A Set of Tools Developed for the Analysis of Open Reading Frames of Yeast (S.cerevisiae) Chromosome VI

Motoe Sasanuma Zhong-qing Wang Kazuhiro Shibata Syun-ichi Sasanuma Masashi Ozawa
Eiko Kohriki Toshihiko Eki Akinori Sarai Hideaki Sugawara Yasufumi Murakami

Tsukuba life Science Center, The Institute of Physical and Chemical Research (RIKEN),
3-1-1 Koyadai, Tsukuba, Ibaraki 305, Japan

1 Introduction

The DNA sequence of chromosome VI (270kb) of the budding yeast S. cerevisiae, has been determined. The sequence data were then analyzed with the software GENETYX (Software Development Co.) to find out the candidates of the coding region. As there is little number of intron in budding yeast genome, it is relatively easy to find the candidates of the coding region. Open reading frames identified by the software were generally the same as the coding sequence with little exception. To carry out further analysis of the open reading frames, we tried to automate the process required for similarity search.

The processes automated are

1) Extracting the DNA sequence from the target sequence to be analyzed according to the position data of ORF produced by GENETYX.
2) Changing the data format of the sequence data to those suitable for the analysis with FASTA
3) Carrying out automatic batch analysis with FASTA
4) Extracting the most homologous 5 entries according to the optimum score of the FASTA analysis
5) Changing the data format of 4) to the format readable by the spreadsheet program of Macintosh.

2 Flowchart of the data analysis and the program developed

The flowchart of the sequence data analysis is shown in the following page. The program developed are as follows:

1) **ORF**
This program extracts the DNA sequence from the target sequence analyzed by GENETYX according to the position data of ORF produced by GENETYX and produces sequence files in SUN workstation.

2) **BatchToGem**
This program reformats the sequence data to the GCG format and carries out batch analysis with FASTA against Genbank and EMBL DNA sequence database. This program works on SUN workstation.

3) **BatchToPIR**
This program reformats the amino acid sequence data to the GCG format and carries out batch analysis with FASTA against PIR protein database. This program works on SUN workstation.

笹沼 基恵，王 忠清，柴田 一浩，笹沼 俊一，小澤 雅司，高力 英子，浴浴 俊彦，皿井 明倫，菅原 秀明，村上 康文 ：
理化学研究所ライフサイエンス筑波センター，〒305 茨城県つくば市高野台 3-1-1

4) **BatchToSWS**

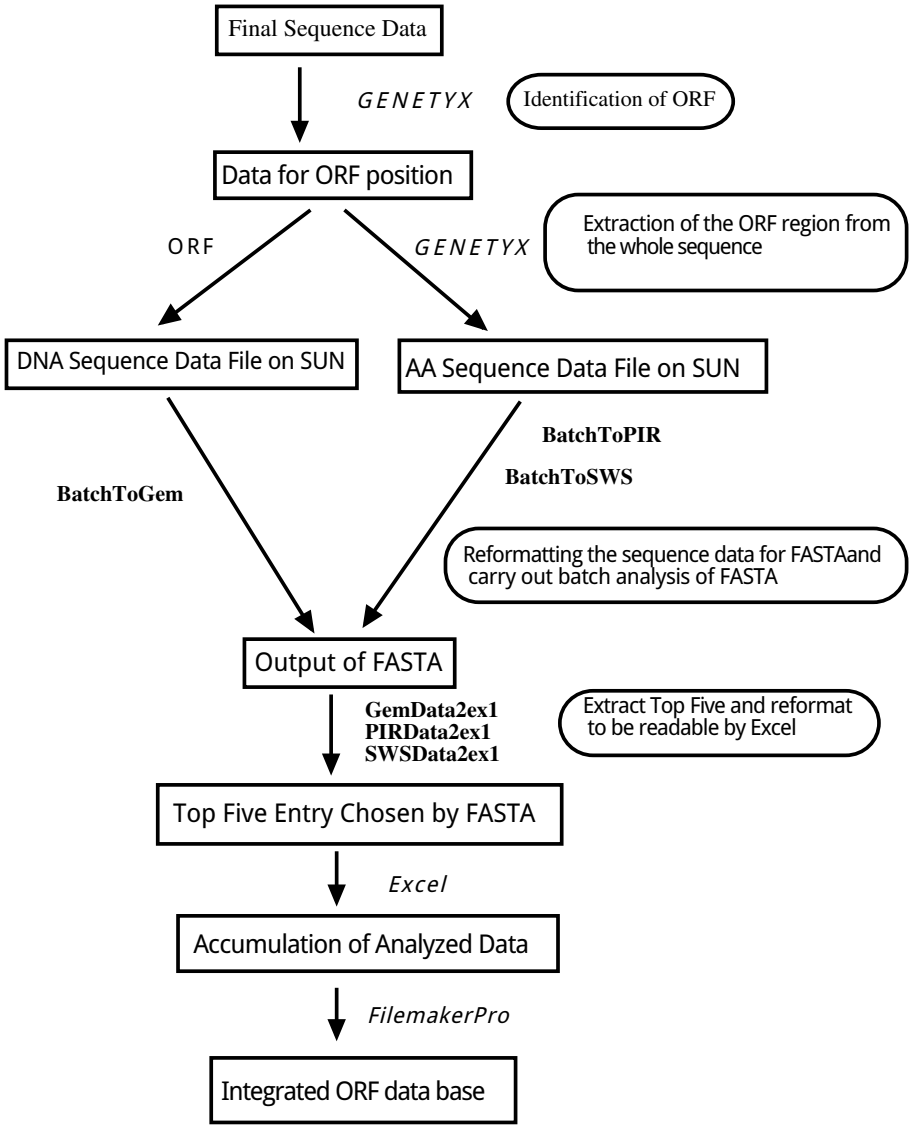
This program reformats the sequence data to the GCG format and carries out batch analysis with FASTA against SWISS-PROT protein database. This program works on SUN workstation.

5) **CountAlpha**

This program goes through the sequence data files and determines the number of A, C, G, T as well as N, R e.t.c.. This tool has been developed to check whether the data handling is properly being carried out. This program works on SUN workstation.

6) **GemData2ex1, PIRData2ex1, SWSData2ex1**

These programs find out the top five hits of the FASTA analysis against each database, extract the top five data and reformat the data to be readable by Excel (Microsoft corporation) on Macintosh. This program works on SUN workstation.

3 Development of the database for the ORF analysis

To summarize the results of these analysis, a database has been developed using FileMaker(Claris corporation). To incorporate all analyzed results to this database, analyzed results were accumulated with Excel and then transferred to the database. Excel and FileMaker Pro were currently used because they are convenient for the rapid and flexible development of the prototype system.