users participating as both clients and servers in a common search space; queries are launched into the space and, when one is satisfied, two peers communicate directly. The SEED has somewhat different requirements. Instead of an anonymous group, annotation teams of SEED users are well known to each other and, in fact, desire to carefully circumscribe the membership. Data exchanges are not based on generic queries; instead, queries for synchronization are directed at a specific SEED instance and user. To facilitate this model, the system architecture has a Seed Rendezvous Point where a distributed annotation team can meet in cyberspace and exchange data. The first version of the Seed Rendezvous Point is implemented as a simple Web service at a well-known address and is really just a place to register presence and act as a third-party store-and-forward facility to enable communication in the face of local firewall issues.

The architecture for future versions of the SEED will be similar in principle but more robust in terms of security, general data exchanges, and collaboration capabilities. Using a peer-to-peer model allows distributed annotation teams to be formed where each member works in an area of expertise and the team, and only the team, can share that work. By controlling access to data items, researchers using restricted data can still participate in the exchange of data with larger annotation teams while conducting their own research.
These capabilities allow independent groups to extend the functionality of the SEED, either cooperatively or by introducing branches in code releases. It relieves any dependence on central repositories, although de facto archives may emerge. The intended output produced by users of the SEED is a set of annotated subsystems. Each annotated subsystem is a collection of functional roles, along with exactly which genes implement those roles in the existing set of sequenced genomes.

As the number of available genomes increases, effective high-throughput annotation will be based on developing consistent annotations of specific subsystems across hundreds of genomes simultaneously. A user of the SEED with knowledge of a specific subsystem can develop a detailed understanding of exactly what variants of the subsystem exist, which functional roles make up each variant, and which variant applies to each of the sequenced organisms.

The SEED represents an ambitious effort to stimulate and support education, research, and collaboration relating to the analysis of genomic data. The project's primary objective is to dramatically improve the ability of biologists to construct and propagate large-scale integrations containing hundreds of genomes, expression data, metabolic data, and other forms of biological data needed to support the analysis of organisms. We anticipate that the successful development of the SEED will present numerous benefits to the scientific community. The most notable will be enabling analysis of newly sequenced genomes within the context of comparative data; such analysis will inevitably lead to far superior initial characterizations. The SEED leverages a network communications design and a philosophical functional design that incorporate the notion of communicating groups of peers. This design is in keeping with current directions in network services and at the same time facilitates collaboration and cooperation in the biological community,

including the active participation of students and researchers at institutions without a large information technology infrastructure.

Acknowledgements

We thank Mike Papka for his help with the special issue. This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract W-31-109-ENG-38.

Ross Overbeek (ross@thefig.info) is co-founder of the Fellowship for Interpretation of Genomes and is a senior member of the Mathematics and Computer Science Division at Argonne National Laboratory.

Terry Disz (disz@mcs.anl.gov, http://www.mcs.anl.gov/~disz/) co-directs the collaborative tools effort in the Futures Lab of Argonne National Laboratory's Mathematics and Computer Science Division.

Rick Stevens (stevens@mcs.anl.gov; http://www.mcs.anl.gov/~stevens/) is division director of the Mathematics and Computer Science Division, Argonne National Laboratory and professor of computer science at the University of Chicago.