DOE/ER/61848

Construction of an Integrated Database to Support Genomic Sequence Analysis

Patrick Gillevet and Ross Overbeek

GenoBase Developments

One central goal of our effort is to develop an integrated database to support comparative analysis of genomes. We now call this logic-programming-based system GenoBase (the previous acronym had tobe changed because another project had already used it). In Phase I of the current proposal, the goal was to produce an initial integration of DNA sequence data, protein sequence data, available data on expression of genes within Escherichia coli (from the Eco2dbase project), and currently available data on metabolism. In fact, we have achieved a somewhat broader integration of available data, in large part because of the assistance from collaborators at NIH, George Mason University, members of the DBEMP project in Russia, and researchers at the Swedish Institute of Computer Science.

The central goal of an integration following the architecture that we have proposed is to make a wide variety of biological data available through convenient access for users. Two issues need to be directly addressed:

1. It must be possible to easily include new forms of data as they become available. For example, during the period since the original proposal was written, substantial amounts of data in the form of the Blocks database, the DBEMP data relating to metabolism, and newly developed phylogenetic trees have become available, and all are directly relevant to interpretation of sequence data. Anyone familiar with the effort normally required to integrate diverse categories of data, especially if a commitment is made to cast the data in a relational form (which we do not), will realize that most commonly-used technologies require substantial resources. We have explicitly attempted to develop a technology that reduces the cost of accessing and operating data, without expending the resources required to achieve a completely consistent, normalized representation of the diverse data items.

2. It must be possible to easily navigate through the ensemble of objects described within the database. In this respect, our effort is based on the same intellectual foundations that similar object-oriented systems utilize. Most of those systems have focused directly on producing extremely interactive, GUI-based navigation systems. Ours has focused on a complementary issue -- effective operation on the ensemble of data (rather than just display and maintenance of objects). We feel that this is an important capability and that we have unique resources at our disposal to address this issue. For example, it should be possible to rapidly answer questions like

   "What patterns occur an unusually large number of times in upstream regions of genes expressed under heat shock?"

   "Given a new class of promoters (such as those recently described in Science for the E. coli genome), what genes include instances in their -40 to -60 upstream regions?"

   "Given a pathway under study, which of the enzymes in the pathway correspond to known protein sequences? Do any of these sequences have a known crystal structure? Which of the sequences correspond to known Block Groups?"

DISCLAIMER

DISCLAIMER

To effectively answer such questions, one requires not only the ability to navigate between collections of diverse objects, but also to be able to apply a rich set of operators to extract the relevant information.

The current GenoBase system has been enhanced as follows:

1. We have worked on increasing our ability to handle larger volumes of data within the flexible framework demonstrated on our earlier prototypes. This effort has culminated in our use of a standard set of database routines developed by the Swedish Institute of Computer Science and supported under the Sicstus Prolog Project at SICS. We have created a single large database at NIH that includes all of the data from EMBL, the Swiss Protein Data Bank, the Enzyme Data Bank, and some limited data on metabolism from the DBEMP. Our intent is to work on developing a stable system including a richer set of objects, and then to make the data available on an object-server accessible through the World Wide Web. This effort has been done in collaboration with Ron Taylor at NIH and is progressing rapidly.

2. We have worked closely with E. Selkov of the DBEMP to encode his data on metabolism into objects that can be used to integrate many forms of data in existing databases. Indeed, the wealth of data provided within the DBEMP will almost certainly play a central role in any effective integration in the future, and our initial efforts should be of use to many international groups in their efforts to achieve higher levels of integration.

3. We have developed a number of Windows-based user interfaces that allow relatively convenient access to the objects within GenoBase. The first effort was based on technology developed within the GDE effort and allowed us to explore generalized navigation tools, as well as a framework for applying operators and collecting the generated results (Appendix A). A second effort has built upon this experience and has produced an initial interface based on TCL/TK. We plan on making this version available at a number of sites; it will be used to gain experience in what new operators are required and what new categories of objects will be needed to address the needs of practicing biologists.

Achievements of the Mycoplasma Project

At the termination of the project at Harvard, we had accumulated over a million raw bases of Mycoplasma capricolum sequence (1,039,095 bp) with a total of over a quarter of a million linear bases (267,686 bp). We sequenced 1,505 random clones from the organism, producing 187,309 raw bases of sequence. These assembled into 1,032 unique starting points of 137,372 linear bases.

We are currently "walking on" 381 contigs into which 1198 of the original random clones have assembled. During the last year of the project we accumulated 901,723 bases of raw walking data, which has assembled into 215,236 bases of linear walking sequence with an average of 4. 3 fold coverage. We have some 52,450 bases comprising the remaining 308 unused starting points.

We have done some preliminary analysis of the random sequences using the Blastx search algorithm against the Non-Redundant Genbank Database available at the National Library of Medicine. Blastx results indicate that 14% of the random sequences have similarity to proteins in the database (Pvalue < 10-6). On the other hand, of the 292 open reading frames longer than 30 amino acids found in the 381 contigs, fully 53% of them have similarities to proteins with Pvalue < 10-6 (see Appendix A). These results indicate that random one-pass sequencing may

have little identifiable information content; thus, obtaining high-quality, accurate sequences becomes extremely important.

We have roughly classified the types of similarities found using Blast. In summary, we found the following:

1. Similarities to proteins involved the **replication apparatus** and DNA repair enzymes, such DNA binding proteins, gyrase, ligase, and polymerase. These similarities were expected.

2. Similarities to proteins of the **translational** and **transcriptional apparatus,** such as many of the tRNA synthetases, RNA polymerase, and initiation/elongation factors. We note that over half of the tRNA synthetases have been identified. This work suggests that a significant portion of the genome has already been sequenced (see below).

3. Several examples of what appear to be **regulatory** proteins. Many similarities are to higher organisms, however; thus, their function in Mycoplasma is unknown.

4. Similarities to **transport** proteins. These are to be expected because this organism is an extracellular parasite and must import most of its metabolic precursors.

5. Similarities to both **catabolic** enzymes and **anabolic** enzymes. Most of these are predictable a priori knowing the biochemistry of the organism.

6. Similarities to proteins involved in pathogenesis. Some, like the P1 adhesion protein, are expected. Others, like hemolysin, are unexpected.

7. Numerous anomalous similarities, many to higher eukaryotes. These are totally unexpected and need further investigation to determine the validity of the similarity score and the validity of the alignment.

We analyzed the DNA of the organism on CHEF gels, exploring the rare cutting pattern to identify large regions of the organism. The Mycoplasma DNA has an apparent genome size of 1 megabase on pulse field gels using yeast chromosomes as molecular weight makers, but we believe these DNA sizes to be exaggerated because of the high AT content of the organism. We have recently enumerated the rare restriction cuts that have been sequenced (Table I). Interestingly, we appear to have identified about 35% of the known rare restriction sites in the organism in 215,000 bases; this extrapolates to a genome size of only 765 kb, much smaller than the estimates determined from the pulse field gels.

Table I Rare Cutting Sites

	Recognition Site	Expected	Observed
Fsp I	TGCGCA	5	2
Bgl I	GCCNNNNGCC	6	0
Apa I	GGGCCC	2	1
BssH II	GCGCGC	1	0
Sal I	GTCGAC	2	2
Sma I	CCCGGG	2	1
Xho I	CTCGAG	2	1
TOTAL		20	7

Graphic User Interface to Support GenoBase Queries

One of the major obstacles to the routine use of Logic programming environments by biologists is the difficulty in understanding the underlining data structures of the environment. GenoBase is a prolog based environment that links many different databases using the concept of typed objects. We found that we neede a versatile querying environment to eficiently integrate the data from the Mycoplasma Genome Project at Harvard University into GenoBase. We have proceeded to develop a graphic user interface at Harvard University based on EZshelltool to generate prolog predicates to query the integrated database environment. We will described the basic format of the GUI and give several examples of typical GenoBase queries. The overview that is described consists of opening a local shelltool and running the GenoBase environment on a remote machine.

EZshelltool

EZshelltool is an X Windows-based graphic user interface which allows the seamless integration of functions into a shelltool. This environment is based the linkage of external programs into the shelltool by a user-expandable menu system and is supported on Sun™ and DEC™ workstations. There is no limitation to the number of external functions that can be linked to the interface. This user-defined menu system allows the customization of an environment with very little effort. We have used this development tool to prototype a graphic user interface to generate prolog predicates that query GenoBase.

Overview of GenoBase Interface

One initially starts ezshelltool on their local machine and obtains a window with various menus and a standard Xwindows based shell (figure 1). One then telnets to the remote site and uses the **Startup** menu to start prolog (figure 2) and load GenoBase (figure 3). In all instances, clicking on the menu items writes a prolog predicate to standard out (the shelltool) and the predicate is evaluated.

The main window to start a query is evoked by clicking on the **Object** menu (figure 4). In the first example we will search E.coli for all coding sequences. One selects the **By Name, Type, Genome** button and then selects the **Type** of object from the pop up list (figure 5) as cds (coding sequence) and the **Genome** from another pop up list. as E.coli (figure 6). One now has all the information to generate a valid GenoBase query and clicks **OK** at the top of the window (figure 7). As stated previously, this then generates a prolog query and writes it to standard out to be interpreted by the prolog process running on the remote machine. The resulting output is directed back to standard in which is the ezshelltool display.

In the second example, we select **Special Objects** as the method to pick objects and select which from the pop up list (figure 8). We also choose to save the set of objects retrieved by saving the results into the **HeatShock** variable (figure 9). As before the predicate is generated and evaluated by the prolog process.

The system contains a full **GenoBase Help** function, an example of which is given on figure 10. We finally demonstrate the utility of both the GUI environment and the GenoBase system itself. Here we have asked to enumerate all the **sequence_fragments** from **M. capricolum** that have a link from **DNA to peptide to enzyme to enzyme pathway** (figure 11). The example shows the results of the first item in the return list that has a link to the Electron Transport Chain pathway (**ETC_1**). It can be seen that it would be difficult to remember the syntax of the predicate as well as difficult to type it in without ant errors. The GUI completely

alleviates these problems and make the system relatively trivial to use. We also include a printout of the entire menu system (Appendix A).

New GUI Interface

The main problem with the EZshell interface is the lack of running menus, that is the pop up lists are too large to be accommodated on the screen. This is especially true when attempting to load in all genomes in GenBank. It is in this light that the GUI prototyped here is being converted to TCL/TK interface. This should address the major problem with the present system and result in a very facile tool that will allow biologist to generate ad hoc GenoBase queries.

Future Analysis and Annotation of Raw Sequence

We have developed the GenoBase system to support interpretation of genomic sequence data. To effectively use the system to analyze data like that produced in the Mycoplasma capricolum sequencing effort, one must now produce an initial set of annotations that identify putative CDSs, regulatory signals, and so forth.

To this end, we have begun an extensive effort to create a system that wouldfunction as follows:

1. First, the sequence is automatically submitted to a suite of available tools (such as Blast, Fasta, Blocks, Genmark, and Blaize). This process involves a combination of locally maintained tools and access to available servers over the network; it is all achieved without manual intervention. The results from the tools are translated into Prolog facts asserting specific properties (such as similarities to known sequence and putative CDSs from tools like Genmark).

   This effort clearly requires building on the rich set of tools that have been developed by other researchers to address precisely this problem. We have had contact with a number of the groups offering such services and have received several useful suggestions. In the case of Genmark, we have formed a collaboration in which we exchange initial analysis in order to gain more insight into the capabilities of each available tool.

2. The encoded output from the tools, much of which is quite irrelevant, must then be analyzed and used to construct a coherent set of annotations. This work, we believe, is best done within the context of high-level tools and requires direct access to the capabilities offered by a system such as GenoBase. Specifically, not only does such an annotation system produce input for storage and analysis within GenoBase, it also depends on the flexible access provided by GenoBase to develop an effective integration of the output of the available suite of tools.

We expect that this system will be fully operational by the end of Phase II of the proposal. We believe that it effectively complements many aspects of our efforts in developing GenoBase, and directly supports the interpretation of sequence produced by the Mycoplasma capricolum sequencing effort.