Genomatica: an integrated data management and analysis tool for genome sequencing projects

Yutaka Akiyama, Hirotada Mori †, Satoru Kuhara ††,
Naoki Ogasawara †, Nobuyuki Miyajima †,
Tetsuya Furukawa ††, Kenji Sato * and Yasufumi Murakami **

Institute for Chemical Research, Kyoto University
†Advanced Institute for Science and Technology, Nara
††School of Genetic Resource and Technology, Kyushu University
‡Kazusa DNA Institute
††Computer Center, Kyushu University
* Department of Information Engineering, Kyushu University
** RIKEN Tsukuba Life Science Center

E-mail: genomatica@kuicr.kyoto-u.ac.jp

Abstract

Genomatica is an integrated software tool designed for helping systematic management of a large number of DNA sequence fragments obtained through a genome sequencing project.

Its graphic user-interface also allows users to look, with any magnifying factor, into any position of the specified chromosome and to browse various kinds of collected information altogether (including: DNA sequence itself, related gene descriptions, bibliographic references, corresponding GenBank entries, confirmed or putative coding regions, results from homology analysis for the expected protein, RNA genes, clone information, enzyme restriction maps, comments from administrator, private memorandums by user).

We are planning to use Genomatica in E. coli (local data compilation mainly managed by Mori), B.subtilis (by Ogasawara), and S.cerevisiae (by Murakami) genome sequencing projects.

The Genomatica project was started on 1992 as one of the advanced genome database projects sponsored by Human Genome Center, University of Tokyo. In June 1993, ver.2.0 which was fully re-designed with NCBI vibrant library was released. Further augmented version Genomatica 2.1 (with several sequence analysis functions and network communication modules) will be released on Nov. 1993 and will be distributed through anonymous ftp services. The Genomatica system is currently available for X11 window system on Unix workstations, but Macintosh and IBM-PC versions will be also announced soon.

*秋山泰, 向原哲, 浦原哲, 小笠原直毅, 宮崎伸行, 芳川哲也, 佐藤賢二, 村上康文
** 〒611 京都府京都市東山区 京都大学化学研究所
1 Introduction

It is important and very useful to construct a sophisticated genome database of particular species, on which all the information are precisely placed according to its physical location on the chromosome. With such a database we can easily browse relations between neighboring genes and regulatory factors, and also the database itself will greatly support users in making contig sequences, or an united image of the genome, in a full sequencing project on the way.

There have already been many attempts to construct such an unified map on computer systems. Efforts are especially made in E. coli chromosome research. K. Rudd and his colleagues have been provided their EcoSeq database [1] which contains physical map and GenBank sequences with physical positions determined. They have also presented several tools useful in deciding positions of the sequences[2, 3]. GeneScape[4] is a concise and user-friendly interface on Macintosh for accessing their EcoSeq database contents.

Shin et.al [5] are constructing more integrated database using an object-oriented system and Rouxel et.al [6] have developed a graphical system showing metabolic links among Escherichia coli genes.

We have been developing an integrated data management tool, Genomatica.

The features of Genomatica system are:

• Functions for supporting sequencing project. (making contig sequence, etc.)
• Powerful graphic interface and finder system that allow users to use the Genomatica system like an ‘encyclopedia’ for specific chromosome.
• Cross-linking to several databases (planned)
• Automatic data update service via Internet (planned)
• Collaboration or unifying with sequence analysis software (planned)

In this paper we will present current progress in the development of Genomatica system and show our future plans utilizing Genomatica in genome sequencing projects.

2 Constructing a database in Genomatica - E. coli example

We are planning to use Genomatica in managing sequence data from E. coli, B. subtilis and S. cerevisiae genome projects (mainly in Japan). In this section, we would like to provide brief introduction of the progress in Japanese E. coli genome sequencing project as an example and show what kinds of information we have been prepared in Genomatica.

In Japan, the project “Systematic sequencing in Escherichia coli” is in progress (1992). Sequences between contigs at 0 - 7 min region were determined [7] and this brings the total length of contiguous finished sequence to 274kb. Genes and putative ORFs, along with the similarities with protein database are summarized in Table 1. Some predicted putative genes in newly sequenced regions showed interesting similarities with the protein database, such as transaldorase, NADH dehydrogenase, iron transport system, nitrogen fixation systems, etc.
According to the number of genes and putative ORFs, the average length of the gene is about 1.2kb.

We have prepared the following *E. coli* data in Genomatica:

1. Linkage map published by Barbara J. Backmann
2. Ordered clones by Kohara, et.al.
3. Physical map of *Escherichia coli* chromosome
4. Contig sequence deduced from sequence data
5. Genes and putative ORF data predicted from contig sequence

Ordered clones are a well organized bank of recombinant clones containing every part of the dissected chromosome. 476 out of 3600 clones can cover virtually the whole genome with ample overlapping. Since the position and the length of these clones were determined, we incorporated those data into Genomatica. Physical map is a map of cutting sites by restriction nuclease and that of the *Escherichia coli* chromosome by eight restriction nuclease was also determined by Kohara et.al. Those data were also incorporated into Genomatica. Through elaborate gene manipulation and sequencing, more than 2300 entries and $4 \times 10^6$ bp of *Escherichia coli* can be found in databank such as GenBank. These, however, include the data from different strains and mutants and also include duplicates or overappings. To construct the *Escherichia coli* database in Genomatica, we selected *E. coli* K12 data from GenBank excluding other strain's data. As a result, the number of entries is 2121 and the total base pairs is $3.8 \times 10^6$ bp. To sort the sequence data into groups using overappings, we ran the computer program, blast developed by NCBI. When overappings were found in IS sequence, these groups were kept for further analysis. As a result, the sequence data were divided into 800 groups out of 2121 data. In the group containing more than two sets of data we made contig sequence using an ATSQ computer program developed by SDC Inc. At the same time to make contig sequence, those precise positions on the chromosome were determined using EcoMap computer program developed by K. Rudd (1990) or other genetic analysis obtained through various literatures. About 500 contig's positions are determined at present. ORFs translated from contigs and data around ORFs, such as position on the chromosome, homology against PIR protein database, etc. must be also prepared in Genomatica and is now in progress.

3 Implementation

Genomatica system (currently version 2.0.14) is written in C language (about 6000 lines). In order to ensure its portability we employed the CoreLib routines and the Vibrant graphic libraries both developed by NCBI, NIH.

The Genomatica system is running on X11+Motif window systems on SPARCstations (SunOS 4.1.2). Because the system is fully written in CoreLib and Vibrant graphic libraries, porting to Macintosh and IBM-PC are expected to be relatively easy and we are planning to have versions for the two platforms soon.
Table 1:
putative genes whose products show sequence homology with known proteins

| Orf | start | end | Number of codon | max(b) | pluss(c) | Gene(d) | PIR entry(e) | Description | Homology score | %match |
|-----|-------|-----|-----------------|--------|----------|---------|-------------|-------------|---------------|-----------|--------|
| 8   | 7727  | 797 | 455             | 670    | (set)    | YVCTCV  | *YSCTAL1    | Tetracycline resistance protein - E. coli | 150        | 54.4   |
| 9   | 7827  | 207 | 183             |        |          | *YSCTAL1 |             | Saccharomyces cerevisiae TALI gene for transaldolase | 497        | 51.9   |
| 10  | 8365  | 162 | 95              |        |          | YSCTAL1 |             | Saccharomyces cerevisiae TALI gene for transaldolase | 393        | 51.9   |
| 18  | 15583 | 374 | 370             |        |          | QEC47   |             | Hypothetical protein - E. coli | 1766       | 96.2   |
| 29  | 15830 | 105 | 0               |        |          | S08881  |             | Hypothetical protein II (Insertion seq. IS421) | 537        | 99.0   |
| 54  | 33956 | 183 | 194             |        |          | S06477  |             | - Escherichia coli | 220        | 32.1   |
| 55  | 35927 | 194 | 194             |        |          | S06477  |             | Enol-CoA hydratase precursor, mitochondrial | 120        | 37.3   |
| 56  | 37923 | 531 | 522             |        |          | S01667  |             | Enol-CoA hydratase precursor, mitochondrial | 531        | 29.3   |
| 57  | 38017 | 320 | 248             | (baliP) |          | F37844  |             | 4-Coenzyme-CoA ligase - Parsley | 514        | 30.5   |
| 58  | 38701 | 114 | 108             | (baliP) |          | F37844  |             | 4-Coenzyme-CoA ligase - Escherichia coli | 186        | 44.2   |
| 59  | 40647 | 381 | 380             |        |          | C32452  |             | Isovaleryl-CoA dehydratase precursor, hepatic - Rat | 461        | 29.2   |
| 60  | 41199 | 502 | 504             | (setT) |          | S10987  |             | BetT protein - Escherichia coli | 592        | 25.7   |
| 61  | 41296 | 284 | 268             | (setA) |          | S10707  |             | BetA protein - Escherichia coli | 209        | 26.2   |
| 62  | 42207 | 212 | 76              | (setB) |          | S14071  |             | BetB protein - Escherichia coli | 259        | 46.5   |
| 63  | 43287 | 453 | 420             | (setC) |          | S14072  |             | BetC protein - Escherichia coli | 788        | 37.0   |
| 64  | 43520 | 159 | 55              | (setZ) |          | S14973  |             | BetX protein - Escherichia coli | 221        | 36.3   |
| 65  | 45460 | 168 | 166             | (set)  |          | YTSBG6  |             | Tetracycline resistance protein - Bacillus cereus plasmid pNC16 | 147        | 20.2   |
| 66  | 46108 | 234 | 220             |        |          | A25949  |             | Glutamate transporter protein - Rat | 172        | 22.4   |
| 67  | 46893 | 181 | 176             |        |          | A25929  |             | NAD(P)H dehydrogenase (quinone) - Human | 196        | 20.0   |
| 68  | 50596 | 297 | 271             |        |          | A23048  |             | Heat shock protein hsp70/translation | 325        | 36.8   |
| 69  | 50958 | 271 | 232             | (setA) |          | XMEXAD  |             | Heat shock protein hsp70/translation | 325        | 36.8   |
| 100 | 27207 | 239 | 232             | (setC) |          | QREAC   |             | Heat shock protein hsp70/translation | 325        | 36.8   |
| 101 | 27436 | 557 | 516             | (setB) |          | QSREBD  |             | Heat shock protein hsp70/translation | 325        | 36.8   |
| 102 | 27513 | 341 | 322             | (setA) |          | QRSUC   |             | Heat shock protein hsp70/translation | 325        | 36.8   |
| 103 | 27718 | 419 | 392             |        |          | YTECQ   |             | Heat shock protein hsp70/translation | 325        | 36.8   |
| 105 | 70156 | 216 | 201             | (1estC) |          | S07866  |             | leuT protein - Salmonella typhimurium | 946        | 92.0   |
| 106 | 80567 | 481 | 668             | (1estC) |          | S10711  |             | Leucine-tRNA synthetase chain leuC - Salmonella typhimurium | 1988       | 86.0   |
| 107 | 81781 | 418 | 363             | (1estB) |          | DEHSIC  |             | Serine-tRNA synthetase chain leuC - Salmonella typhimurium | 900        | 53.0   |
| 108 | 82304 | 502 | 522             | (1estA) |          | S08431  |             | Serine-tRNA synthetase chain leuC - Salmonella typhimurium | 1812       | 76.4   |

In this table, novel orfs that showed homology scores of more than 100 (optimized) with one or more proteins are presented. The nucleotide number of the first codon (a) was counted from the A of the initiation codon (ATG) of thrA, clockwise along the E. coli genetic map. The maximum number (b) of contiguous sense codons, namely, the number of sense codons spanned by two adjacent nonsense codons, and the plausible number (c) of sense codons counted from the first ATG codon are indicated. Gene symbols (d) of E. coli or other organisms together with PIR entries (e) of the highest homology scores are shown. Asterisks (*) before the entry names indicate the GenBank entries of DNA sequence that were used to detect homology in deduced amino acid sequence.
3.1 Data Access Layer

The most basic layer in the Genomatica data management system is a set of data management functions those are originally written for Genomatica data structures. We have had many intensive discussions on the tradeoff between the merit and disadvantage of using commercial DBMS (database management systems), regardless of whether relational DB or object-oriented DB. As a result, we chose to build a simple but very efficient dedicated routines for the data access tasks, instead of using commercial general-purpose DBMS. Using such an original data access facility, we can make the system fast and keep money cost of the total system very low and allow the Genomatica system remaining to be a free public domain software, while the accessing protocol cannot be so general (i.e. cannot accept full SQL query, etc.)

3.2 Screen overview

![Image of Genomatica main window]

Figure 1: Genomatica main window

Figure 1 shows main window of Genomatica which comprises four parts: 1) menu bar [upper most], 2) operation panel [upper], 3) label panel [left most], and 4) data panel [the main graphic]
3.3 Zooming and Positioning

User can change zooming factor and position of the browsing window by one of the three actions: 1) keyboard input, 2) button action, or 3) pointing by mouse.

3.4 Popup windows

Various kinds of popup window can be displayed with user's double clicking on any active object in the data panel. The popped up window may have several operation buttons on it, allowing a user to see further analysis report or an administrator to change/add/delete some property of the object.

In the case of current version of our *E. coli* database, variation of popup windows are as follows: gene description and references, physical map information, clone, contig sequence description, GenBank entry linking, etc.

![Figure 2: Popup windows](image)

3.5 Finder

New Genomatica ver. 2.1 will support various finder functions, though former ver.2.0 has only simple finder menu shown in Figure 3. With the simple finder, a user can scan, for example, all genes having name "trp?" or all GenBank entries "ECODNA???". The next version of finder will allow users to make a complicated query and to obtain graphical map for the matching objects on the chromosome.
3.6 User Preferences

Many user preference setting menu are provided. With the viewselect menu and color pallet function, each user can easily change the appearance of the main display. Default settings for these features can be specified in a special file named "genomatica.conf". In future, user preference menu for various sequence analysis functions should be also provided.

3.7 Future work

We are now building new modules to provide the following functionalities:

- Contig editing (currently this is implemented as a graphical editing command for pre-processing result from the semi-automatic alignment routine in the Genetyx commercial program)
- ORF predictor and ORF homology search (by BLAST, FASTA)
- More intelligent finder. Graphical output of finder results.
- Linkage to GenBank, PIR, and PDB (planned) databases.
- Automatic data update service modules using network protocols, etc.
4 Conclusion

We have been developing an integrated sequence data management tool called Genomatica. We are planning to utilize Genomatica in *E. coli*, *B. subtilis*, and *S. cerevisiae* sequencing projects, and for *E. coli* project we have already prepared many kinds of related data in the Genomatica system. We are currently augmenting Genomatica functions especially in terms of supporting administrative tasks and ORF analysis. On November 1993, we will release Genomatica 2.1 and the software along with newly compiled *E. coli* data, will be distributed via anonymous ftp services.

Acknowledgement

This work was supported by a grant-in-aid for scientific research on the priority area 'Genome Informatics' from the Ministry of Education, Science and Culture, Japan.

Genomatica software development project was organized and supported by Human Genome Center, Institute of Medical Science, Tokyo University.

References


