Arabidopsis genome analysis as exemplified by analysis of chromosome 4

Klaus F. X. Mayer, Kai Lemcke, Christine M. Schüller, Stephen Rudd and Paolo Zaccaria

Date received (in revised form): 11th August 2000

Klaus F. X. Mayer, Kai Lemcke, Stephen Rudd and **Paolo Zaccaria** are at the Munich Information Centre for Protein Sequences (MIPS). They are involved in the analysis of the *Arabidopsis thaliana* genome.

Christine M. Schüller is at Biomax Informatics GmbH, Lochhamer Str. 11, 82152 Martinsried.

Keywords: *genome analysis, bioinformatics, annotation, plant genome*

Abstract

During the last decade the small cruciferous plant *Arabidopsis thaliana* has become a model organism for flowering plants. Sequencing and analysis of the *Arabidopsis* genome is nearing completion. Beside an overview on methods and strategies for *Arabidopsis* genome analysis, a summary of the results from the first analysis is presented. This includes an overview on chromosomal organisation and topological features as well as a first comparison with other genomes.

INTRODUCTION

Arabidopsis is a close relative of rape, mustard, broccoli and cauliflower. Besides the advantages that make it highly valuable for laboratory use, it has one of the smallest plant genome sizes known so far. This along with high gene density and a comparably low repetitive sequence content account for its use as a model genome. The estimated *Arabidopsis* genome size of 130 Mb is approximately 10-fold larger than the size of the unicellular eukaryotic yeast genome (12.5 Mb)[1] and is in the same range as other multicellular model systems such as *Caenorhabditis elegans* (97 Mb)[2] and *Drosophila melanogaster* (180 Mb).[3] This is of special importance since other plant genomes are orders of magnitudes larger (eg corn 6,000 Mb, rye 9,000 Mb, wheat 15,000 Mb).[4] A major component of these genomes, however, is thought to be composed of non-coding sequence repeats and transposable elements. Nevertheless, owing to the close evolutionary relationship among angiosperms and the well-conserved gene function and structure, the *Arabidopsis* gene repertoire is generally regarded as a basic toolbox for higher plant genomes.

ARABIDOPSIS THALIANA GENOME SEQUENCING

The Arabidopsis Genome Initiative (AGI) is composed of scientists and sequencing consortia from Europe, Japan and the USA. The approach being undertaken involves sequencing contigs that have been mapped to the genome. This means sequencing is performed on clones anchored within a minimal tiling path, seed clones or that have been identified as overlapping clones on the basis of BAC (bacterial artificial chromosome) end sequence identity. AGI's goal is to retrieve the complete sequence of the five *Arabidopsis* chromosomes in 10 gap-free contigs. This corresponds to a sequence from each chromosome arm ranging from the telomeres to the highly repetitive centromeres. Chromosomes 2 and 4 are, however, tipped by nearly homogenous 10.5 kb rDNA repeats, the so-called nuclear organiser regions (NOR), extending for 3.5–4 Mb towards the telomere. For these two chromosome arms completion was assigned by reaching the rDNA repeats rather than the telomeric repeats.

Essential resources such as high-quality BAC clone libraries,[5] a BAC fingerprint database which allows BAC contig identification based on the specific restriction patterns,[6] BAC clone contigs for the complete genome[6,7] and a genome survey sequencing program which cover more than 47,000 BAC end sequences were developed.

Klaus Mayer
GSF Forschungszentrum f. Umwelt und Gesundheit,
Max Planck Institut für Biochemie,
Am Klopferspitz 18a,
82152 Martinsried,
Germany

Tel: +49 89 8578 2453
Fax: +49 89 8578 2655
E-mail: mayer@mips.biochem.mpg.de

complete *Arabidopsis* sequence and analysis available by end of 2000

At the date of writing more than 90 per cent of the *Arabidopsis* genome has been sequenced and it is planned to have the complete genome sequence available and analysed by the end of 2000. The first milestones for the *Arabidopsis* genome project were the completion of the sequence and the subsequent analyses for chromosomes 2 and 4. These chromosomes cover 37 Mb and contain 7,800 genes which account for approximately 30 per cent of the protein coding genes from *A. thaliana*.[8,9]

extraction of genetic elements from the nucleotide sequence is key in genome analysis

Subsequently the sequencing and analysis approach undertaken for chromosome 4 will serve as the main focus and example for the *Arabidopsis* genome analysis. Although the individual sequencing consortia independently run sequence analysis and gene extraction on the sequence, comparable analysis standards have been established (Mayer and White, unpublished) to ensure analysis consistency.

SEQUENCING OF CHROMOSOME 4

A minimal tiling path consisting of 135 BAC and P1 clones and 56 cosmids was constructed and the remaining gaps were closed using polymerase chain reaction (PCR). Seventy sequenced markers have been anchored to these sequences. The relative BAC clone order has been compared to the known order from the genetic recombinant inbred (RI) map[10] to ensure tiling path integrity. Sequencing of the clones was performed by the *European Union Arabidopsis Genome Sequencing Consortium* which consisted of more than 20 individual sequencing laboratories throughout Europe and the *Cold Spring Harbor, Washington University in St Louis and PE Biosystems Arabidopsis Sequencing Consortium* (CSH/WU). Sequence analysis and the assembly of the chromosome arms was carried out by MIPS, and in part by CSH/WU. All sequences and analysis data have been integrated into the MIPS *Arabidopsis thaliana* Database (MATDB).[11] The top arm sequence ranges from the distal NOR to the characteristic pAL1 repeats and spans 3.05 Mb. The bottom arm sequence is 14.5 Mb long and ranges from the highly repetitive centromeric region to the distal telomere.

INFORMATION EXTRACTION/SEQUENCE ANALYSIS AND ANNOTATION

A key issue in genome analysis is the extraction of genetic elements from the nucleotide sequence, eg identification of transcribed regions, protein coding genes, transposons, tRNAs and regulatory regions. Successful and reliable identification of these features is influenced by the amount and quality of data on the elements to be analysed and on the availability and reliability of programs that can be applied to computational analyses.

In *Arabidopsis* approximately 30 per cent of the genes have either already been detected by experimental procedures or have closely related homologues detected by experimental procedures. This portion of the genome in principle can be analysed and genes can be extracted by similarity- and homology-based methods. In addition a large collection of *Arabidopsis* expressed sequence tags (ESTs) are publicly available, and there is an even larger collection of ESTs from other plants which can be used for heterologous similarity searches.

Approximately 45 per cent of the *Arabidopsis* genes have at least one cognate EST match. However, the EST libraries in use are not normalised and EST distribution is therefore strongly biased towards highly expressed genes. Roughly 75 per cent of the ESTs stem from only 6 per cent of the genes.[12] ESTs derived from lowly expressed genes; non-constitutively expressed genes and tissue-specific genes are largely under-represented. This, together with the short length of ESTs, restricts their use in gene-finding and modelling. Nevertheless ESTs have proven to be very helpful in the detection of gene borders and for the editing and confirmation of splice sites.

Table 1: Prediction sofware suitable for the analysis of the *Arabidopsis* genome

Program	Web site	Type of prediction
Netstart	http://www.cbs.dtu.tk/services/NetStart/	Start codon
NetGene2	http://www.cbs.dtu.dk/services/Netgene2/	Splice sites
Genscan	http://ccr-081.mit.edu/GENSCAN.html	Complete gene models
GeneMark.hmm	http://dixie.biology.gatech.edu/GeneMark/eukhmm.cgi/	Complete gene models
Genefinder		Complete gene models
Grail v1.3	http://compbio.ornl.gov/Grail-1.3/	Complete gene models
MZEF	http://sciclio.cshl.org/genefinder/	Splice sites and exons

to assist gene finding, gene prediction software is used

Only a subset of genes can be detected and modelled by using extrinsic data. To assist gene finding and modelling, gene prediction software is used (see Table 1).

Software for the prediction of splice sites, exons and complete gene structures has improved considerably during the past few years.[13] Genome analysis and gene identification in prokaryotes and lower eukaryotes such as yeast have now become a standard procedure with a variety of reliable gene prediction tools available. Gene prediction in higher eukaryotes is more complicated. A major difference and complicating factor between prokaryotes and higher eukaryotes is the split gene structure. Whereas the presence of introns is an exceptional case in prokaryotes, more than 80 per cent of the *Arabidopsis* genes contain at least one intron. Although the *Arabidopsis* genome is comparably dense, genes can be composed from as many as 79 exons and can be dispersed over as much as 30 kb of genomic sequence.[14] Consequently the challenge of *in silico* gene finding in higher eukaryotes involves the modelling of complete genes from the combination of predicted introns, exons and splice sites.

***in silico* gene finding in higher eukaryotes involves the modelling of complete genes from the combination of predicted introns, exons and splice sites**

Owing to the complexity of this process, failures and deviations from the genuine gene structure are frequent. The fusion of exons, missing exons, an aberrant splice site choice, addition of additional exons and concatenation of genes are frequently observed deviations. To optimise *in silico* gene recognition the usage of different algorithms in parallel is common. The combination of outputs and knowledge of the strengths and weaknesses of the individual programs has to be taken into consideration for a combinational weighting, which increases in accuracy/performance. Consequently automated methods to combine the analyses from different gene-finding programs have been proven to be successful.[15,16]

MIPS ANALYSIS PIPELINE

the analysis pipeline uses both extrinsic and intrinsic data

The MIPS analysis pipeline takes into account the topics discussed above and uses extrinsic and intrinsic data for the analyses (Figure 1). Several quality checks are carried out, namely checks for vector contamination, the presence of *Escherichia coli* contamination, coherence with the restriction map and consistency with overlapping clones. The sequence is then subjected to various analytical tools, which fall into three main categories. The first includes calculation of features such as tRNAs and repeats. The two remaining branches consist of what is designated as the calculation of extrinsic and intrinsic values. The extrinsic branch comprises calculations against a series of databases, such as the EMBL nucleotide database, databases containing *Arabidopsis* and other plant-derived ESTs, non-redundant protein databases and a transposon database. The intrinsic branch involves gene prediction calculations

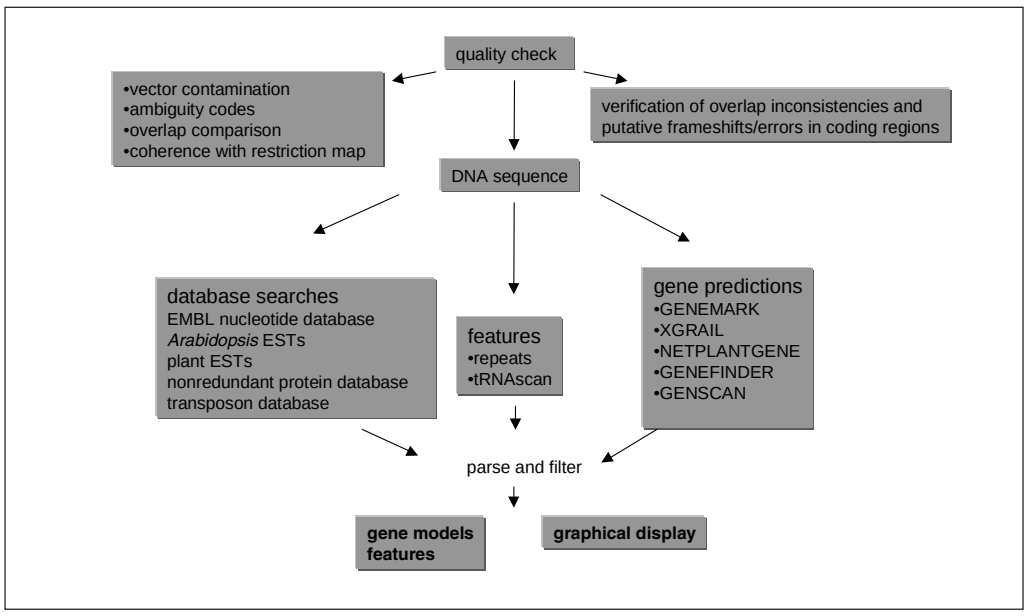

Figure 1: Overview of the MIPS analysis pipeline

visualisation is an essential component of the annotation procedure

using several algorithms (Genscan,[17] GeneMark.hmm[18] and Genefinder[19]) and splice site predictions (NetGene2[20]). Finally the output of the individual analysis steps are parsed and summarised graphically. Visualisation of the complete analysis data in parallel has been proven to be an essential and highly valuable component of the annotation procedure. It allows the trained user to assess the data almost at first glance. For this purpose a Genome Viewer written in JAVA (unpublished) is used at MIPS (Figure 2). Using the graphical display, trained annotators can integrate, correlate and assess the raw analysis data from all gene predictions and database searches almost instantly and can adjust gene models very rapidly. Parallel usage and integration of several gene prediction programs on one hand and further adjustments to database matches on the other hand lead to a high-accuracy gene retrieval.

ADDITION OF KNOWLEDGE

Ideally the outcome from a genome project should be comparable to the body of data available for all well-studied genes. This outcome should include data on all genetic elements and their nucleotide sequence, the amino acid sequence for all proteins encoded, along with an in-depth description of the functional role for each element, their interactions and regulation.

Around 30 per cent of the *Arabidopsis* genes either have already been detected experimentally or have strong homologies to experimentally detected genes from other species which allows the (restricted) deduction of functional roles. This relies on the notion that proteins similar in sequence are also similar in function, which has been experimentally shown to be true in numerous cases. Note that the experimental detection of genes does not necessarily coincide with, or include, the analysis of biochemical function. This further lowers the percentage of genes with a clear biochemical role attributable by similarity comparisons. Functional categorisation can be expanded by using calculations against domain and family signature databases which can give useful hints towards the functional role of proteins. Its strength has been impressively shown for the functional categorisation of the *Drosophila* genome[21] (for a detailed description of pattern databases and their use see Attwood[22]). Taking into account the data from similarity searches and calculations performed using domain and family signature databases, each new annotated protein is assigned to a functional category using the MIPS functional catalogue and literature citations are added.

The Genome Viewer serves as a workbench for annotation and serves for visualisation of extrinsic and intrinsic data

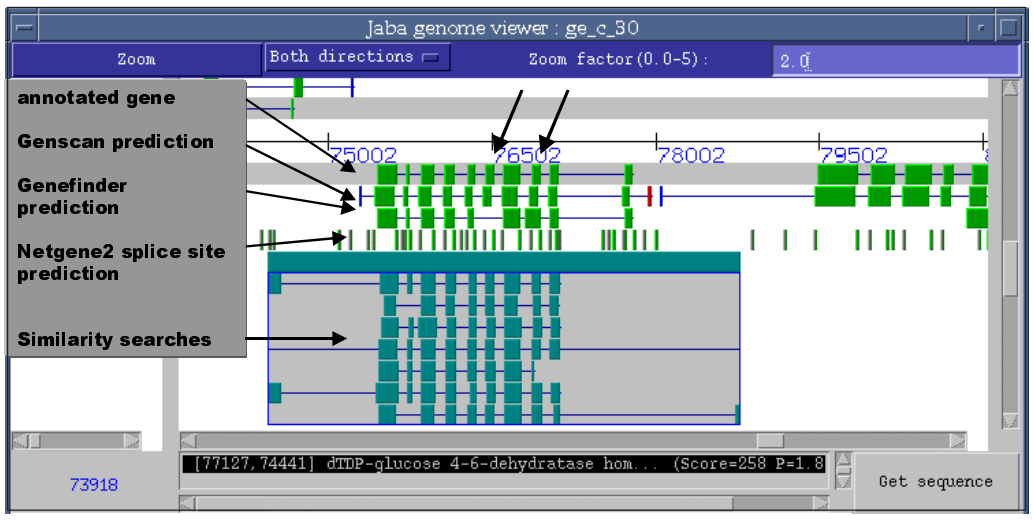

Figure 2: Gene predictions performed by Genscan and Genefinder, predicted splice sites as well as the annotated gene model are depicted. Matches detected by similarity searches are shown in parallel. Prediction discrepancies between Genscan and Genefinder are present (arrows). Taking into account similarity matches these discrepancies can be resolved

an exhaustive collection of methods and databases is used for functional and structural characterisation of protein sequence

PEDANTIC BIOINFORMATIC ANALYSIS

As mentioned above, for in-depth analysis apart from sequence similarity comparisons it is of great value to apply a large variety of bioinformatic tools and databases to the proteins for further analysis. The PEDANT system[23] uses a exhaustive collection of methods and databases important for completely automated functional and structural characterisation of protein sequences (a list of used methods and databases is available.[24,25]) PEDANT generates a comprehensive summary page for each individual protein with retrievable detailed results. To facilitate interpretation a graphical overview on pattern matches, structure prediction and protein alignments and their position on individual proteins are generated.

various navigational axes are provided for database access

DATABASE ACCESS AND DATA RETRIEVAL

The MATDB[26] contains data stemming not only from the EU-funded project but from all other *Arabidopsis* genome sequencing groups. It can be accessed using various navigational axes. Besides keyword, gene name, clone name, functional category searches, etc., the MATDB provides opportunity to navigate using positional information. Chromosome overviews are supplied with the option of zooming into regions of interest and then to individual features. For each feature analysis data are summarised on a single page and numerous links for in-depth data retrieval and further analysis data are provided.

Similar to this, the PEDANT database[27] also provides numerous navigational axes. Besides the retrieval of analysis data for individual genes/proteins, lists of proteins containing protein structures of interest (SCOP domains, transmembrane domains, etc.) can be retrieved. PEDANT also supports questions focused towards function (PROSITE patterns, functional categorisation, clusters of orthologous genes (COGs), etc.) and more specialised searches such as similarity to human proteins can be addressed.

TOPOLOGICAL CHROMOSOME ANALYSIS

The approximate 3,800 genes on *Arabidopsis* chromosome 4 are relatively evenly distributed along the two chromosome arms. The average gene density of 1 gene/4.6 kb exemplifies the

the *Arabidopsisi* genome is densely packed

dense packaging of genes on the *Arabidopsis* genome. Density of transposable elements increases towards the centromere whereas gene density drops considerably throughout the heterochromatic subcentromeric regions. Although there is a clear relationship between the proximity to the genetic centromere and an increase in transposon density on one hand, and a decrease of gene density on the other hand, the 'transcriptional profile' as depicted by the number of ESTs matching to individual genes shows remarkable peaks within the heterochromatic regions thereby exemplifying the presence of comparably highly expressed genes within this region (Figure 3). An in-depth analysis of the heterochromatic so-called 'knob'-region on chromosome 4 sheds some light on the very complex composition of this region.[28] It largely consists out of repeat types, in part arranged in long tandem arrays and a complex distribution pattern of various types of transposons although interdispersed genes are also present.

Orthologous clusters

orthologous genes arranged in tandem arrays are prominent in the *Arabidopsis* genome

Another striking finding is the presence of numerous clusters of orthologous genes. On chromosome 4 more than 230 clusters with orthologous gene units arranged in tandem arrays are present. The number of similar units in a row ranges from 2 to 20 individual units. The categories of genes arranged in tandem arrays include, among others, protein kinases, cytochrome P450 proteins and proteins of currently unknown function. Whereas the yeast genome is largely devoid of multigene families organised as gene clusters of more than two related units, such clusters have been observed in higher eukaryotes (homeobox cluster, immunoglobulin cluster, major histocompatibility complex (MHC) cluster). In plants clustered orthologous genes have been described in the context of resistance to pathogens. There is evidence that these clusters serve as templates for illegitimate recombination to create new specificities in pathogen response.[29] Creation of new specificities and rapid adaptation to changing environmental constraints using this mechanism indeed is an attractive interpretation for the high occurrence of multigene families arranged in tandem arrays. However, alternative and/or concurrent interpretations such as co-regulation of the individual units or tissue or developmental timing specificity of individual units cannot be ruled out. This is of particular importance since only the protein coding section of the genes has been under investigation so far. Important structures such as regulatory elements so far have not been subjected to an analysis.

Segmental duplications

Besides similar genes located in the close vicinity, large-scale duplications ranging over several megabases have been

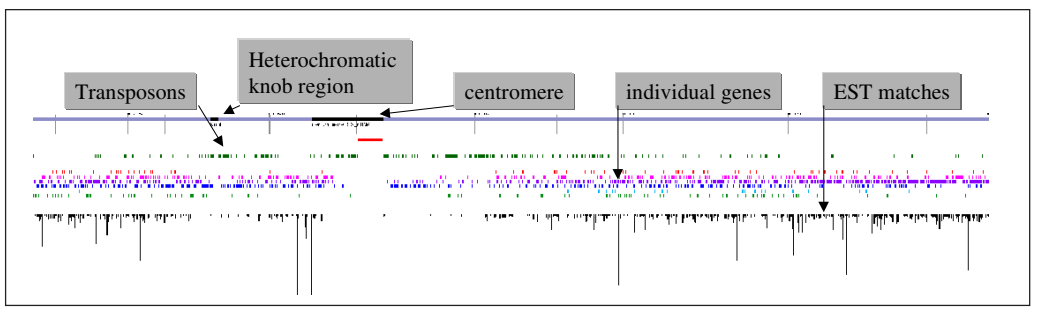

Figure 3: Overview on part of *Arabidopsis* chromosome 4. Approximately 8 Mb of *Arabidopsis* chromosome IV are depicted. Transposon density, gene density and ESTs matching to individual genes are shown. Increasing transposon density and decreasing gene density towards the centromere are readily visible. Highly expressed genes (as measured by the number of EST matches) close to the centromere are observable

similarity of the *Arabidopsis* genome to non-plant depends on cellular process

detected between chromosomes 4 and 2 as well as between chromosomes 2 and 1.[8,9] As an explanation it has been suggested that *A. thaliana* had a tetraploid ancestor. It has been hypothesised that extensive rearrangement by breakage and dispersion across the genome may have an important role in the stabilisation of polyploid plant genomes.[30]

A question to address in this context is to what extent the *Arabidopsis* genome is redundant. Two recently described examples of largely overlapping gene functions may serve to illustrate this issue.[31,32] Functional redundancy on a large scale indeed would interfere with future functional genomics projects by complicating forward genetic screening considerably. We will be able to answer this question at least in part once the complete *Arabidopsis* genome has been completely sequenced and analysed. Data on pairs of highly similar and probably redundant genes will be available to provide fodder for more refined forward and reverse genetic screens.

INTERGENOME COMPARISON

Although chromosome 4 contains only a subset of the *Arabidopsis* genome it is intriguing to ask about the relationship and similarity to other genomes, though one must assume that the chromosome 4 composition is representative of the complete *Arabidopsis* genome.

For this purpose we performed a comparison to other eukaryotic genomes whose sequence had been finished at that timepoint, namely the yeast genome[1] and the *C. elegans* genome.[2] In addition a non-redundant set of human genes and the genomes of *E. coli*[33] and of the photosynthetic cyanobacterium *Synechocystis* sp.,[34] a representative of the probable progenitor of chloroplasts was compared with the chromosome 4 proteome. The chromosome 4 dataset was divided along the assigned functional categories and calculated (using BLAST) against the respective datasets. The percentage of genes with at least one hit below a conservative threshold BLAST *P* value of 1×10^{-30} against the respective datasets was plotted (Figure 4).

Although the very complex outcome of this analysis deserves a much more broad discussion and needs to undergo a more refined and in-depth analysis, some correlations need to be mentioned.

The similarity rate of genes belonging to certain functional categories, eg

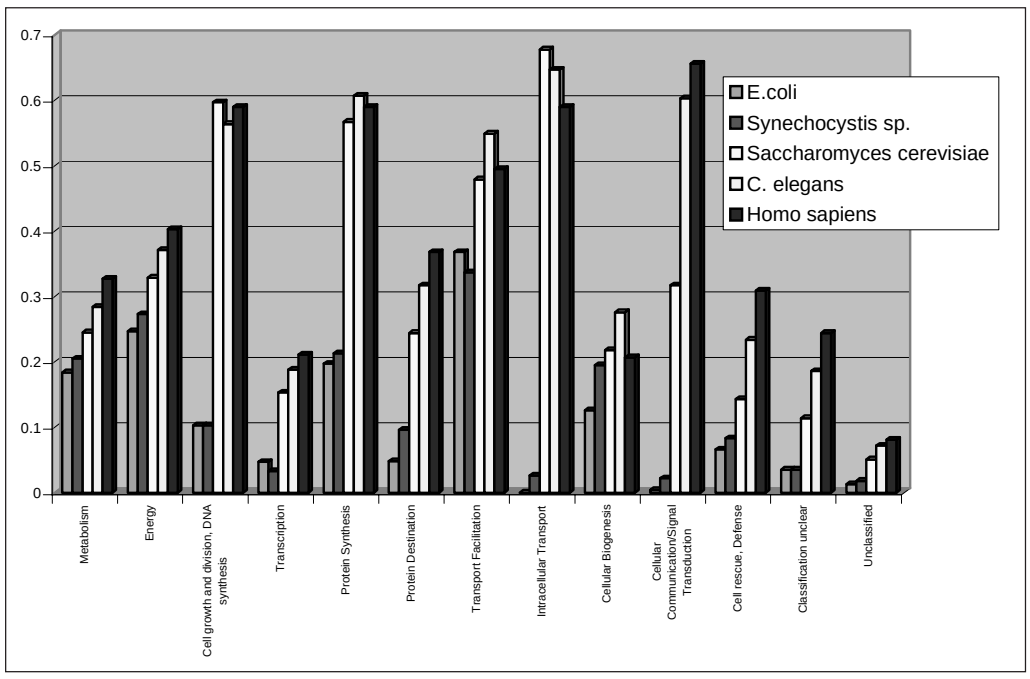

Figure 4: Intergenome comparison between *Arabidopsis* chromosome 4 and various datasets. All chromosome 4 genes with associated functional categories were searched against the complete set of translations from *E. coli*, *Synechocystis* sp., *C. elegans*, *Saccharomyces cerevisiae* and a *Homo sapiens* non-redundant protein database. The proportion of genes with at least on BLAST match of $\leq10^{-30}$ to the respective proteome is shown (0.5 = 50%)

metabolism and energy, between *Arabidopsis* and the eukaryotic datasets was considerably lower than in categories such as intracellular traffic and protein synthesis for example. This might be explicable by the well-known extensive secondary metabolism activities in plants and that plants photosynthesise. This might explain a large *Arabidopsis* gene set that fulfils its task in these categories and that does not have a counterpart within the other (non-plant) eukaryotes.

For some categories (eg transcription, protein synthesis, cell growth, cell division and DNA synthesis) there is a very clear distinction between the prokaryotic and eukaryotic comparisons which is attributable to the well-known basic differences in the underlying basic molecular machinery involved in this processes.

In other categories, eg cellular communication and signal transduction, the boundary between prokaryotes and eukaryotes becomes somewhat indistinct. Whereas 60–65 per cent of the *Arabidopsis* genes have at least one match to *C. elegans* or the human dataset, only ~30 per cent of the genes within this category have a counterpart within yeast. This reflects very well the need of multicellular organisms to have a more pronounced toolkit for communication and signal transduction in between cells and different tissues.

multicellular organisms have a more pronounced toolkit for communication and signal transduction in between cells and different tissues

Human homologues

It was surprising to see that even in an evolutionary distant species such as humans a large variety of closely related genes could be detected. Among the most prominent examples are genes with similarities to the *BRCA2* breast cancer gene and the genes whose defects cause the Werner syndrome and Niemann-Pick C disease. These findings underline the remarkable sequence conservation of proteins involved in a variety of cellular functions which may permit the transfer of biological knowledge even between different phyla.

FUTURE PROSPECTS

With the forthcoming completion of the first complete genome sequence of a higher plant a milestone in plant molecular biology will have been reached. The *Arabidopsis* genome will give a complete list of a plant's gene repertoire. The data will give detailed molecular insight into structure and composition of plant genomes and will shed a light on genome dynamics and redundancy. It will also be the basis for forthcoming functional genomic projects which will involve forward and reverse genetic experiments on a large scale. Combined with new bioinformatic methods, these data will help to elucidate biochemical and signal transduction pathways. It will be feasible in the future to get complete catalogues of functional interactions at transcriptional and protein level for individual cells and complete organisms.

In addition we will be able to address syntheny-related questions, which will enable us to transfer the knowledge gained from studying a weed to plants of agricultural importance.

References

1. Mewes, H. W. *et al.* (1997), 'Overview of the yeast genome', *Nature*, Vol. 387(6632 Suppl.), pp. 7–65.
2. The *C. elegans* Sequencing Consortium (1998), 'Genome sequence of the nematode *C. elegans*. A platform for investigating biology', *Science,* Vol. 282, pp. 2012–2018.
3. Adams, M. D. *et al.* (2000), 'The genome sequence of *Drosophila melanogaster*', *Science*, Vol. 287, pp. 2185–2195.
4. Heslop-Harrison, J. S. (2000), 'Comparative genome organization in plants: From sequence and markers to chromatin and chromosomes', *Plant Cell*, Vol. 12, pp. 617–635.
5. Mozo, T. *et al.* (1998),'Use of the IGF BAC library for physical mapping of the *Arabidopsis thaliana* genome', *Plant J.*, Vol. 16(3), pp. 377–384.
6. Marra, M. *et al.* (1999), 'A map for sequence analysis of the *Arabidopsis thaliana* genome', *Nature Genetics*, Vol. 22(3), pp. 265–270.
7. Mozo, T. *et al.* (1999),'A complete BAC-based physical map of the *Arabidopsis*

thaliana genome', *Nature Genetics*, Vol. 22(3), pp. 271–275.

8. Mayer, K. *et al.* (1999), 'Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*', *Nature*, Vol. 402(6763), pp. 769–777.
9. Lin, X. *et al.* (1999), 'Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*', *Nature*, Vol. 402(6763), pp. 761–768.
10. Jarvis, P., Lister, C., Szabo, V. and Dean, C. (1994), 'Integration of CAPS markers into the RFLP map generated using recombinant inbred lines of *Arabidopsis thaliana*', *Plant Mol. Biol.*, Vol. 24(4), pp. 685–687.
11. http://www.mips.biochem.mpg.de/proj/thal/
12. Parnell, L., Schüller, C., Mayer, K. and Zaccaria, P. (1999), 'The nature of the EST: Distribution across chromosome IV', Poster presentation, 10th international conference on *Arabidopsis research*, Melbourne, Australia.
13. Pavy, N. *et al.* (1999), 'Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences', *Bioinformatics*, Vol. 15 (11), pp. 887–899.
14. Garcia, V., Salanoubat, M., Choisne, N. and Tissier, A. (2000), 'An ATM homologue from *Arabidopsis thaliana*: complete genomic organisation and expression analysis', *Nucleic Acids Res.*, Vol. 28 (8), pp. 1692–1699.
15. Murakami, K. and Takagi, T. (1998), 'Gene recognition by combination of several gene-finding programs', *Bioinformatics*, Vol. 14(8), pp. 665–675.
16. Mathe, C. *et al.* (2000), 'Gene prediction and gene classes in *Arabidopsis thaliana*', *J. Biotechnol.*, Vol. 78(3), pp. 293–299.
17. Burge, C. and Karlin, S. (1997), 'Prediction of complete gene structures in human genomic DNA', *J. Mol. Biol.*, Vol. 268(1), pp. 78–94.
18. Lukashin, A.V. and Borodovsky, M. (1998), 'GeneMark.hmm: new solutions for genefinding', *Nucleic Acids Res.*, Vol. 26, pp. 1107–1115.
19. Green, P. and Hillier, L., unpublished software.
20. Tolstrup, N., Rouze, P. and Brunak, S. (1997), 'A branch point consensus from *Arabidopsis* found by non-circular analysis allows for better prediction of acceptor sites', *Nucleic Acids Res.*, Vol. 25, pp. 3159–3163.
21. Rubin, G. M. *et al.* (2000), 'Comparative genomics of the eukaryotes', *Science*, Vol. 287, pp. 2204–2215
22. Attwood ,T. K. (2000), 'The role of pattern databases in sequence analysis', *Briefings Bioinformatics*, Vol.1(1), pp. 45–59.
23. Frishman, D. and Mewes, H.-W. (1997), 'PEDANTic genome analysis', *Trends Genet.*, Vol. 13, pp. 415–416.
24. http://pedant.mips.biochem.mpg.de/databank.html
25. http://pedant.mips.biochem.mpg.de/methods.html
26. http://www.mips.biochem.mpg.de/proj/thal
27. http://pedant.mips.biochem.mpg.de
28. The Cold Spring Harbor Laboratory, Washington University Genome Sequencing Center and PE Biosystems Arabidopsis Sequencing Consortium (2000) 'The complete sequence of a heterochromatic island from a higher eucaryote', *Cell*, Vol. 100, pp. 377–386.
29. Parniske, M. *et al.* (1997), 'Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato', *Cell*, Vol. 91(6), pp. 821–832.
30. Bancroft, I. (2000), 'Insights into the structural and functional evolution of plant genomes afforded by the nucleotide sequences of chromosomes 2 and 4 of *Arabidopsis thaliana*', *Yeast*, Vol. 17, pp. 1–5.
31. Pelaz, S. *et al.* (2000), 'B and C floral organ identity functions require SEPALLATA MADS-box genes', *Nature*, Vol. 405, pp. 200–203.
32. Liljegren, S. J. *et al.* (2000), 'SHATTERPROOF MADS-box genes control seed dispersal in A*rabidopsis*', *Nature*, Vol. 404, pp. 766–770.
33. Blattner, F. R. *et al.* (1997), 'The complete genome sequence of *Escherichia coli* K-12', *Science*, Vol. 277, pp. 1453–1462.
34. Kotani, H. and Tabata, S. (1998), 'Lessons from the sequencing of the genome of a unicellular cyanobacterium, Synechocystis SP. PCC6803', *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, Vol. 49, pp. 151–171.