ANL/MCS/CP--91847 CONF-9507246--9

The Role of Integrated Databases In Microbial Genome Sequence Analysis and Metabolic Reconstruction

Terry Gaasterland Natalia Maltsev Ross Overbeek

Argonne National Laboratory
Mathematics and Computer Science Division
Argonne, IL 60439
gaasterland, maltsev, overbeek@mcs.anl.gov

We have built a dual integrated database environment based on the notion of interconnected objects with properties that exist as derived views on top of existing data repositories. The search engine in the integrated environment answers queries that travel from one type of data to another, constraining data objects on the way. The WWW based version of the system, called PUMA is intended as an environment to support the interpretation and presentation of genomes. It is currently being used actively to support the presentation of the Methanococcus jannaschii genome, which is being sequenced by TIGR and the Sulfolobus sulfotaricus genome, which is being sequenced at the Canadian National Research Council Institute for Marine Biosciences. Our goal is to build a suitable framework to support effective access and use of the data for whole genomes in the context of other relevant data. More than merely being a navigation tool for molecular biological data, PUMA serves as the delivery vehicle for the MAGPIE (Multipurpose Automated Genome Project Investigation Environment) automated sequence interpretation system.

The PUMA system provides access to data about metabolic pathways, enzymes, compounds, organisms, encoded activity and assay condition information for enzymes in particular organisms, and multiple sequence alignments. The integrated data objects include the following:

- >700 encoded metabolic pathways (from the EMP Enzyme and Metabolic Pathway database). For example, see Glycolysis.
- the compounds in the 700 encoded metabolic pathways (from EMP). For example, see phospho`enol`pyruvate.
- >865 enzymes (from EMP). For example, see 2.7.1.11.
- >2700 organisms in a general phylogenetic tree built from alignments of their 16S rRNA sequences (from the ribosomal database project).
- >4500 multiple sequence alignments of all protein sequences with their phylogenetic trees (in collaboration with Randy Smith, Baylor College of Medicine, nonredundant PIR and SwissProt). For example, see 448.29 or see A3176 for an alternative format.
- the PIR and SwissProt protein sequences.
- the 80 organisms for which there are more than 50 known protein sequences in the public domain (computed from PIR and SwissProt). For example, see ORGANISM INDEX .

QUERYING DATA OBJECTS

Data objects are interconnected in two-way relationships. For example, enzymes are ``in" metabolic pathways; pathways contain enzymes; pathways contain compounds; organisms have pathways; enzymes have sequences; sequences appear in alignments; alignments have trees; instances of enzymes have assay conditions; and so on. Operators over the data objects provide tools for inserting new sequences in the alignments, pattern matching algorithms, sequence extraction algorithms, and the usual database operations (e.g. sets, counting, bags).

Perhaps the easiest way to describe how the integrated system can be used to answer queries in this integrated domain is to consider several examples.

EXAMPLE 1:

Suppose one wanted to know amino acid patterns that uniquely identify enzymes with particular ranges of activity values. One would pose the following query:

- Start with instances of enzymes that have activities within the specified ranges and that operate on particular compounds.
- Find the enzyme numbers of those enzymes - for each enzyme number, find the set of alignments that contain a sequence with that number
- For each alignment, obtain a ``pattern" that uniquely identifies the alignment

EXAMPLE 2:

Suppose one wanted a set of close organisms that grow under particular conditions and contain particular metabolic pathways:

- Start with an organism with the desired growth properties.
- Inspect the general phylogenetic tree to find neighboring organisms with the same growth properties.
- Take the (possibly proper) subset that contain the metabolic pathway of interest (based on substrate and product of the entire pathway --- without caring about the intermediate compounds).

Now, suppose one wanted a list of all the enzymes involved in those pathways in those organisms:

- For each pathway in each organism, gather its reaction equations (specific to the organism) and get the associated enzyme number.

Next, suppose one wanted the associated alignments and their patterns. Also request that any generality in an alignment's pattern be grounded with specific data from sequences in organisms close to the organisms of interest:

- For each enzyme number, gather the associated alignments.
- For each alignment pattern, find the protein sequence whose organism is closest to the set of organisms of interest.

```
    o For any "disjunction" (e.g. ``valine or isoleucine"), pick
      the particular amino acid that appears in this closest
      sequence.
```

Both of these examples show ways that one would navigate through the space of connected data objects to find alignments that are associated with enzymes and organisms with properties of interest.

CONNECTIVITY

The key to the navigation is the richness of connectivity that is provided by the metabolic pathways, the phylogenetic tree, and the multiple sequence alignments. Connections are maintained as relational tables holding links between objects. For example, the following excerpt links EC numbers with the anabolic pathway in which it appears and indicates whether or not sequences for the enzyme appear in a multiple sequence alignment.

```
EC      AAC.MPW 2OGGLUNADPH.ANA 1.4.1.13        y
EC      AAC.MPW 2OGLYS2OG.ANA   1.1.1.155       n
EC      AAC.MPW 2OGLYS2OG.ANA   1.2.1.31        n
EC      AAC.MPW 2OGLYS2OG.ANA   1.5.1.10        n
EC      AAC.MPW 2OGLYS2OG.ANA   1.5.1.8         n
```

Tables such as this are constructed off-line (via Perl scripts that inspect objects) and updated regularly. Separate tables give the local and/or remote locations of objects and accessible attributes. Our experience is that the data changes slowly enough that off-line compilation of connections is sufficient.

FUNCTIONAL OVERVIEWS OF ORGANISMS

We use the connectivity tables together with a general outline of metabolic functions to derive specialized functional overviews for individual organisms. For organisms with a substantial number of sequenced proteins, the specialized functional outline provides an emerging picture of what is known to date about the organisms metabolism. The connectivity inherent in PUMA provides access to the organism's known metabolism. Rather than us describing an overview, we ask you to click below to access the automatically generated functional overview of E. coli:

Functional Overview for Escherichia coli.

An overview is generated by connecting keywords to entries in the general functional overview (when appropriate, keywords include EC numbers). The SwissProt descriptions for the set of known protein sequences in the organism are then matched against the keywords. A match constitutes a connection between the function and the sequence. Any function without a connected sequence is eliminated. Any sequence without a connected function is put into a miscellaneous set (the general function overview is still under construction --- currently we are focusing on archae functions).

IMPLEMENTATION

A set-based query environment based on the same object+connections+attributes paradigm called GenoBase currently runs for E. coli data. GenoBase traverses a series of object types and provides sets of objects as answers to queries. GenoBase runs locally at Argonne with a TCL interface.

World-Wide-Web access to the integrated environment (PUMA) also traverses series of object types, but answers are provided one object at a time. The data objects described above are currently supported in the WWW query environment. Work is underway to load the objects into the GenoBase query environment. To check out the system yourself, Click here.