B.Sc. iv sem zoology

Data summarizing frequency distribution

The Comprehensive Yeast Genome Database (CYGD) and the MIPS Functional Catalogue

Since the release of its complete genome in 1996, yeast has been firmly established as the most investigated eukaryotic organism. The MIPS Yeast Genome Database (MYGD) provides detailed information on open reading frames (ORFs), RNA genes and other genetic elements. Additionally, systematic functional genome analysis has been attempted by applying techniques such as gene disruption in conjunction with powerful expression analysis and two-hybrid techniques. These methods generate information on how proteins cooperate in complexes, pathways and cellular networks. MYGD forms the scaffold for a number of intensively used catalogues exemplified by the MIPS Functional Catalogue, the protein complexes, the protein–protein interactions and the special review and pathway sections. MYGD is perpetually maintained and updated.

In early 2001 the data from the systematic functional analysis project in yeast (EUROFAN II) became public and joined the EUROFAN I and SCDEGEN results. To cope with queries across this diverse functional data and the PEDANT analysis of all 14 available yeast species the BioRS retrieval system (Biomax Informatics) has been implemented for CYGD. With BioRS it is now possible to query different types of database located at different sites using a single query. This results in the simple all-encompassing functionality essential for a real comprehensive genome database.

As the amount of specialist yeast related data continues to grow, we are exploring a model to integrate additional data collections and knowledge into the Comprehensive Yeast Genome Database (CYGD). CYGD is built upon collaboration with several yeast laboratories and includes specialized databases. 20 000 newly identified genes from 13 hemiascomycetous yeasts, generated by the Genolevure project (1), have already been integrated. In addition, detailed information on transcription factors and their binding sites, transport proteins and metabolic pathways are being included or interlinked to the core data. All data is searchable using a scalable graphical view of the *Saccharomyces cerevisiae* chromosomes.

THE MIPS *NEUROSPORA CRASSA* DATABASE (MNCDB)

Neurospora crassa is a model organism for the lifestyle and development of mycelial fungi and is therefore interesting for the comparative analysis of fungi. Several genome sequencing projects have been launched; the German Neurospora Sequencing Project comprises the sequencing of Chromosomes II (~4.6 Mb) and V (~9.2 Mb) and is close to completion. The MNCDB database at MIPS contains ~17 Mb of genomic sequences. Of these, 8 Mb could be assigned to Chromosomes II and V, 9 Mb cannot yet be assigned to any chromosome. The MNCDB page allows blast searches against this dataset and additional sequence data derived from shotgun sequencing.

The dataset was analysed for potential coding regions using the gene prediction program FGENESH which has been recently trained for *Neurospora*. The coding sequences were subsequently submitted to the PEDANT genome analysis suite (see below). The results of this analysis can be viewed from the *Neurospora* project page. Moreover, detailed manually supervised gene modeling and annotation has been performed for most of the sequences belonging to chromosomes II and V with 2400 proteins recently processed manually. These proteins are classified according to the MIPS Functional Catalogue, EST hits are referenced, cross-references and literature citations are also included in the annotation. PEDANT delivers additional information on possible protein function and structure.

PEDANT: FUNCTIONAL AND STRUCTURAL GENOMICS ON THE WEB

The main purpose of the PEDANT (2) genome database is to quickly disseminate well-organised information on completely sequenced and unfinished genomes. Each gene product undergoes exhaustive automatic characterisation using a large variety of bioinformatics tools. The genome browser enables the user to navigate through a number of pre-computed catalogues and select proteins belonging to a specific functional and structural class. Thus, typical

queries are: ‘Which proteins have the TIM-barrel fold?’, ‘Which proteins are involved in nucleotide metabolism?’ or ‘Which proteins have more than five transmembrane regions?’. For each ORF in the dataset an integrated, hypertext-linked protein report is provided. The advanced DNA viewer represents contigs in graphical form and allows the user to navigate, zoom, produce six-frame translations and show DNA features such as restriction sites and genetic elements (genes, ORFs, exons, tRNAs, etc.). The protein viewer visualises similarity hits, sequence motifs and predicted protein features.

In order to facilitate keeping track of the data we have introduced release numbers for the database. The release 1.0.0 available at the time of writing contains the total of 141 genomic sequences. The PEDANT Web site is split in three major divisions.

Genomes that are being annotated and published by MIPS

This section currently includes *Arabidopsis thaliana*, *N.crassa*, and *Thermoplasma acidophilum*. These datasets include extensive manual annotation.

Completely sequenced and published genomes

In most of the cases the sequence data and ORF nomenclature as provided by the NCBI genomes division are employed, and the ORF descriptions supplied by the original authors are preserved. This section currently contains four eukaryotic, 54 eubacterial and eight archaeal genomes as well as 12 plasmids.

Unfinished and/or unpublished genomic sequences

Gene prediction is conducted by ORPHEUS (3) in a completely automatic fashion, usually allowing for large overlaps between ORFs. This leads to many overpredicted ORFs, but ensures that fewer real ORFs are missed. In many cases, the PEDANT database is the only source of annotation for such datasets. A total of 60 genomes can be found here, one eukaryotic, 55 eubacterial and four archaeal, respectively.

We would like to particularly stress the role of the PEDANT server as a structural genomics resource. Structural assignments and predictions for over 508 820 genomic proteins have been made so far, which makes PEDANT the most comprehensive resource of this kind on the Web. PEDANT is currently being actively used to support target selection in a number of structural genomics projects.

Complementary to the PEDANT genomes database, ‘mini-PEDANT’ has been set up as a service to assign a wide variety of attributes to protein sequences. mini-PEDANT returns a comprehensive view to the functional properties of the submitted sequence. Users do not need specific knowledge to use optimize complex parameter settings, as with other servers available. The system was developed using methods of the PEDANT system (2) to assign functional and structural properties such as multiple alignments to homologous sequences, presence of sequence motifs and domains, secondary structures and membrane spanning segments. Three-dimensional structures are assigned whenever reliable alignments to known structures are found. Integration of generic MIPS databases, such as the MIPS Functional Catalogue (4), expand the scope of mini-PEDANT, thus becoming a comprehensive and easy-to-use sequence analysis tool. The system is part of the ‘Helmholtz Network of Bioinformatics’ (HNB) and ‘Genomanalyse im Biologischen System Pflanze’ (GABI) bioinformatics infrastructures. A CORBA server allows for access to mini-PEDANT results by remote applications.

Genome analysis in plants

The German national plant genome network (GABI) network joins over 30 experimental groups and MIPS (GABIinfo) provides a common bioinformatics infrastructure. The complete sequence of the *A.thaliana* genome is used as a reference model for information to be transferred to other plants, including crops and as framework for the

compilation of experimental results. The database of the *Arabidopsis* genome (MATDB; 5), contains detailed information while integrating external expert annotation and databases. GABIinfo incorporates data from other plants deposited in the public databases to provide a useful resource for the intergenome analysis of plant sequences (e.g. 6).

The predominantly available data resource from plants other than *Arabidopsis* and rice are ESTs. Although the analysis of EST sequences is facing the problems of low sequence quality, cloning and processing artifacts and often unreliable information about source and orientation, much information can be uncovered by clustering procedures including various automated quality control steps. We have used HarvESTer (Biomax Informatics) to generate consensus clusters that are further processed and compared to the information available for *Arabidopsis*. Although consensus clusters can be used to compare different datasets by homology searches, this approach is only feasible for closely related species, since the sequence conservation in coding regions and especially in the 5′- and 3′-UTRs drops in part dramatically. Thus, an additional step has been introduced for the analysis of EST clusters to extract open reading frame information. We use the analysis of codon frequencies within different species for partial gene prediction. Resulting amino acid sequences are subjected to the PEDANT analysis procedure. Our first focus is ESTs from barley where the analysis of >80 000 ESTs has resulted in the detection of approximately 20 000 individual genes.

The HumanInfoBase (HIB) and the cDNA database of the DHGP

The HumanInfoBase is a database of putative human gene transcripts (7). UniGene clusters are assembled, and the resulting consensus sequences are submitted to the PEDANT software system. HIB is a database of automatically annotated putative human transcripts together with a functional classification based on systematic homology searches and pattern analysis. The primary source of the data is the UniGene database (8). Predicted proteins are classified according to several distinct criteria. Each putative transcript is assigned to a category in the MIPS Functional Catalogue. Each entry in the database contains the primary information generated by the sequence analysis; a comprehensive graphical display of the data associated to each putative transcript is provided by a protein viewer.

Although the genomic sequence of human genome has been released, cDNA sequencing continues to be crucial for the accurate identification of gene structures. The main goal of the German cDNA Project, as a part of the German Human Genome Project (DHGP), is the isolation, analysis and application of novel full-length cDNAs. From 64 067 randomly sequenced ESTs (35 292 711 bp) 4559 cDNAs (11 628 418 bp) were completely sequenced by a consortium of eight laboratories. After exhaustive analysis and both automatic and manual annotation, the sequences pass several client specific steps of publication, before they are released to the public databases. Published sequences can be accessed from the MIPS site.