Data analysis and presentation of large-scale nucleotide sequence information

Yasukazu Nakamura

Doctor of Science

Department of Genetics
School of Life Science
The Graduate University for Advanced Studies

2000

Contents

Summary 3

Chapter 1 Codon-anticodon assignment and detection of codon usage trends in microbial genomes 9

1-1 Introduction 10

1-2 Materials and methods 12

1-3 Results and discussion 16

1-4 Figures and Tables 22

Chapter 2 Construction of the WWW databases for the complete nucleotide sequence of the genome of *Synechocystis* sp. strain PCC6803 38

2-1 Introduction 39

2-2 Material and Methods 40

2-3 Description of the WWW sites 41

2-4 Figures and Tables 47

Chapter 3 High-throughput genome annotation of *Arabidopsis thaliana* 55

3-1 Introduction 56

3-2 Materials and methods 60

3-3 Results and discussion 62

3-4 Figures and Tables 67

References 85

Acknowledgements 94

Summary

Rapid, automated sequencing technologies with related advances in computational analysis and informatics have transformed the nature of biological research. The huge amounts of sequence data challenge the scientific community to understand and use this new information effectively. In this thesis, I describe the construction of data analysis and presentation systems for sequence information; these systems will assist in the identification of gene and implementation of a high-throughput genome sequencing era.

Chapter 1

CUTG (codon usage tabulated from GenBank) is a comprehensive database of codon usage. To generate an electronic data set for codon usage for each gene and for codon choice trends in each genome, Ikemura *et al.* have compiled codon usage in genes encoding proteins contained within the international DNA sequence database. The data files are available on ftp sites at Kazusa DNA Research Institute, National Institute of Genetics and European Bioinformatics Institute.

The compilation is synchronized with major releases of GenBank. The latest data source available during the preparation of this thesis was NCBI-GenBank Flat File Release 120.0. The frequencies of each of the 382,241 complete protein-coding sequences (CDSs) was compiled from the taxonomic divisions of the DNA sequence database. The sum of the codons used by 11,388 organisms has also been calculated. A list of the codon usage of genes and the sum of the codons used by each organism can be viewed at http://www.kazusa.or.jp/codon/. A new WWW interface has been developed to provide data in a format compatible with that of the CodonFrequency output in the GCG Wisconsin

Package™. Also, for each species, there is a query box to search for information in the comments for each gene. The user can choose CDSs by keyword and then generate codon usage tables from the selected genes. This tool provides researchers with the ability to examine intra-species variations in codon usage.

As an application of codon usage-based analysis for microbial genome sequencing efforts, I used only the sequences of the ribosomal protein genes as standards for calculation when I performed a modified codon adaptation index (CAI) analysis. This is in contrast to the traditional method of analysis, which relies on prior knowledge of the sequences of the most highly expressed genes.

To begin, I tabulated the patterns of codon-anticodon recognition in the following microorganisms whose genomes have been sequenced completely: *Haemophilus influenzae* Rd, *Methanococcus jannaschii,* and *Synechocystis* sp. strain PCC6803. For *Escherichia coli*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae,* and *Saccharomyces cerevisiae,* the previously adopted codon-anticodon combination was used.

I then used a modified CAI (Sharp and Li, 1987) as a measure of synonymous codon bias. The original CAI value for each gene was measured with the codon preferences of the genes for highly expressed proteins such as ribosomal proteins and elongation factors, as a basis. To generalize this method to organisms for which only sequence information exists, I modified the procedure of extraction by simply taking into account the sequences of the ribosomal protein-coding genes, and the codon usage biases of the ribosomal protein genes of each of the seven microbial genomes was recalculated. With these values, CAI_{rp}, a CAI that depended on the codon biases of the ribosomal protein genes, was calculated for all of the protein-coding genes of the genome.

Of the seven genomes examined, a clear correlation between the CAI_{rp} score and

the level of protein-coding gene expression was observed for all but the genes of *M. genitalium*. For the six genomes, elongation factors, and chaperonins, and ribosomal proteins had high CAI scores In contrast, genes for transposases and genes of prophage origin, which are expressed at lower levels, had the lowest CAI scores. This result indicates that codon usage analysis based on ribosomal protein gene sequences may be useful for predicting the expression levels of unknown genes. This method would be particularly useful for microbes where the entire genomes is being sequenced, since the DNA sequences of most, if not all, genes would be available.

Chapter 2

A WWW database system that provides information for deduced protein-coding genes was constructed for the cyanobacteria sequencing project.

Cyanobacteria are prokaryotic microorganisms that carry a complete set of genes for oxygenic photosynthesis. In 1996, Kaneko *et al.* reported the complete 3.57 megabase (Mb) sequence of the genome of *Synechocystis* sp. strain PCC6803, which contains 3,168 potential protein-coding genes.

CyanoBase (http://www.kazusa.or.jp/cyano/) is an online resource for accessing genomic data for the cyanobacterium. The core portion of CyanoBase contains annotations for each of the 3,168 protein-coding genes deduced from the entire nucleotide sequence of the *Synechocystis* sp. strain PCC6803 genome. The annotation for each protein-coding gene is accessible through three menus on the main page of this database: map image, gene classification lists, and keyword and similarity search engines. The aim of this database is to provide detailed information on potential protein-coding genes through a user-friendly interface that includes clickable genome maps and a hypertext classification list.

The database also contains repository facilities that store and offer experimental information and proposed function of each gene. Of the 3,168 deduced genes on the *Synechocystis* genome, 1,722 are annotated as functionally unassigned, which included 1,270 putative genes, 418 genes similar to hypothetical ones, and 34 genes similar to expressed sequence tags (ESTs) of other genomes. To analyze the functions of these genes, systematic disruption of each gene and characterization of the resulting mutants is thought to be a promising strategy.

CyanoMutants (http://www.kazusa.or.jp/cyano/mutants/) is a cumulative database that allows users to stores and access mutant information through the WWW. Each entry in CyanoMutants contains three sections: identification of the mutated gene, information about the phenotype, and person to whom correspondence should be addressed. Each entry is linked to the corresponding annotation in CyanoBase. The corresponding page in CyanoBase contains a link to the page in CyanoMutants that provides mutant information. These linked information will prevent unnecessary overlaps in experiments and promote communication among scientists to elucidate the functions of putative genes in cyanobacteria.

As of December 2000, CyanoMutants contained 431 mutant entries, 134 of which have phenotype description. The number of genes registered is expected to increase continuously since a large number of gene disruption experiments have been carried out since the release of the genomic sequence of *Synechocystis* sp. strain PCC6803.

Chapter 3

A protocol to automate the execution of similarity searches and gene prediction programs was developed for the *Arabidopsis thaliana* genome sequencing project. High-

throughput annotation of 27 Mb genomic sequences of *A. thaliana* has been carried out with the assistance of the system.

The 125 Mb genome of *A. thaliana* is organized into five chromosomes and contains an estimated 25,500 genes. To understand the entire genetic system in this plant, an international sequencing project of the *A. thaliana* genome has been initiated 1996, and currently it is in completion phase. Our research group is participating in sequencing the entire bottom arm and portions of the top arm of chromosome 5 and also the top arm of chromosome 3. During the process of annotating the genomic sequences of clones on the chromosomes, I have constructed a computer-aided system for high-throughput gene identification.

In this system, nucleotide sequences are translated in six frames with use of the universal codon table, and each frame is subjected to a similarity search against the non-redundant protein database, nr, with use of the BLAST program. Each local alignment, that shows an E-value < 0.001 to known protein sequences, is extracted and stored. Potential exons for protein-coding genes are predicted with the computer programs Grail and GENSCAN. For localization of exon-intron boundaries, donor/acceptor sites for splicing are predicted by NetGene2 and SplicePredictor. To identify transcribed regions and structural RNA genes, the BLAST program is used to compare nucleotide sequences with the EST and RNA gene data sets. For assignment of transfer RNA genes and transfer RNA structures, tRNA-scanSE is used.

All outputs are then parsed and stored in the General Feature Format (GFF). When required, the results are parsed and loaded into a WWW-based information display system called *Arabidopsis* Genome Displayer. This display system shows the positional relation of genome features along a genomic sequence. Simultaneously, an annotation

composing interface allows manual editing of the gene model showing tentative nucleotide and protein sequences and exon-intron organization. The annotator performs similarity searches as needed on the working model during the gene-modeling process. After careful editing, the most reasonable model of a genomic region is saved in the in-house database as a deduced gene.

In conclusion, 6,124 potential protein-coding genes were assigned to the 27 Mb regions of *Arabidopsis* chromosomes 3 and 5 covered by 461 clones and gap-closing units. The average density of genes was estimated to be 1 gene per 4.4 kb. One hundred twenty-seven RNA genes were deduced by similarity searches and computer predictions. Of 6,124 deduced protein-coding genes, 2,808 carried EST sequences, indicating that 46% of the total genes in *A. thaliana* may be represented in the current EST databases.

Chapter 1

Codon-anticodon assignment and detection of codon usage trends in microbial genomes

1-1 Introduction

1-1-1 Codon usage tabulated from the international DNA databases

The choice among synonymous codons within a genome is not random. After analyzing the codon usage patterns of 90 protein-coding genes in various organisms, Grantham *et al.* proposed in 1980 the "genome hypothesis" which suggested the existence of a consistent and genome-specific trend in codon usage. Among genes in each unicellular organism, there is a major trend of codon choice patterns, regardless of gene function; "codon dialect" found for individual unicellular organisms (Ikemura, 1985). By measuring the transfer RNA content of a cell, Ikemura revealed that the codon usage trend is highly correlated to the isoaccepting transfer RNA population of individual organisms in *Escherichia coli* (Ikemura, 1981a) and *Saccharomyces cerevisiae* (Ikemura, 1982).

It has also been found that the extent of codon bias for each gene is related to the protein production level of each gene (Ikemura, 1981b; Bennetzen and Hall, 1982). Codon usage in genes encoding abundant proteins is almost always much more dependent on transfer RNA content (strong accent) than that in moderately or poorly expressed genes (moderate accent). It has also been found that foreign-type genes such as those of transposons, plasmids and viruses often have different codon usage patterns from the respective host dialect.

In higher organisms, such as mammals, codon usage among genes is highly variable. Codon choice patterns mainly reflect the G+C content of the whole genome or local characteristics, namely GC mosaic or isochore (Ikemura, 1985; Bernardi *et al.*, 1985). Research of the intra-species variations of codon usage may provide an interesting line of investigation regarding the evolution of the genome.

To calculate codon usage for each gene and trends of codon choice for each genome, in 1986 it was began to compile codon usage of protein-coding genes contained within the

international DNA sequence database (Maruyama *et al.*, 1986). The database has been called CUTG (Codon Usage Tabulated from GenBank). The basic aim of the database is to provide an electronic data set for codon usage-based analyses. Since each codon usage for a protein-coding gene is compiled as a simple double-lined entry, it is easy to import worksheets or to parse and calculate with computer languages such as C or Perl.

1-1-2 Codon-anticodon assignment and detection of codon usage trends in seven microbial genomes

On the basis of sequence information derived from the large scale sequencing of entire genomes (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995; Bult *et al.*, 1996; Kaneko *et al.*, 1996; Goffeau *et al.*, 1996; Himmelreich *et al.*, 1997), the transfer RNA gene complements of various genomes have been assigned based on sequence similarity to known transfer RNA sequences or their genes or by prediction using computer programs, such as tRNAscan-SE (Lowe and Eddy, 1997). However, to determine the actual relationship between transfer RNA molecules and the codons used to encode the amino acid sequences of an organism's entire complement of protein-coding genes, a comprehensive analysis of the codon usage pattern has to be carried out using a codon-anticodon table created for each organism.

To date, the recognition patterns between the codon and the corresponding anticodon have been reported for four organisms. In *E. coli*, an exhaustive survey of transfer RNA genes was performed and the pattern of codon-anticodon recognition has been studied intensively (Komine *et al.*, 1990). Andachi *et al.* (1989) sequenced 29 transfer RNA species that represent the entire transfer RNA complement of the organism *Mycoplasma caplicolum* and deduced a codon recognition pattern that turned out to be similar to that of mitochondria. They attributed this similarity to two types of evolutionary constraint, the small size and high AT content of the *M. caplicolum* genome. The same recognition pattern was observed in *Mycoplasma pneumoniae* (Simoneau *et al.*, 1993). In *S. cerevisiae*, 274 transfer RNA genes were extracted from the complete genome sequence

(Percudani *et al.,* 1997), and they were assigned to 42 classes of distinct codon specificity.

To quantify the bias in synonymous codon usage, several measures have been proposed. The frequency of optimal codons (F_{op}) was adopted as an index of translational efficiency in *E. coli* and *S. cerevisiae* (Ikemura, 1981b, 1982). The optimal codon designates the most preferred codon estimated from the cell's content of isoaccepting transfer RNA and its preference for the codon-anticodon interaction. Another measure of synonymous codon usage bias is the codon adaptation index (CAI) (Sharp and Li, 1987). CAI indicates the degree of adaptation to the codon preference pattern of the highly expressed genes in a given species, which has been shown to be mostly optimized for translation. Although, both F_{op} and CAI rely on the same biologic phenomena, CAI is more suitable for an automated computational approach because its derivation requires only the DNA sequences of genes.

I tabulated the patterns of codon-anticodon recognition in seven microbial organisms whose genomes have been completely sequenced. Using these codon-anticodon tables, I performed a novel CAI analysis employing only the sequences of the ribosomal protein genes as a standard for calculation instead of the usual method, which relies on prior knowledge of the sequences of the most highly expressed genes. I examined the potentiality of this simplified method for organisms for which only sequence information was available.

1-2 Materials and methods

1-2-1 Compilation

The codon usage in individual protein-coding genes has been calculated using the nucleotide sequence obtained from GenBank sequence database. The compilation is synchronized with major release of GenBank. Sum of the codon use of protein-coding genes for each organism has also been calculated. Divisions `pri` (primate sequence

entries), rod (rodent sequence entries), mam (other mammalian sequence entries), vrt (other vertebrate sequence entries), inv (invertebrate sequence entries), pln (plant sequence entries), bct (bacterial sequence entries), vrl (viral sequence entries) and phg (phage sequence entries) were used. Other files such as est (expressed sequence tag sequence entries), pat (patent sequence entries) and rna (structural RNA sequence entries), for example, were not used, since they were not taxonomical collections and consisted of only a small number of full-length protein-coding genes. Partially sequenced protein-coding genes were not compiled in this database. Codons containing ambiguous bases were excluded from the compilation.

1-2-2 Description of CUTG files

Files named gb***.codon list the codon usage of each gene registered in the selected GenBank Flat Files. The LOCUS names given in GenBank were used to designate individual genes. Each LOCUS name is followed by fields of information extracted from the FEATURES part of the CDS used to define the open reading frames analyzed here. The order of the codons in the table is described in CODON_LABEL file on the ftp site.

To reveal the characteristics of the codon usage of a wide range of organisms, as well as viruses and organella, the frequency (per thousand) of codon use in each organism was calculated by summing up the numbers of codons used. Files named gb***.spsum list the sum of numbers of codon usage in each species as well as in viruses and organella.

1-2-3 Codon-anticodon assignment for the microbial genomes

The sequence files of the genomes of *Haemophilus influenzae* Rd, *Methanococcus jannaschii,* and *Synechocystis* sp. PCC6803 were retrieved from GenBank (ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria/), and information on the transfer RNA genes was extracted according to the feature tables. Based on the nucleotide sequences of

the complete set of transfer RNA genes deduced from the complete sequences of seven microbial genomes, information about codon - anticodon recognition was deduced and tabulated. For *E. coli, M. genitalium, M. pneumoniae* and *S. cerevisiae,* the previously adopted codon-anticodon combination was employed (Komine *et al.,* 1990, Simoneau, *et al.,* 1993, Percudani *et al.,* 1997).

1-2-4 Codon usage table construction for the microbial genomes

The codon usage of the entire protein-coding gene complement of seven microbial genomes was calculated and re-compiled using the programs for CUTG (Nakamura *et al.,* 1997a, 1998b, 1999b, 2000b). The annotated flat sequence files of the genomes of *E. coli* (accession number: U00096), *H. influenzae* Rd (L42043), *M. jannaschii* (L77117), *M. genitalium* (L43967), and *M. pneumoniae* (U00089) were retrieved from the repository of NCBI/GenBank, and that of *Synechocystis* sp. PCC6803 was taken from CyanoBase (http://www.kazusa.or.jp/cyano/). The nucleotide sequences of genes coding for proteins were extracted from the genome sequences using the CDS description in the feature tables. CDS's annotated as either `partial' or `pseudo', or those containing any ambiguous bases were excluded.

The protein-coding gene sequences of *S. cerevisiae* were retrieved from SGD (*Saccharomyces* Genome Database, http://genome-www.stanford.edu/saccharomyces/). The numbers of full-length protein-coding genes used for the codon usage analysis were 4,283, 1,717, 1,680, 467, 677, 3,166, and 6,217 for *E. coli, H. influenzae* Rd, *M. jannaschii, M. genitalium, M. pneumoniae, Synechocystis,* and *S. cerevisiae,* respectively, and are listed in Table 1-1. To evaluate the codon usage in the ribosomal protein genes, the nucleotide sequences annotated as "a ribosomal protein gene" were extracted from the flat sequence files of the six bacterial genomes. For *S. cerevisiae,* the DNA sequences of the ribosomal proteins were extracted based on information from the ORF table constructed by MIPS (ftp://ftp.mips.embnet.org/yeast/tables.dir/mips_orfs_table.ascii). The genes which were annotated as mitochondrial ribosomal protein genes were excluded, and the redundancy of

multi-copy genes was removed. The numbers of ribosomal protein gene sequences extracted and used for the codon usage analysis were 55, 54, 59, 46, 41, 53, and 60 for *E. coli, H. influenzae, M. jannaschii, M. genitalium, M. pneumoniae, Synechocystis,* and *S. cerevisiae,* respectively, and are listed in Table 1-2.

1-2-5 Codon usage bias analysis of the microbial genomes

The relative adaptiveness of each codon to the ribosomal protein genes, W_{rp}, was calculated according to the method by Sharp and Li (1987). W is the frequency of use of a codon compared to that of the most frequently used codon for an amino acid. Thus,

$$W_{ij} = \frac{X_{ij}}{X_{imax}}$$

where X_{ij} is the number of occurrences of the jth codon for the ith amino acid, and X_{imax} is the X values for the most frequently used codon for the ith amino acid.

As illustration, in the leucine codon box for *E. coli* ribosomal proteins, (see Table 1-4), W_{rp} of CUG (the most frequently used codon) is calculated as:

$$W_{LeuCUG} = \frac{52.7}{52.7} = 1.000$$

W_{rp} of CUA is the frequency of use of that codon compared to the frequency of the optimal codon for that amino acid (CUG):

$$W_{LeuCUA} = \frac{3.9}{52.7} = 0.002$$

CAI_{rp}, the codon adaptation index (ribosomal protein), for each of the full length protein genes was then calculated as the geometric mean of W_{rp} values for each genome.

$$CAI = \left(\prod_{k=1}^{L} W_k \right)^{\frac{1}{L}}$$

In order to avoid underflow, this equation was computed as follows.

$$CAI = \exp\left(\frac{1}{L}\sum_{k=1}^{L}\ln W_k\right)$$

where W_k is the W value for the kth codon in the gene and L is the number of codons in the gene. The AUG and UGG codons were not taken into account, because they do not have synonymous codons and did not contribute to the codon bias measurement, except for those in two *Mycoplasma* species, where the UGG codon is translated as tryptophan.

1-3 Results and discussion

1-3-1 Codon usage database

The latest data source available during the preparation of this thesis was NCBI-GenBank Flat File Release 120.0 (15 October 2000). The frequencies of each of the 382,241 complete protein coding sequences (CDS's) have been compiled from the taxonomical divisions of the GenBank DNA sequence database. The sum of the codons used by 11,388 organisms has also been calculated. The data files are available by anonymous ftp from Kazusa DNA Research Institute, National Institute of Genetics and European Bioinformatics Institute sites (Table 1-3).

The database can be easily accessed through the World Wide Web server (Fig. 1-1) which provides a user-friendly interface for interactive process. A list of the codon usage of genes and the sum of the codons used by each organism can be viewed at http://www.kazusa.or.jp/codon/. Access statistics logged the web site for CUTG has been accessed average 15,000 per month from all over the world.

The database displays codon usage in a format compatible with that of CodonFrequency output in the GCG Wisconsin Package™. Thus, users, who have the GCG package in their local environment, can do further analyses with the files generated by the database. Also, for each species, there is a query box to search for information in the

comments for each gene. The user can choose CDSs by keyword and then generate codon usage tables from the selected genes. This tool provides researchers with the ability to examine intra-species variations in codon usage. For example, protein production levels can be predicted from the complete genome sequences of microbes using the codon usage biases compiled from ribosomal protein genes as described in this part.

1-3-2 Codon-specificity of the transfer RNA complements of seven microbial genomes

On the basis of structural information derived from the entire protein and transfer RNA gene constituents of the genomes of *H. influenzae, M. jannaschii,* and *Synechocystis,* the recognition patterns of the codons and the corresponding anticodons were deduced. The results are listed in Table 1-4 along with those reported for *E. coli,* two *Mycoplasma* species, and *S. cerevisiae.* The number of transfer RNA genes present in the entire genome of each organism and the degree of representation of the different classes of anticodon are summarized in Table 1-1.

In addition to the reported 56 transfer RNA genes representing 33 species of distinct anticodon, a gene for selenocysteine tRNA-UCA was newly identified in the *Haemophilus* genome. The gene lies between positions 753,200 - 753,290 on the complementary strand and was identified through its similarity to *E. coli* Sec-tRNA and by a computer search using the tRNAscan-SE program. Out of a total of 57 transfer RNA genes assigned to the *Haemophilus* genome, no gene for tRNAP-GGG was uncovered even after careful reexamination of the registered sequence data. This might suggest that the CCY codon is recognized by tRNAP-UGG containing an unmodified U, as reported for mitochondria (Barrel *et al.,* 1980, Bonitz *et al.,* 1980, Heckman *et al.,* 1980) and *M. caplicolum* (Andachi *et al.,* 1989). Alternatively, some unknown type(s) of modification could confer ability to recognize the CCY codon of one of the two tRNAP-UGG species present in *Haemophilus.*

One of the characteristic features of the anticodon sets of *M. jannaschii, H.*

influenzae Rd and the two *Mycoplasma* species compared with those from *Synechocystis* and *E. coli* was that limited numbers of transfer RNA genes containing CNN anticodons were used. For example, *M. jannaschii* lacked transfer RNA genes with the anticodons L-CAA, L-CAG, P-CGG, T-CGU, A-CGC, R-CCU and G-CCC which *Synechocystis* possessed. In *M. jannaschii,* one third of the NNG codons seemed to be recognized only by transfer RNAs with UNN anticodons. One possibility is that AT pressure on the genomes of *M. jannaschii* and *M. genitalium* might resulted in a restricted use of NNG codons. This is especially evident in the case of codon boxes for S, P, T, and A, and this may underlie the reduction in the number of transfer RNA genes containing CNN anticodons. Another possibility is that the relatively large DNA contents of the genomes of *Synechocystis* and *E. coli* may have permitted the occurrence of transfer RNA molecules with nonobligate anticodons.

Both *M. genitalium* and *M. pneumoniae* use the same set of 33 classes of anticodon. They lack the transfer RNA genes for the recognition of codons, L-CUN, N-GUN, P-CCN, and A-GCN (Simoneau *et al.,* 1993, Bult *et al.,* 1996). It has been reported that this limited set of anticodons is sufficient for the recognition of the 62 codon classes of both *M. caplicolum* and *M. pneumoniae* (Andachi *et al.,* 1989, Simoneau *et al.,* 1993), and on this basis we derived the codon-anticodon table for *M. genitalium* shown in Table 1-4

1-3-3 Codon usage and its trend in seven microbial genomes

Based on the nucleotide sequences of all the protein-coding genes of the seven microbial genomes that have been completely sequenced, the codon usage frequency in each organism was calculated as described in the section 1-2-4, and the results are listed in Table 1-4. The frequency of codon usage of the ribosomal protein genes was calculated separately using the sequences of the genes listed in Table 1-2, and the results are shown in Table 1-4. Ikemura (1980b, 1981) reported the most preferred codons among the synonymous codons of *E. coli* and *S. cerevisiae* based on an analysis of their contents of isoaccepting transfer RNA molecules and on the nature of the codon-anticodon interaction. Using this approach, the optimal codons indicated by # in Table 1-4 can be seen to coincide

well with those that are preferentially used by ribosomal protein genes, which are highlighted in the table.

Sharp and Li (1987) proposed the codon adaptation index (CAI) as a measure of synonymous codon bias, using as a basis the codon preference of genes for highly expressed proteins, such as ribosomal proteins and elongation factors, whose high degrees of expression have been checked experimentally. Using this index, they demonstrated correlation between increasing codon bias and the increasing level of gene expression in *E. coli* and *S. cerevisiae.* To generalize this method to organisms for which only sequence information exists, we modified the procedure of extraction by simply taking into account uniquely the sequences of the ribosomal protein genes. In this way, we calculated an index for the codon bias of ribosomal genes. The codon usage bias in the ribosomal protein genes of each of the seven microbial genomes was recalculated according to the method of Sharp and Li (1987). The resulting W_{rp} value indicates the frequency of usage of a given codon compared to that of the most preferred codon among the synonymous set of codons for ribosomal protein genes. As shown in Table 1-4 (W_{rp} column), codon bias was evaluated according to the frequency of the usage, with the maximum value being set at 1. Using these values, a CAI, which depended on the codon bias of the ribosomal protein genes, was calculated for all of the protein-coding genes of the genome, and the values were designated as CAI_{rp} to avoid any confusion with the original term, CAI. Table 1-5 lists the 20 genes with the highest CAI_{rp} score and the 10 genes with the lowest score in each genome. Of the seven genomes examined, six possessed genes for proteins known to be abundant in the cells. Among these, ribosomal proteins, elongation factors, and chaperonins appeared in the highest 20. In *E. coli,* a close correlation between CAI_{rp} and the logarithm of protein content in rich medium (Van Bogelen *et al.,* 1992) was observed (Figure 1-2). In *Synechocystis,* several photosynthetic genes also gained high scores. In contrast, genes for transposases and genes of prophage origin, which were expected to be poorly expressed, were among the bottom 10. The result indicates that a CAI analysis based on ribosomal protein gene sequences may be useful for the prediction of the

expression levels of unknown genes.

In contrast to *M. pneumoniae,* a clear correlation between the CAI_{rp} score and the level of gene expression seemed to be absent in the case of the genes of *M. genitalium.* As both species use the same set of anticodons, the high AT pressure in the genome of *M. genitalium* may be responsible for the bias in codon usage.

The distribution of CAI_{rp} scores for the genes of each organism was in accordance with the copy number of the transfer RNA genes of their genomes (Table 1-1). *S. cerevisiae* contains 274 transfer RNA genes and showed a high CAI_{rp} score of 0.938 and a low score of 0.132, with a standard deviation (SD) of 0.122. In contrast, *M. genitalium* which contains 33 copies of transfer RNA genes, showed a high score of 0.827 and a low score of 0.581 with a standard deviation of 0.038. In *E. coli* and *S. cerevisiae*, the redundancy of a particular species of transfer RNA gene was found to directly result in the abundance of the corresponding transfer RNA molecule in the cell. Under such circumstances, it is plausible that the translation efficiency of genes is considerably influenced by the bias in codon usage, the effect of which would be exaggerated by variations in the copy numbers of the transfer RNA genes.

1-3-4 Conclusion

In the present study, I modified CAI analysis to take into account the DNA sequences of ribosomal protein genes instead of employing the usual method, which relies on an analysis of the experimentally derived expression levels of highly expressed genes. CAI_{rp} enables us to correlate the codon usage of genes with their protein production levels. I used this method as the basis for our calculations and demonstrated the feasibility of the procedure for genomes whose entire DNA sequences are known. Despite the shortcomings observed for genomes subjected to maintaining a small genome size or extreme GC content or both, this simplified method should be useful for the prediction of the expression levels of unknown genes. An estimation of the production level, along with other bioinformatic analyses, should provide information useful for the prediction of gene function and

designing biochemical experiments to determine function. This method would be particularly useful for microbes where the entire genomes is being sequenced, since the DNA sequences of most, if not all, genes would be available.

The codon-anticodon tables and the CAI_{rp} scores for the entire genes in seven organisms are available through the WWW at http://www.kazusa.or.jp/comparative/.

1-4 Figures and Tables

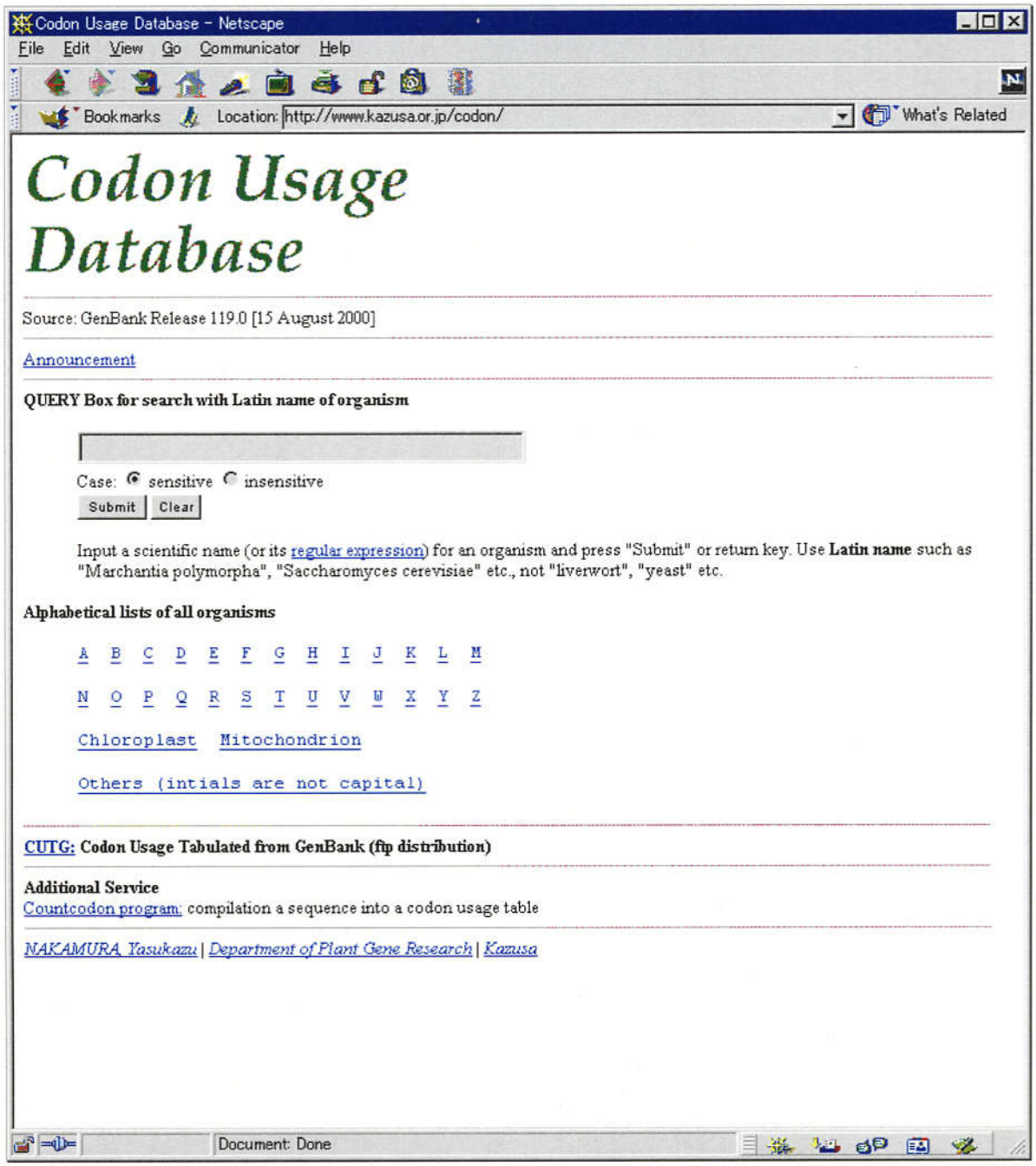


Figure 1-1 CUTG on the WWW

A user-friendly interface to use interactively with CUTG is to access the WWW server at Kazusa. A dataset for each organism is made seachable with its Latin name.

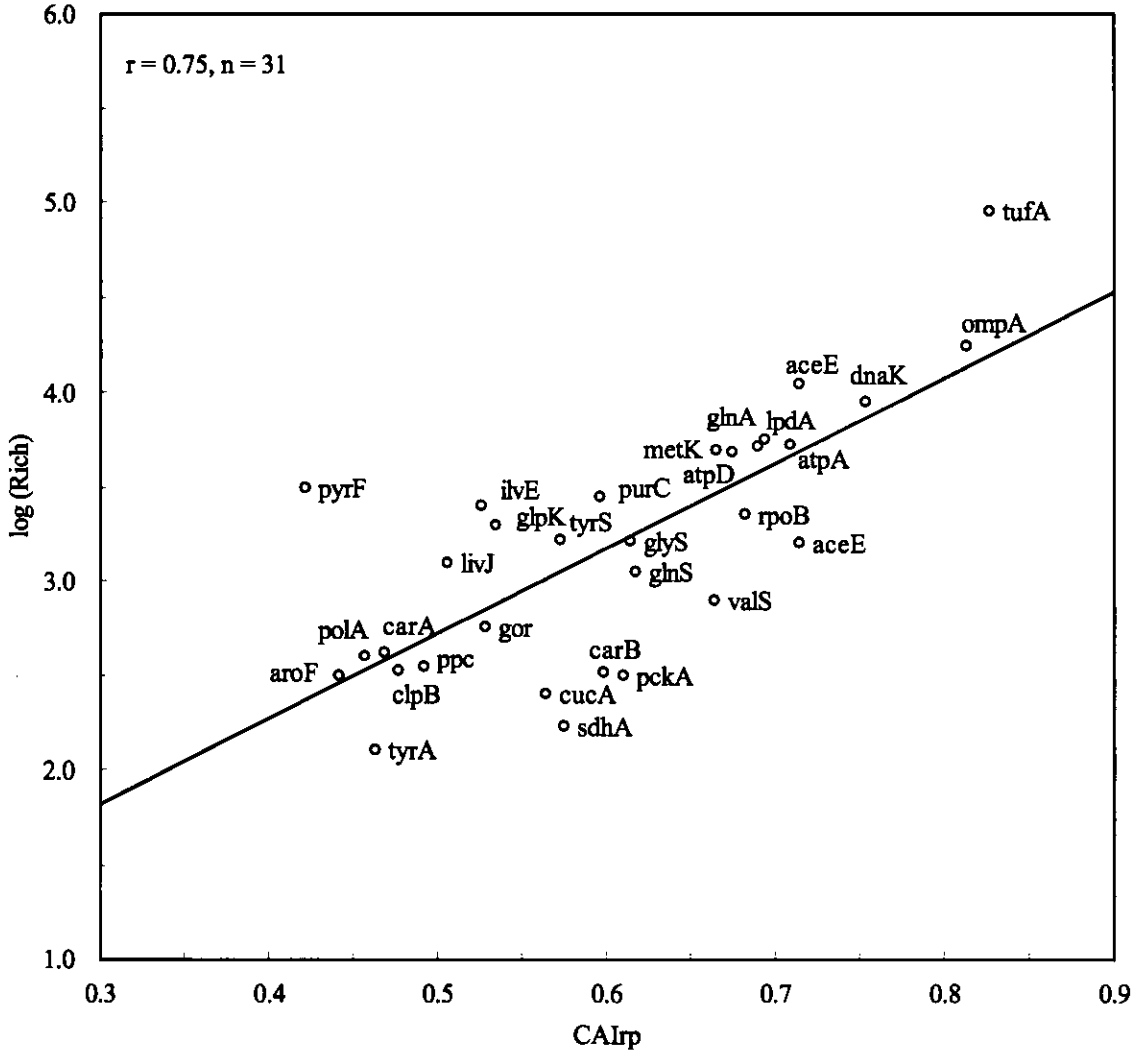


Figure 1-2 Correlation between CAI_{rp} and the amount of proteins in *E. coli* cells grown in a rich medium

Log (Rich) represents the common logarithm of the amount of protein molecules per genome in cells grown in rich medium, and r and n represent correlation coefficient and the number of samples, respectively.

Table 1-1 Number of proteins and transfer RNA genes and distribution of CAI_{rp} scores in seven organisms

Species	No. of protein genes	No. of tRNA genes	No. of anticodon classes	CAI_{rp}		
				minimum	maximum	SD
E coli [a].	4283	83	41	0.154	0.873	0.100
H. influenzae Rd [a]	1717	56	33	0.247	0.868	0.086
M. jannaschii	1680	37	35	0.396	0.853	0.058
M. genitalium	467	34	33	0.581	0.827	0.038
M. pneumoniae	677	34	33	0.438	0.809	0.044
Synechocystis PCC6803	3166	42	41	0.405	0.852	0.060
S. cerevisiae [a]	6217	274	42	0.074	0.929	0.122

[a]Genes for selenocysteine transfer RNA were excluded for those organisms

Table 1-2 List of ribosomal protein genes used for codon usage analysis

Species		Gene name	No. of genes
E. coli	*rpl*	A-F, I-Y	55
	rpm	A-J	
	rps	A-V	
H. influenzae Rd	*rpl*	1-6, 9-25, 27-36	54
	rps	1-21	
M. jannaschii	*rpl*	1-7, 11-14, 14B, 15, 15B, 18-19, 21-24, 24E, 29, 29E, 30-32, 34, 37, 37a, 40, 44, 46, X	59
	rps	3, 3a, 4, 4E, 5-8, 8E, 9-13, 15A, 17, 17B, 18-19, 19S, 24, 27, 27A, 33, HG12, HS6-type	
M. genitalium	*rpl*	1-6, 9-11, 13-24, 27-29,	46
	rps	31-362-14, 17-19	
M. pneumoniae	*rpl*	1-7, 9-11, 13-21, 23-24, 27	41
	rps	2-20	
Synechocystis PCC6803	*rpl*	1-6, 9-24, 27-29, 31-36	53
	rps	1-21	
S. cerevisiae	*rpl*	3-19, 21, 23-24, 26-27, 30-39, 44', 47	60
	rps	3-19, 21, 23-30	

Table 1-3 URL's for CUTG distribution

Site	URL
Kazusa (primary)	ftp://ftp.kazusa.or.jp/pub/codon/current/
NIG (mirror)	ftp://ftp.nig.ac.jp/pub/db/codon/current/
EBI (mirror)	ftp://ftp.ebi.ac.uk/pub/databases/cutg/

Table 1-4 Codon-anticodon recognition pattern and codon usage for the entire and ribosomal protein gene contents

Escherichia coli

		Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA
UUU	F	22.3	7.2	0.325		UCU	S	8.4	17.3	1.000		UAU	Y	16.2	4.3	0.323		UGU	C	5.2	1.2	0.333	
UUC	F	16.6	22.1 #	1.000	GAA 2	UCC	S	8.6	11.9	0.688	GGA 2	UAC	Y	12.2	13.3 #	1.000	GUA 3	UGC	C	6.5	3.7	1.000	GCA 1
UUA	L	13.9	2.1	0.036	UAA 1	UCA	S	7.2	1.1	0.064	UGA 1	UAA	*	2.0	7.1	-		UGA	*	0.9	0.6	-	
UUG	L	13.7	2.9	0.050	CAA 1	UCG	S	8.9	0.7	0.040	CGA 1	UAG	*	0.2	0.0	-		UGG	W	15.3	5.5	1.000	CCA 1
CUU	L	11.0	3.5	0.060		CCU	P	7.0	4.8	0.219		CAU	H	12.9	5.8	0.385		CGU	R	20.9	58.4 #	1.000	
CUC	L	11.1	2.8	0.048	GAG 1	CCC	P	5.5	0.4	0.019	GGG 1	CAC	H	9.8	15.1	1.000	GUG 1	CGC	R	22.0	27.1 #	0.464	ACG 4
CUA	L	3.9	0.1	0.002	UAG 1	CCA	P	8.4	4.4	0.200	UGG 1	CAA	Q	15.3	7.6	0.306	UUG 2	CGA	R	3.6	0.3	0.005	
CUG	L	52.7	58.0 #	1.000	CAG 4	CCG	P	23.2	22.1 #	1.000	CGG 1	CAG	Q	28.8	24.9 #	1.000	CUG 2	CGG	R	5.4	0.4	0.007	CCG 1
AUU	I	30.3	14.1	0.349		ACU	T	8.9	24.2 #	1.000		AAU	N	17.7	4.8	0.158		AGU	S	8.8	2.1	0.120	
AUC	I	25.1	40.4 #	1.000	GAU 3	ACC	T	23.4	22.0 #	0.909	GGU 2	AAC	N	21.6	30.7 #	1.000	GUU 3	AGC	S	16.0	11.6	0.672	GCU 1
AUA	I	4.4	0.3	0.007	CAU 1	ACA	T	7.1	1.9	0.080	UGU 1	AAA	K	33.6	67.0 #	1.000	UUU 6	AGA	R	2.1	0.7	0.012	UCU 1
AUG	M	27.9	24.4	1.000	CAU 3	ACG	T	14.4	2.5	0.103	CGU 1	AAG	K	10.3	28.0	0.417		AGG	R	1.2	0.0	0.000	CCU 1
	fM				CAU 3																		
GUU	V	18.2	48.0 #	1.000		GCU	A	15.3	50.7 #	1.000		GAU	D	32.1	16.9	0.616		GGU	G	24.7	47.7 #	1.000	
GUC	V	15.3	8.7	0.182	GAC 2	GCC	A	25.5	9.4	0.187	GGC 2	GAC	D	19.1	27.4	1.000	GUC 3	GGC	G	29.6	30.4 #	0.638	GCC 4
GUA	V	10.9	24.9 #	0.519	UAC 5	GCA	A	20.1	29.1 #	0.579	UGC 3	GAA	E	39.4	48.4 #	1.000	UUC 4	GGA	G	8.0	0.7	0.014	UCC 1
GUG	V	26.4	11.8 #	0.245		GCG	A	33.7	17.7 #	0.353		GAG	E	17.8	16.2	0.334		GGG	G	11.1	1.0	0.020	CCC 1

Genomic GC% 50.8, GC% for coding region 51.8, GC% for 3rd letter of codon 55.9

***Haemophilus influenzae* Rd**

		Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA
UUU	F	32.2	10.5	0.568		UCU	S	16.7	19.1	1.000		UAU	Y	24.3	9.0	0.941		UGU	C	6.8	3.4	1.000	
UUC	F	12.3	18.5	1.000	GAA 1	UCC	S	4.4	1.0	0.051	GGA 1	UAC	Y	6.9	9.5	1.000	GUA 1	UGC	C	3.5	1.3	0.375	GCA 1
UUA	L	49.9	45.1	1.000	UAA 2	UCA	S	12.6	12.6	0.662	UGA 2	UAA	*	2.5	7.4	-		UGA	*	0.4	0.1	-	
UUG	L	18.3	7.6	0.168	CAA 1	UCG	S	4.1	0.6	0.029		UAG	*	0.5	0.0	-		UGG	W	11.2	5.6	1.000	CCA 1
CUU	L	20.1	11.9	0.265		CCU	P	12.4	9.3	0.500		CAU	H	13.4	7.0	0.476		CGU	R	23.8	56.1	1.000	
CUC	L	5.3	0.8	0.019	GAG 1	CCC	P	2.9	0.1	0.008		CAC	H	7.1	14.7	1.000	GUG 1	CGC	R	10.1	14.3	0.217	ACG 2
CUA	L	6.7	4.1	0.090	UAG 1	CCA	P	16.8	18.5	1.000	UGG 2	CAA	Q	38.7	31.6	1.000	UUG 2	CGA	R	5.3	0.7	0.011	
CUG	L	4.5	1.4	0.031		CCG	P	5.0	2.8	0.152		CAG	Q	7.5	3.2	0.102		CGG	R	1.2	0.1	0.002	CCG 1
AUU	I	50.5	25.5	0.871		ACU	T	16.1	33.7	1.000		AAU	N	36.5	14.2	0.631		AGU	S	12.0	6.0	0.316	
AUC	I	14.3	29.3	1.000	GAU 3	ACC	T	11.3	5.8	0.171	GGU 1	AAC	N	12.2	22.5	1.000	GUU 2	AGC	S	8.5	8.0	0.419	GCU 1
AUA	I	6.0	0.6	0.019	CAU 1	ACA	T	15.7	10.7	0.317	UGU 1	AAA	K	56.3	80.7	1.000	UUU 3	AGA	R	3.7	2.0	0.030	UCU 1
AUG	M	24.0	23.4	1.000	CAU 3	ACG	T	8.7	1.5	0.046		AAG	K	6.8	11.1	0.137	CUU 1	AGG	R	0.5	0.0	0.000	
	fM																						
GUU	V	20.7	48.1	1.000		GCU	A	21.2	38.9	0.772		GAU	D	42.2	29.8	1.000		GGU	G	28.7	59.8	1.000	
GUC	V	6.6	2.9	0.061	GAC 1	GCC	A	10.9	3.1	0.061	GGC 1	GAC	D	7.5	10.1	0.340	GUC 3	GGC	G	19.5	15.0	0.251	GCC 3
GUA	V	18.8	33.0	0.685	UAC 4	GCA	A	32.9	50.4	1.000	UGC 3	GAA	E	54.1	54.2	1.000	UUC 3	GGA	G	10.9	3.1	0.052	UCC 1
GUG	V	20.9	12.6	0.262		GCG	A	16.9	14.3	0.284		GAG	E	10.5	10.9	0.202		GGG	G	7.2	0.6	0.009	

Genomic GC% 38.2, GC% for coding region 38.8, GC% for 3rd letter of codon 29.1

Table 1-4 *continued*

Methanococcus jannaschii

		Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA
UUU	F	33.2	15.1	1.000		UCU	S	10.4	2.6	0.167		UAU	Y	33.3	14.1	0.943		UGU	C	8.6	8.3	1.000	
UUC	F	8.7	9.6	0.637	GAA 1	UCC	S	2.7	0.6	0.040	GGA 1	UAC	Y	9.9	15.0	1.000	GUA 1	UGC	C	4.1	2.7	0.324	GCA 1
UUA	L	52.3	49.6	1.000	UAA 1	UCA	S	14.4	15.4	1.000	UGA 1	UAA	*	2.7	7.1	-		UGA	*	0.4	0.1	-	
UUG	L	19.0	13.2	0.265		UCG	S	0.7	0.2	0.016	CGA 1	UAG	*	0.3	0.0	-		UGG	W	7.1	7.4	1.000	CCA 1
CUU	L	9.1	3.3	0.066		CCU	P	8.5	9.3	0.242		CAU	H	10.2	5.5	0.413		CGU	R	0.2	0.0	0.000	
CUC	L	3.2	2.7	0.054	GAG 1	CCC	P	1.3	0.6	0.016	GGG 1	CAC	H	4.1	13.3	1.000	GUG 1	CGC	R	0.1	0.5	0.006	ACG 1
CUA	L	8.4	1.3	0.027	UAG 1	CCA	P	22.6	38.3	1.000	UGG 1	CAA	Q	9.1	12.5	1.000	UUG 1	CGA	R	0.3	0.2	0.003	
CUG	L	2.2	0.9	0.017		CCG	P	1.4	0.5	0.013		CAG	Q	5.3	10.5	0.835		CGG	R	0.1	0.0	0.000	CCG 1
AUU	I	48.2	36.9	1.000		ACU	T	14.6	9.3	0.317		AAU	N	36.8	13.4	0.769		AGU	S	10.9	6.8	0.444	
AUC	I	10.6	12.5	0.340	GAU 2	ACC	T	4.3	3.7	0.125	GGU 1	AAC	N	15.5	17.4	1.000	GUU 1	AGC	S	5.3	4.6	0.302	GCU 1
AUA	I	45.4	26.8	0.726	CAU 1	ACA	T	19.9	29.2	1.000	UGU 1	AAA	K	72.4	78.3	1.000	UUU 1	AGA	R	27.8	79.4	1.000	UCU 1
AUG	M	22.1	26.1	1.000	CAU 1	ACG	T	1.7	0.2	0.008		AAG	K	30.9	41.8	0.533	CUU 1	AGG	R	9.9	8.0	0.101	
	fM				CAU 1																		
GUU	V	43.6	55.1	1.000		GCU	A	24.8	33.0	0.996		GAU	D	45.4	26.2	1.000		GGU	G	13.1	20.8	0.437	
GUC	V	4.7	7.7	0.139	GAC 1	GCC	A	5.5	1.9	0.059	GGC 1	GAC	D	9.6	10.0	0.381	GUC 1	GGC	G	4.3	1.7	0.036	GCC 1
GUA	V	14.8	14.7	0.268	UAC 2	GCA	A	22.6	33.1	1.000	UGC 1	GAA	E	51.5	54.2	1.000	UUC 1	GGA	G	36.3	47.6	1.000	UCC 1
GUG	V	5.7	2.8	0.051		GCG	A	2.4	1.5	0.044		GAG	E	34.9	27.3	0.503		GGG	G	10.4	7.4	0.156	

Genomic GC% 31.4, GC% for coding region 32.0, GC% for 3rd letter of codon 24.8

Mycoplasma genitalium

		Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA
UUU	F	52.6	28.8	1.000		UCU	S	12.4	9.1	0.557		UAU	Y	23.9	14.8	1.000		UGU	C	6.6	5.2	1.000	
UUC	F	8.2	3.8	0.134	GAA 1	UCC	S	4.0	6.3	0.387	GGA 1	UAC	Y	8.3	8.8	0.594	GUA 1	UGC	C	1.6	2.2	0.412	GCA 1
UUA	L	50.1	39.4	1.000	UAA 1	UCA	S	16.4	13.6	0.830	UGA 1	UAA	*	2.0	4.9	-		UGA	W	6.3	3.2	1.000	UCA 1
UUG	L	14.2	9.9	0.250		UCG	S	1.1	1.2	0.075	CGA 1	UAG	*	0.7	2.2	-		UGG	W	3.4	2.8	0.857	CCA 1
CUU	L	19.9	15.7	0.398		CCU	P	14.6	15.1	1.000		CAU	H	10.3	9.7	0.887		CGU	R	6.9	14.5	0.580	
CUC	L	5.0	7.5	0.191		CCC	P	3.6	7.1	0.469		CAC	H	5.5	10.9	1.000	GUG 1	CGC	R	3.0	7.5	0.302	GCG 1
CUA	L	12.7	11.2	0.285	UAG 1	CCA	P	10.8	10.6	0.704	UGG 1	CAA	Q	38.3	36.0	1.000	UUG 1	CGA	R	1.3	1.7	0.068	UCG 1
CUG	L	4.4	4.2	0.105		CCG	P	0.9	0.5	0.031		CAG	Q	8.9	8.8	0.244		CGG	R	1.0	2.2	0.086	
AUU	I	51.6	39.9	1.000		ACU	T	25.4	28.2	1.000		AAU	N	45.9	29.4	0.901		AGU	S	25.7	16.3	1.000	
AUC	I	17.9	24.5	0.614	GAU 1	ACC	T	10.3	17.1	0.607	GGU 1	AAC	N	28.9	32.6	1.000	GUU 1	AGC	S	6.7	5.2	0.321	GCU 1
AUA	I	12.6	8.8	0.220	CAU 1	ACA	T	16.6	13.4	0.475	UGU 1	AAA	K	70.4	93.0	1.000	UUU 1	AGA	R	14.0	24.9	1.000	UCU 1
AUG	M	15.2	20.8	1.000	CAU 1	ACG	T	1.6	2.5	0.087	CGU 1	AAG	K	24.4	43.3	0.465	CUU 1	AGG	R	4.6	9.1	0.364	
	fM				CAU 1																		
GUU	V	37.7	39.0	1.000		GCU	A	27.5	29.1	1.000		GAU	D	42.3	27.4	1.000		GGU	G	22.9	29.6	1.000	
GUC	V	3.4	5.7	0.146		GCC	A	4.1	6.5	0.222		GAC	D	6.9	4.2	0.152	GUC 1	GGC	G	5.0	8.0	0.271	GCC 1
GUA	V	13.1	12.9	0.332	UAC 2	GCA	A	21.4	22.2	0.762	UGC 1	GAA	E	45.6	33.0	1.000	UUC 1	GGA	G	11.5	13.2	0.448	UCC 1
GUG	V	7.1	11.5	0.296		GCG	A	2.6	4.5	0.153		GAG	E	11.2	13.1	0.397		GGG	G	6.8	11.9	0.401	

Genomic GC% 31.7, GC% for coding region 31.6, GC% for 3rd letter of codon 23.1

Table 1-4 *continued*

Mycoplasma pneumoniae

		Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA
UUU	F	43.0	23.6	1.000		UCU	S	8.2	5.1	0.432		UAU	Y	14.3	8.4	0.558		UGU	C	5.4	2.5	1.000	
UUC	F	12.7	7.5	0.315	GAA 1	UCC	S	9.5	9.7	0.824	GGA 1	UAC	Y	17.9	15.1	1.000	GUA 1	UGC	C	2.1	1.6	0.625	GCA 1
UUA	L	39.2	33.3	1.000	UAA 1	UCA	S	8.7	8.7	0.743	UGA 1	UAA	*	2.0	4.6	-		UGA	W	6.0	2.2	0.560	UCA 1
UUG	L	21.5	17.1	0.514		UCG	S	6.4	4.6	0.392	CGA 1	UAG	*	0.8	1.9	-		UGG	W	5.8	4.0	1.000	CCA 1
CUU	L	10.1	7.5	0.224		CCU	P	8.3	9.7	0.762		CAU	H	6.2	4.3	0.267		CGU	R	9.7	21.7	0.951	
CUC	L	12.2	13.6	0.410		CCC	P	9.0	7.1	0.562		CAC	H	11.9	16.0	1.000	GUG 1	CGC	R	10.7	22.8	1.000	GCG 1
CUA	L	10.6	8.1	0.243	UAG 1	CCA	P	10.9	12.7	1.000	UGG 1	CAA	Q	37.9	34.2	1.000	UUG 1	CGA	R	2.5	3.2	0.139	UCG 1
CUG	L	9.5	5.5	0.167		CCG	P	6.6	7.6	0.600		CAG	Q	15.6	7.8	0.227		CGG	R	5.0	9.0	0.396	
AUU	I	46.0	48.8	1.000		ACU	T	19.3	17.3	0.623		AAU	N	25.1	14.3	0.331		AGU	S	21.0	11.7	1.000	
AUC	I	14.4	14.3	0.292	GAU 1	ACC	T	21.9	27.7	1.000	GGU 1	AAC	N	37.0	43.1	1.000	GUU 1	AGC	S	10.6	7.6	0.649	GCU 1
AUA	I	5.5	1.9	0.039	CAU 1	ACA	T	10.4	7.8	0.280	UGU 1	AAA	K	46.3	60.3	0.931	UUU 1	AGA	R	4.0	4.4	0.194	UCU 1
AUG	M	15.6	21.6	1.000	CAU 1	ACG	T	7.9	6.3	0.229	CGU 1	AAG	K	39.0	64.7	1.000	CUU 1	AGG	R	2.8	1.3	0.056	
	fM				CAU 1																		
GUU	V	21.2	24.4	1.000		GCU	A	25.2	30.6	1.000		GAU	D	30.4	19.8	1.000		GGU	G	27.9	37.7	1.000	
GUC	V	11.0	8.2	0.338		GCC	A	16.5	17.1	0.560		GAC	D	19.2	15.9	0.800	GUC 1	GGC	G	11.8	12.4	0.328	GCC 1
GUA	V	13.7	21.2	0.870	UAC 2	GCA	A	13.8	14.3	0.466	UGC 1	GAA	E	42.0	40.7	1.000	UUC 1	GGA	G	6.4	7.8	0.206	UCC 1
GUG	V	18.7	22.7	0.929		GCG	A	11.0	13.0	0.425		GAG	E	14.7	11.3	0.276		GGG	G	8.9	9.0	0.239	

Genomic GC% 40.0, GC% for coding region 40.7, GC% for 3rd letter of codon 41.9

***Synechocystis* PCC6803**

		Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA
UUU	F	29.4	14.0	1.000		UCU	S	8.6	8.2	0.448		UAU	Y	17.0	9.5	0.758		UGU	C	6.2	3.6	1.000	
UUC	F	10.6	9.3	0.667	GAA 1	UCC	S	16.1	18.4	1.000	GGA 1	UAC	Y	12.0	12.5	1.000	GUA 1	UGC	C	3.8	2.5	0.692	GCA 1
UUA	L	26.1	10.6	0.527	UAA 1	UCA	S	3.9	0.8	0.045	UGA 1	UAA	*	1.3	4.5	-		UGA	*	0.6	0.0	-	
UUG	L	29.5	20.0	1.000	CAA 1	UCG	S	4.0	2.5	0.134	CGA 1	UAG	*	1.1	2.7	-		UGG	W	15.5	5.4	1.000	CCA 1
CUU	L	9.8	7.3	0.363		CCU	P	9.8	11.1	0.455		CAU	H	11.5	8.6	0.778		CGU	R	10.2	28.6	1.000	
CUC	L	14.0	16.3	0.815	GAG 1	CCC	P	25.3	24.4	1.000	GGG 1	CAC	H	7.2	11.1	1.000	GUG 1	CGC	R	12.4	27.3	0.957	ACG 1
CUA	L	13.9	8.4	0.418	UAG 1	CCA	P	7.9	3.6	0.146	UGG 1	CAA	Q	34.1	27.2	1.000	UUG 1	CGA	R	5.2	2.9	0.101	
CUG	L	20.4	16.3	0.815	CAG 1	CCG	P	8.4	4.0	0.163	CGG 1	CAG	Q	21.3	14.6	0.535		CGG	R	13.6	25.7	0.899	CCG 1
AUU	I	40.3	33.2	1.000		ACU	T	13.8	11.0	0.288		AAU	N	25.2	16.2	0.819		AGU	S	15.0	8.6	0.470	
AUC	I	17.9	25.7	0.773	GAU 2	ACC	T	26.6	38.2	1.000	GGU 1	AAC	N	15.1	19.8	1.000	GUU 1	AGC	S	10.3	11.1	0.604	GCU 1
AUA	I	4.5	1.0	0.029	CAU 1	ACA	T	6.6	3.7	0.097	UGU 1	AAA	K	29.4	53.7	1.000	UUU 1	AGA	R	4.4	4.9	0.173	UCU 1
AUG	M	19.5	24.4	1.000	CAU 1	ACG	T	7.9	5.1	0.133	CGU 1	AAG	K	12.3	28.8	0.537		AGG	R	4.7	3.0	0.106	CCU 1
	fM				CAU 1																		
GUU	V	16.5	19.8	0.702		GCU	A	20.0	26.9	0.626		GAU	D	32.2	23.5	1.000		GGU	G	19.9	30.9	1.000	
GUC	V	11.2	17.6	0.624	GAC 1	GCC	A	38.6	43.0	1.000	GGC 1	GAC	D	17.9	19.1	0.813	GUC 1	GGC	G	23.0	24.4	0.791	GCC 1
GUA	V	10.5	11.3	0.400	UAC 1	GCA	A	10.5	7.3	0.169	UGC 1	GAA	E	44.7	50.1	1.000	UUC 1	GGA	G	12.7	9.3	0.302	UCC 1
GUG	V	29.1	28.1	1.000		GCG	A	15.6	11.8	0.275	CGC 1	GAG	E	15.5	14.0	0.279		GGG	G	17.9	12.8	0.413	CCC 1

Genomic GC% 47.7, GC% for coding region 48.6, GC% for 3rd letter of codon 49.8

Table 1-4 ***continued***

Saccharomyces cerevisiae

		Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA			Total	rp	Wrp	tRNA
UUU	F	26.8	9.8	0.399		UCU	S	23.5	30.9 #	1.000		UAU	Y	19.0	3.7	0.168		UGU	C	8.1	5.6	1.000	
UUC	F	18.3	24.5 #	1.000	GAA 10	UCC	S	14.1	20.1 #	0.650	GGA 11	UAC	Y	14.4	22.2 #	1.000	GUA 8	UGC	C	5.0	0.6	0.107	GCA 4
UUA	L	26.4	14.8	0.295	UAA 7	UCA	S	19.0	3.8	0.124	UGA 3	UAA	*	1.0	4.9	-		UGA	*	0.6	0.6	-	
UUG	L	26.8	50.2 #	1.000	CAA 10	UCG	S	8.8	1.5	0.049	CGA 1	UAG	*	0.5	0.5	-		UGG	W	10.5	6.7	1.000	CCA 6
CUU	L	12.6	3.0	0.060		CCU	P	13.4	5.6	0.177		CAU	H	13.8	6.0	0.388		CGU	R	6.3	15.3	0.245	
CUC	L	5.7	0.5	0.010	GAG 1	CCC	P	6.9	1.2	0.038	GGG 2	CAC	H	7.7	15.3	1.000	GUG 7	CGC	R	2.7	0.8	0.013	ACG 6
CUA	L	13.5	6.5	0.129	UAG 3	CCA	P	17.6	32.0	1.000	UGG 10	CAA	Q	26.6	35.0	1.000	UUG 9	CGA	R	3.2	0.2	0.003	
CUG	L	10.8	1.5	0.030		CCG	P	5.4	1.5	0.047		CAG	Q	12.3	2.4	0.069	CUG 1	CGG	R	1.9	0.4	0.006	CCG 1
AUU	I	30.1	24.2 #	0.896		ACU	T	20.0	25.1 #	1.000		AAU	N	36.2	8.3	0.266		AGU	S	14.6	2.9	0.095	
AUC	I	16.9	27.0 #	1.000	GAU 13	ACC	T	12.5	22.0 #	0.876	GGU 11	AAC	N	24.5	31.1 #	1.000	GUU 10	AGC	S	10.1	2.0	0.065	GCU 4
AUA	I	18.4	3.2	0.119	CAU 2	ACA	T	17.9	4.0	0.161	UGU 4	AAA	K	42.4	23.4	0.277	UUU 7	AGA	R	21.0	62.7 #	1.000	UCU 11
AUG	M	20.8	15.2	1.000	CAU 5	ACG	T	8.2	1.5	0.060	CGU 1	AAG	K	30.4	84.4 #	1.000	CUU 14	AGG	R	9.6	4.3	0.069	CCU 1
	fM				CAU 5																		
GUU	V	21.6	43.6 #	1.000		GCU	A	20.1	63.3 #	1.000		GAU	D	37.7	15.6	0.842		GGU	G	22.5	59.0 #	1.000	
GUC	V	11.3	29.9 #	0.685	GAC 14	GCC	A	12.2	20.4 #	0.322	GGC 11	GAC	D	20.1	18.6	1.000	GUC 15	GGC	G	9.8	5.0 #	0.085	GCC 16
GUA	V	12.2	3.4	0.079	UAC 2	GCA	A	16.3	5.6	0.089	UGC 5	GAA	E	45.4	52.3 #	1.000	UUC 14	GGA	G	11.2	3.0	0.051	UCC 3
GUG	V	10.9	2.8	0.065	CAC 2	GCG	A	6.3	1.8	0.029		GAG	E	19.5	4.7	0.091	CUC 2	GGG	G	6.2	1.5	0.026	CCC 2

Genomic GC% 38.3, GC% for coding region 39.7, GC% for 3rd letter of codon 38.1

The frequency of the usage of each codon is shown in permillage of overall counts. The most preferred codons in the synonymous codon box are highlighted. For *E. coli* and *S. cerevisiae,* the optimal codons (Ikemura, 1981b, 1982) are indicated by #. The Wrp column shows the relative adaptability of each codon, as calculated from the codon usage of the ribosomal protein genes.

Table 1-5. List of genes with the highest and the lowest CAI_{rp} scores in 7 organisms.

Escherichia coli

CAI_{rp}	L /total	LOCUS#No.	product
Highest			
0.875	67/ 75	ECOLI#1634	major outer membrane lipoprotein precursor *lpp*
0.860	104/ 117	ECOLI#3877	50S ribosomal subunit protein L7/L12 *rplL*
0.854	375/ 418	ECOLI#2710	enolase *eno*
0.852	286/ 320	ECOLI#1736	glyceraldehyde 3-phosphate dehydrogenase A *gapA*
0.838	319/ 359	ECOLI#2164	outer membrane protein C precursor *ompC*
0.833	58/ 67	ECOLI#3476	*cspA*
0.829	337/ 383	ECOLI#3257	*tufA*
0.828	159/ 181	ECOLI#586	alkyl hydroperoxide reductase c22 protein *ahpC*
0.824	111/ 126	ECOLI#3153	30S ribosomal subunit protein S9 *rpsI*
0.815	291/ 335	ECOLI#923	outer membrane protein A *ompA*
0.810	332/ 382	ECOLI#3871	elongation factor EF-Tu duplicate gene *tufB*
0.810	78/ 91	ECOLI#2134	50s ribosomal protein l25 *rplY*
0.808	299/ 347	ECOLI#2847	fructose 1,6-bisphosphate aldolase *fba*
0.806	631/ 721	ECOLI#869	formate acetyltransferase 1 *pflB*
0.806	199/ 230	ECOLI#169	30s ribosomal protein s2 *rpsB*
0.804	452/ 525	ECOLI#4030	GroEL protein *mopA*
0.802	469/ 544	ECOLI#877	30S ribosomal protein S1 *rpsA*
0.796	58/ 68	ECOLI#3835	50S ribosomal protein L31 *rpmE*
0.793	230/ 273	ECOLI#170	elongation factor ts *tsf*
0.791	192/ 227	ECOLI#3875	50S ribosomal subunit protein L1 *rplA*
Lowest			
0.189	35/ 81	ECOLI#2032	
0.188	40/ 98	ECOLI#2950	
0.184	173/ 409	ECOLI#3542	*rfaL*
0.179	49/ 105	ECOLI#644	
0.175	34/ 86	ECOLI#311	
0.173	26/ 64	ECOLI#2559	prophage cp4-57 regulatory protein AlpA *alpA*
0.166	12/ 31	ECOLI#1191	protamine-like protein *tpr*
0.165	11/ 31	ECOLI#3679	ilvGMEDA operon leader peptide *ilvL*
0.163	20/ 51	ECOLI#4134	
0.152	36/ 99	ECOLI#676	hypothetical 12.6 kD protein in RhsC 3'region *ybfB*

Haemophilus influenzae Rd

CAI_{rp}	L /total	LOCUS#No.	gene ID	product
Highest				
0.868	51/ 57	HIL42023#956	HI0965	ribosomal protein S20 *rpS20* GB:M10428_1
0.835	78/ 87	HIL42023#1446	HI1468	ribosomal protein S15 *rpS15*
0.826	196/ 222	HIL42023#507	HI0516	ribosomal protein L1 *rpL1*
0.825	106/ 121	HIL42023#634	HI0641	ribosomal protein L7/L12 *rpL7/L12*
0.822	129/ 147	HIL42023#535	HI0544	ribosomal protein L9 *rpL9*
0.814	46/ 53	HIL42023#941	HI0950	ribosomal protein L33 *rpL33*
0.813	76/ 87	HIL42023#1307	HI1328	ribosomal protein S15 *rpS15*
0.810	132/ 152	HIL42023#371	HI0381	peptidoglycan-associated outer membrane lipoprotein *pal*
0.807	63/ 74	HIL42023#152	HI0154	acyl carrier protein *acpP*
0.802	46/ 54	HIL42023#156	HI0158	ribosomal protein L32 *rpL32*
0.801	466/ 536	HIL42023#1203	HI1220	ribosomal protein S1 *rpS1*
0.796	58/ 67	HIL42023#1348	HI1370	molybdenum-pterin binding protein *mopI*
0.792	236/ 273	HIL42023#904	HI0914	elongation factor EF-Ts *tsf*
0.784	35/ 42	HIL42023#990	HI0998	ribosomal protein L34 *rpL34*
0.782	299/ 349	HIL42023#1149	HI1164	outer membrane protein P5 *ompA*
0.780	361/ 421	HIL42023#922	HI0932	enolase *eno*
0.775	91/ 107	HIL42023#773	HI0782	ribosomal protein L22 *rpL22*
0.773	107/ 125	HIL42023#16	HI0017	hypothetical
0.772	163/ 195	HIL42023#906	HI0916	export factor homolog *skp*
0.769	96/ 114	HIL42023#198	HI0201	ribosomal protein L19 *rpL19*
Lowest				
0.325	136/ 250	HIL42023#526	HI0535	urease protein *ureH*
0.320	27/ 49	HIL42023#1545	HI1569	*H. influenzae* predicted coding region HI1569
0.316	31/ 61	HIL42023#1487	HI1510	*H. influenzae* predicted coding region HI1510
0.311	33/ 56	HIL42023#1167	HI1183	*H. influenzae* predicted coding region HI1183
0.301	121/ 230	HIL42023#347	HI0355	hypothetical
0.300	47/ 84	HIL42023#1472	HI1495	*H. influenzae* predicted coding region HI1495
0.297	119/ 230	HIL42023#346	HI0354	nitrate transporter ATPase component *nasD*
0.287	15/ 26	HIL42023#1541	HI1564	*H. influenzae* predicted coding region HI1564
0.274	29/ 60	HIL42023#1252	HI1270	*H. influenzae* predicted coding region HI1270
0.247	4/ 10	HIL42023#1439	HI1460	invasin precursor outer membrane adhesin *yopA*

Methanococcus jannaschii

CAI_{rp}	L /total	LOCUS#No.	gene ID	product
Highest				
0.853	38/ 42	L77117#707	MJ0707	ribosomal protein L40
0.850	89/ 100	L77117#508	MJ0508	ribosomal protein L12
0.845	79/ 89	L77117#593	MJ0593	ribosomal protein L37a
0.843	70/ 79	L77117#657	MJ0657	ribosomal protein L14B
0.836	50/ 56	L77117#1152	MJ1153	*M. jannaschii* predicted coding region MJ1153
0.811	162/ 186	L77117#1046	MJ1047	ribosomal protein S7
0.797	125/ 146	L77117#1045	MJ1046	ribosomal protein S12
0.797	114/ 133	L77117#746	MJ0746	hypothetical protein GP:U21086_2
0.795	49/ 57	L77117#98	MJ0098	ribosomal protein L37
0.790	84/ 98	L77117#394	MJ0394	ribosomal protein S24
0.789	98/ 115	L77117#467	MJ0467	ribosomal protein L24
0.788	151/ 174	L77117#270	MJ0269	ferredoxin oxidoreductase, gamma subunit
0.785	446/ 522	L77117#998	MJ0999	chaperonin
0.782	142/ 166	L77117#543	MJ0543	Wilm's tumor suppressor homolog
0.782	51/ 59	L77117#246	MJ0245	ribosomal protein S17B
0.781	53/ 61	L77117#401	MJ0401	*M. jannaschii* predicted coding region MJ0401
0.779	56/ 66	L77117#932	MJ0932	archaeal histone
0.778	210/ 247	L77117#178	MJ0177	ribosomal protein L4 human
0.777	97/ 115	L77117#465	MJ0465	ribosomal protein S17
0.776	55/ 66	L77117#169	MJ0168	archaeal histone
Lowest				
0.468	132/ 195	L77117#844	MJ0844	methyl coenzyme M reductase operon, protein C
0.464	89/ 135	L77117#897	MJ0897	*M. jannaschii* predicted coding region MJ0897
0.461	57/ 90	L77117#366	MJ0366	*M. jannaschii* predicted coding region MJ0366
0.450	50/ 81	L77117#353	MJ0353	*M. jannaschii* predicted coding region MJ0353
0.442	137/ 205	L77117#1465	MJ1466	*M. jannaschii* predicted coding region MJ1466
0.442	59/ 93	L77117#464	MJ0464	hypothetical protein SP:P14022
0.425	77/ 118	L77117#877	MJ0877	hemin permease
0.413	134/ 203	L77117#12	MJ0012	*M. jannaschii* predicted coding region MJ0012
0.410	35/ 55	L77117#16	MJ0016	*M. jannaschii* predicted coding region MJ0016
0.396	65/ 106	L77117#905	MJ0905	*M. jannaschii* predicted coding region MJ0905

Mycoplasma genitalium

CAI_{rp}	L /total	LOCUS#No.	gene ID	product
Highest				
0.827	118/ 135	MYCCG#336	MG337	*M. genitalium* predicted coding region MG337
0.805	194/ 226	MYCCG#443	MG445	tRNA guanine-N1-methyltransferase *trmD*
0.801	163/ 189	MYCCG#376	MG377	*M. genitalium* predicted coding region MG377
0.798	1369/1590	MYCCG#385	MG386	cytadherence-accessory protein *hmw1_2*
0.793	151/ 178	MYCCG#350	MG351	inorganic pyrophosphatase *ppa*
0.792	254/ 297	MYCCG#20	MG020	proline iminopeptidase *pip_1*
0.791	163/ 190	MYCCG#285	MG286	*M. genitalium* predicted coding region MG286
0.791	142/ 166	MYCCG#457	MG459	surface exclusion protein *prgA* Plasmid pCF10
0.788	134/ 158	MYCCG#345	MG346	hypothetical protein GB:M65289_3
0.785	130/ 154	MYCCG#407	MG408	pilin repressor *pilB_1*
0.784	214/ 254	MYCCG#366	MG367	ribonuclease III *rnc*
0.781	308/ 366	MYCCG#216	MG217	bifunctional endo-1,4-beta-xylanase xyla precursor *xynA*
0.781	128/ 151	MYCCG#452	MG454	osmotically inducible protein *osmC*
0.779	245/ 291	MYCCG#431	MG433	elongation factor Ts *tsf*
0.778	135/ 160	MYCCG#244	MG245	hypothetical protein GB:M12965_1
0.777	248/ 294	MYCCG#349	MG350	*M. genitalium* predicted coding region MG350
0.777	72/ 86	MYCCG#41	MG041	phosphohistidinoprotein-hexose phosphotransferase *ptsH*
0.776	99/ 119	MYCCG#361	MG362	ribosomal protein L7/L12 'A' type *rpL7/L12*
0.774	340/ 406	MYCCG#51	MG051	thymidine phosphorylase *deoA*
0.774	224/ 267	MYCCG#437	MG439	*M. genitalium* predicted coding region MG439
Lowest				
0.634	130/ 174	MYCCG#163	MG163	ribosomal protein L5 *rpL5*
0.629	94/ 127	MYCCG#283	MG284	*M. genitalium* predicted coding region MG284
0.628	101/ 138	MYCCG#195	MG196	translation initiation factor IF3 *infC*
0.622	62/ 85	MYCCG#444	MG446	ribosomal protein S16 *BS17*
0.622	47/ 66	MYCCG#173	MG173	initiation factor 1 *infA*
0.620	338/ 466	MYCCG#170	MG170	preprotein translocase secY subunit *secY*
0.609	79/ 112	MYCCG#434	MG436	*M. genitalium* predicted coding region MG436
0.597	224/ 313	MYCCG#189	MG189	membrane protein *msmG*
0.597	87/ 125	MYCCG#176	MG176	ribosomal protein S11 *rpS11*
0.581	221/ 318	MYCCG#188	MG188	membrane protein *msmF*

Mycoplasma pneumoniae

CAI_{rp}	L /total	LOCUS#No.	product
Highest			
0.809	123/ 141	U00089#190	PTS system mannitol-specific component IIA EIIA-MTL *mtlF*
0.797	200/ 235	U00089#665	ribosomal protein L23 *rplW*
0.781	82/ 98	U00089#512	ribosomal protein L21 *rplU*
0.780	99/ 118	U00089#304	ribosomal protein L7/L12 'A' type *rplL*
0.767	98/ 118	U00089#640	ribosomal protein L17 *rplQ*
0.765	171/ 206	U00089#666	ribosomal protein L4 *rplD*
0.763	115/ 138	U00089#131	DNA-directed RNA polymerase delta subunit *rpoE*
0.762	147/ 178	U00089#315	inorganic pyrophosphatase *ppa*
0.761	154/ 187	U00089#111	thymidine kinase *tdk*
0.758	47/ 58	U00089#39	ribosomal protein L35 *rpmI*
0.752	439/ 535	U00089#270	heat shock protein GroEL *groEL*
0.750	175/ 211	U00089#604	ribosomal protein S6 *rpsF*
0.749	94/ 116	U00089#250	
0.748	255/ 311	U00089#88	transcription antitermination factor *nusG*
0.748	154/ 188	U00089#288	
0.748	88/ 108	U00089#656	ribosomal protein L24 *rplX*
0.747	119/ 146	U00089#601	ribosomal protein L9 *rplI*
0.745	351/ 432	U00089#271	nonspecified aminopeptidase
0.745	268/ 329	U00089#412	glycerladehyde-3-phosphate dehydrogenase *gap*
0.745	148/ 182	U00089#126	elongation factor P *efp*
Lowest			
0.548	130/ 203	U00089#374	
0.542	167/ 256	U00089#471	
0.542	82/ 124	U00089#189	
0.537	92/ 141	U00089#379	
0.534	136/ 214	U00089#24	
0.534	82/ 124	U00089#65	
0.524	57/ 88	U00089#265	
0.503	93/ 153	U00089#341	
0.489	101/ 159	U00089#52	
0.438	15/ 26	U00089#238	ATP synthase protein I *atpI*

Synechocystis PCC6803

CAI_{rp}	L /total	LOCUS#No.	gene ID	product
Highest				
0.852	54/ 62	synecho#1498	ssr2194	hypothetical protein
0.830	136/ 160	synecho#653	sll1578	phycocyanin a subunit *cpcA*
0.825	107/ 125	synecho#838	sll1746	50S ribosomal protein L12 *rpl12*
0.819	65/ 75	synecho#734	ssl2084	acyl carrier protein *acp*
0.818	167/ 195	synecho#1720	sll1444	3-isopropylmalate dehydratase *leuD*
0.814	387/ 453	synecho#477	slr1756	glutamate-ammonia ligase *glnA*
0.813	348/ 408	synecho#1501	sll1234	S-adenosylhomocysteine hydrolase *ahcY*
0.813	132/ 156	synecho#1297	slr1986	allophycocyanin b chain *apcB*
0.813	92/ 109	synecho#1103	sll1767	30S ribosomal protein S6 *rps6*
0.811	450/ 529	synecho#833	slr2076	60kD chaperonin 1 *groEL*
0.810	26/ 32	synecho#2794	smr0003	cytochrome b6-f complex subunit PetM *petM*
0.808	87/ 104	synecho#832	slr2075	10kD chaperonin *groES*
0.808	45/ 54	synecho#1123	ssr3189	hypothetical protein
0.807	80/ 96	synecho#2228	ssl0020	ferredoxin *petF*
0.803	94/ 112	synecho#2152	sll0767	50S ribosomal protein L20 *rpl20*
0.803	61/ 73	synecho#288	ssr1604	50S ribosomal protein L28 *rpl28*
0.803	29/ 35	synecho#750	sml0006	50S ribosomal protein L36 *rpl36*
0.801	130/ 154	synecho#1296	slr2067	allophycocyanin a chain *apcA*
0.800	376/ 448	synecho#2221	slr0009	ribulose bisphosphate carboxylase large subunit *rbcL*
0.800	268/ 318	synecho#1479	slr2120	hypothetical protein
Lowest				
0.438	69/ 115	synecho#387	sll0650	transposase
0.438	69/ 115	synecho#1471	slr2112	transposase
0.436	204/ 340	synecho#1456	slr1522	transposase
0.436	192/ 321	synecho#2661	slr0099	transposase
0.434	91/ 153	synecho#3108	slr0460	transposase
0.428	171/ 294	synecho#2006	slr1931	hypothetical protein
0.423	53/ 91	synecho#2770	ssr1175	transposase
0.420	121/ 210	synecho#2782	sll0667	transposase
0.411	131/ 224	synecho#2686	sll0092	transposase
0.405	75/ 132	synecho#1793	slr1682	transposase

Saccharomyces cerevisiae

CAI_{rp}	L /total	gene ID	product
Highest			
0.938	44/ 46	YJL189W	questionable ORF
0.925	303/ 322	YGR192C	glyceraldehyde-3-phosphate dehydrogenase 3
0.917	511/ 543	YLR044C	pyruvate decarboxylase, isozyme 1
0.912	300/ 321	YJR009C	glyceraldehyde-3-phosphate dehydrogenase 2
0.904	454/ 488	YAL038W	pyruvate kinase
0.904	392/ 422	YHR174W	enolase II 2-phosphoglycerate dehydratase
0.902	414/ 444	YBR118W	translation elongation factor eEF1 alpha-A chain, cytosolic
0.902	396/ 426	YGR254W	enolase I 2-phosphoglycerate dehydratase
0.902	325/ 348	YKL060C	fructose-bisphosphate aldolase
0.900	414/ 444	YPR080W	translation elongation factor eEF1 alpha-A chain, cytosolic
0.889	109/ 119	YLR061W	ribosomal protein
0.888	94/ 103	YGL030W	ribosomal protein L30.e
0.887	296/ 322	YJL052W	glyceraldehyde-3-phosphate dehydrogenase 1
0.886	217/ 234	YBR181C	ribosomal protein S6.e
0.886	204/ 221	YJR123W	ribosomal protein S5.e
0.886	81/ 89	YPR043W	ribosomal protein L37a.e
0.884	233/ 251	YLL045C	questionable ORF
0.883	218/ 234	YPL090C	ribosomal protein S6.e
0.882	127/ 138	YOR369C	acidic ribosomal protein S12
0.881	94/ 102	YOR293W	ribosomal protein S10.e
Lowest			
0.168	48/ 123	YKL086W	hypothetical protein
0.166	93/ 232	YOL110W	RAS suppressor
0.165	43/ 126	YKL097C	hypothetical protein
0.161	39/ 103	YOR343C	hypothetical protein
0.159	93/ 238	YER153C	translational activator of cytochrome c oxidase subunitIII
0.159	61/ 149	YLR281C	hypothetical protein
0.158	50/ 126	YLR279W	questionable ORF
0.143	38/ 113	YLR280C	questionable ORF
0.133	45/ 151	YML010W-A	
0.132	44/ 146	YLR391W	weak similarity to hypothetical proteins YAR068w and YHR214w-a

L: number of codons that have synonymous codons / total: number of total codons.

Chapter 2

Construction of the WWW databases for the complete nucleotide sequence of the genome of *Synechocystis* sp. strain PCC6803

2-1 Introduction

2-1-1 Cyanobacteria

Cyanobacteria are prokaryotic microorganisms which carry a complete set of genes for oxygenic photosynthesis. They are distinct from other photosynthetic bacteria such as purple and green bacteria, because cyanobacteria utilize H_2O as an electron donor but the others do not. The importance of cyanobacteria as model organisms for the study of photosynthesis is increasing due to confluence of findings from many lines of research. As bacteria, they are amenable to a variety of genetic and molecular biological manipulations. While the striking similarity between cyanobacterial and chloroplast photosynthesis has been recognized for some time, the full extent of this relationship is only now being appreciated. The accumulated cytological, physiological, and biochemical evidence over the past century gave rise to the endosymbiont hypothesis for the origin of chloroplasts; this hypothesis states that chloroplasts are descended from cyanobacteria that became endosymbionts in an originally non-photosynthetic host.

2-1-2 *Synechocystis* sp. PC6803 genome sequencing project

Although cyanobacteria comprise one of the largest constituents of the gram-negative bacteria, only a limited number of strains have been used for physiological and genetic studies. *Synechocystis* sp. strain PCC6803 is a unicellular cyanobacterium. It has been widely used for the study of the mechanism of photosynthesis because this strain has the ability to grow both photoautotrophically and photoheterotrophically, allowing disruption of the photosynthetic protein-encoding genes without lethality. Also it is naturally transformable (Grigorieva and Shestakov, 1982).

Resent progress in DNA sequencing technology has allowed the generation of large quantities of nucleotide sequence data in a short period of time. In 1996, the nucleotide

sequence of the entire genome of *Synechocystis* sp. strain PCC6803 was determined by Kazusa DNA research institute. This was the first report of the whole genome in a photoautotrophic organism. By taking a close look at the structure of each gene and the organization of the entire genome, genetic information characteristic of photoautotrophic bacteria will be better understood.

2-1-3 Application of the internet technology for genomic research

The current version of CyanoBase then released (in 1996), and have been maintained to offer public access to support research in the genomic information on the cyanobacterium. The basic aim of this database is to supply detailed information on potential protein-encoding genes with user-friendly interfaces.

Of 3,168 deduced genes of *Synechocystis*, 1,722 were annotated as functionally unassigned, which included 1,270 putative genes, 418 genes similar to hypothetical ones, and 34 genes similar to expressed sequence tags (ESTs) of other genomes. To analyze the functions of these genes, systematic disruption of each gene and characterization of the resulting mutants is thought to be a promising strategy. CyanoMutants (http://www.kazusa.or.jp/cyano/mutants/) is a cumulative database that allows users to stores and access mutant information through the WWW.

2-2 Material and Methods

2-2-1 Nucleotide sequence and annotation

The complete nucleotide sequence of the complete nucleotide sequence of genome of *Synechocystis* sp. strain PCC6803, was determined by author's group (accession number of DDBJ: AB001339). The total length of the circular genome is 3,573,471 bp. A total of 3,168 potential protein coding genes were assigned by computer assisted analyses, including similarity searches and predictions by computer programs (Kaneko *et al.*, 1996).

2-2-2 Nomenclature of identifier for protein-encoding genes

The standard name consists of a three-letter code where the first letter represents the species name (s: *Synechocystis*), the second, the length and/or the method of identification of the open reading frame (ORF) (l: longer than 300 bp, s: 150 - 297 bp, m: shorter than 150 bp, g: a gene predicted only by the computer program GenMark (Borodovsky and McIninch, 1993), and the third, the reading direction (l or r: leftward or rightward). The three letter code is followed by a four-digit number. The standard name has been added to each CDS in flat files of the DDBJ/EMBL/GenBank databases according to the format `/note=standard_name:` tag, and all the genes are represented by their standard names in CyanoBase.

2-2-3 Implementation

Map interfaces and data retrieving of the annotation to each gene were implemented in the JMGD system (Nobuyuki Miyajima, Personal communication) with the Sybase SQL server system 11 as a DBMS. The ACeDB (Durbin and Mieg, 1991) formatted database file which was the core portion of CyanoBase was prepared separately. The data on ACeDB can also be constructed and distributed through anonymous ftp server at Kazusa.

2-3 Description of the WWW site

2-3-1 Description of CyanoBase

The URL of CyanoBase is: http://www.kazusa.or.jp/cyano/ (Fig. 2-1). Main structure of CyanoBase consist in positional information and annotation of deduced protein-encoding genes in the entire nucleotide sequence of PCC6803 genome. The annotation for each gene can be accessed through three menus on the main page of CyanoBase; i.e. Physical maps of the genome, Gene category list and Keyword-Search.

Map Image I and II show restriction maps of the circular genome in Java and GIF

formats, respectively. When a given position on the map is clicked, a local map covering the corresponding 90 kb area appears (Fig. 2-2). In each local map, the positions of the cosmid and lambda phage clones used for sequence determination, and the long-PCR products used to close gaps between clones are shown by blue bars, and the assigned protein-encoding genes are designated by color-coded boxes under the bars. Each bar and box provides a link to detailed information on the corresponding clone or gene, as described in the next section.

One thousand eight hundred and sixty four genes out of the 3,168 potential protein-encoding genes assigned to the *Synechocystis* genome showed similarity to sequences already registered in the public DNA or protein databases. These genes have been classified according to their predicted biological function (Kaneko *et al.*, 1995, Kaneko *et al.*, 1996). On the web page, a table of gene classification is presented in a hierarchical manner, with a link to the annotation to each gene (Fig. 2-3).

A keyword search box is provided to allow a search for information about the gene. This is achieved by either submitting the gene name (a three-letter genetic name), the name of the gene product, or a standard name for protein-encoding gene.

A sample image of an annotation page to a potential protein gene is shown in Fig. 2-4. An annotation page consist of information of location, links to sequence, results of similarity searches and classification. Nucleotide positions of the initiation and termination points of a coding region and the coding direction on the physical map are indicated. A reverse-link to a linear physical map is provided which enables users to obtain information about genes in adjacent regions. Both the nucleotide and amino acid sequences of a gene can be obtained through the links. For nucleotide sequence, two input boxes are provided to allow specification of the initiation and termination positions of the sequence to be retrieved. The positions of the first and the last nucleotides of a coding region are given as default numbers. A link to the results of a BLAST search for each sequence in the protein sequence database and summary information on the most similar sequence are provided. The gene category to which a gene of interest belongs is presented in a hierarchical manner.

A link to a gene category list provides information on other genes with the same or similar biological function.

On the main page, additional services are furnished. An input form for a similarity search is provided. Users can choose one out of two combinations of program and reference sequence: a BLASTP program with the library for 3,168 deduced protein sequences and a BLASTN program with the library for the complete nucleotide sequence of the entire genome of *Synechocystis*. A link to the main page of the proteome project of *Synechocystis* (Sazuka and Ohara, 1997; Sazuka *et al.*, 1999) is provided. The following files are provided for distribution by ftp; two files containing the nucleotide and deduced amino acid sequences of the assigned 3,168 protein-encoding genes, a file in either flat file or Macintosh Excel format containing a table summarizing the annotations to 3,168 potential protein genes, a file containing the complete nucleotide sequence of the genome, and the file "cyanoace" which contains both the genomic sequence and annotated information in ACeDB format. There is a link to the nucleotide sequences that have been submitted to the international DNA databases. The sequence was divided into 27 files and each file was separately registered. Information can be obtained for each entry through the DBGET integrated database retrieval system on the GenomeNet WWW Server. Bibliography link provides information on related publications through links to the PubMed publication information service at NCBI/NLM.

2-3-2 Description of CyanoMutants

CyanoMutants is a repository database which stores and provides mutant information through the WWW. An entry to CyanoMutants contains three essential sections; (i) identification of the mutated gene, (ii) information about phenotype and (iii) to whom correspondence should be addressed.

Each database entry links to a corresponding annotation in CyanoBase (Fig. 2-5). A user can follow the link and browse the annotation page of the protein-coding gene, which consists of information concerning the location of the genome, sequence retrieval links,

similarity search results and classification. When CyanoMutants stores a mutant for a protein-coding gene, the corresponding annotation page in CyanoBase shows a link to a page which provides mutant information in CyanoMutants. A link to a mutant page is not shown in the CyanoBase annotation page if mutant data for the gene has not been submitted to CyanoMutants.

A sample image of a submission page is shown in Figure 2-6. An entry is divided into the following sections. (1) information about the gene. In order to link to CyanoBase, an identifier of the mutated gene must be included. There are optional input boxes for gene name, product name and the function of the product. (2) mutant information. This section consists of a mutant name, a mutation type (interruption, deletion or site-or domain-directed mutation), a phenotype if observed, phenotype details, segregation (complete or incomplete) and a storage type (DNA and/or Mutant). (3) correspondence address and other information. MEDLINE ID of publication (if available), correspondence (name, e-mail and web site address) and additional information.

Adding a new mutant from a WWW submission form has been simplified as much as possible in order to facilitate the process for the researcher. Although there are several input and check boxes on the page, only three sections are required to be completed in order to submit an entry, ie. a gene identifier, phenotype (yes or no) and an address for correspondence (author's name and e-mail). A researcher who wants to submit detailed data may use optional sections. Using the additional information box, any further format-free information such as an author's postal address, whether it is distributable or not, and experimental details, can be included.

Information in the database is accessed through two ways; a list browser and a keyword search box. The list browser shows the first 20 entries sorted by the gene identifier as a default (Fig. 2-7). Sorting on the other keys can be done by clicking `sort by' on the top of the columns. The identifier is a link to detailed information for each mutant. An entry of a mutant page shows each submitted data section listed in the list table. A search box allows a simple keyword search against the entire database of information

including gene names, gene ids, researchers' names and any text-based contents. Returned contents have the same format as the list browser described above. The searched keyword is shown by highlight in the entry page.

As of December 2000, CyanoMutants contained 431 mutant entries including 134 phenotype descriptions. The number of genes registered is expected to increase continuously since a large number of gene disruption experiments have been carried out since the release of the genomic sequence of *Synechocystis* sp. strain PCC6803. The information registered in CyanoMutants will prevent unnecessary overlaps in experiments and promote communication among scientists to elucidate the functions of putative genes in cyanobacteria.

2-3-3 Conclusion

Cyanobacteria carry a complete set of genes for higher plant-like oxygenic photosynthesis. The strain *Synechocystis* sp. PCC6803 has the advantage of being easily transformed by DNA allowing research into the function of unknown genes through gene disruption and insertion mutational analysis. Since the strain can carry out photoheterotrophic growth without loss of viability, the photosynthetic-related protein genes can be disrupted to investigate their contribution to the photosynthetic pathway.

In 1996, author's research group reported the complete nucleotide sequence of the genome of *Synechocystis* sp. strain PCC6803, in which 3,168 potential protein-coding genes were deduced. This has been the only report of the complete structural analysis of the genome of a photoautotrophic organism until report of *Arabidopsis* genome sequencing project, which is described in chapter 3 of this thesis. The CyanoBase was released and has since been maintained to offer researchers access to information present on the genome of the cyanobacterium. The basic aim of this database is to supply detailed information about potential protein encoding genes through user-friendly interfaces. In 1998, CyanoMutants was developed as a repository database for the storage and distribution of information about *Synechocystis* sp. PCC6803 mutants. The repository database stores and provides

submitted information of an experimental nature or conjectures about function of each gene in CyanoBase. These system serves an effective tool that is used for retrieve both annotation or mutant information for deduced genes for further experiments of functional analysis, especially analyses of photoautotrophic-related genes.

2-4 Figures and Tables

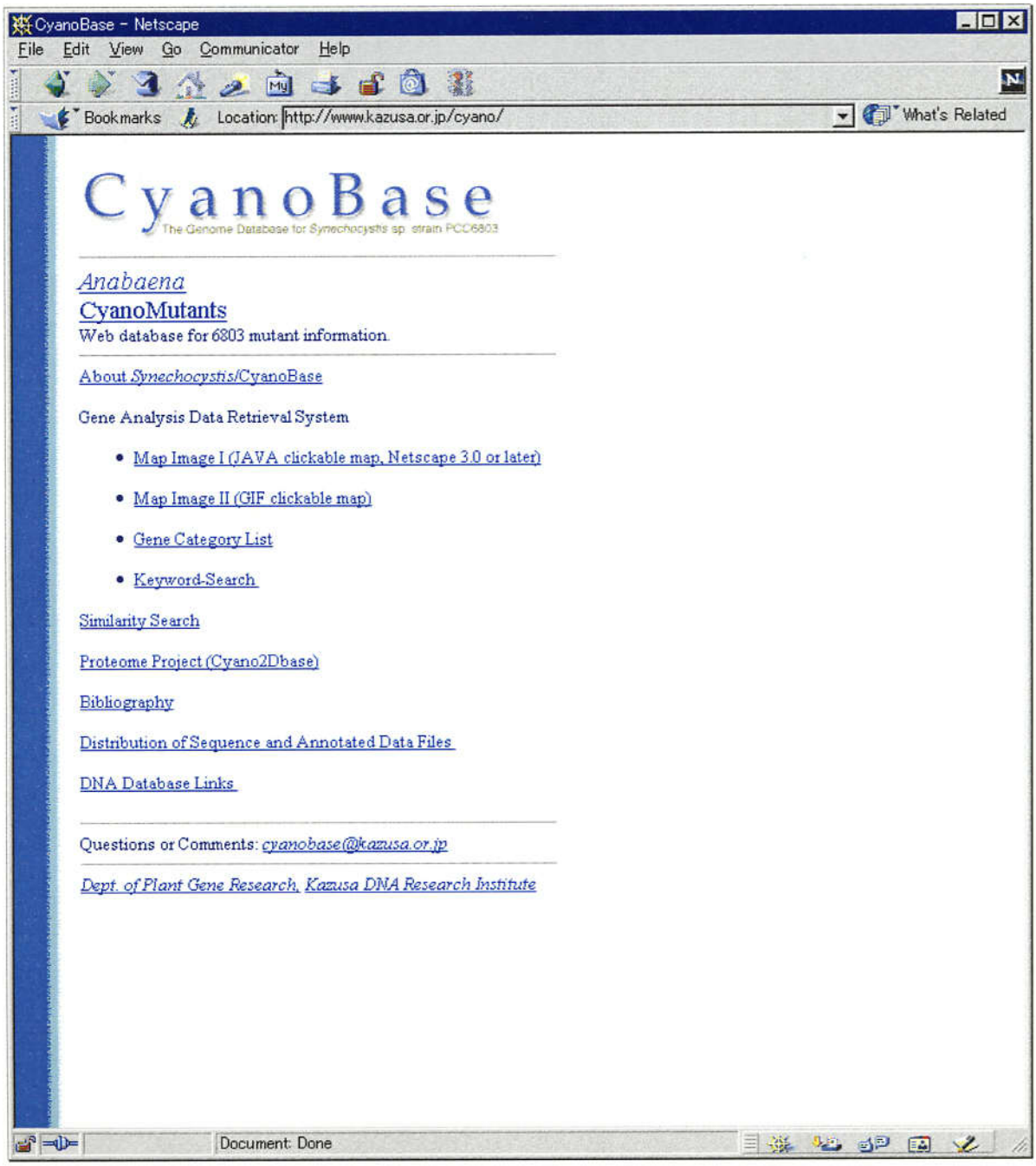


Figure 2-1 Top page of CyanoBase

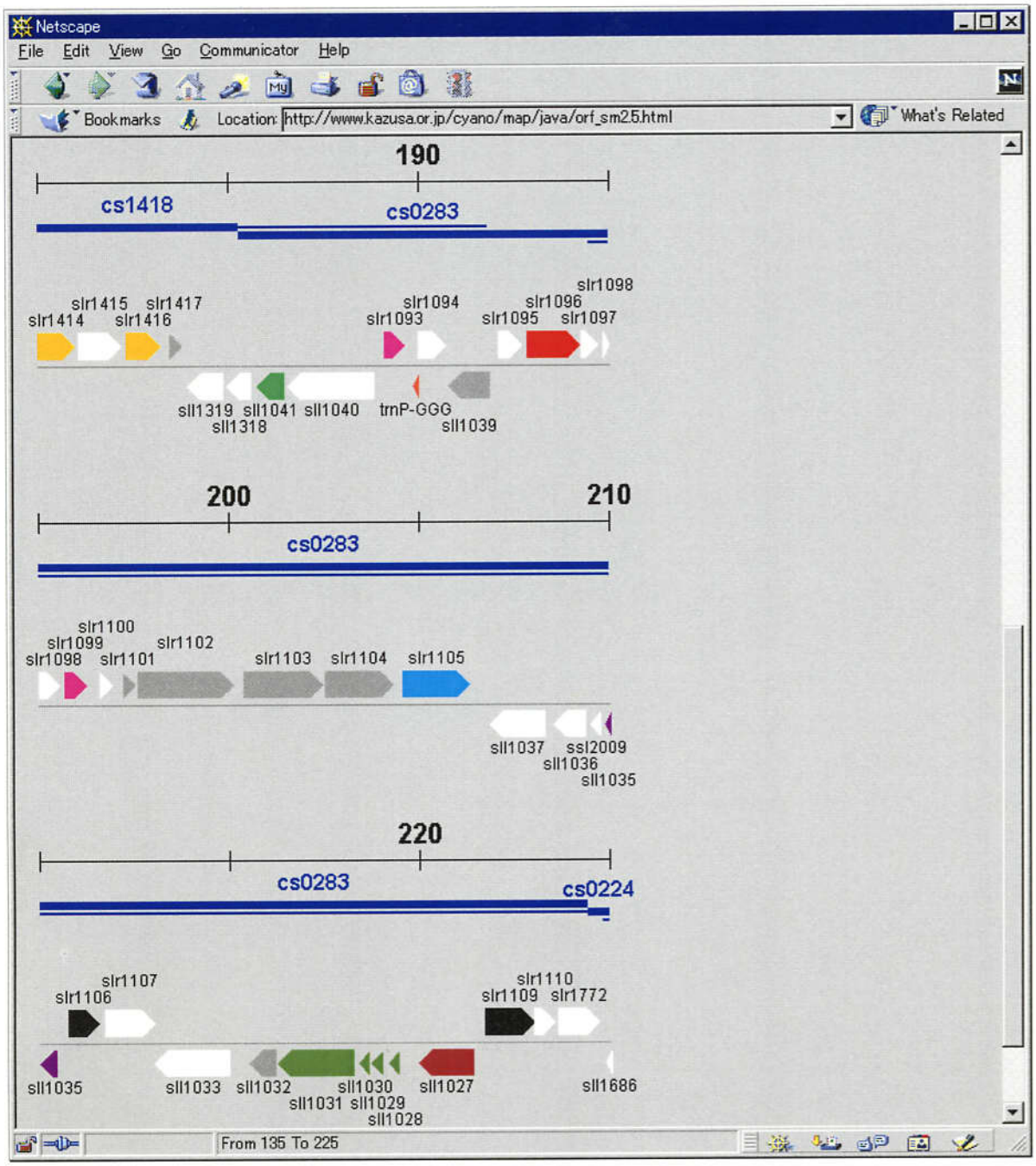


Figure 2-2 Java based image map of a local area

Each blue bar (a clone) or color-coded box (assigned gene) is a hyperlink to information on that area.

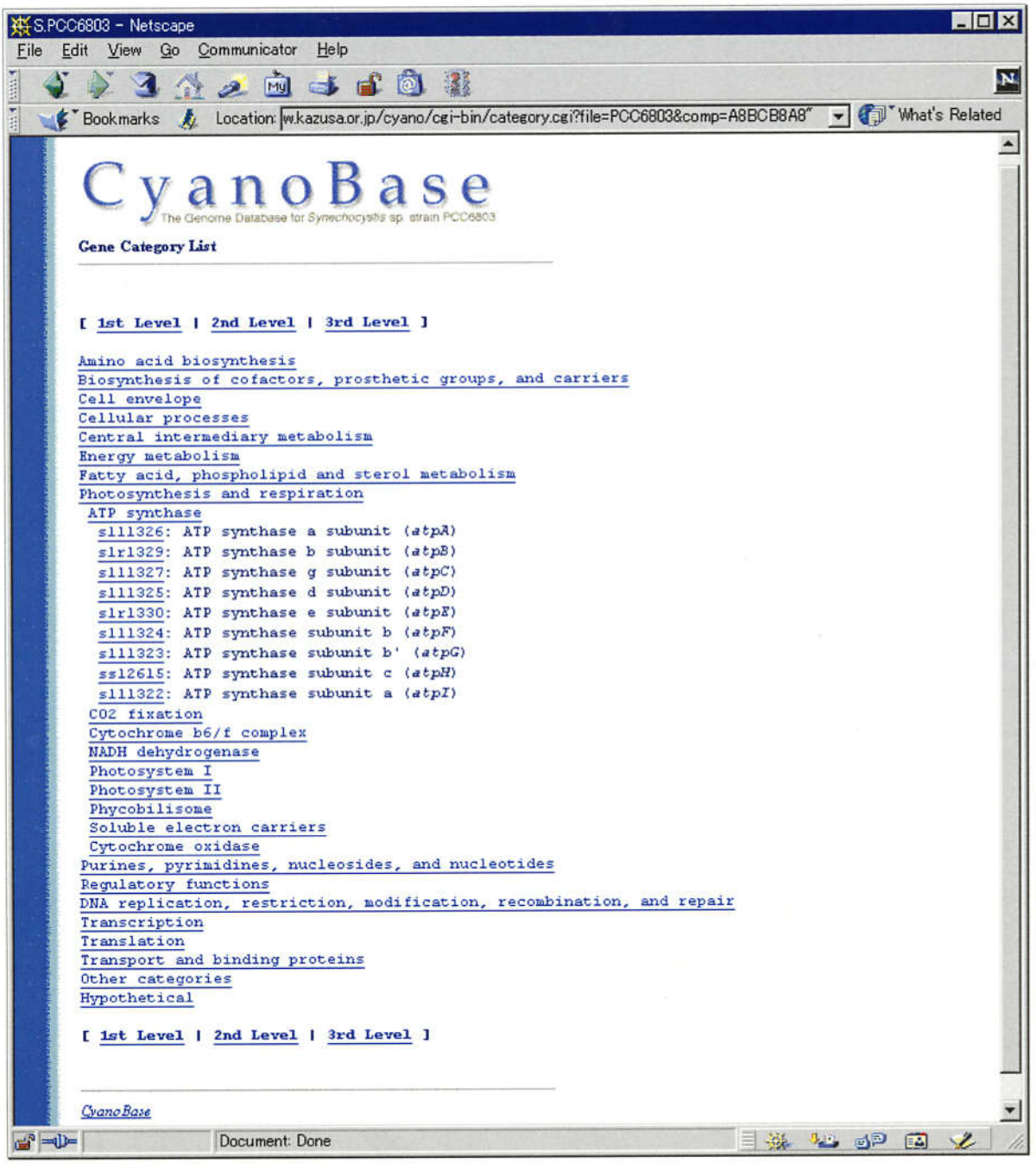


Figure 2-3 The hypertext-based gene classification list

A class (Photosynthesis and respiration) and a subclass (ATP synthase) are depicted as examples.

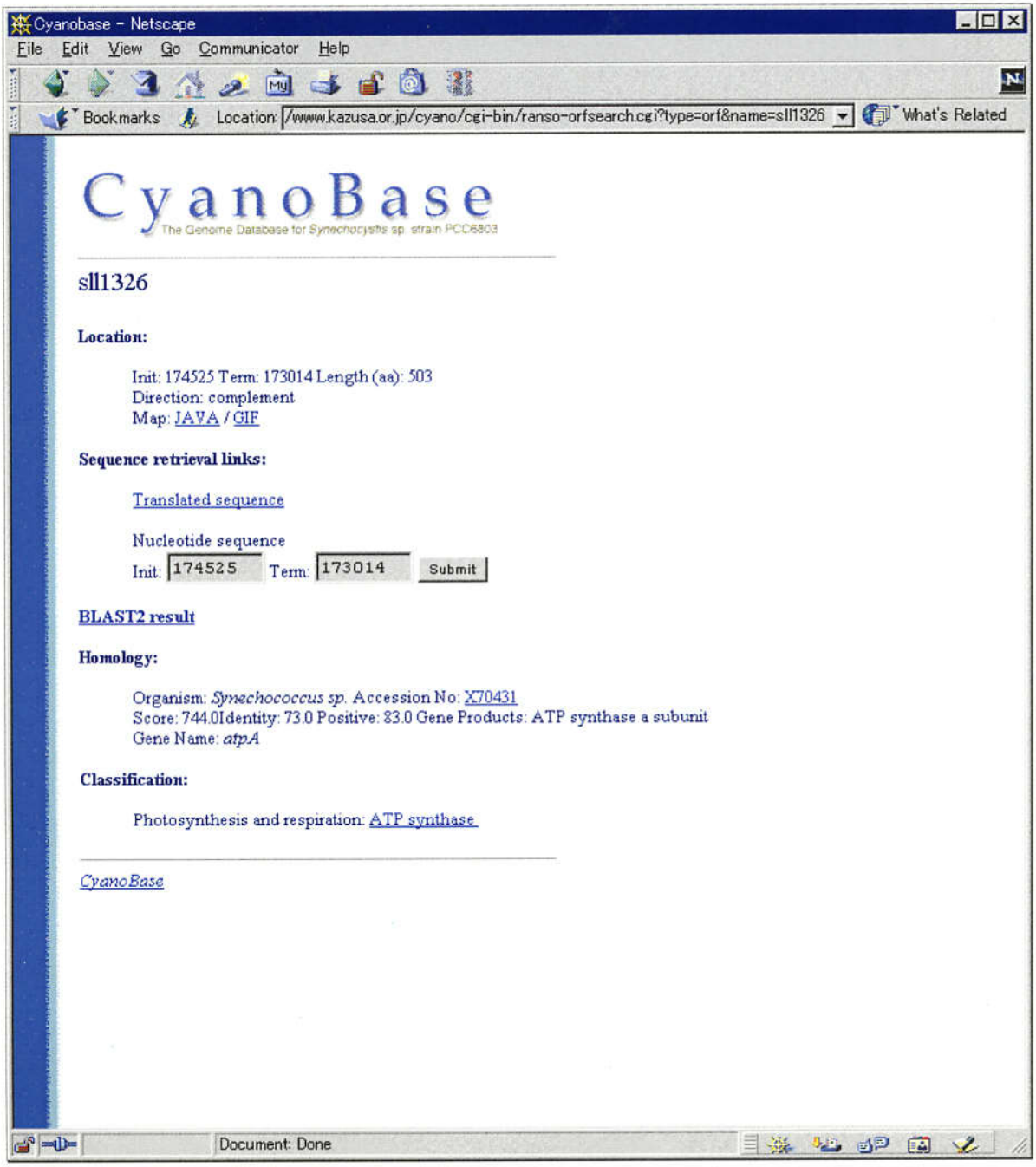


Figure 2-4 An annotation page of CyanoBase

An anntoation for sll1326 (*atpA* gene) is shown as an example.

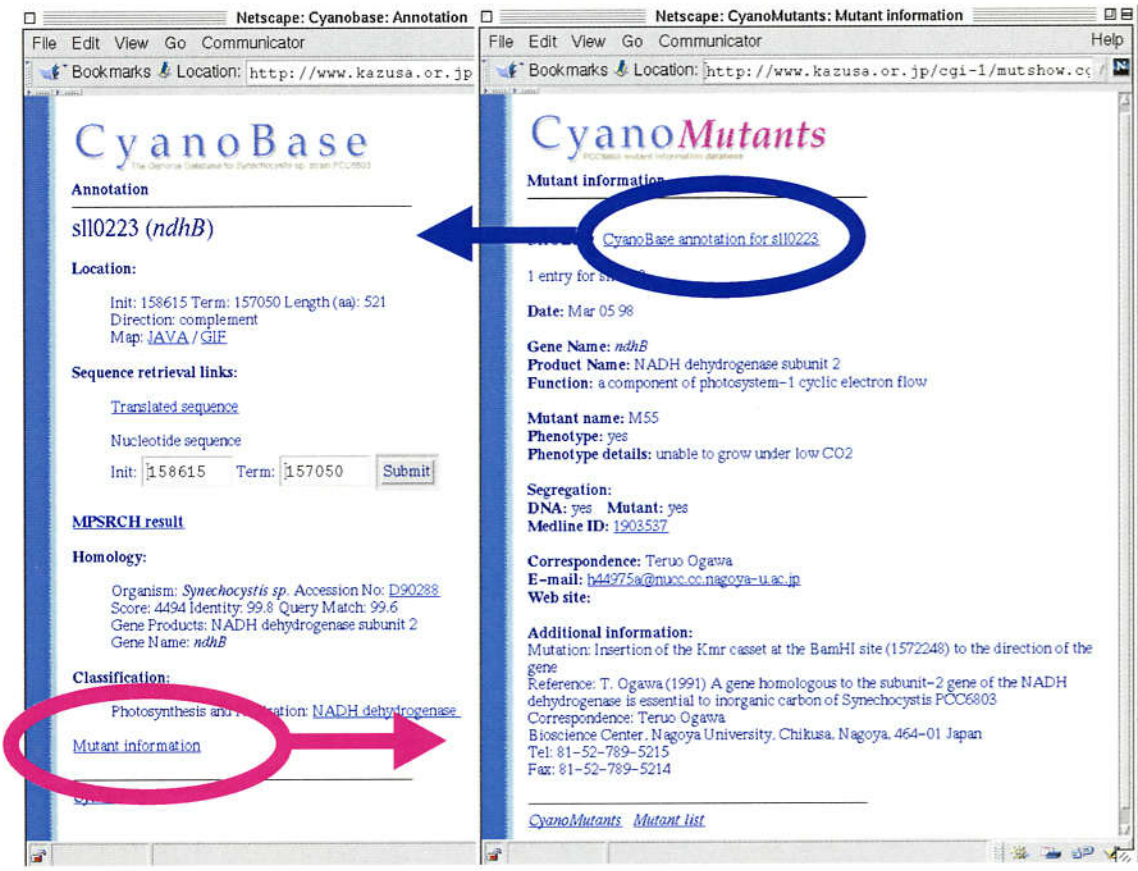


Figure 2-5 CyanoBase/Mutants link

Example of linked entries between CyanoBase and CyanoMutants. Corresponding entries are automatically linked and shown by checking a gene identifier on each database.

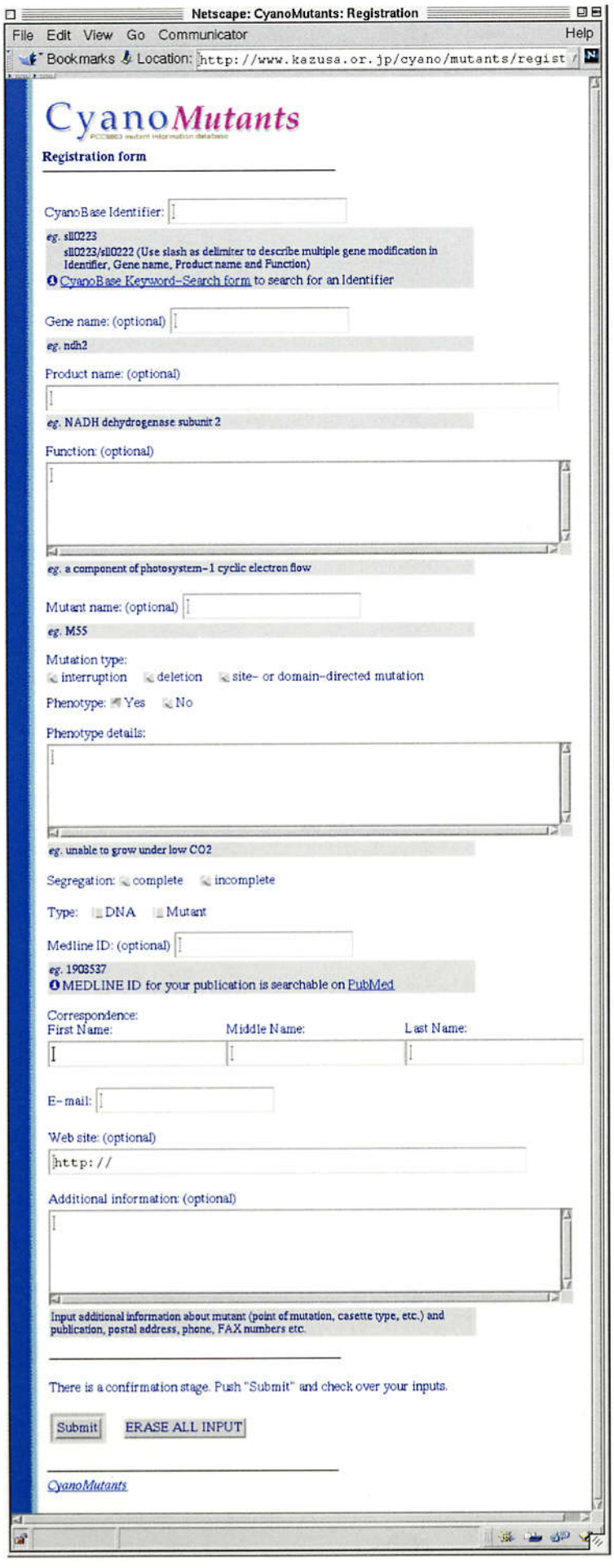

Netscape: CyanoMutants: Registration
File Edit View Go Communicator Help
Bookmarks Location: http://www.kazusa.or.jp/cyano/mutants/regist /
CyanoMutants
Registration form
CyanoBase Identifier:
eg. sll0223
sll0223/sll0222 (Use slash as delimiter to describe multiple gene modification in Identifier, Gene name, Product name and Function)
CyanoBase Keyword-Search form to search for an Identifier
Gene name: (optional)
eg. ndh2
Product name: (optional)
eg. NADH dehydrogenase subunit 2
Function: (optional)
eg. a component of photosystem-1 cyclic electron flow
Mutant name: (optional)
eg. M55
Mutation type:
interruption deletion site- or domain-directed mutation
Phenotype: Yes No
Phenotype details:
eg. unable to grow under low CO2
Segregation: complete incomplete
Type: DNA Mutant
Medline ID: (optional)
eg. 1908537
MEDLINE ID for your publication is searchable on PubMed
Correspondence:
First Name: Middle Name: Last Name:
E-mail:
Web site: (optional)
http://
Additional information: (optional)
Input additional information about mutant (point of mutation, casette type, etc.) and publication, postal address, phone, FAX numbers etc.
There is a confirmation stage. Push "Submit" and check over your inputs.
Submit
ERASE ALL INPUT
CyanoMutants

Figure 2-6 CyanoMutants submission form

A CyanoMutants submission form. The minimal information required for a submission is a gene identifier, phenotype and correspondence name and Email address of submitter. Upon submission, automatic error-checking is carried out. Then if the data contains three sections required to a minimum submission, a confirmation step is given. The submission data is then stored as an entry of CyanoMutants.

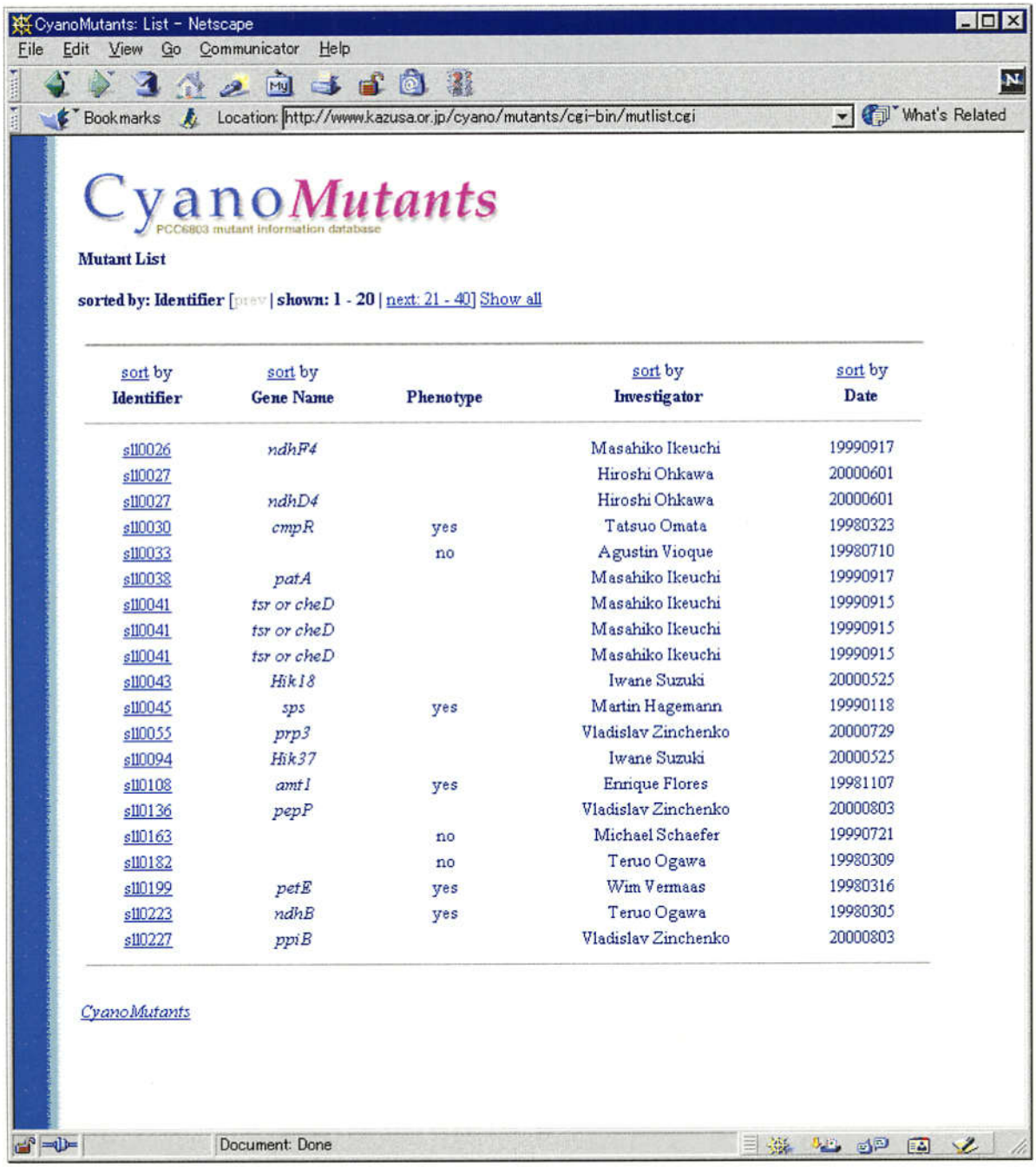


Figure 2-7 List of CyanoMutants

The entry list of CyanoMutants. As a default, the first 20 entries, alphabetically sorted by gene identifier, are shown. Sorting on the other keys can be performed clicking the `sort by' links at the top of columns. Each identifier is a link to detailed information for each mutant.

Chapter 3

High-throughput genome annotation of *Arabidopsis thaliana*

3-1 Introduction

3-1-1 *Arabidopsis thaliana*

A. thaliana is a small flowering plant that is used widely by researchers as a model organism of plant biology.

The genus *Arabidopsis* belongs to the Brassicaceae (mustard or crucifer) family. *Arabidopsis* is found in temperate regions of Asia, Europe, and North Africa and has been widely introduced to other areas of the world. Several species belong to the *Arabidopsis* genus, the most well known and extensively used in research being *A. thaliana* (L.) Heynh (2n=10; common name, thale cress or mouse-eared cress).

Although *A. thaliana* is not of major agricultural significance, it does have several important advantages for research in the genetics and molecular biology of plants.

- In comparison to other angiosperms, *Arabidopsis* has a relatively small genome. The haploid content was estimated at 120 Mb (Goodman *et al.,* 1995; Meinke *et al.,* 1998). Rice, tobacco, and pea have haploid genome sizes of 450 Mb, 1,600 Mb, and 4,500 Mb, respectively. One reason (that the genome is so small) is that there is comparatively little repetitive DNA, with nearly 50% of the nuclear DNA encoding proteins (Lin *et al.,* 1999; Mayer *et al.,* 1999). This feature is convenient for adopting saturation mutagenesis strategies or map-based cloning strategies to identify and clone genes of interest.
- *Arabidopsis* has a short life cycle. It can germinate from seed in 6 to 8 weeks, and growth of the plant is non-seasonal. Therefore, unlike other model systems such as maize or rice, several generations can be produced in a single year. Moreover, its small size makes cultivation in restricted spaces, and thousands of seeds are produced per plant. These features facilitate rapid genetic and mutagenesis experiments with the plant.

- It is diploid, unlike many crop plants, which are polyploid. The identification of recessive traits is easier, and complications due to gene dosage effects do not occur.
- It can be transformed easily by *Agrobacterium tumefaciens*.
- It is self-fertile, which makes maintenance of homozygous lines straightforward. Additionally crosses can be performed if necessary.
- A large number of mutant lines that have been generated through various mutagenesis strategies and of natural populations that have been collected in the wild, and these mutants can be used as tools in the elucidation of the biology of this model system.
- The yeast artificial chromosome (YAC)-based physical maps of chromosomes 2, 4, and 5 have been reported (Zachgo *et al.*, 1996; Schmidt *et al.*, 1996, Schmidt *et al.*, 1997). Sato *et al.* (1998b) have covered more than 80% of chromosome 3 with YAC clones and more than 90% of the long arm of chromosome 5 with YAC and P1 clones. Kotani *et al.* (1997b) assembled a sequence-ready contig map of chromosome 5 for the genome sequencing project.
- EST information for more than 110,000 sequences has been accumulated (Hofte *et al.*,1993; Cooke *et al.*, 1996; Newman *et al.*, 1994; Asamizu *et al.*, 2000).

These features have led to *Arabidopsis* becoming the "model organism" for studies of the molecular genetics of flowering plants.

3-1-2 *Arabidopsis thaliana* genome project

By taking advantages of the above features and circumstances, we initiated large-scale sequencing of the *Arabidopsis* genome in early 1996. The Arabidopsis Genome Initiative (AGI) was then established to facilitate coordinated international *Arabidopsis* genome sequencing projects (Kaiser, 1996). Our group was participating in sequencing of the entire bottom arm and portions of the top arm of chromosome 5 and also the top arm of chromosome 3 as a part of AGI. AGI sequencing group was consist of The Institute for Genomic Research (TIGR), a consortium of Cold Spring Harbor and Washington University

Genome Sequencing Center, and SPP consortium (Stanford University, Plant Science Institute, University of Pennsylvania and Plant Gene Expression Center/USDA-U.C. Berkeley) in United Status, Two European Union consortiums for Chromosome 4 and 5, and Chromosome 3, and Kazusa DNA Research Institute from Japan.

AGI reported the *Arabidopsis* genome sequences, including annotation of predicted genes and assignment of functional categories (The Arabidopsis Genome Initiative, 2000). The reported regions covered 115.4 Mb of an estimated 125-megabase of genome and the sequenced regions contained ca. 25,500 protein genes. In total, Kazusa DNA Research Institute participated 24% (27.6/115.4 Mb) of genomic sequencing and author made 24% (6,124/25,498) of protein gene assignment in the report. Since at the time of publication, several sequence gaps in chromosome arms were being sequenced, the complete genomic sequences and deduced information on the unique regions on the genome will be available in early the year 2001.

3-1-3 Computer-aided annotation procedure

From data collection and analysis to data management, computer systems play essential roles in genome research. Without high-performance computers and appropriately designed data processing systems, high-throughput genome sequencing projects would not be possible.

The primary computational tools available for analysis of nucleotide sequences are database search programs such as BLAST and coding-region prediction. Because not all protein-coding genes are known or stored in databases, prediction of protein-coding regions is necessary based on statistical information. Such gene identification efforts that use prediction tools are important in large-scale eukaryotic genome sequencing projects.

Annotation, which is the elucidation and description of the biologically relevant features of a sequence, is an essential and time-consuming process in current genome sequencing projects. Annotation is necessary to increase the usefulness of the sequence data, and the quality with which annotation is done directly affects the value of the sequence.

Especially, gene-modeling procedure is complex due to the exon-intron structure of eukaryotic genes.

GENSCAN (Burge and Karlin, 1997) uses a probabilistic model of gene structure and compositional properties to predict complete gene structures, including exons, introns, promoter, and poly(A) signals, in genomic sequences. Forward and backward recursions are performed which allow determination of the most likely gene structure in the sequence, probability of each predicted exon, and (optionally) all suboptimal exons with probability above a given cutoff.

GRAIL (Uberbacher and Mural, 1991) uses a neural network that combines a series of coding prediction algorithms to provide analysis of the protein coding potential of a DNA sequence. GRAIL can search for poly(A) sites, CpG islands, repetitive DNA, and frameshift errors in DNA sequences of human, mouse, *Arabidopsis, Drosophila* and *E. coli* genomes.

There are two splice site prediction algorithms specially trained for *Arabidopsis* in public. NetGene2 (Hebsgaard *et al.,* 1996) uses artificial neural networks combined with a rule-based system to predict intron splice sites in *Arabidopsis*, human and *C. elegans*. SplicePredictor (Brendel and Kleffe, 1998) is another algorithm for splice site prediction. It implements models based upon the two variables of splice site signal strength and compositional contrast for splice site prediction, trained on reliable sets of maize and *Arabidopsis thaliana* genomic sequences.

tRNAscan-SE program (Lowe and Eddy, 1997) identifies transfer RNA genes in genomic DNA or RNA sequences. It combines the specificity of the Cove probabilistic RNA prediction package (Eddy and Durbin, 1994) with the speed and sensitivity of tRNAscan 1.3 (Fichant and Burks, 1991) plus an implementation of an algorithm described by Pavesi *et al.* (1994), which searches for eukaryotic pol III transfer RNA promoters (EufindtRNA).

Although good results have been obtained with a variety of computational algorithms, no perfect gene-modeling program has been available. The problem of gene

structure prediction has not been completely solved. Therefore, a manual combination of information must be used to deduce eukaryotic gene structure precisely.

3-2 Materials and methods

3-2-1 Sequencing strategy

Sequence analysis was started from multiple initiation points of the genome with the use of P1, TAC or BAC clones containing known markers as templates. PCR was used to examine the authenticity of the selected clones by anchoring both ends of the clones onto YAC contigs. The nucleotide sequence of each clone was determined with a shotgun-based strategy.

3-2-2 Sequencing data assembly

Chromatograms from the sequencer were transferred from Macintosh to Unix workstations via ftp and then base-called into sequences with the phred program (Ewing *et al.,* 1998; Ewing and Green, 1998). They were assembled and edited with the phrap (Phil Green, University of Washington, Seattle, USA) and consed (Gordon *et al.,* 1998) program or the AutoAssembler software from PE Applied Biosystems. With single-pass random sequences corresponding to approximately 12 times the equivalent of a P1, TAC or BAC insert, most of the inserts were assembled into a single contig with more than 80% coverage of both strands. The minimum requirement for confirmation of the sequences in the single-stranded gaps was to sequence the complementary strand by either dye-terminator or dye-primer sequencing or to sequence the same strand by both dye-terminator and dye-primer sequencing. This strategy secured enough accuracy for further analysis of gene structure. Comparison of independent reassembly sequenced clones revealed accuracy rates between 99.99 and 99.999%.

3-2-3 Automated search and prediction process

An overview of the data flow in this analysis engine is illustrated in Figure 3-2. Completed nucleotide sequences were subjected to similarity searches against the non-redundant protein database, nr, with use of the BLASTX algorithm in the BLAST2 program (Altschul *et al.*, 1997). Information for local alignments, that showed E-values of < 0.001 to known protein-coding sequences were extracted and stored. Potential exons of protein-coding genes were predicted by the computer programs GRAIL and GENSCAN. For localization of exon-intron boundaries, donor/acceptor sites for splicing were predicted by NetGene2 and SplicePredictor. To identify transcribed regions and structural RNA genes, the BLAST program was used to compare nucleotide sequences with the EST and RNA gene databases. Transfer RNA genes and their structures were predicted by tRNA-scanSE.

All the outputs were parsed and stored in GFF (General Feature Format; formerly, Gene-Finding Format described in a document at http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml). The results were then parsed and loaded into a web-based display system called *Arabidopsis* Genome Displayer. Figure 3-3 shows sample images from *Arabidopsis* Genome Displayer. Gene structures proposed in the annotated sequences as well as those predicted by computer programs, are presented, and each graphic item is a hyperlink to detailed information for the corresponding area.

The *Arabidopsis* Genome Displayer shows the relation of the features in the database along a genomic sequence. Simultaneously, an annotation-making interface allows manual editing of the gene model showing tentative nucleotide and protein sequences and exon-intron organization. Figure 3-4 illustrates a sample procedure for the gene-modeling process. To predict exact donor/acceptor sites for splicing, gapped alignments made by nap in the AAT package (Huang *et al.*, 1997) were examined. Also, alignments made by gap were examined to fit EST sequences on the genomic sequence. After the editing process, the most reasonable model of a region is saved in our in-house database system as a deduced gene structure.

3-2-4 Gene function assignment

Modeled genes were annotated by similarities to known functionally annotated genes. Table 3-2 shows the criteria used for annotation of the *A. thaliana* protein-coding genes. The deduced genes whose functions could be predicted were classified according to the Riley system with modify by Kaneko *et al.* for the genome of photosynthetic organisms (Riley, 1993; Kaneko, 1996) into 22 categories with respect to different biological roles. The function of each gene was predicted through sequence comparisons with genes with known functions.

3-3 Results and discussion

3-3-1 Sequenced regions and clones

Thirteen contigs of 27,622,037 bp were assembled from 472 P1, BAC, TAC and gap-closing units. Average insert size of P1 and TAC, and BAC were 80 kb and 100 kb respectively. A YAC and a lambda clone and a direct PCR product from genomic DNA (CIC5B3, LA522 and GA469 in Table 3-1b, respectively) were used to cover regions not found by cloned P1, TAC or BAC. Figure 3-1 shows Kazusa-allocated regions on chromosome 3 (1 contig) and 5 (12 contigs) for genomic sequencing. The relative positions of the sequenced clones are listed in Table 3-1 as sequence contigs.

3-3-2 Construction of automated sequence processing and gene modeling systems

I developed a system to automate execution of similarity searches and gene prediction programs (above part of Figure 3-2). When a sequence unit (an insert of P1, TAC or BAC clone) is finished to be nucleated, the sequence information was subjected to similarity searches against protein and EST databases and several prediction algorithms automatically. Then the results are parsed and stored in a database. Dataset for each

clone can be called and visualized with a web-based display system. The display system: *Arabidopsis* Genome Displayer has been open as a WWW database to public users to browse sequence information deposited from the international sequencing teams through a user-friendly graphic interface and search engines as shown in Figure 3-3.

For gene identification on *Arabidopsis thaliana* genomic sequences, I constructed a manual gene-modeling system that combines and displays the outputs of search and prediction algorithms. The display system described above has made possible high-speed gene composition as illustrated in Figure 3-4. In the process of gene-modeling, the results from programs are parsed and loaded into a WWW-based information display system. Since this interface visualize the positional relationships of the searched or predicted features along the sequence, an annotator can construct working gene-models easily. For each working model, the annotator performs similarity searches to known sequences with BLAST during the gene-modeling process. If a BLAST alignment to known gene has one or more gaps to be edited, another model must be composed and checked. Regions without BLAST hits were composed only with use of information from gene-finding programs. After careful editing process, the most reasonable gene model on a region was stored as a deduced gene structure in the GFF-based annotation data file with version number and the name of annotator. A high-throughput gene-modeling process of Kazusa-allocated region of *A. thaliana* sequences was carried out with the assistance of the system. Throughput of the gene-modeling process was carried out at 225k/day/person (200 exons or 50 protein genes) at the end of the annotation works.

The three annotation centers (Kazusa, TIGR and MIPS) in AGI constructed separately similar annotation approaches involving *in silico* gene-finding methods, comparison to EST and protein databases, and manual adjustment of the data. All of the gene-finding algorithms were trained and optimized with parameters based on known *Arabidopsis* gene structures. Eighty per cent of the gene structures predicted by the three informatics centers were completely consistent, 93% of ESTs matched gene models, and less than 1% of ESTs matched predicted non-coding regions, indicating that most potential genes

were identified by the gene-finding process.

3-3-3 Gene content and feature

In conclusion, 6,124 potential protein-coding genes and 127 RNA genes were deduced in the 27,061,818 bp regions of chromosomes 3 and 5 of *Arabidopsis* that were covered by 461 P1, TAC, and BAC clones and a YAC, a direct PCR and a lambda clone as gap-closing units. Since 11 clones on chromosome 5 (shown with asterisk in contig 9 or contig 10 in Table 3-1b) were originally allocated EU *Arabidopsis thaliana* genome sequencing project and were annotated by MIPS (Munich Information center for Protein Sequences), annotations for these clones were excluded in this work. The average density of protein-coding genes in the annotated regions in Kazusa was estimated to be 1 gene per 4.4 kb. Structural features and deduced functions of all the protein and RNA genes are available from DDBJ with each accession on Table 3-1 or via anonymous ftp on Kazusa DNA Research Institute (ftp://ftp.kazusa.or.jp/arabidopsis/chr3 or chr5).

Of the 6,251 protein and RNA genes deduced, 5,196 (83%) were assigned either contains clear or partial similarity to genes in the DNA databases and 1,055 (17%) were allocated as previously unknown genes (Figure 3-5). From a viewpoint of the function, of the 6,251 genes, 3,284 (53%) were assigned either a definite or putative function and 2,947 (47%) were labeled as functionally unknown genes (Figure 3-6). In addition, 769 pseudogenes, most of which are related to proteins found in retrotransposons and are located near the centromere, were found. One hundred twenty-seven genes that encode structural RNAs were also identified: 108 transfer RNAs, 1 rRNA, 17 snRNAs, and 1 7SL RNA. Of identified protein genes, 2,808 carried EST sequences, indicating that 46% of the total genes in *A. thaliana* may be represented in the current EST databases.

Classification of the known and putative genes by biological role or biochemical function shows that most major cellular processes appear to be represented, with genes involved in regulatory functions (e.g., DNA-binding proteins/transcription factors), protein fate (e.g., protein trafficking, folding, degradation), and signal transduction (e.g., protein

kinases) comprising the largest functional groups (Figure 3-7).

3-3-4 Conclusion

In the present study, I developed a gene-modeling system including automated gene finding system and WWW-based data visualization system for eukaryotic genome sequencing projects. With use of the high-throughput annotation system, 6,124 potential protein-coding genes and 127 RNA genes were deduced to the 27 Mb regions of *Arabidopsis* chromosomes 3 and 5 along the line of the international agreement of AGI. In the year 2000, AGI reported the *Arabidopsis* genome sequences, including annotation of predicted genes and assignment of functional categories (The Arabidopsis Genome Initiative, 2000). The reported regions covered 115.4 Mb and the sequenced regions contained about 25,500 protein genes. In total, author's group was participate one fourth of genomic sequencing and author made one fourth of protein genes' assignment in the report.

The complete sequences of chromosomes 2 and 4 of *Arabidopsis* were reported previously (Lin *et al.,* 1999; Mayer *et al* 1999), and then the reports for chromosomes 1, 3, and 5 were published (Theologis *et al.,* 2000; European Union Chromosome 3 *Arabidopsis* Sequencing Consortium, The Institute for Genomic Research & Kazusa DNA Research Institute, 2000; The Kazusa DNA Research Institute, The Cold Spring Harbor and Washington University in St Louis Sequencing Consortium & the European Union *Arabidopsis* Genome Sequencing Consortium, 2000). The sequence information and annotations by this international collaboration will greatly simplify functional analyses and genetics of higher plants. The information must impact in many aspects of current scientific fields.

Although, rapid functional prediction for protein genes were performed with use of automated annotation systems for the final publication (The Arabidopsis Genome Initiative, 2000), refinements of annotations and functional reassignments of the 25,000 genes must be proceeded with use of various bioinformatics and massive and systematic experimental efforts. Through the WWW-based data visualization system described in this chapter, any

investigator can examine bases of gene-modeling and functional assignment for each deduced gene. I hope the user-friendly data visualization system described here supports high-throughput refinement of predicted gene structures and functional assignments.

3-4 Figures and Tables

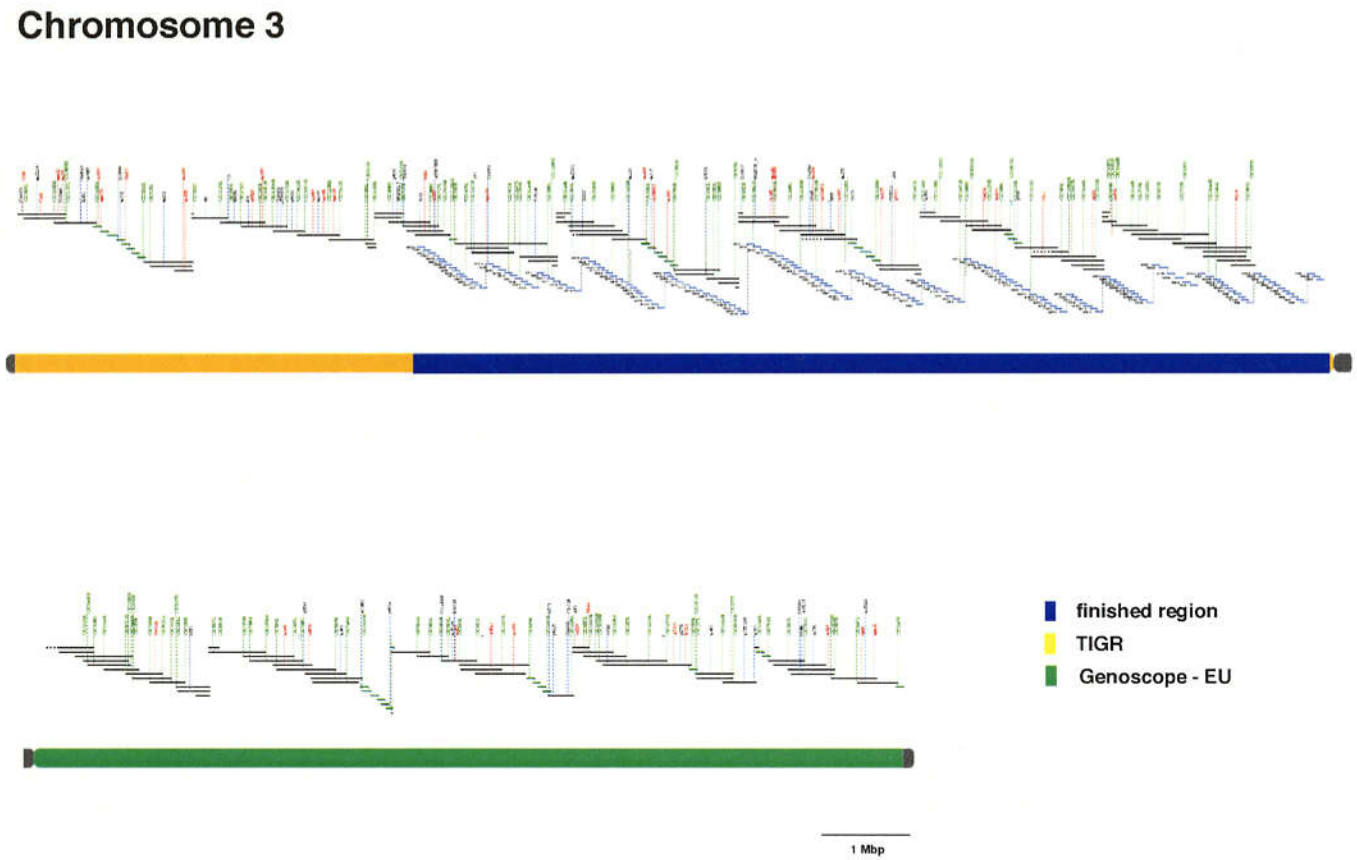


Figure 3-1a Kazusa allocated regions on chromosome 3

Relative locations of the sequenced clones (blue bar) and the associated markers on the YAC-based physical map of chromosome 3.

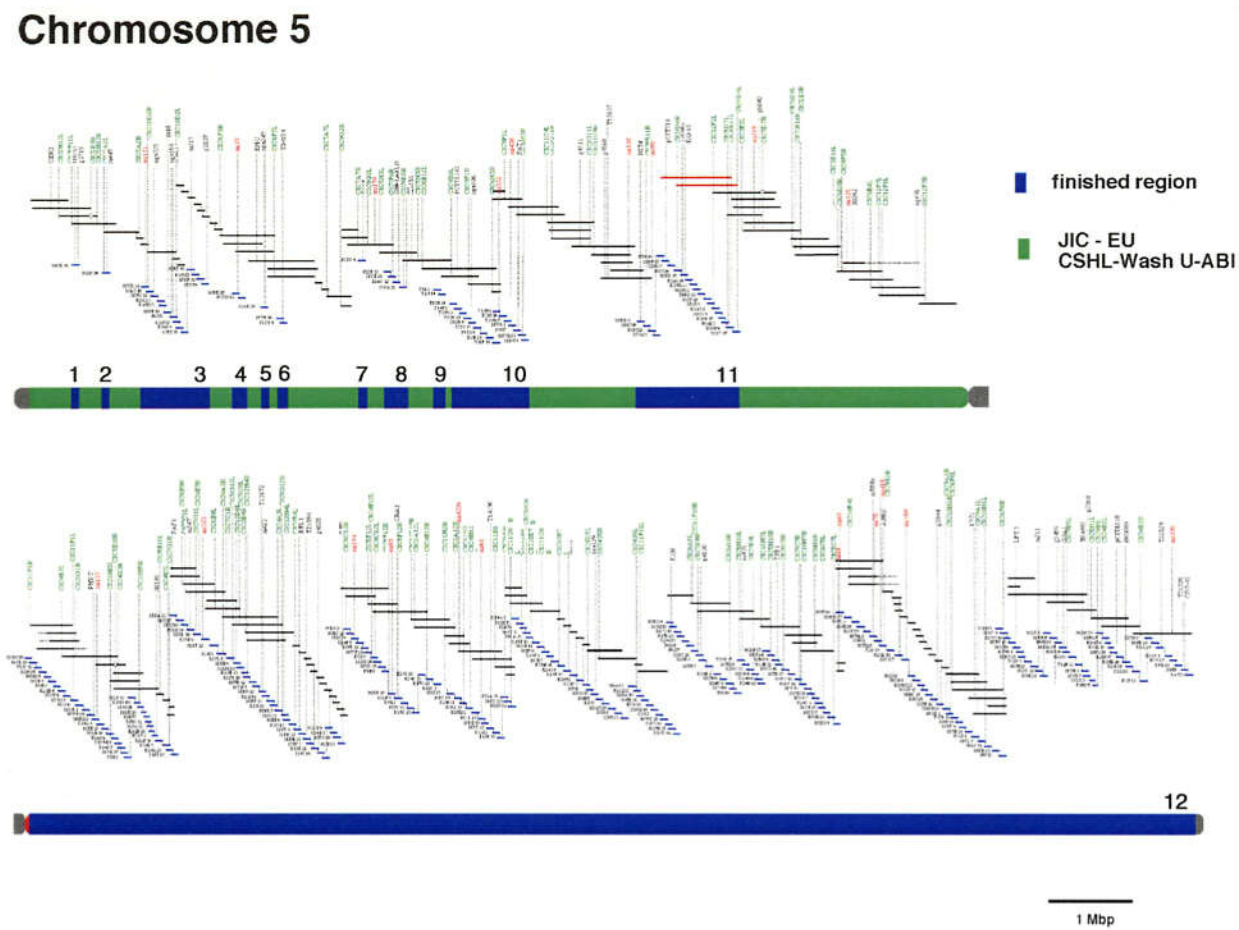


Figure 3-1b Kazusa allocated regions on chromosome 5

Relative locations of the sequenced clones (blue bar) and the associated markers on the YAC-based physical map of chromosome 5. Numbers on each blue bar corresponds to contig number in Table 3-3b.

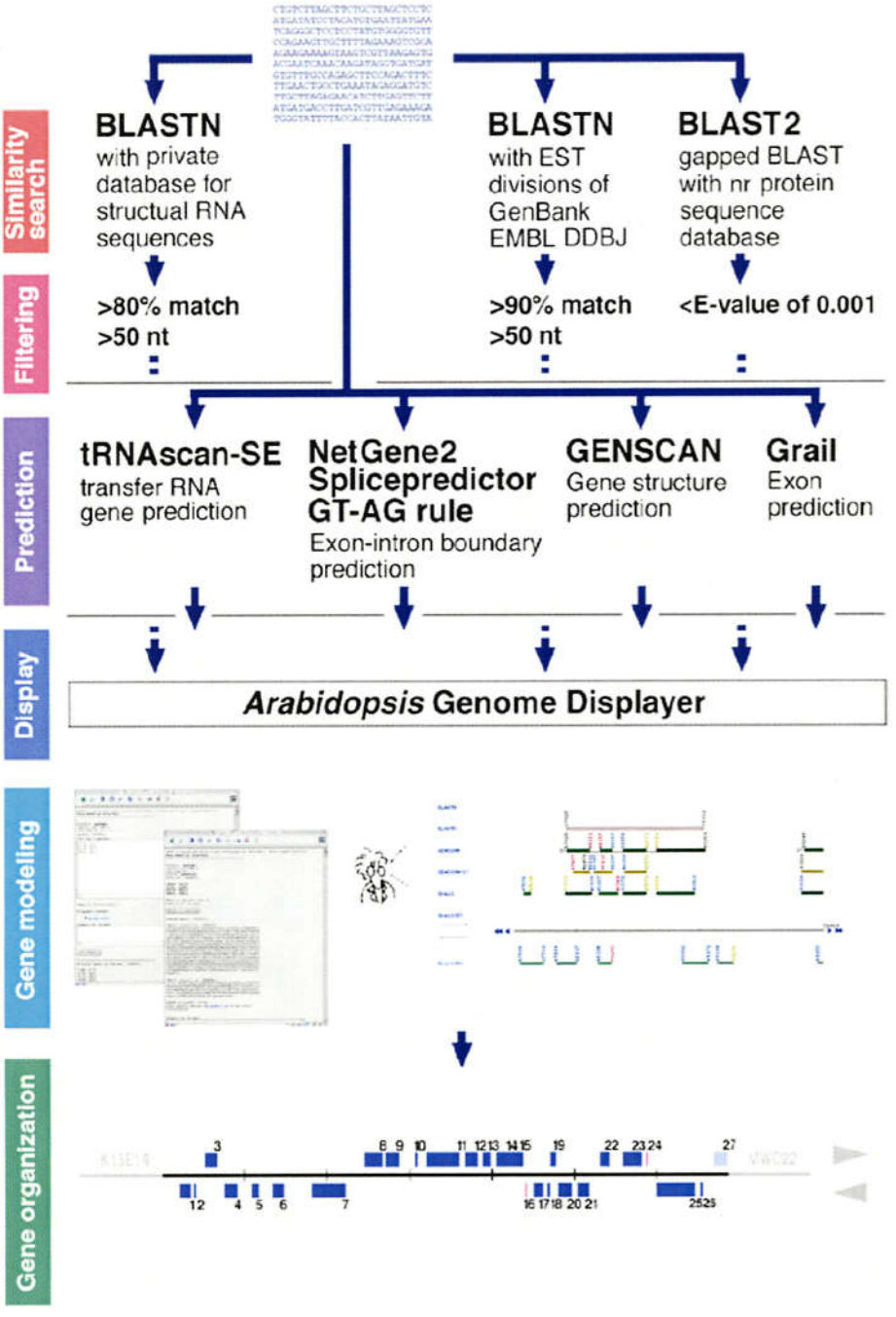


Figure 3-2 An overview of computer-aided annotation Procedure at the Kazusa *Arabidopsis* genome sequencing project

Processes in similarity search, filtering, performing prediction tools and data storage to display, are automated to be executed for a newly determined sequence. Manual gene-modeling process is detailed in Figure 3-4.

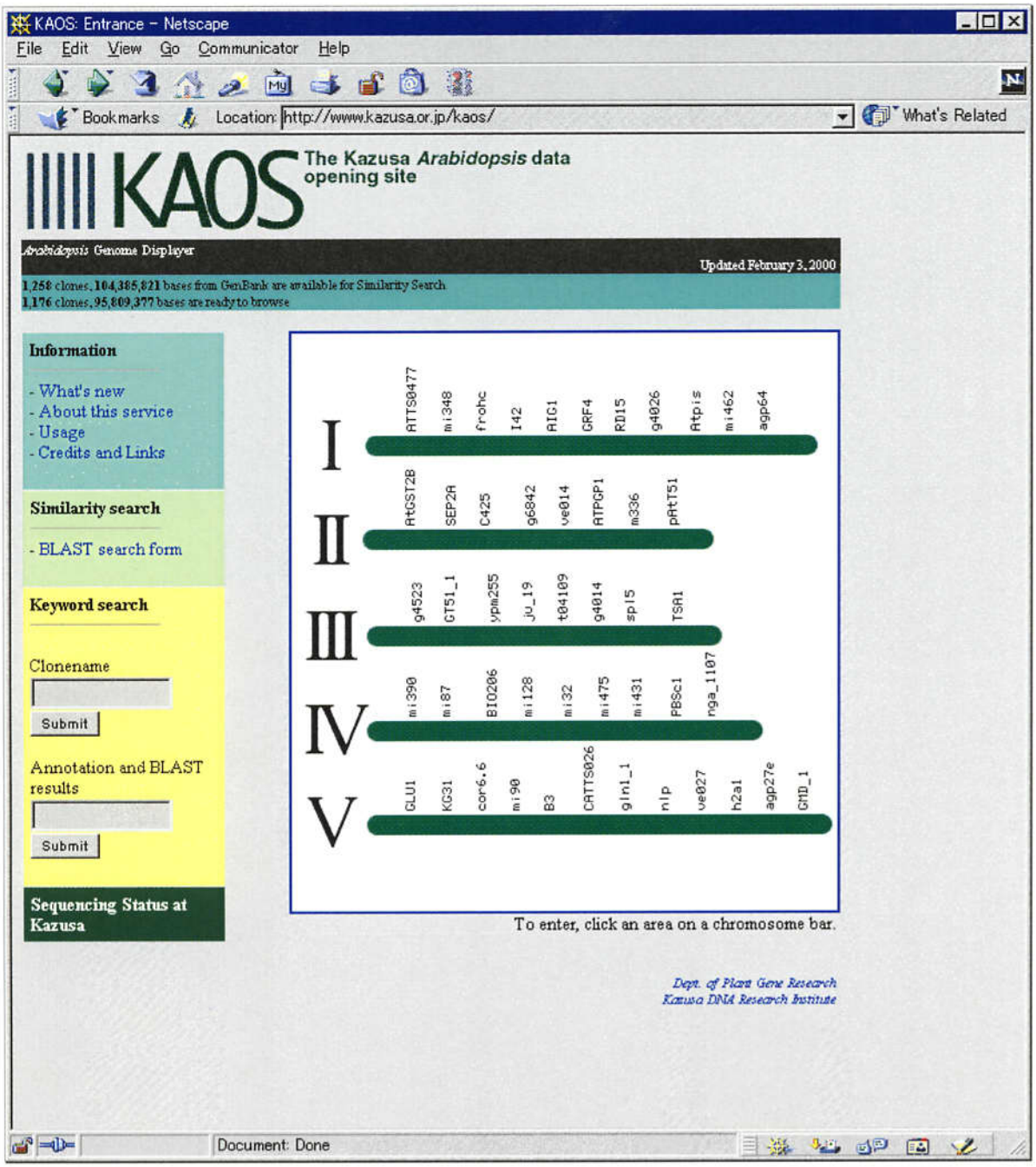


Figure 3-3 *Arabidopsis* Genome Displayer

Figure 3-3a Top page

Image of five chromosomes shown in the right side is a client-side clickable map. If a certain chromosome number is clicked, a list of all the sequence-finished clones on the requested chromosome is shown. If a region between genetic markers placed on the chromosome images shown in green bars is clicked, sequenced clones within the region are listed.

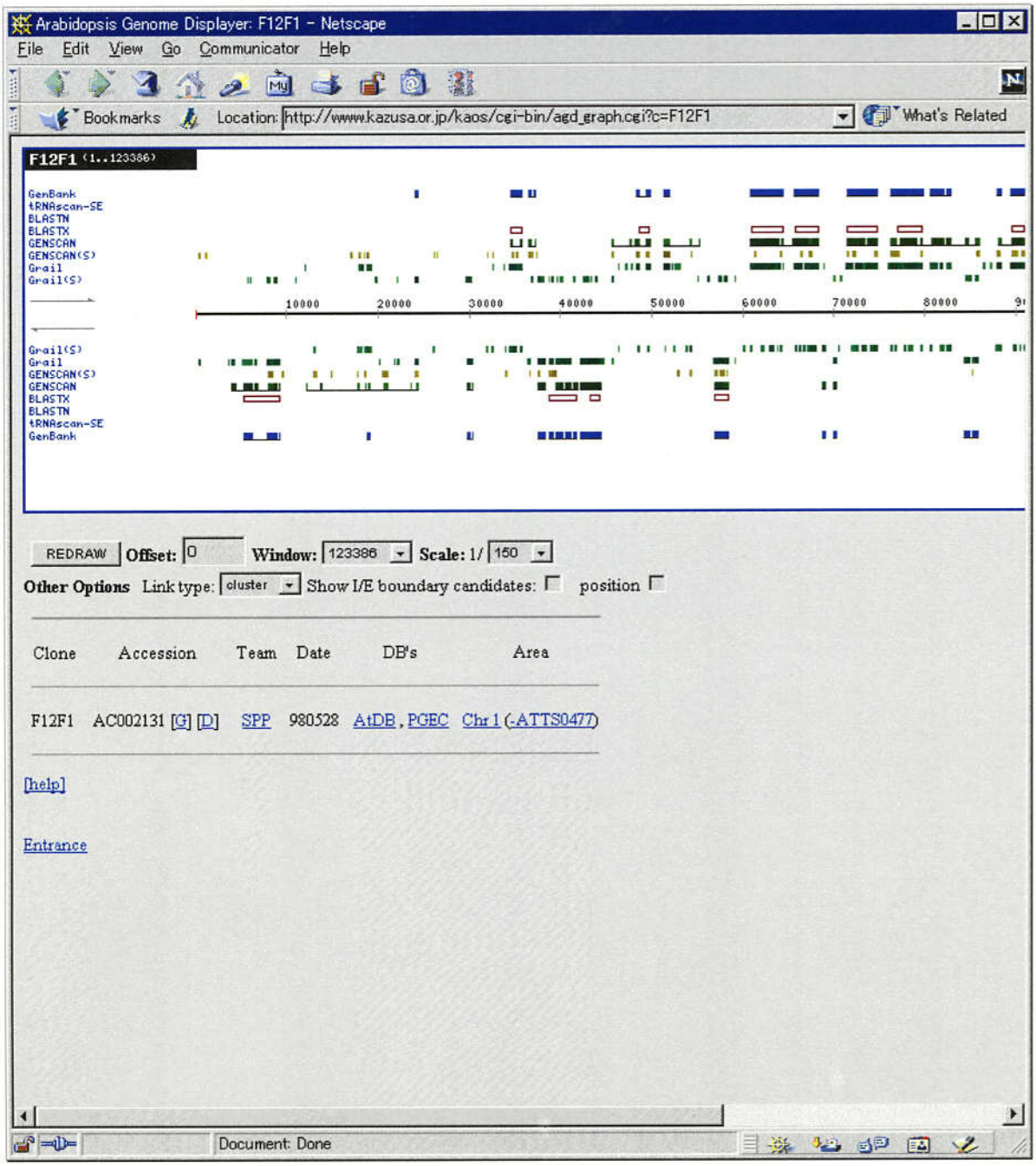


Figure 3-3b Graphics along a clone

When the clone name of a status ready clone is clicked, its whole image is shown in a different window. On the image, annotated CDS's (protein genes), BLASTN/BLASTX alignment clusters and GENSCAN gene models are clickable at the default setting, "cluster".

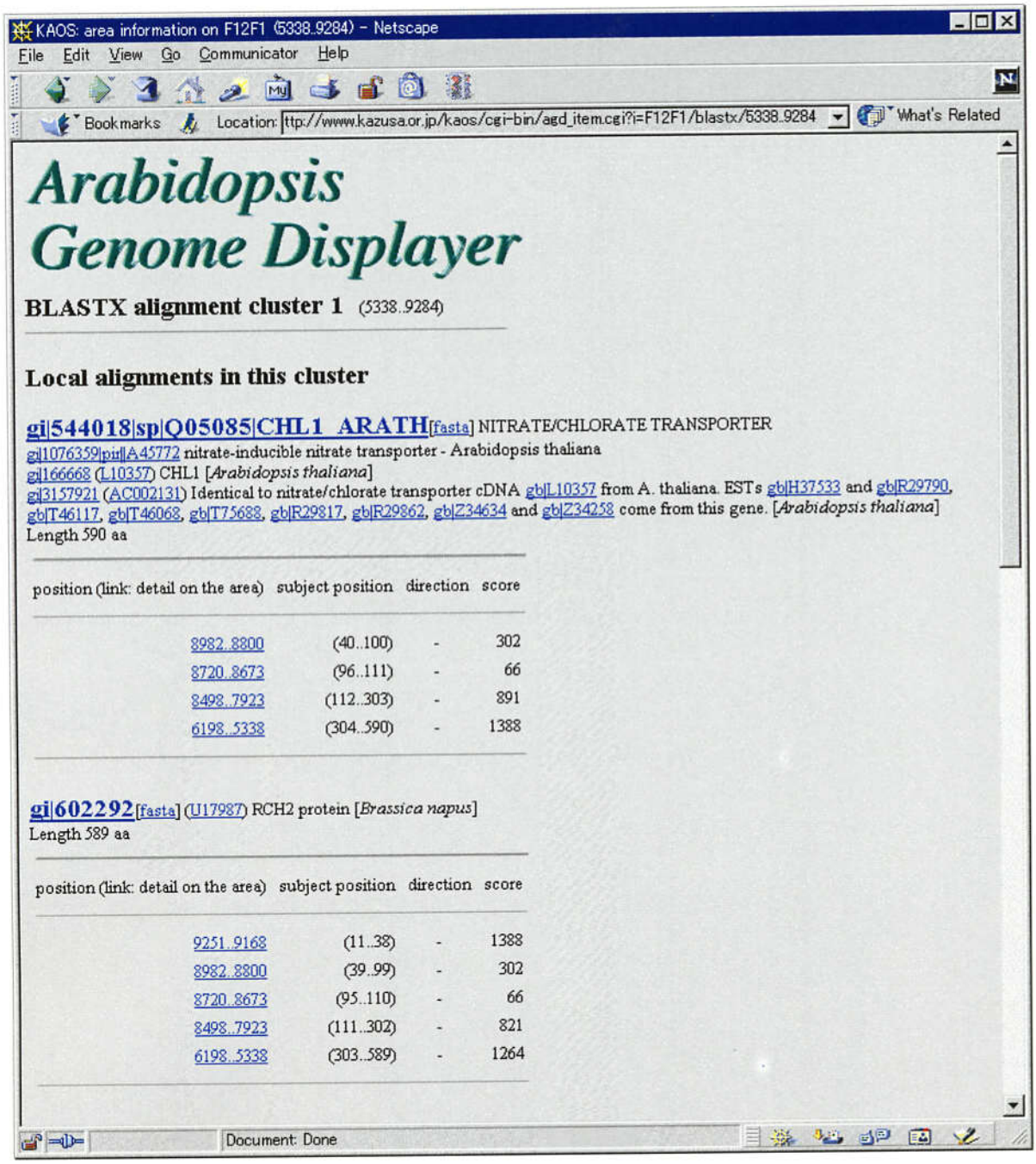


Figure 3-3c List of a cluster

When "cluster" is selected for Link Type, clickable images are linked here. Displayed here is the collection of same directional BLAST local alignments, CDS and gene model by GENSCAN, which means, clustered exons within the same gene are listed.

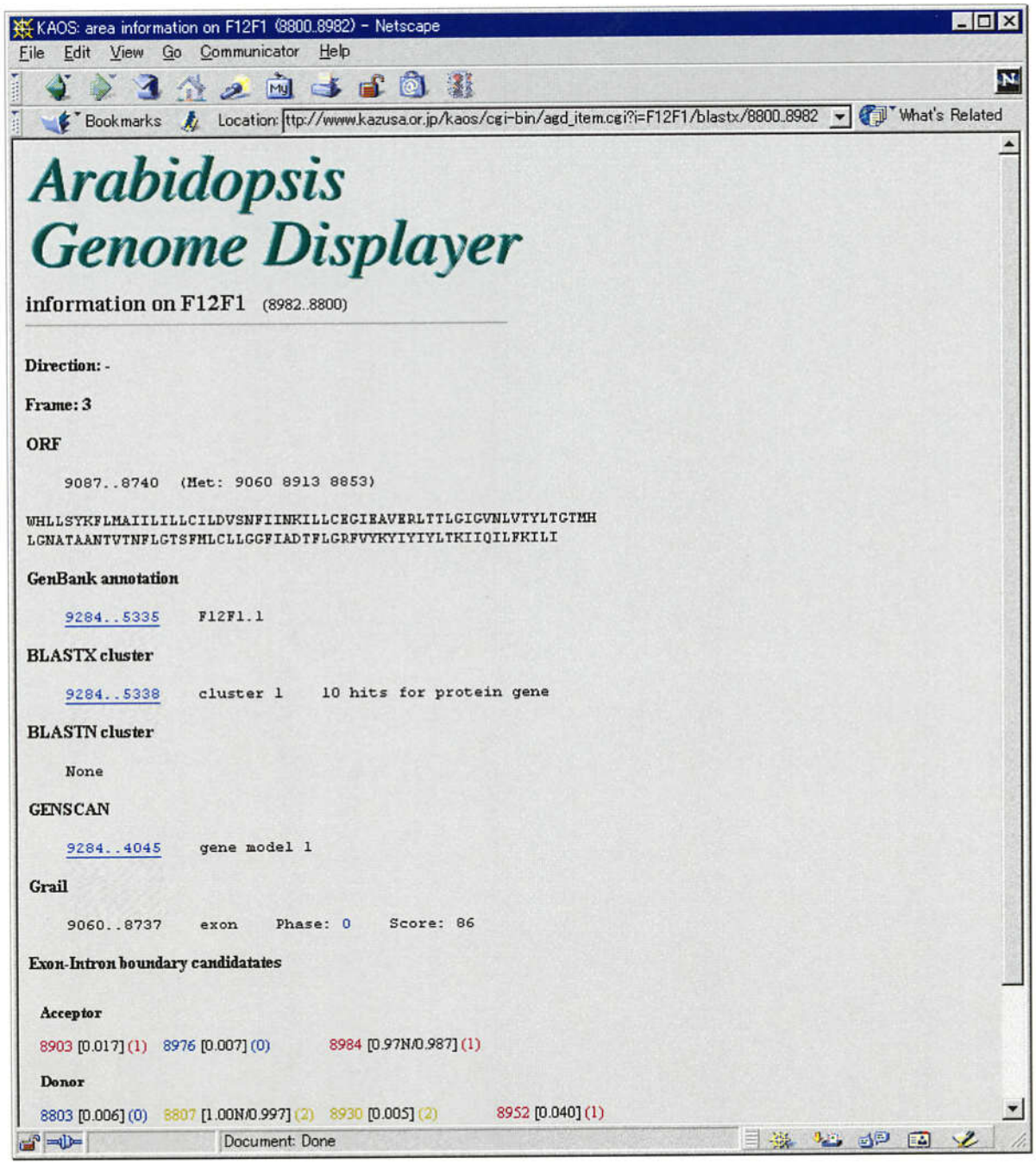


Figure 3-3d Information on a region

It is composed of ORF(s). Exons in annotation, Grail and GENSCAN within the same region are listed. Their positions are linked to the "cluster" page. Links for a single exon like Grail exon and GENSCAN suboptimal exon are only mentioned. Intron/exon boundary candidates and their phases on the frame are also shown. Phases are colored in the same way as graphics are, that is, phase 0 (intron/exon boundary is at the end of the codon) is blue, phase 1 (at the first) is red and phase 2 (at the second) is yellow.

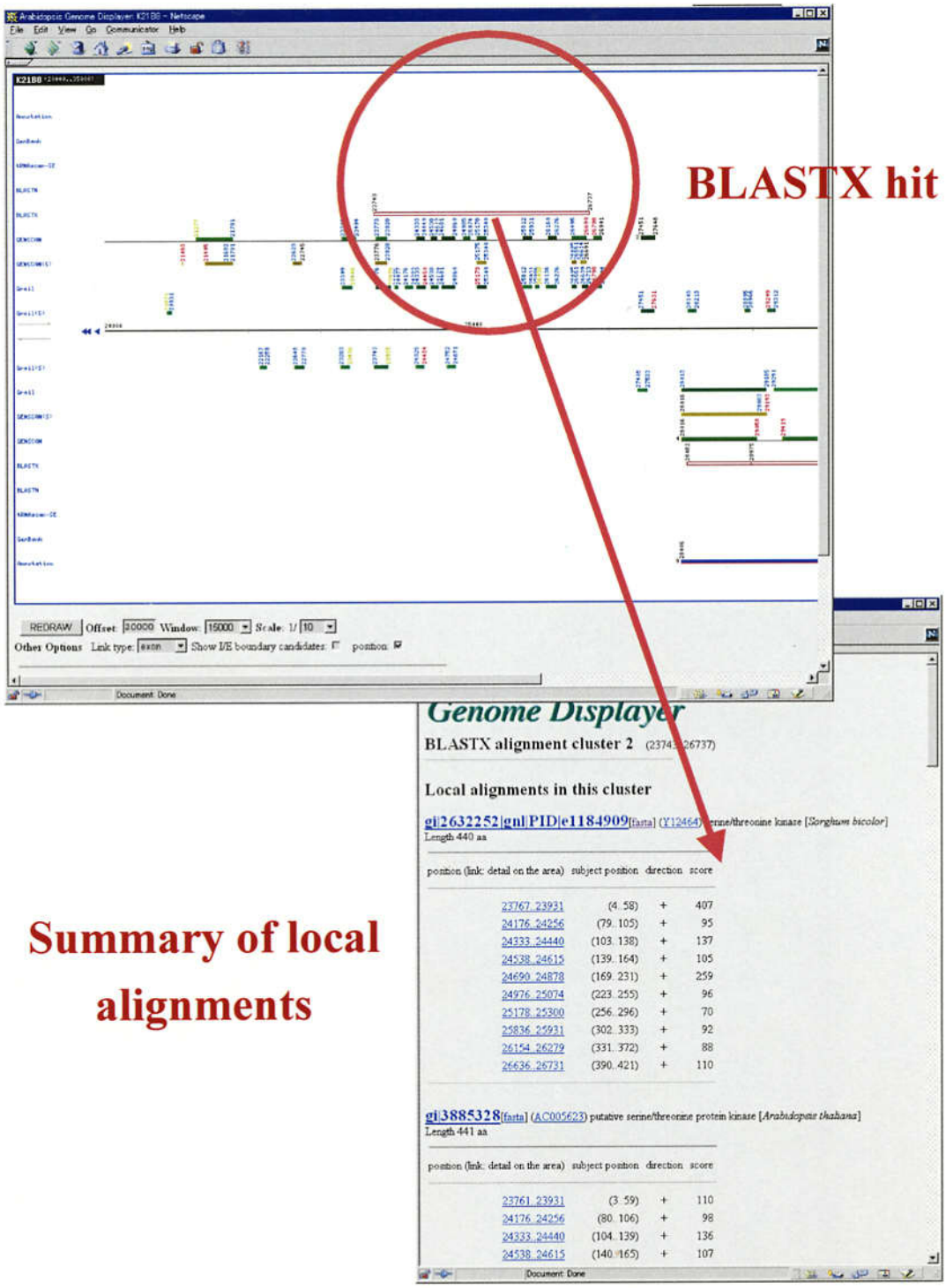


Figure 3-4a Gene modeling procedure (1)

When one or more BLAST hits are found, an annotator open a summary of local alignments in a region interested in, using *Arabidopsis* genome displayer.

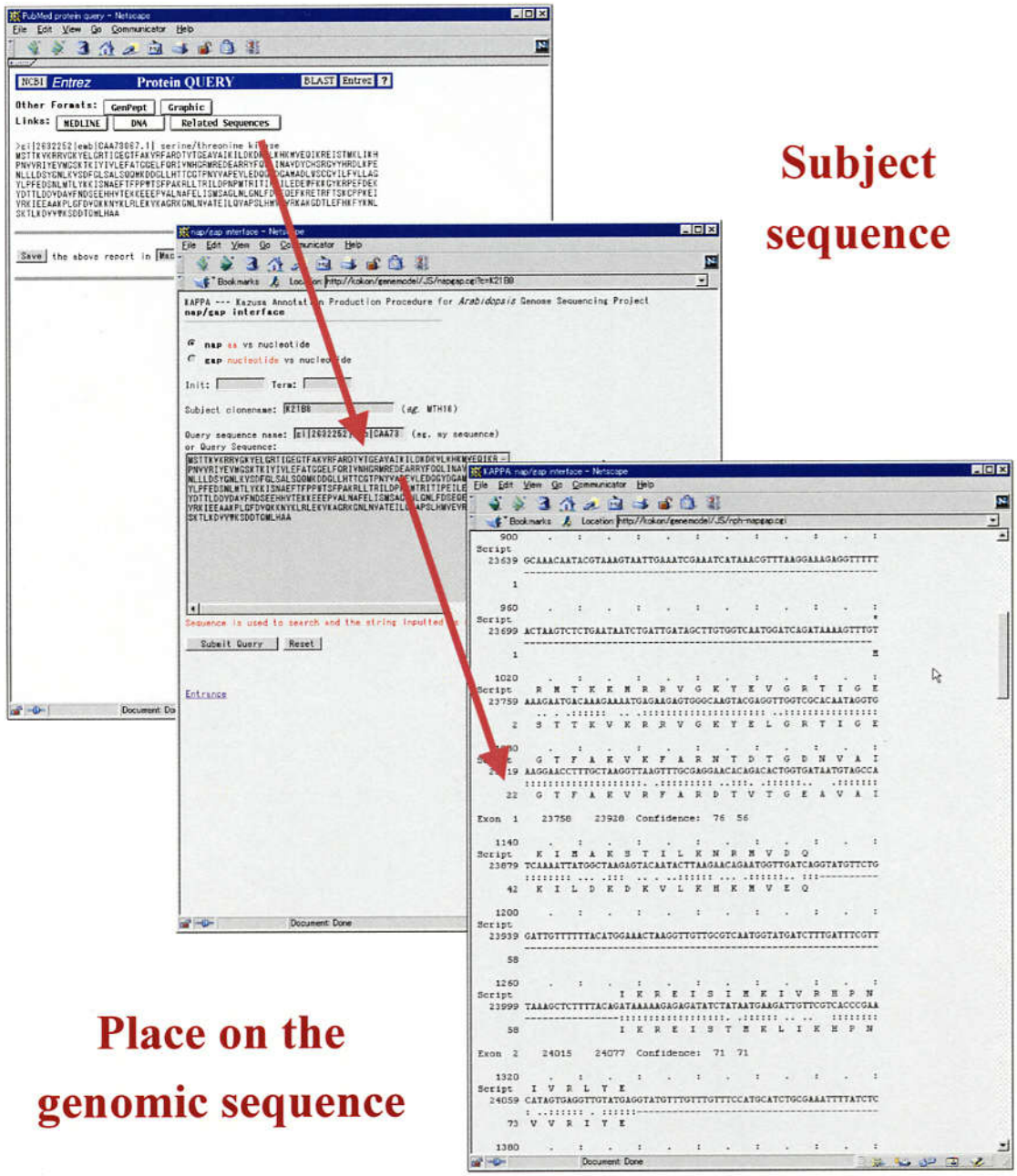


Figure 3-4b Gene modeling procedure (2)

Then an annotator selects a subject sequence in order to fit the sequence on the genomic sequence. Gap and nap programs (Huang *et al.*, 1997) are used for a subject of nucleotide sequence and amino acid sequence, respectively.

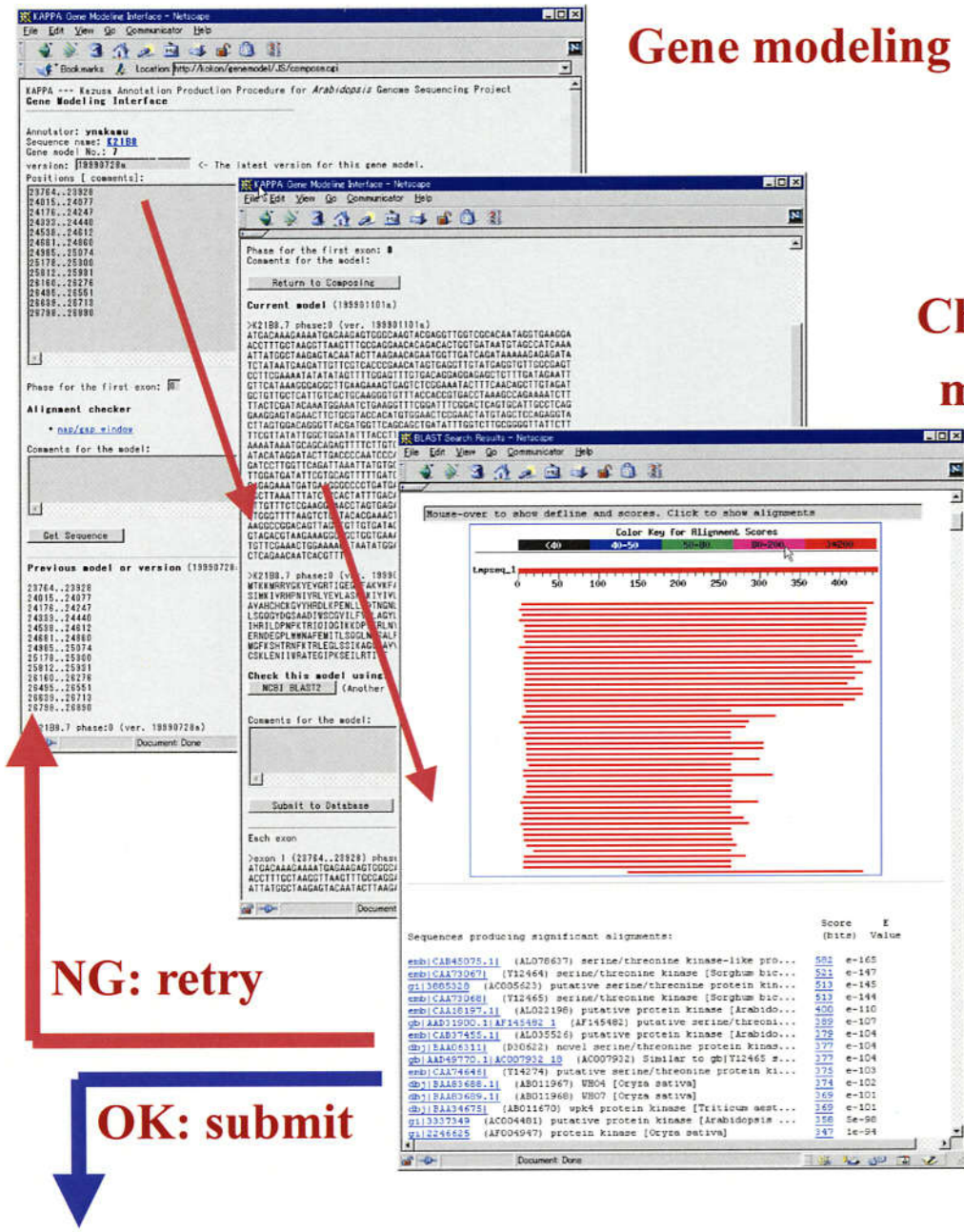


Figure 3-4c Gene modeling procedure (3)

An annotator edits a gene model on the region manually, considering exon-intron organization. Then an annotator performs similarity searches by BLAST on each working model. If a model shows a good alignment to a known sequence (or no similar sequence is found but a seems to be a good model), it is submitted as a deduced gene. If an alignment has one or more gap(s) to be edited, another model must be composed and checked. Regions without any BLAST hits must to be edited using information from gene-finding programs.

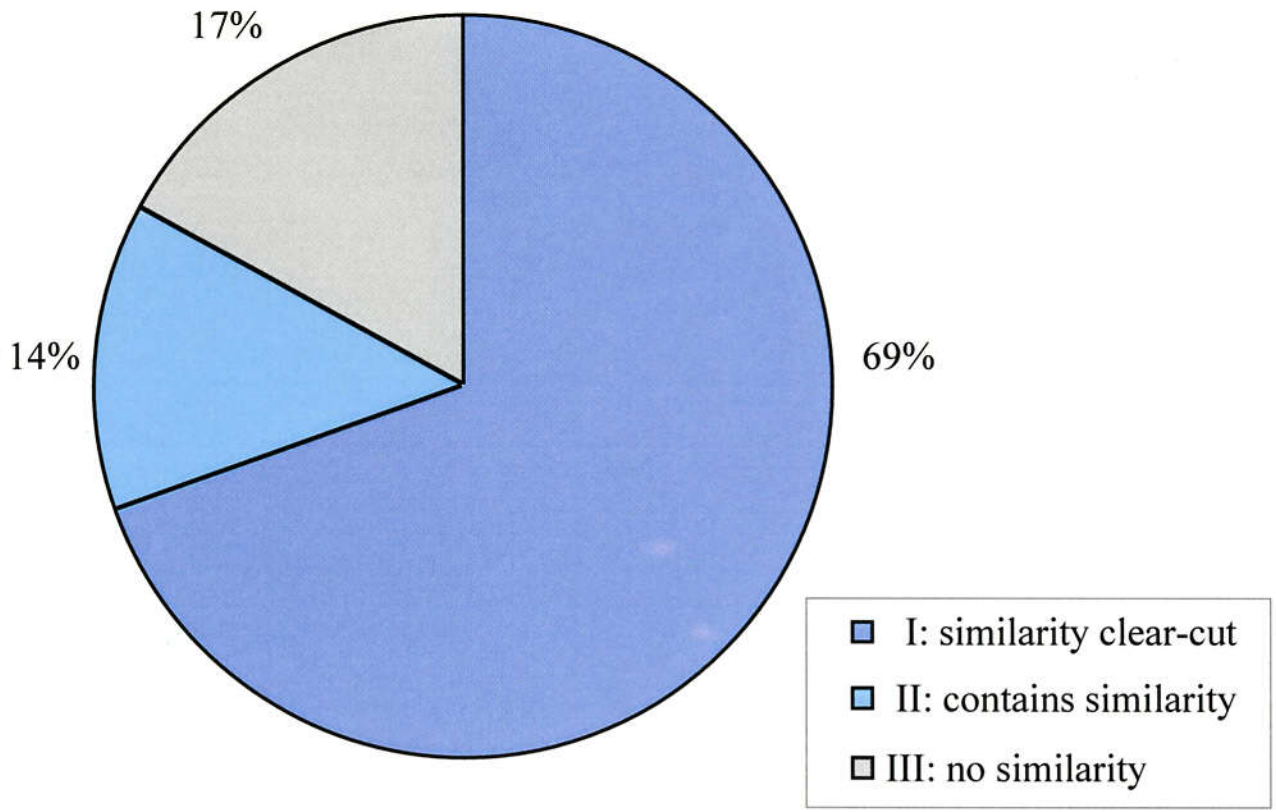


Figure 3-5 Classification of deduced proteins according to similarity level to known protein genes

Of the 6,251 protein and RNA genes deduced, 5,196 (83%) were assigned either clear or contains similarity to genes in the DNA databases and 1,055 (17%) were labeled as previously unknown genes.

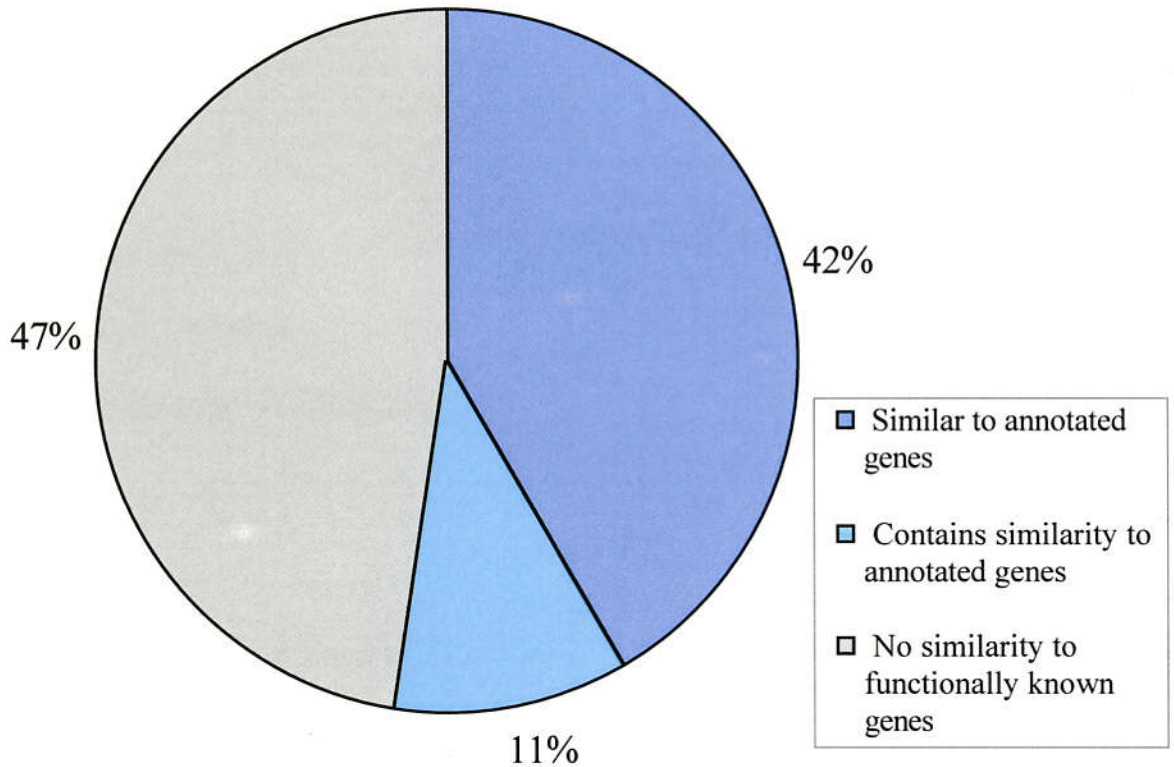


Figure 3-6 Classification of deduced proteins according to similarity level to functionally known protein genes

Of the 6,251 protein and RNA genes deduced, 3,284 (53%) were assigned either a definete or putative function and 2,947 (47%) were labeled as functionally unknown genes.

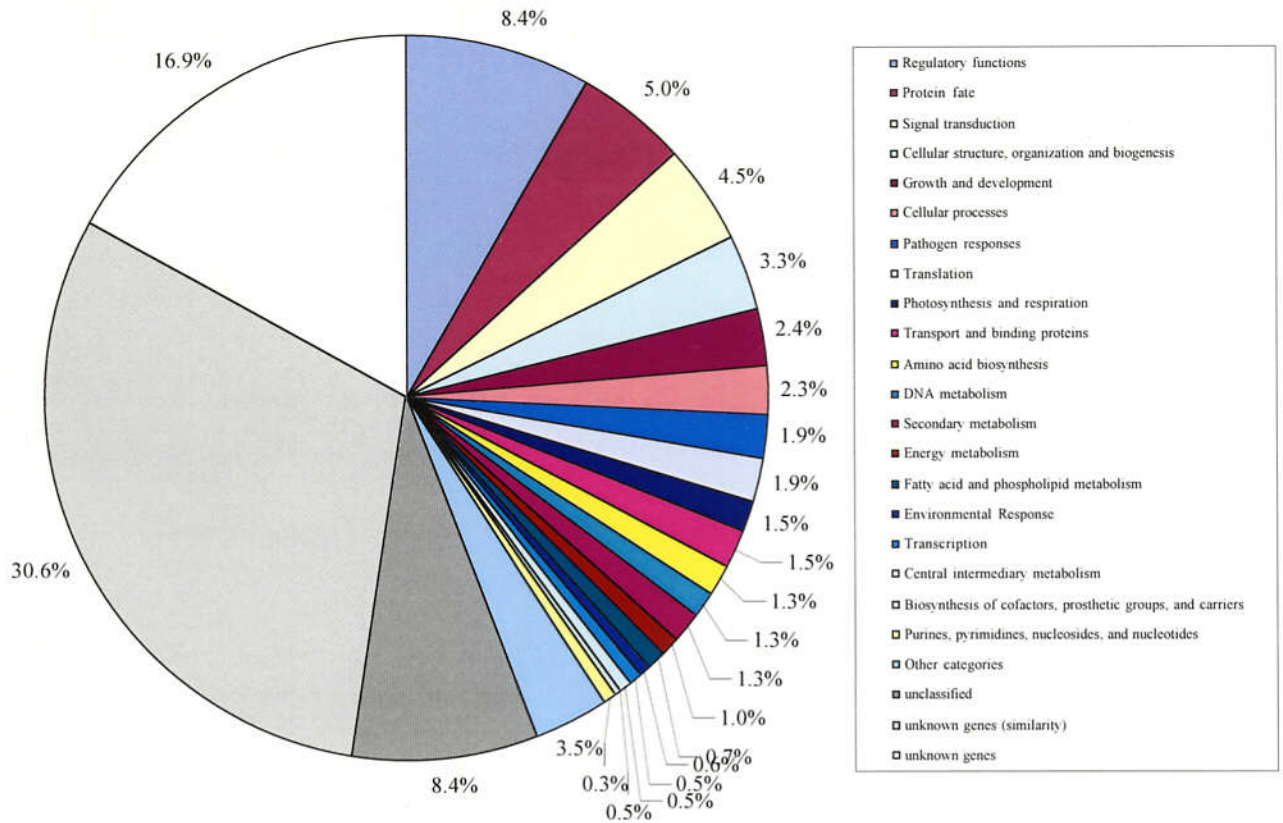


Figure 3-7 Classification of annotated proteins according to deduced role or function

The deduced genes whose function could be anticipated were grouped into 22 categories with the respect to different biological roles, according to modified Riley system.

Table 3-1a Positional information of the sequenced P1, TAC and BAC clones on chromosome 3.

clonename	ACCESSION	length	position in the contig
MEC18	AP002040	80099	1
T23B7	AP002063	40167	80100
MQC3	AP002047	86536	120267
T16H11	AP002055	21024	206803
MMF12	AP002044	21200	227827
MBK21	AB024033	79125	249027
MJM20	AC023838	15401	328152
MGH6	AC024128	81020	343553
MJG19	AP000375	62948	424573
MJH23	AP002042	6255	487521
MDC11	AB024034	86233	493776
MRP15	AP000603	83287	580009
K20M4	AP002038	25790	663296
MMM17	AP001307	80733	689086
MCP4	AB028610	38810	769819
MDC16	AB019229	84294	808629
MAG2	AP000600	73859	892923
MLE3	AP000416	3059	966782
MLN21	AB022220	83906	969841
MOA2	AB028617	52232	1053747
MIE1	AB023038	84462	1105979
T21E2	AP002061	47466	1190441
K15M2	AP000370	79279	1237907
F4B12	AP001299	37262	1317186
K7L4	AC023839	65047	1354448
MJK13	AC024081	82893	1419495
MQD17	AB028619	22199	1502388
MSJ11	AB017071	81370	1524587
MVC8	AB026653	35762	1605957
MSL1	AB012247	79355	1641719
MYA6	AB023046	75289	1721074
MDC8	AP000373	71521	1796363
MGL6	AB022217	79459	1867884
K20I9	AB028608	43500	1947343
MUH15	AP001308	11291	1990843
K14A17	AB026636	92620	2002134
MCE21	AP000384	18231	2094754
MGD8	AB022216	80167	2112985
MTO12	AB028620	19801	2193152
MKP6	AB022219	83233	2212953
MIG5	AB026646	27194	2296186
MEB5	AB019230	74968	2323380
MBG14	AB026641	6588	2398348
MRC8	AB020749	76816	2404936
MIE15	AP000414	34067	2481752
MYF24	AB026658	88989	2515819
K24M9	AP001303	45292	2604808
MVE11	AB026654	75508	2650100
MCB22	AP002039	29694	2725608
K13E13	AP000735	79186	2755302
MHP21	AP002041	4345	2834488
MVI11	AP000419	81875	2838833
MLD14	AB025624	83097	2920708
T31J18	AP002065	23377	3003805
MMB12	AP000417	73977	3027182
MPN9	AB025631	81035	3101159
MZE19	AP002050	31522	3182194
MAL21	AP000383	83253	3213716
MQC12	AB024036	87799	3296969
K10D20	AP000410	55161	3384768
F3H11	AP002034	33442	3439929
MOE17	AB025629	81677	3473371
MFD22	AP001304	33330	3555048
MSA6	AP000604	63748	3588378
MXL8	AB023045	85599	3652126
MHC9	AP001305	58510	3737725
MIL23	AB019232	80818	3796235
MSD21	AB025634	59793	3877053
MEK6	AP000739	6184	3936846
MZN24	AB028622	82348	3943030
MKA23	AP001306	70100	4025378
MMP21	AP002046	18211	4095478
MCB17	AB022215	83205	4113689
F16J14	AP000731	47827	4196894
MWI23	AB022223	77273	4244721
F5N5	AP001300	71327	4321994
MXC7	AB026655	79284	4393321
K13C10	AP000734	2569	4472605
K14B15	AB025608	83689	4475174
F28F4	AP000733	2490	4558863
MLM24	AB015474	87487	4561353
MEE5	AP000374	20724	4648840
MDB19	AB023036	90349	4669564
MYM9	AP000377	71178	4759913
F14O13	AP001297	106203	4831091
MUJ8	AB028621	78921	4937294
K13K6	AP002037	10480	5016215
K7M2	AP000382	80393	5026695
MXP5	AP002048	35471	5107088
MOB24	AB020746	79706	5142559
MSD24	AP000740	40018	5222265
K7P8	AB028609	78529	5262283
K3G3	AP000412	24057	5340812
MJL12	AB026647	81542	5364869
MTE24	AP000376	611	5446411
MWL2	AB025639	84896	5447022
T5M7	AP001313	65177	5531918
K13N2	AB028607	77483	5597095
K9I22	AP000599	18284	5674578
MPE11	AB023041	83650	5692862
MJL14	AP000601	9642	5776512
MTC11	AB024038	83859	5786154
F20C19	AP001298	71184	5870013
MFE16	AB028611	82646	5941197
MLJ15	AB026648	43481	6023843
MDJ14	AB016889	87630	6067324
MQP17	AP000602	12106	6154954
MOJ10	AB026649	73686	6167060
MYF5	AP001312	29176	6240746
K17E12	AP000381	63604	6269922
K1G2	AB024028	70952	6333526
MMJ24	AB025626	54452	6404478
MGF10	AB018114	84702	6458930
K16N12	AP000371	58864	6543632
K24A2	AP001302	30648	6602496
MMG15	AB028616	86139	6633144
MIG10	AP000415	31712	6719283
T19D11	AP002056	10900	6750995
MZF16	AP002051	67776	6761895
MFJ20	AB026644	85430	6829671
T20D4	AP002059	16123	6915101
MZN14	AP000420	82890	6931224
T19N8	AP002057	73391	7014114
MLD15	AP000386	82430	7087505
MYI13	AP002049	14454	7169935
K5K13	AB025615	83544	7184389
MRI12	AP000388	18212	7267933
MXE2	AB018121	86854	7286145
MUO22	AP001310	15145	7372999
MXO21	AB026657	77401	7388144
MMF24	AP002045	18605	7465545
MUO10	AP001309	77589	7484150
T13B17	AP002459	30780	7561739
MWE13	AP002457	24294	7592519
MTO24	AP000606	82360	7616813
T13J10	AP002052	46739	7699173
MOD1	AB028618	85690	7745912
T26G12	AP002064	84785	7831602
K17E7	AP000736	81580	7916387
T20F20	AP002060	56287	7997967
MIL15	AB028615	84157	8054254
T6J22	AP001314	75842	8138411
MVA11	AP001311	83339	8214253
MSJ3	AP000389	55739	8297592
MQP15	AB016878	85567	8353331
MED16	AP000738	2581	8438898
MED5	AB026642	77287	8441479
F21A17	AP000732	64714	8518766
T4A2	AP002066	18364	8583480
MIF6	AB028614	82347	8601844
F11I2	AP001296	36034	8684191
K11J14	AP000411	90627	8720225
MJI6	AP002043	83912	8810852
T22P15	AP002461	10607	8894764
T22B15	AP002062	112406	8905371
T22C2	AP002458	67517	9017777
T1O13	AP002058	71990	9085294
F8N14	AP001301	68639	9157284
T8O3	AP002068	54386	9225923
F1M23	AP002033	112684	9280309
F9K1	AP002036	1116	9392993
F6H5	AP002035	95681	9394109
T8N9	AP002462	36329	9489790
F1D9	AP002460	111554	9526119
T7B9	AP002067	84711	9637673
T13O13	AP002053	11845	9722384
T15D2	AP002054	87219	9734229

Table 3-1b Positional information of the sequenced P1, TAC and BAC clones on chromosome 5.

clonename	ACCESSION	length	position in the contig
contig 1			
MOK16	AB005240	80770	1
contig 2			
MED24	AB005235	75475	1
contig 3			
MUK11	AB008271	79073	1
MLG18	AB025625	1225	79074
MUG13	AB005245	86630	80299
K2A11	AB018111	29292	166929
K18I23	AB010692	72691	196221
MOP10	AB005241	57892	268912
MJJ3	AB005237	87835	326804
K18J17	AB017060	54252	414639
K16F4	AP002030	34867	468891
MBL20	AP002544	35301	503758
MHF15	AB006700	83865	539059
F15M7	AP002543	74350	622924
MPH15	AP002032	71277	697274
MOJ9	AB010697	86380	768551
contig 4			
MBK20	AB010070	78172	1
MXM12	AB005249	83599	78173
contig 5			
MAH20	AB006697	80970	1
contig 6			
MTH16	AB020752	68098	1
MYH9	AB016893	82390	68099
contig 7			
MXC9	AB007727	79590	1
contig 8			
MSH12	AB006704	79259	1
MXE10	AB011484	13474	79260
MAC12	AB005230	74613	92734
MUA22	AB007650	65465	167347
contig 9			
T9L3*	AL391149	89370	1
F2G14*	AL391146	85992	89371
contig 10			
T20K14*	AL391143	90176	1
F14F8*	AL391144	96892	90177
F1N13*	AL391145	74687	187069
T21H19*	AL391148	77311	261756
MQK4	AB005242	82001	339067
MTG13	AB008270	50641	421068
F5E19*	AL391147	77976	471709
F2K13*	AL391141	98101	549685
MKP11	AB005238	75188	647786
T10B6*	AL391142	33563	722974
K3M16*	AL391150	26604	756537
K10A8*	AL391151	43387	783141
MVA3	AB006706	81701	826528
MPI7	AB011480	40548	908229
MCM23	AB015473	36495	948777
MRG7	AB012246	83948	985272
contig 11			
MWD9	AB007651	85421	1
MQJ16	AB012244	57031	85422
MDJ22	AB006699	77363	142453
K5A21	AB024030	13874	219816
K8E10	AB025618	10757	233690
MRN17	AB005243	86065	244447
T20O7	AB026660	18446	330512
MYJ24	AB006708	78844	348958
MKD15	AB007648	76329	427802
T32G24	AB025642	13186	504131
K19M13	AB018110	42563	517317
MQM1	AB025633	81365	559880
MRO11	AB005244	82415	641245
MZF18	AB009056	65958	723660
MLE8	AB010696	15871	789618
K12G2	AB016883	45878	805489
MOP9	AB006701	84194	851367
K16H17	AB016884	45529	935561
T31K7	AB025641	23476	981090
K18P6	AB010068	74589	1004566
MXC17	AB016881	27263	1079155
contig 12			
MGG23	AB028613	35313	1
MSK10	AF104920	81414	35314
T13C12	AB024037	72800	116728
K3D20	AF058825	81208	189528
MOI20	AP000421	18272	270736
MVP2	AB025636	80402	289008
F6I13	AP000369	11049	369410
K21B8	AB025611	50015	380459
MOK9	AB015477	87459	430474
K2K18	AB023031	41465	517933
MJE4	AB013393	63989	559398
MXH1	AB011485	87210	623387
MWP19	AB020753	11026	710597

clonename	ACCESSION	length	position in the contig
MIK22	AB005236	82035	721623
F14A1	AB025602	55790	803658
MEE13	AB026643	84710	859448
MAB16	AB018112	70475	944158
T30G6	AB026661	76994	1014633
CIC5B3	AP002549	57286	1091627
F24C7	AP002029	73663	1148913
MPK17	AP000418	16898	1222576
F5H8	AB025605	70098	1239474
MLF18	AB016877	74842	1309572
K15O15	AB024026	23026	1384414
MJG14	AB017068	86121	1407440
MSK20	AP000605	7150	1493561
MNJ8	AB017069	88128	1500711
T25O11	AP000607	48312	1588839
MPA22	AB025630	51436	1637151
K12B20	AB018107	78874	1688587
T31G3	AB026662	77	1767461
K22F20	AB016873	64827	1767538
K18L3	AB012241	78239	1832365
K19A23	AB025610	9579	1910604
F16F17	AB028606	61510	1920183
MXA21	AB005247	87841	1981693
MSI17	AB011481	26624	2069534
MXI10	AB005248	83646	2096158
MBB18	AB005231	79537	2179804
MKD10	AB011478	47460	2259341
K15E6	AB009048	71736	2306801
MXF12	AB016892	66237	2378537
K3K3	AB010694	68889	2444774
MUL8	AB009054	83450	2513663
MIJ24	AB012243	68697	2597113
MKM21	AB016876	44499	2665810
K13H13	AB024023	6395	2710309
MYH19	AB010077	77380	2716704
MUD12	AB022222	22601	2794084
MSN9	AB010699	61001	2816685
MPO12	AB006702	86263	2877686
K21I16	AB017062	30578	2963949
MNF13	AB009052	85992	2994527
K1B16	AB015470	14323	3080519
MHK7	AB011477	78181	3094842
MMG1	AB023040	6478	3173023
MEE6	AB010072	83698	3179501
K1O13	AB019225	25275	3263199
MYC6	AB006707	82315	3288474
MPK23	AB020748	3544	3370789
MBK23	AB005233	79837	3374333
MUF8	AB025635	13776	3454170
K16L22	AB016871	79109	3467946
MJC20	AB017067	83689	3547055
K5J14	AB023032	59762	3630744
MDH9	AB016888	85791	3690506
K16E1	AB022210	33963	3776297
MFO20	AB013391	51737	3810260
MJB21	AB007647	78369	3861997
MBD2	AB008264	79976	3940366
MRD20	AB020750	3007	4020342
MMG4	AB008267	79046	4023349
K24F5	AB023030	5880	4102395
MNL12	AB017070	45911	4108275
MWF20	AB025638	91193	4154186
K9D7	AB016875	60476	4245379
MQO24	AB026652	8541	4305855
MQD19	AB026651	87286	4314396
F6B6	AP000368	37511	4401682
MRH10	AB006703	71522	4439193
MLN1	AB005239	84544	4510715
K9L2	AB011475	60583	4595259
MFC16	AB017065	61290	4655842
K15C23	AB024024	73921	4717132
K23L20	AB016874	81729	4791053
K21C13	AB010693	75709	4872782
K17O22	AB019224	67720	4948491
K18C1	AB012240	71618	5016211
K9E15	AB020744	62052	5087829
MFC19	AB018113	85020	5149881
K2N11	AB022213	30340	5234901
MRA19	AB012245	86722	5265241
K15I22	AB016870	59648	5351963
MCL19	AB006698	84510	5411611
MDE13	AB025620	14139	5496121
MPL12	AB010698	87434	5510260
K1I1I	AB019223	51529	5597694
F10E10	AB028605	38089	5649223
MZA15	AB016882	79995	5687312
MSD23	AB022221	33479	5767307
MQD22	AB013394	85661	5800786
K14A3	AB025609	34498	5886447
MQL5	AB018117	88398	5920945

clonename	ACCESSION	length	position in the contig
MNJ7	AB025628	80117	6009343
MGC1	AB028612	2593	6089460
MCA23	AB016886	82503	6092053
K16F13	AB024025	19742	6174556
MDN11	AB017064	83373	6194298
MIF21	AB023039	59372	6277671
K23F3	AP000372	36824	6337043
MJE7	AB020745	74298	6373867
K15N18	AB015468	66084	6448165
K24G6	AB012242	78973	6514249
K19E20	AB017061	61712	6593222
K20J1	AB023028	36243	6654934
K21P3	AB016872	86001	6691177
K7J8	AB023034	56963	6777178
K6M13	AB023033	77129	6834141
MNI5	AB025627	21011	6911270
K2I5	AB025613	71807	6932281
K21G20	AB025612	35238	7004088
K9P8	AB024032	70670	7039326
MPF21	AB026650	64569	7109996
K6A12	AB024031	64136	7174565
MXI22	AB012248	80115	7238701
MBA10	AB025619	46323	7318816
MFB16	AB023037	66087	7365139
K7B16	AB025617	26721	7431226
K16E14	AB026637	8850	7457947
K3K7	AB017063	68726	7466797
MWD22	AB023044	87180	7535523
MFG13	AB025621	48008	7622703
K17N15	AB018109	81293	7670711
K10D11	AB025607	13379	7752004
MIO24	AB010074	86212	7765383
MJM18	AB025623	16203	7851595
MSG15	AB015478	81347	7867798
F17P19	AB025603	48420	7949145
K24M7	AB019226	73999	7997565
T4M5	AP000378	18384	8071564
F6N7	AB025606	74282	8089948
MXC20	AB009055	81575	8164230
MNB8	AB018116	46872	8245805
MFH8	AB025622	62420	8292677
K19E1	AB013388	73428	8355097
MYN8	AB020754	54528	8428525
MNC6	AB015476	82952	8483053
MGN6	AB017066	61380	8566005
K6O8	AB025616	2163	8627385
K19P17	AB007644	73840	8629548
MJP23	AB018115	31827	8703388
K18G13	AB013387	31309	8735215
MDK4	AB010695	78596	8766524
GA469	AP000380	7717	8845120
F24B18	AB026634	50524	8852837
MRB17	AB016879	75744	8903361
K5F14	AB022214	31178	8979105
MBG8	AB005232	80315	9010283
K13P22	AB017059	19971	9090598
MCO15	AB010071	82918	9110569
MTE17	AB015479	80675	9193487
MWC10	AB023043	3542	9274162
MDF20	AB009050	86699	9277704
MWJ3	AB018120	42356	9364403
MYN21	AB026659	28207	9406759
MDA7	AB011476	82033	9434966
K24C1	AB023029	29498	9516999
MXK23	AB026656	9670	9546497
MCD7	AB009049	87685	9556167
MKN22	AB019234	27229	9643852
MIK19	AB013392	84129	9671081
MPI10	AB020747	29605	9755210
MHM17	AB024035	78423	9784815

clonename	ACCESSION	length	position in the contig
MUL3	AB023042	82010	9863238
MJB24	AB019233	58589	9945248
MSF19	AB016891	34508	10003837
MUA2	AB011482	83478	10038345
MRI1	AB018118	50700	10121823
MTI20	AB013396	79676	10172523
F2C19	AB026635	2872	10252199
K21L19	AB024029	41087	10255071
MCK7	AB019228	87090	10296158
MQJ2	AB025632	51440	10383248
MZN1	AB020755	81672	10434688
K19M22	AB016885	77758	10516360
K18B18	AB024027	35896	10594118
MNC17	AB016890	75803	10630014
F2O15	AB025604	75125	10705817
MTH12	AB006705	74877	10780942
MMN10	AB015475	84325	10855819
MGO3	AB019231	43570	10940144
F15L12	AB026632	54987	10983714
K9B18	AB015471	10085	11038701
MUF9	AB011483	77298	11048786
MUP24	AB005246	77999	11126084
MAE1	AB015472	54632	11204083
MSL3	AB008269	61634	11258715
MAF19	AB006696	78379	11320349
MFB13	AB010073	80376	11398728
MCI2	AB016887	15607	11479104
K11J9	AB012239	69142	11494711
MAC9	AB010069	57246	11563853
K22G18	AB022212	45453	11621099
MTG10	AB016880	81284	11666552
MMI9	AB019235	81736	11747836
K19B1	AB015469	69927	11829572
MRG21	AB020751	55151	11899499
MQB2	AB009053	78145	11954650
MJH22	AB009051	27856	12032795
MDC12	AB008265	81662	12060651
K9H21	AB023035	15319	12142313
MLE2	AB007649	58527	12157632
MBK5	AB005234	89779	12216159
MGI19	AB007646	37225	12305938
MBM17	AB019227	52717	12343163
MHJ24	AB008266	48622	12395880
MSJ1	AB008268	83594	12444502
T12B11	AB025640	22022	12528096
MUB3	AB010076	82188	12550118
MVP7	AB025637	39645	12632306
MXK3	AB019236	81494	12671951
F15O5	AB026633	17878	12753445
MQN23	AB013395	86064	12771323
MNA5	AB011479	88356	12857387
K19O4	AB026638	7240	12945743
K21L13	AB026639	63921	12952983
MPA24	AB010075	84440	13016904
K22J17	AB020743	11211	13101344
K14B20	AB018108	40251	13112555
K2A18	AB011474	79899	13152806
K1L20	AB022211	47665	13232705
K1F13	AB013389	83511	13280370
MSN2	AB018119	62927	13363881
MUD21	AB010700	69850	13426808
K8A10	AB026640	31132	13496658
K21H1	AB020742	74342	13527790
K3G17	AB025614	15290	13602132
K8K14	AB007645	72698	13617422
K9I9	AB013390	51860	13690120
LA522	AP000737	2808	13741980

Clonename with asterisk in contig 9 or 10 on chromosome 5 are clones originally allocated to EU *Arabidopsis thaliana* genome project then transferred to Kazusa DNA Research Institute. Since these 11 clones were annotated by EU annotation group, annotations on these clones were excluded from statistics in this thesis.

Table 3-2 Annotation criteria in the Kazusa *Arabidopsis* Genome sequencing project

Identical (100% match) to a subject sequence
The original gene name is preserved with its identifier.
Strong (E-value <e-100) and overall match
Named directory after the database entries or "PRODUCT homolog"
Named as "strong similarity to unknown protein (identifier)"
Good (E-value e-100~e-20) and over 50% Query sequence match
Named after the database entries as "PRODUCT-like (protein)"
Named as "similar to unknown protein (identifier)"
Partial match (E-value <e-20)
Named as "contains similarity to PRODUCT"
Named as "contains similarity to unknown protein (identifier)"
Models without any defined database matches (E-value >e-20)
Named as "unknown protein".

Table 3-3 Structural features of deduced protein coding genes in the kazusa allocated regions of *Arabidopsis thaliana* genome

Features	
Gene length (bp) including introns	78-17,203 (1,918)
Product length (amino acids)	25-4,706 (427)
Number of intron/gene	0-48 (4.0)
Coding exon length (bp)	2-5,966 (256)
Intron length (bp)	23-2,989 (157)
GC content of exons	44 %
GC content of introns	32 %

Structural statistics of the potential protein coding genes deduced so far are listed.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.

Andachi, Y., Yamao, F., Muto, A. and Osawa, S. (1989) Codon recognition patterns as deduced from sequences of the complete set of transfer RNA species in *Mycoplasma capricolum*. Resemblance to mitochondria. *J. Mol. Biol.*, **209**, 37-54.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature,* **408**, 796-815.

Asamizu, E., Sato, S., Kaneko, T., Nakamura, Y., Kotani, H., Miyajima, N. and Tabata, S. (1998) Structural analysis of *Arabidopsis thaliana* chromosome 5. VIII. Sequence features of the regions of 1,081,958 bp covered by seventeen physically assigned P1 and TAC clones. *DNA Res.*, **5**, 379-391.

Asamizu, E., Nakamura, Y., Sato, S., Fukuzawa, H. and Tabata, S. (1999) A large scale structural analysis of cDNAs in a unicellular green alga, *Chlamydomonas reinhardtii.* I. Generation of 3433 non-redundant expressed sequence tags. *DNA Res.*, **6**, 369-373.

Asamizu, E., Nakamura, Y., Sato, S. and Tabata, S. (2000) A large scale analysis of cDNA in *Arabidopsis thaliana:* generation of 12,028 non-redundant expressed sequence tags from normalized and size-selected cDNA libraries. *DNA Res.*, **7**, 175-180.

Barrell, B. G., Anderson, S., Bankier, A. T., De Bruijn, M. H., Chen, E., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R. and Young, I. G. (1980) Different pattern of codon recognition by mammalian mitochondrial tRNAs. *Proc. Natl. Acad. Sci. U. S. A.*, **77**, 3164-3166.

Bennetzen, J. L. and Hall, B. D. (1982) Codon selection in yeast. *J. Biol. Chem.*, **257**, 3026-3031.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. and Wheeler, D. L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15-18.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) The mosaic genome of warm-blooded vertebrates. *Science,* **228**, 953-958.

Boguski, M. S., Tolstoshev, C. M. and Bassett, D. E. Jr, (1994) Gene discovery in dbEST. *Science,* **265**, 1993-1994.

Bonitz, S. G. and Tzagoloff, A. (1980) Assembly of the mitochondrial membrane system. Sequences of yeast mitochondrial tRNA genes. *J. Biol. Chem.*, **255**, 9075-9081.

Borodovsky, M. and McIninch, J. (1993) Recognition of genes in DNA sequence with ambiguities. *Biosystems,* **30**, 161-171.

Brendel, V. and Kleffe, J. (1998) Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. *Nucleic Acids Res.*, **26,** 4748-4757.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H. P., Fraser, C. M., Smith, H. O., Woese, C. R. and Venter, J. C. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science,* **273,** 1058-1073.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268,** 78-94.

Burset, M. and Guigo, R (1996) Evaluation of gene structure prediction programs. *Genomics,* **34,** 353-367.

Cooke, R., Raynal, M., Laudie, M., Grellet, F., Delseny, M., Morris, P. C., Guerrier, D., Giraudat, J., Quigley, F., Clabault, G., Li, Y. F., Mache, R., Krivitzky, M., Gy, I. J., Kreis, M., Lecharny, A., Parmentier, Y., Marbach, J., Fleck, J., Clement, B., Philipps, G., Herve, C., Bardet, C., Tremousaygue, D., Höfte, H. *et al.* (1996) Further progress towards a catalogue of all *Arabidopsis* genes: analysis of a set of 5000 non-redundant ESTs. *Plant J,* **9,** 101-124.

Creusot, F., Fouilloux, E., Dron, M., Lafleuriel, J., Picard, G., Billault, A., Le Paslier, D., Cohen, D., Chaboue, M. E., Durr, A. *et al.* (1995) The CIC library: a large insert YAC library for genome mapping in *Arabidopsis thaliana. Plant J.,* **8,** 763-770.

Durbin, R. and Mieg, J. T. (1991-) ACeDB. ftp://ncbi.nlm.nih.gov/repository/acedb/

Dzelzkains, V. A. and Bogorad, L. (1988) Molecular analysis of a mutant defective in photosynthetic oxygen evolution and isolation of a complementing clone by a novel screening procedure., *EMBO J.,* **7,** 333-338.

Eddy, S. R. and Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22,** 2079-2088.

European Union Chromosome 3 Arabidopsis Sequencing Consortium, The Institute for Genomic Research & Kazusa DNA Research Institute. (2000) Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana. Nature,* **408,** 820-822.

Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8,** 175-185.

Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8,** 186-194.

Fichant, G. A. and Burks, C. (1991) Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Biol.*, **220,** 659-671.

Fickett, J. W. and Tung, C. S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **20**, 6441-6450.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L., Glodek, A., Kelly, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Frazer, C. M., Smith H. O. and Venter, J. C. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* **269**, 496-512.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J.M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J. F., Dougherty, B. A., Bott, K. F., Hu, P. C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison III, C. A. and Venter, J. C. (1995) The minimal gene complement of *Mycoplasma genitalium. Science,* **270**, 397-403.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996) Life with 6000 genes. *Science,* **274**, 563-567.

Goodman, H. M., Ecker, J. R. and Dean, C. (1995) The genome of *Arabidopsis thaliana. Proc. Natl. Acad. Sci. U. S. A.,* **92**, 10831-10835.

Gordon, D., Abajian, C. and Green, P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.,* **8**, 195-202.

Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pave, A. (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.,* **8**, r49-r62.

Grigorieva, G. and Shestakov, S. (1982) Transformation in the cyanobacterium *Synechocystis* 6803. *FEMS Microbiol. Lett.,* **13**, 367-370.

Huang, X., Adams, M. D., Zhou, H. and Kerlavage, A. R. (1997) A tool for analyzing and annotating genomic sequences. *Genomics,* **46**, 37-45.

Hebsgaard, S. M., Korning, P. G., Tolstrup, N., Engelbrecht, J., Rouze, P. and Brunak, S. (1996) Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucleic Acids Res.,* **24**, 3439-3452.

Heckman, J. E., Sarnoff, J., Alzner-Deweerd, B., Yin, S. and Rajbhandary, U. L. (1980) Novel features in the genetic code and codon reading patterns in *Neurospora crassa* mitochondria based on sequences of six mitochondrial tRNAs. *Proc. Natl. Acad. Sci. U. S. A.,* **77**, 3159-3163

Himmelreich, R., Hilbert, H., Plagens, H., Pirki, E., Li, B.-C. and Herrmann, R. (1997) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae. Nucleic Acids Res.,* **24**, 4420-4449.

Hirst, J. D. and Sternberg, M. J. E. (1992) Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural network. *Biochemistry,* **31,** 7211-7218.

Höfte, H., Desprez, T., Amselem, J., Chiapello, H., Rouze, P., Caboche, M., Moisan, A., Jourjon, M. F., Charpenteau, J. L., Berthomieu, P. *et al.* (1993) An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana. Plant J.,* **4,** 1051-1061.

Ikemura, T. (1981a) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.,* **146,** 1-21.

Ikemura, T. (1981b) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.,* **151,** 389-409.

Ikemura, T. (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J. Mol. Biol.,* **158,** 573-597.

Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.,* **2,** 13-34.

Kaiser, J. (1996) First global sequencing effort begins. *Science,* **274,** 30.

Kaneko, T., Tanaka, A., Sato, S., Kotani, H., Sazuka, T., Miyajima, N., Sugiura, M. and Tabata, S. (1995) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92% of the genome. *DNA Res.,* **2,** 153-166.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M. and Tabata, S. (1996) Sequence analysis of the genome of the unicellular Cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.,* **3,** 109-136.

Kaneko, T., Kotani, H., Nakamura, Y., Sato, S., Asamizu, E., Miyajima, N. and Tabata, S. (1998) Structural analysis of *Arabidopsis thaliana* chromosome 5. V. Sequence features of the regions of 1,381,565 bp covered by twenty one physically assigned P1 and TAC clones. *DNA Res.,* **5,** 131-145.

Kaneko, T., Katoh, T., Sato, S., Nakamura, Y., Asamizu, E., Kotani, H., Miyajima, N. and Tabata, S. (1999) Structural analysis of *Arabidopsis thaliana* chromosome 5. IX. Sequence features of the regions of 1,011,550 bp covered by seventeen P1 and TAC clones. *DNA Res.,* **6,** 183-195.

Kaneko, T., Katoh, T., Sato, S., Nakamura, A., Asamizu, E. and Tabata, S. (2000) Structural analysis of *Arabidopsis thaliana* chromosome 3. II. Sequence features of the 4,251,695 bp regions covered by 90 P1, TAC and BAC clones. *DNA Res.*, **7**, 217-221.

The Kazusa DNA Research Institute, The Cold Spring Harbor and Washington University in St Louis Sequencing Consortium & the European Union Arabidopsis Genome Sequencing Consortium. (2000) Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana. Nature,* **408,** 823-826.

Komine, Y., Adachi, T., Inokuchi, H. and Ozeki, H. (1990) Genomic organization and physical mapping of the transfer RNA genes in *Escherichia coli* K12. *J. Mol. Biol.*, **212,** 579-598.

Kotani, H., Nakamura, Y., Sato, S., Kaneko, T., Asamizu, E., Miyajima, N. and Tabata, S. (1997a) Structural analysis of *Arabidopsis thaliana* chromosome 5. II. Sequence features of the regions of 1,044,062 bp covered by thirteen physically assigned P1 clones. *DNA Res.*, **4,** 291-300.

Kotani, H., Sato, S., Fukami, M., Hosouchi, T., Nakazaki, N., Okumura, S., Wada, T., Liu, Y. G., Shibata, D. and Tabata, S. (1997b) A fine physical map of *Arabidopsis thaliana* chromosome 5: construction of a sequence-ready contig map. *DNA Res.*, **4,** 371-378.

Kotani, H., Nakamura, Y., Sato, S., Asamizu, E., Kaneko, T., Miyajima, N. and Tabata, S. (1998) Structural analysis of *Arabidopsis thaliana* chromosome 5. VI. Sequence features of the regions of 1,367,185 bp covered by 19 physically assigned P1 and TAC clones. *DNA Res.*, **5,** 203-216.

Lin, X., Kaul, S., Rounsley, S., Shea, T. P., Benito, M. I., Town, C. D., Fujii, C. Y., Mason, T., Bowman, C. L., Barnstead, M., Feldblyum, T. V., Buell, C. R., Ketchum, K. A., Lee, J., Ronning, C. M., Koo, H. L., Moffat, K. S., Cronin, L. A., Shen, M., Pai, G., Van, Aken, S., Umayam, L., Tallon, L. J., Gill, J. E., Venter, J. C. *et al.* (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana. Nature,* **402,** 761-768.

Liu, Y.-G., Mitsukawa, N., Vazquez-Tello, A. and Whittier, R. F. (1995) Generation of a high-quality P1 library of *Arabidopsis* suitable for chromosome walking. *Plant J.*, **7,** 351-358.

Liu, Y.-G., Mitsukawa, N., Lister, C., Dean, C. and Whittier, R. F. (1996) Isolation and mapping of new set of 129 RFLP markers in *Arabidopsis thaliana* using recombinant inbred lines. *Plant J.*, **10,** 733-736.

Lowe, T. M. and Eddy, S. R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25,** 955-964.

Maruyama, T., Gojobori, T., Aota, S. and Ikemura, T. (1986) Codon usage tabulated from the GenBank genetic sequence data. *Nucleic Acids Res.*, **14 (Suppl.),** r151-197.

Mayer, K., Schuller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Dusterhoft, A., Stiekema, W., Entian, K. D., Terryn, N., Harris, B., Ansorge, W., Brandt, P., Grivell, L., Rieger, M., Weichselgartner, M., de Simone, V., Obermaier, B., Mache, R., Muller,

M., Kreis, M., Delseny, M., Puigdomenech, P., Watson, M., McCombie, W. R. *et al.* (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana. Nature,* **402,** 769-777.

Meinke, D. W., Cherry, J. M., Dean, C., Rounsley, S. D. and Koornneef, M. (1998) *Arabidopsis thaliana:* a model plant for genome analysis. *Science,* **282,** 679-682.

Mewes, H. W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S. G., Pfeiffer, F. and Zollner, A. (1997) Overview of the yeast genome, *Nature,* **387 (suppl.),** 7-65.

Nakamura, Y., Gojobori, T. and Ikemura, T. (1997a) Codon usage tabulated from the international DNA sequence databases. *Nucleic Acids Res.,* **25,** 244-245.

Nakamura, Y., Sato, S., Kaneko, T., Kotani, H., Asamizu, E., Miyajima, N. and Tabata, S. (1997b) Structural analysis of *Arabidopsis thaliana* chromosome 5. III. Sequence features of the regions of 1,191,918 bp covered by seventeen physically assigned P1 clones. *DNA Res.,* **4,** 401-414.

Nakamura, Y., Kaneko, T., Hirosawa, M., Miyajima, N. and Tabata, S. (1998a) CyanoBase, a www database containing the complete nucleotide sequence of the genome of *Synechocystis* sp. strain PCC6803. *Nucleic Acids Res.,* **26,** 63-67.

Nakamura, Y., Gojobori, T. and Ikemura, T. (1998b) Codon usage tabulated from the international DNA sequence databases. *Nucleic Acids Res.,* **26,** 334.

Nakamura, Y., Sato, S., Asamizu, E., Kaneko, T., Kotani, H., Miyajima, N. and Tabata, S. (1998b) Structural analysis of *Arabidopsis thaliana* chromosome 5. VII. Sequence features of the regions of 1,013,767 bp covered by sixteen physically assigned P1 and TAC clones. *DNA Res.,* **5,** 297-308.

Nakamura, Y., Kaneko, T., Miyajima, N. and Tabata, S. (1999a) Extension of CyanoBase. CyanoMutants: repository of mutant information on *Synechocystis* sp. strain PCC6803. *Nucleic Acids Res.,* **27,** 66-68.

Nakamura, Y., Gojobori, T. and Ikemura, T. (1999b) Codon usage tabulated from the international DNA sequence databases; its status 1999. *Nucleic Acids Res.,* **27,** 292.

Nakamura, Y. and Tabata, S. (1999c) Codon-anticodon assignment and detection of codon usage trends in seven microbial genomes. *Microb. Comp. Genomics,* **2,** 299-312.

Nakamura, Y., Kaneko, T. and Tabata, S. (2000a) CyanoBase, the genome database for *Synechocystis* sp. strain PCC6803: status for the year 2000. *Nucleic Acids Res.,* **28,** 72.

Nakamura, Y., Gojobori, T. and Ikemura, T. (2000b) Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.,* **28,** 292.

Newman, T., de Bruijn, F. J., Green, P., Keegstra, K., Kende, H., McIntosh, L., Ohlrogge, J., Raikhel, N., Somerville, S., Thomashow, M. *et al.* (1994) Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. *Plant Physiol.,* **106,** 1241-1255.

Nikaido, I., Asamizu, E., Nakajima, M., Nakamura, Y., Saga, N. and Tabata, S. (2000) Generation of 10,154 expressed sequence tags from a leafy gametophyte of a marine red alga, *Porphyra yezoensis. DNA Res.*, **7**, 223-227.

Pavesi, A., Conterio, F., Bolchi, A., Dieci, G. and Ottonello, S. (1994) Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. *Nucleic Acids Res.*, **22**, 1247-1256.

Percudani, R., Pavesi, A. and Ottonello S. (1997) Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae. J. Mol. Biol.*, **268**, 322-330.

Riley, M. (1993) Functions of the gene products of *Escherichia coli. Microbiol. Reviews*, **57**, 862-952.

Sato, S., Kotani, H., Nakamura, Y., Kaneko, T., Asamizu, E., Fukami, M., Miyajima, N. and Tabata, S. (1997) Structural analysis of *Arabidopsis thaliana* chromosome 5. I. Sequence features of the 1.6 Mb regions covered by twenty physically assigned P1 clones. *DNA Res.*, **4**, 215-230.

Sato, S., Kaneko, T., Kotani, H., Nakamura, Y., Asamizu, E., Miyajima, N. and Tabata, S. (1998a) Structural analysis of *Arabidopsis thaliana* chromosome 5. IV. Sequence features of the regions of 1,456,315 bp covered by nineteen physically assigned P1 and TAC clones. *DNA Res.*, **5**, 41-54.

Sato, S., Kotani, H., Hayashi, R., Liu, Y. G., Shibata, D. and Tabata, S. (1998b) A physical map of *Arabidopsis thaliana* chromosome 3 represented by two contigs of CIC YAC, P1, TAC and BAC clones. *DNA Res.*, **5**, 163-168.

Sato, S., Kaneko, T., Kotani, H., Hayashi, R., Liu, Y. G., Shibata, D. and Tabata, S. (1999a) A sequence-ready contig map of the top arm of *Arabidopsis thaliana* chromosome 3. *DNA Res.*, **6**, 117-121.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E. and Tabata, S. (1999b) Complete structure of the chloroplast genome of *Arabidopsis thaliana. DNA Res.*, **6**, 283-290.

Sato, S., Nakamura, Y., Kaneko, T., Katoh, T., Asamizu, E., Kotani, H. and Tabata, S. (2000a) Structural analysis of *Arabidopsis thaliana* chromosome 5. X. Sequence features of the regions of 3,076,755 bp covered by sixty P1 and TAC clones. *DNA Res.*, **7**, 31-63.

Sato, S., Nakamura, Y., Kaneko, T., Katoh, T., Asamizu, E. and Tabata, S. (2000b) Structural analysis of *Arabidopsis thaliana* chromosome 3. I. Sequence features of the regions of 4,504,864 bp covered by sixty P1 and TAC clones. *DNA Res.*, **7**, 131-135.

Sazuka, T. and Ohara, O. (1997) Towards a proteome project of cyanobacterium *Synechocystis* sp. strain PCC6803: linking 130 protein spots with their respective genes. *Electrophoresis*, **18**, 1252-1258.

Sazuka, T., Yamaguchi, M. and Ohara, O. (1999) Cyano2Dbase updated: linkage of 234 protein spots to corresponding genes through N-terminal microsequencing. *Electrophoresis*, **20**, 2160-2171.

Schmidt, R., West, J., Cnops, G., Love, K., Balestrazzi, A. and Dean, C. (1996) Detailed

description of four YAC contigs representing 17 Mb of chromosome 4 of *Arabidopsis thaliana* ecotype Columbia. *Plant J.*, **9**, 755-765.

Schmidt, R., Love, K., West, J., Lenehan, Z. and Dean, C. (1997) Description of 31 YAC contigs spanning the majority of *Arabidopsis thaliana* chromosome 5. *Plant J.*, **11**, 563-572.

Sharp, P. M. and Li, W.-S. (1987) The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281-1295.

Simoneau, P., Li, C. M., Loechel, S., Wenzel, R., Herrmann, R. and Hu, P. C. (1993) Codon reading scheme in *Mycoplasma pneumoniae* revealed by the analysis of the complete set of tRNA genes. *Nucleic Acids Res.*, **21**, 4967-4974.

Solovyev, V. V., Salamov, A. A. and Lawrence, C. B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.*, **22**, 5156-5163.

Theologis, A., Ecker, J.R., Palm, C.J., Federspiel, N.A., Kaul, S., White, O., Alonso, J., Altafi, H., Araujo, R., Bowman, C.L., Brooks, S.Y., Buehler, E., Chan, A., Chao, Q., Chen, H., Cheuk, R.F., Chin, C.W., Chung, M.K., Conn, L., Conway, A.B., Conway, A.R., Creasy, T.H., Dewar, K., Dunn, P., Etgu, P., Feldblyum, T.V., Feng, J., Fong, B., Fujii, C.Y., Gill, J.E., Goldsmith, A.D., Haas, B., Hansen, N.F., Hughes, B., Huizar, L., Hunter, J.L., Jenkins, J., Johnson-Hopson, C., Khan, S., Khaykin, E., Kim, C.J., Koo, H.L., Kremenetskaia, I., Kurtz, D.B., Kwan, A., Lam, B., Langin-Hooper, S., Lee, A., Lee, J.M., Lenz, C.A., Li, J.H., Li, Y., Lin, X., Liu, S.X., Liu, Z.A., Luros, J.S., Maiti, R., Marziali, A., Militscher, J., Miranda, M., Nguyen, M., Nierman, W.C., Osborne, B.I., Pai, G., Peterson, J., Pham, P.K., Rizzo, M., Rooney, T., Rowley, D., Sakano, H., Salzberg, S.L., Schwartz, J.R., Shinn, P., Southwick, A.M., Sun, H., Tallon, L.J. *et al.* (2000) Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana. Nature,* **408**, 816-820.

Uberbacher, E. C. and Mural, R. J. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. U. S. A.*, **88**, 11261-11265.

Uberbacher, E. C., Xu, Y. and Mural, R. J. (1996) Discovering and understanding genes in human DNA sequence using GRAIL. *Methods in Enzymology,* **266**, 259-281.

Van Bogelen, R. A., Sankar, P., Clark, R. L., Bogan, J. A. and Neighthardt, F. C. (1992) The gene-protein database of *Escherichia coli:* Edition 5. *Electrophoresis,* **13**, 1014-1054.

Williams, J. G. K. (1988) Construction of specific mutations in photosystem II photosynthetic reaction center by genetic engineering methods in *Synechocystis* 6803. *Methods in Enzymol.*, **167**, 766-778.

Xu, Q., Jung, Y. S., Chitnis, V. P., Guikema, J. A., Golbeck, J. H. and Chitnis, P. R. (1994) Mutational analysis of photosystem I polypeptides in *Synechocystis* sp. PCC 6803. Subunit requirements for reduction of NADP+ mediated by ferredoxin and

flavodoxin. *J. Biol. Chem.*, **269**, 21512-21518.

Zachgo, E. A., Wang, M. L., Dewdney, J., Bouchez, D., Camilleri, C., Belmonte, S., Huang, L., Dolan, M. and Goodman, H. M. (1996) A physical map of chromosome 2 of *Arabidopsis thaliana. Genome Res.*, **6**, 19-25.

Acknowledgements

I wish to express my sincere gratitude to Prof. Toshimichi Ikemura, Laboratory of Evolutionary Genetics, National Institute of Genetics, for his kind and excellent guidance and generous participation in his advisory roles for the author's doctorate.

I also thank Prof. Takashi Gojobori, Prof. Satoshi Kuhara, Prof. Toshihiko Shiroishi, Prof. Asao Fujiyama and Prof. Tetsuji Kakutani for serving on my supervisory committee.

My heatleft appreciation is addressed to Dr. Satoshi Tabata, head of Department of Plant Gene Research, Kazusa DNA Research Institute, for providing the oppotunity to propose these significant and valuable research projects. This work could not have been accomplished without his courteous guidance and continuous encouragement.

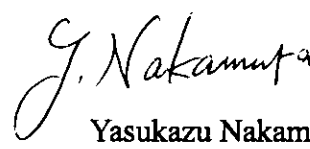
Special thanks must be given to my colleagues Ms. Erika Asamizu, Drs. Takakazu Kaneko, Tomohiko Kato and Shusei Sato, The First Laboratory for Plant Gene Research, for their countless and thoughtful arrangements for accomplishment of these works and valuable suggestions, discussions and advices. I extends my sincere thanks to Mr. Takaharu Kimura, Ms. Mitsuyo Kohara, Ms. Atsuko Kubota, Ms. Shinobu Nakayama, Ms. Sayaka Shinpo and Mr. Manabu Yamada for excellent technical assistance in areas of computer-aided nucleotide sequence analyses.

I wish to thank Dr. Yoshihiro Ugawa, presently at Environmental Education Center, Miyagi University of Education, for his help in constructing and distributing the CUTG at DNA Information Stock Center, from 1996 to 1999.

I grately appreciate Drs. Yutaka Akiyama (presently at Real World Computing Pertnership) and Susumu Goto, Institute for Chemical Research, Kyoto University, for providing a browsing program for the gene category list used in CyanoBase. The author also thank Drs. Teruo Ogawa, Masahiko Ikeuchi, Tatsuo Omata and Wim Vermass for helpful suggestions in construction of CyanoMutants.

I would like to express my sincere gratitude to the director Dr. Michio Oishi and the former director Dr. Mitsuru Takanami, Kazusa DNA Research Institute for providing this opportunity, and their valuable suggestions and encouragement for this work.

March 2001



Yasukazu Nakamura