Meeting report

Genomics meets informatics

Katsumi Isono

Address: Department of Biology, Faculty of Science, Kobe University, 1-1 Rokkodai, Kobe 657-8501, Japan. E-mail: isono@biol.kobe-u.ac.jp

Published: 6 March 2001

Genome ***Biology*** 2001, **2(3)**:reports4006.1–4006.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/3/reports/4006

A report on the 11th Workshop on Genome Informatics, the annual meeting on genome informatics and related subjects supported by the Genome Informatics Society of Japan, Tokyo, Japan, 18-19 December, 2000.

The workshop [http://giw.ims.u-tokyo.ac.jp/giw2000/] has been held since 1990, with the main aim of promoting the exchange of ideas between scientists who have an informatics background and experimental biologists. It was divided into oral presentations and poster and software demonstration sessions along with two invited lectures.

Genome sequencing and functional analysis

In light of the completion of the genome sequence of *Arabidopsis thaliana*, the talk by Bal A. Antonio of the Rice Genome Research Program [http://rgp.dna.affrc.go.jp/] at STAFF, Tsukuba, Japan on the current status of the rice (*Oryza sativa*) genome sequence attracted great attention. Whereas *Arabidopsis* is an experimental model plant, rice is a commercial product: about 515 million tons of rice grains are annually produced and consumed by people in 89 countries. *O. sativa* has a small genome (about 430 Mb) compared with many other common cereal plants. Moreover, a recent careful comparison of the physical and genetic maps of chromosome I (*O. sativa* has 12 chromosomes altogether) indicates that perhaps the total genome size of rice might actually be smaller than 400 Mb. If that were the case, then the current report of 9.3% of finished sequence would actually equate to more than 10% of the total genome of rice. In addition, more than 40,000 expressed sequence tag (EST) data have been produced. Antonio reported how the data have been integrated into a compound database and retrieval system termed INE (INtegrated rice genome Explorer; 'ine' means rice in Japanese) that has been made available to the public (accessible through the above-mentioned website). As the genomes of other cereal plants such as sorghum and maize are being sequenced as well, the database should become a useful resource for various types of data mining for basic and applied genomics in the near future.

Gene Myers (Celera Genomics, USA) gave an invited, timely talk on the sequence assembly of the *Drosophila* and human genomes. He explained in detail several key steps in which new approaches were taken to improve both experimental data acquisition and informatics processing of the acquired data. For example, he described the preparation of plasmid libraries with 2 and 10 kb inserts, the pair-wise end-sequencing of bacterial artificial chromosome (BAC) libraries at 10-15 redundancies, and data assembly using four major computer program packages: 'screener', 'overlapper', 'unitigger' and 'scaffolder'. The screener program masks repeat sequences (such as heterochromatic regions and ribosomal DNA) from data assembly and the overlapper program detects shorter repeats of 40 base pairs or longer and analyzes them to determine whether they reflect intrinsic tandem repeats or sequence overlaps. After screening the sequence data in this way, the data are fed into the core part of the assembly program package, unitigger and scaffolder, to extend contigs and assemble them into 'scaffolds' (ordered and oriented contig sets). The *Drosophila* sequence data, when assembled into the final sequence, were said to contain errors as low as 1 per 10 kb. Initially, Celera intended to achieve tenfold coverage of the total genome with 500 bases in each sequence read. The overall coverage was actually 14.6-fold, however, and each sequence read was 551 bases on average. The procedures described by Myers were particularly impressive to both experimental biologists and informaticists at the workshop, because of the difficulties associated with assembling short sequence segments into large contigs and finally into one finished genomic sequence. Whether the accuracy of Celera's data is as high as described by Myers remains to be seen, but the *Drosophila* and human data will certainly be of importance to both biologists and informatics scientists.

Computer algorithms and handling of biological data

The genome sequences of more than 35 microorganisms are already available. One of the most popular ways of using the data is to perform comparative analysis of the genes and gene clusters in these organisms, many of which are not suited to laboratory experiments. Clemens Suter-Crazzolara (LION Bioscience, Heidelberg, Germany) emphasized the importance of gene-to-gene comparisons in organisms by integrating their genomic sequence data through a system termed SRS (sequence retrieval system), and by using analysis software such as genomeSCOUT and bioSCOUT (available for a fee; http://www.lionbioscience.com/). He created a graphical 'comparative neighborhood view' in which 'clusters of orthologous groups' (or COGs) in individual genomes are selected and interrelated with each other by comparing their sequences, relative genomic map positions, and so on. In addition, a graphical view of metabolic pathways of each organism can also be compared with the aid of the 'KEGG (Kyoto encyclopedia of genes and genomes) database' [http://www.genome.ad.jp/kegg/] by adding organism-specific information. Yasubumi Sakakibara (Tokyo Denki University, Saitama, Japan) talked about the results of their experimental-stage trial of so-called 'DNA computing', in which parallel computing was conducted through DNA-DNA hybridization and subsequent analysis using what the authors term 'intelligent DNA chips' with which simple logical operations such as "if A is true and B is false, then..." can be executed. The idea seems very interesting, but whether it will become a practical computing system remains to be explored further.

Biologists are increasingly confronted with floods of information, so a system with which they can extract useful information for their research purposes with minimal effort is an attractive idea. One of the growing trends in this workshop is to devise a system to extract information from the literature. Miyako Tanaka (Ube National College of Technology, Japan) reported on their system to survey topics in the MEDLINE database by implementing a 'characteristic words'-learning mechanism aided by experts' assistance and a 'keyword recommendation' step in which appropriate keywords are provided to reduce the number of unwanted papers. These human intervention steps are efficiently integrated to make the system cleverer. Several posters reported on similar 'intelligent' information-retrieval systems. The systems appeared to be still premature, however, and none seemed to have reached the stage to be of considerable help to biologists.

Expression profile analysis

The availability of genome sequence data has led to the systematic analysis of the expression of genes and likely genes (ORFs) of various organisms at the transcriptional and translational levels; such analysis has included the use of DNA microarrays. One of the immediate problems is how to process the massive quantities of microarray data. Hirohisa Kishino (University of Tokyo, Japan) and Peter J. Waddell (Chugai Research Institute of Molecular Medicine, Ibaraki, Japan) reported on one of the first demonstrations of graphical modeling of two-dimensional correspondence relationships of genes and tissues obtained from DNA microarray data. They intend to analyze DNA microarray data to find genes whose expression is either enhanced or repressed in certain tissues under certain conditions such as cancer. Unlike other methods in which clustering algorithms are mainly used to correlate gene expressions and tissue types, gene expression patterns are correlated in different tissues based on the calculation of 'partial correlation' of genes and tissues in graphical modeling. Importantly, they have introduced two multiple-regression procedures to overcome the initial problems associated with the graphical modeling method. The paper was awarded this year's best by the program committee. Different approaches to the analysis of DNA microarray data were the subject of other talks and many posters.

Protein structure and interactions

In an invited lecture, Shigeyuki Yokoyama (RIKEN and University of Tokyo) reported on the structural genomics initiative at RIKEN [http://www.rsgi.riken.go.jp/]. The project, directed by Yokoyama, is supported by a huge grant from the Japanese government. Its aim is to systematically characterize 1,000 to 2,000 protein folds that are considered fundamental to protein structure and function and would represent those present in 10,000 to 20,000 protein families. The project will be achieved by using arrays of powerful nuclear magnetic resonance (NMR) machines (600, 800 and 900 MHz) and X-ray crystallography, and by forming an international consortium including groups in Canada, Germany, Japan, USA and France. The results obtained will be published in an electronic journal named the 'Journal of Structural and Functional Genomics' [http://www.kihara.or.jp/jsfg/]. The project has just started and includes the structural analysis of the entire proteins of a thermophilic bacterium *Thermus thermophilus* as well as proteins predicted from the genome analysis of mouse, human, *Caenorhabditis elegans* and *A. thaliana*.

The concept introduced by Yokoyama is in a sense similar to the idea of determining the genomic nucleotide sequence without paying attention to individual genes. Keith Dunker (Washington State University, USA) argued that, as quite a high percentage of proteins in prokaryotes, and more notably in eukaryotes, contain stretches that are in an 'intrinsic disorder' state, analysis of the representative protein folds would not help to solve the structure-function relationships in many cases. Dunker mentioned that some proteins do not form the ordered structure until they interact with substrates, cofactors, and so on, and that the percentage of disordered regions in proteins of eukaryotes is as high as 30% on average. He proposed that function can arise from any of the three protein

states - ordered, molten globule and random coil - as well as transitions between them. Whether the project mentioned by Yokoyama will be successful in solving structure-function relationships of proteins in general remains to be seen, but Dunker's talk implies that the analysis would be more complicated than perhaps imagined.

The analysis of biological structures using a newly developed algorithm was described by Kiyoshi Asai (Electro-technical Laboratories, Tsukuba, Japan). The algorithm is used for processing images such as those obtained from electron microscopy, in particular cryo-electron microscopy in which the orientation of the object particles is very random. It is based on single-particle analysis and performs iterative alignment of images in a reference-free manner. Images clustered in bottom-up and top-down clustering procedures are subject to iterative filtering such as removing one image and calculating its effect on the average of the remaining images of the cluster. The procedure reported appeared promising and should, therefore, be applied to electron micrographs of different structures of biological interest.

The meeting was a success, but it was sometimes made apparent that lack of biological knowledge can lead to misinterpretation of the results of computer analyses. The original aim of the meeting to promote the exchange of data and ideas between informaticists and experimental biologists has, however, been steadily realized in the workshop year after year, although the number of experimental biologists who attend the meeting is still far too few.

Acknowledgements

I am grateful to Hirotada Mori (Nara Advanced Institute of Science and Technology, Japan) and Hideo Matsuda (Osaka University, Japan) who helped me in writing this meeting report.