The *Aspergillus oryzae* Project

--- Our Efforts to analyze *Aspergillus oryzae* Genome ---

The filamentous bacterium *Aspergillus oryzae* is one of the most important microorganisms used in food production in Japan. It has been used in the traditional fermentation industry to produce sake, miso and soy sauce for more than a thousand years.
The *Aspergillus oryzae* EST Analysis Consortium, consisting of national research institutes, universities and companies in Japan, had analyzed more than 17,000 EST sequences between the time it was founded in 1998 and by the year 2000. In 2001, the *Aspergillus oryzae* Genome Analysis Consortium was founded with the aim of decoding the entire *Aspergillus oryzae* genome. CBRC has been in charge of sequence analysis, decoding entire genes in the analyzed draft genome sequences as they became available. By the second quarter of 2003, we had elucidated more than 95% of the *Aspergillus oryzae* genome and predicted all of the gene sets for the draft genome sequences. Here we will outline of our efforts in the proposed project.

Members of CBRC: Kiyoshi Asai, Taishin Kin, Toshitaka Kumagai, Goro Terai (Mathematical Model Team), Hideki Nagasaki (Algorithm Team)

Techniques for finding genes

The *Aspergillus oryzae* genome presumably has genes that are highly homologous with protein sequences that are already known, such as those of the yeast or *Neurospora crassa*, and genes that are unique and difficult to find by searches for homology in databases containing known protein sequences. Given that it is usually more difficult to predict sequences of the latter approach, we decided to extract the homologous gene sequences of known proteins first before extending our search to unique genes elsewhere in the genome.
We first identified the homologous genes that code for known proteins "buried" in the *Aspergillus oryzae* genome using software such as BLAST and ALN(1). ALN was developed mainly by CBRC for aligning DNA and amino acid sequences of the eukaryote genome. We used the GeneDecoder (2) software developed by CBRC and GlimmerM (3) from TIGR for finding the genes unique to *Aspergillus oryzae* in those domains that were unpredictable using ALN. We integrated all of the predicted gene sequences such that they did not conflict with each other and identified approximately 14,000 gene candidates.

Annotation of the discovered genes

We then worked on annotating the predicted gene sequences by predicting function and classifying genes. We referred to the COG (Cluster of Orthologous Groups) of the NCGI for classification and assigned COG function codes to each of the predicted genes. We also developed a tool that enabled us to graphically compare these sequences against those in the COG of model organisms such as yeast and *Caenorhabditis elegans*, which have been more extensively analyzed. Intensive draft annotation was undertaken in the 4th quarter of 2003 based on those approaches by the filamentous bacterium research scientists in Japan and a more detailed analysis of each predicted gene is currently underway.
In 2003, it was decided to share the sequence information and analyzed data with the Sanger Centre, TIGR and the Whitehead Institute/MIT who had previously analyzed related filamentous bacteria such as *Aspergillus nidulans* and *Aspergillus fumigatus*, to further promote analysis of the Aspergillus genome internationally. A much clearer understanding of *Aspergillus oryzae* at the level of the genome is likely in the coming months and years.

(Toshitaka Kumagai)

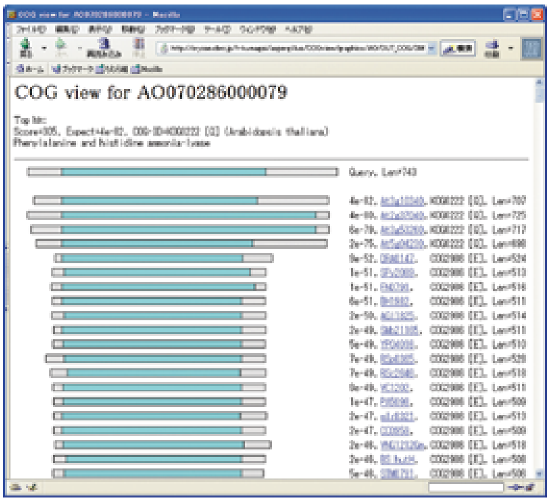

Fig. Comparison between the predicted gene sequences of *Aspergillus oryzae* and other model organisms

(1) Gotoh, O.: "Homology-based gene structure prediction: simplified matching algorithm using a translated codon (tron) and improved accuracy by allowing for long gaps", *Bioinformatics*, **16**, pp.190-202 (2000).
(2) Asai, K., Itou, K., Ueno, Y., Yada, T.: "Recognition of human genes by stochastic parsing", *Pac Symp Biocomput*., pp.228-239 (1998).
(3) Majoros, WH., Pertea, M., Antonescu, C., Salzberg, SL.: "GlimmerM, Exonomy and Unveil: three ab initio eukaryotic gene finders", *Nucleic Acids Res.*, **31**, pp.3601-3604 (2003).