Using Multiple Tools for Automated Genome Interpretation in an Integrated Environment

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

Christoph Sensen
Institute for Marine Biosciences
National Research Council of Canada
1411 Oxford Street, Halifax, N.S.
B3H 3Z1 Canada sensencw@eecs.uic.edu

January 12, 1996

The year 1995 heralded the sequencing of the first whole microbial genomes (Refs 1, 2). The end of 1996 will see at least four more full microbial genomes sequenced. The entire yeast genome will also be submitted to the public databases. At this rate, we can expect to have access to the complete genomes of at least 25-30 organisms within five years.

While most problems of automated sequence generation and sequence assembly (Ref. 3) have been solved, the automated analysis of genomic sequences has scarcely been addressed. Consequently, given the sheer amount of data generated, most genome sequences submitted to the publicly accessible databases are either poorly or not at all annotated. Yet, the success and utility of genomic sequencing depends crucially on the the analysis strategies, the quality of the software algorithms and the tools that are used to manage and interpret the sequence data.

Sequence interpretation "by hand", as it is usually performed today, presents an overwhelming challenge to genome researchers. For each megabase of sequence (the equivalent of 1000 open reading frames in prokaryotes or yeast), one person has to devote an entire year of research to the inital sequence interpretation. Scanning the databases also has to be repeated frequently as new entries are added to the public databases daily. Most genome projects therefore do not have the capacities to analyze the sequence that they generate completely.

To address this situation we are developing a prototype for a UNIX-based integrated environment for automated genome sequence interpretation. The system, called MAGPIE (Multipurpose Automated Genome Investigation Environment), is designed to automatically assign as many different features to the sequence data as possible (Ref. 4). MAGPIE can gather and analyze evidence created by e-mail servers (e.g. BLAIZE, BLAST, BLOCKS and FASTA) in combination with locally installed tools (e.g. PROSEARCH, SAPS and TRNASCAN). Our goal is to reduce user interference with the data gathering and analysis to as low as 5-10 minutes per day. A Prolog-based daemon monitors the sequence as it is generated by the sequencing project (the input sequence can be in more than 15 different file formats), sends requests to toolservers, and sorts the server replies into groups and categories. All replies are translated into Prolog facts, using Perl5 scripts. The MAGPIE analysis module analyzes the information from multiple tools using logical rules represented in Prolog (Ref. 5). Properties having varying levels of confidence are assigned to regions in the genome sequence (Ref. 6, 7). Based on the assigned properties, human-readable reports are generated by the MAGPIE report generation module.

The particular strength of MAGPIE is its flexibility and configurability. All variables (e.g. codon usage of the organism analyzed, e-mail addresses for toolservers, parameters for tools) are stored in ASCII format or in table files that can be modified easily by the user. Tools can be employed for the sequence analysis according to the status of the sequencing project. "Shotgun" sequence, which is in the primary assembly stage and contains a high number of ambiguities, can be analyzed with a different set of tools and parameters than finished sequence, which may have an error rate of 1 in 10000 basepairs. All responses and Prolog facts are stored by MAGPIE as ASCII files. This gives MAGPIE a superior flexibility over relational databases since no (more or less) static database format has to be defined before the first data are collected. MAGPIE can therefore be adapted to new tools or changes in the output format of existing tools with a relatively small programming effort.

MAGPIE generates a variety of different reports, including html reports, which can be accessed through WWW browsers like Mosaic or Netscape. To support distributed, collaborative genome projects, the html reports include information about the status of the sequencing project and enable the current sequence analysis to be exchanged immediately over the Internet. The html reports also offer links to a suite of WWW resources on the Internet that provide additional information about the features identified by MAGPIE. The most important connection is the one to PUMA

(http://www.mcs.anl.gov/home/compbio/PUMA/Production/puma.ht ml) a system that has been developed by Ross Overbeek and his collaborators at the Mathematics and Computer Science Division at Argonne National Laboratory. PUMA interconnects information about phylogenetic relationship of organisms, multiple sequence alignments and metabolic pathways. MAGPIE links this information into the html reports about individual open-reading frames. Other important links are to SRS (Thure Etzold`s Sequence Retrieval System, EMBL, Heidelberg) and the molecular biology-related subset of MEDLINE, served by NCBI. These links offer to the interested researcher a wide variety of information about genes beyond their pure function.

Although MAGPIE automates the data collection and initial analysis, it is clear that the final decision about what an open reading frame encodes can be made only by a human researcher after a careful examination of the MAGPIE results. Therefore the next step in the MAGPIE development will be the implementation of an interface that allows researchers to support or refute MAGPIE results. MAGPIE is still an experimental prototype under development (the official alpha release is planned for July 1996), but it is already used in several microbial genome projects. MAGPIE is beeing applied to the Sulfolobus solfataricus genome project (http://www.imb.nrc.ca/imb/sulfolob/sulhom_e.html), which functions as the alpha text site for MAGPIE (Ref. 8), the public data for Escherichia coli, Haemophilus influenzae, Mycoplasma genitalium and Bacillus subtilis as well as several industrial genome projects. Examples of MAGPIE html reports can be browsed by connecting to the the MAGPIE homepage at http://www.mcs.anl.gov/people/gaast

Acknowledgments

This work was supported by the U.S. Department of Energy, under Contract W-31-109-Eng-38. The Sulfolobus project is supported by the Canadian Genome Analysis and Technology Program, the Canadian Institute for Advanced Research, the National Research Council of Canada, and the Medical Research Council of Canada.

References

1. Fleischmann, R.D., et al. (1995) Science 269, 496-512

2. Fraser, C.M., et al. (1995) Science 270, 397-403

3. Staden, R. (1994) In: Methods in molecular biology (Griffin, A.M. and Griffin, H.G. [eds.]) Vol 22, Humana Press, Totawa, New Jersey, pp. 9-170

4. Gaasterland, T. and Sensen, C.W., (1995) In: Bacterial Genomes: Physical Structure and Analysis (DeBruijn, Lupski and Weinstock [eds.]) Chapman and Hall, New York, London in press

5. Gaasterland, T, Lobo, J., (1994) Proceedings of the International Conference of Very Large Databases, Santiago, Chile

6. Gaasterland, T., et al. (1994) Proceedings of the International Conference on Intelligent Systems for Molecular Biology, Stanford University

7. Gaasterland, T., et al. (1995) Second International Meeting on the Integration of Molecular Biological Databases, Cambridge, England

8. Sensen C. W., et al. (1995) In: Bacterial Genomes: Physical Structure and Analysis (DeBruijn, Lupski and Weinstock [eds.]) Chapman and Hall, New York, London in press