ContigMaker and SAND: Software Tools for Genome Mapping and Sequencing

Akira Suyama[1] suyama@dna.c.u-tokyo.ac.jp

Akira Ohyama[2] akr@kojib.mitsui-knowledge.co.jp

Masami Hagiya[3] hagiya@is.s.u-tokyo.ac.jp

Yoshiaki Furuhata[4]

Takashi Ito[4]

Asao Fujiyama[5] afujiyam@ddbj.nig.ac.jp

Masahira Hattori[4]

Yoshiyuki Sakaki[4]

Toshihisa Takagi[4] takagi@ims.u-tokyo.ac.jp

[1] Institute of Physics, College of Arts and Sciences, The University of Tokyo
3-8-1 Komaba, Meguro-ku, Tokyo 153, Japan

[2] Mitsui Knowledge Industry Co. Ltd.
3-7-4 Kojimachi, Chiyoda-ku, Tokyo 102, Japan

[3] Department of Information Science, Graduate School of Science, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan

[4] Human Genome Center, Institute of Medical Science, The University of Tokyo
4-6-1 Shiroganedai, Minato-ku, Tokyo 108, Japan

[5] Division of Human Genetics, National Institute of Genetics
1111 Yata, Mishima, Shizuoka 411, Japan

Physical mapping and sequencing are two major activities in genome projects. ContigMaker and SAND (Sequence Assembler and editor for the Nested Deletion sequencing method) are software tools to assist genome projects in those activities. They are Motif applications running on UNIX workstations with the X Window System.

ContigMaker is a software tool to construct contig maps from experimental mapping data. It is composed of five major components: map data manager, map analyzer, map viewer, map aid, and project manager. Contig-mapping data obtained by experiments are stored in

[1]陶山 明：東大教養物理、〒153 目黒区駒場 3-8-1

[2]大山 彰：三井情報開発（株）、〒102 千代田区麹町 3-7-4

[3]萩谷昌己：東大理情報、〒113 文京区本郷 7-3-1

[4]古畑芳晶、伊藤隆司、服部正平、榊 佳之、高木利久：東大医科研ヒトゲノム、〒108 港区白金台 4-6-1

[5]藤原秋佐夫：国立遺伝研、〒411 三島市谷田 1111

a database of the map data manager. The stored data are then subjected to analysis by the map analyzer to generate contigs. ContigMaker supports the two strategies for contig construction: the STS (sequence-tagged sites) strategy and the MOF (mapping by oligonucleotide fingerprinting) strategy. The generated contigs are assembled into a contig map according to positions of landmarks falling on the contigs. ContigMaker allows a user to extract landmark information from a public genome database such as the GDB. The contig maps constructed are graphically drawn by the map viewer. The map aid provides miscellaneous small useful tools to finish a contig-mapping task. A repeated task ContigMaker performs can be automated by a macro created by the project manager. The macro will save time and effort for contig map construction.

SAND is a software tool to manage and assemble base sequences in the nested deletion sequencing method developed by M. Hattori et al. The nested deletion method is one of sequencing methods satisfying requirements for large-scale sequencing of genomic DNA, i.e. high accuracy, high speed and low cost. In the method deletion fragments are made with sonication and their sizes are analyzed by means of gel electrophoresis through agarose. The deletion fragments are then sorted in order of size to make a ordered set of nested deletions. Each fragment in the set is sequenced and finally a whole base sequence is reconstructed by assembling base sequences of the nested deletions.

The sequence assembly can be performed in a straightforward manner when a size of each nested deletion fragment is measured accurately. The size measurement in agarose gels, however, is not so accurate that the resultant order of nested deletions of similar lengths are often incorrect. In contrast nested deletions of very different lengths are almost always put into a correct order. SAND takes these facts into account when assembling base sequences of nested deletions. Namely, SAND first assembles base sequences of nested deletions of similar lengths to determine their consensus sequence. Since there is no guarantee that their order is correct, the algorithm for sequence assembly in the shot-gun sequencing is used at this first stage of sequence assembling. SAND then assembles the consensus base sequences determined at the first stage. This assembly can be performed in a straightforward manner because the order of the consensus sequences has been determined accurately. The alignment of base sequences assembled and their consensus sequence are displayed and edited with the multiple sequence alignment editor SAND provides.

SAND and ContigMaker have the same database structure and management system. Data of clone positions managed by ContigMaker and those of base sequences managed by SAND can thus be integrated into a laboratory notebook database essential for genome mapping and sequencing.

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, "Genome Informatics", from the Ministry of Education, Science and Culture of Japan.