APPLICATION OF BIOINFORMATICS IN PLANT BREEDING

D. Vassilev[1], J. Leunissen[2], A. Atanassov[1], A. Nenov[1], G. Dimov[1]
AgroBioInstitute, Sofia, Bulgaria[1]
Wageningen University, The Netherlands[2]

ABSTRACT

The goal of plant genomics is to understand the genetic and molecular basis of all biological processes in plants that are relevant to the specie. This understanding is fundamental to allow efficient exploitation of plants as biological resources in the development of new cultivars with improved quality and reduced economic and environmental costs. This knowledge is also vital for the development of new plant diagnostic tools. Traits considered of primary interest are, pathogen and abiotic stress resistance, quality traits for plant, and reproductive traits determining yield. A genome program can now be envisioned as a highly important tool for plant improvement. Such an approach to identify key genes and understand their function will result in a "quantum leap" in plant improvement. Additionally, the ability to examine gene expression will allow us to understand how plants respond to and interact with the physical environment and management practices. This information, in conjunction with appropriate technology, may provide predictive measures of plant health and quality and become part of future breeding decision management systems. Current genome programs generate a large amount of data that will require processing, storage and distribution to the multinational research community. The data include not only sequence information, but information on mutations, markers, maps, functional discoveries, etc. Key objectives for plant bioinformatics include: to encourage the submission of all sequence data into the public domain, through repositories, to provide rational annotation of genes, proteins and phenotypes, and to elaborate relationships both within the plants' data and between plants and other organisms.

Introduction

Over* the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This deluge of genomic information has, in turn, led to an absolute requirement for computerized databases to store, organize, and index the data and for specialized tools to view and analyze the data.

The publication of the completed *Arabidopsis thaliana* genome sequence (1) and draft sequence for rice genome (10) the plant research and industry has step over the threshold of the genomics era. The numerous applications of genomic information opened wide opportunities to start integrating rich rewards from sub-systems biology, integrative biology and large scale systematic functional genomics projects. With this accumulation of various types of data is possible freely to enter the universe of "genomic understanding". With this un-

* **Abbreviations:** rDNA- recombinant DNA, mRNA – messenger RNA, EST – Expressed Sequence Tag, BLAST – Basic Local Alignment Sequence Tool, NCBI – National Center for Biotechnology Information, cDNA – complementary DNA, dbEST – data base of ESTs, TIGR – The Institute of Genomic Research, QTL – Quantitative Trait Loci, MAS – Marker Assisted Selection.

derstanding it is possible to model and design the amount and sense of changes in gene expression level, or how to localize proteins and assess their interactions with other genes and proteins and finally how are affected the metabolite pools within any given tissue. To reach these horizons there will be a huge scientific undertaking and many aspects will be undoubtedly reliant on bioinformatics (23).

Having in mind the potential power of data hidden within the complete genome scaffolds, or even within the partial transcriptomics data available for more plant species, it is logically to consider that bioinformatics has been integrated as a crucial part of the modern genomics research. Bioinformatics is thoroughly involved with the completion and assessment of a multitude of different complete genome sequences (5). As a science of data management in genomics and proteomics, and as a young discipline in information technology bioinformatics has progressed very fast in the last twenty years. Methods of bioinformatics are practised worldwide to access various databases and to exchange information for comparison, confirmation, storage and analysis of biological data. As on date, there are a number of databases on specific seizes and proteins pertaining to human, animals, plants, bacteria, and other life forms (9).

These databases help in new inventions in biology and medicine that are useful to mankind. Bioinformatics is enabling life sciences to invent novel drug discovery as well as drug delivery systems for greater progress in the field of biotechnology. Such inventions attain importance in the present scenario of patents aid WTO regime. For the future development of biotechnology, bioinformatics will have to play a vital role with the involvement of internet tools and the World Wide Web (WWW). The future rDNA research would be guided largely by the databases available for generic or specific forms. Thus bioinformatics and biotechnology have to move hand in hand for their progress. However, bioinformatics can now be branded as a bonafide discipline within information technology (3).

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. At the beginning of the "genomic revolution", a bioinformatics concern was the creation and maintenance of a database to store biological information, such as raw nucleotide and amino acid sequences. Development of this type of databases involved not only design issues, but the development of complex interfaces, whereby researchers could both access existing data, as well as submit new or revised data (11). The 2005 update of The Molecular Biology Database Collection includes 719 freely available to the public biological databases, 171 more than the year of 2004.

Therefore, the field of bioinformatics has evolved as the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, protein structures and expression patterns. The actual process of analyzing and interpreting data is referred to:

- the development and implementation of tools that enable efficient access to, and use and management of, various types of information
- the development of new algorithms and statistics with which to assess relationships among members of large data sets, such as methods to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences (3).

There are newly developed fields in the science, related to bioinformatics and geno-

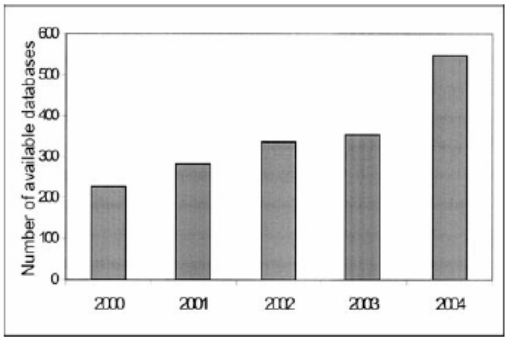


Fig. 1. Growth in number of databases listed in the Molecular Biology Database Collection [2-6].

TABLE 1
Classification of databases in the 2004 edition of the Molecular Biology Database Collection (11)

Category	No of databases
Genomic	164
Protein sequences	87
Human/vertebrate genomes	77
Human genes and diseases	77
Structures	64
Nucleotide sequences	59
Microarray/gene expression	39
Metabolic and signaling pathways	33
RNA sequences	32
Proteomics	6
Other	16

mics, which deals with the visualization of vast amount of non-human-readable primary data and highly complex systems. (28)

Bioinformatics in plant breeding

How important is plant bioinformatics? Plants are the basis of life on earth. They produce the life-supporting oxygen we breathe, they are essential for our nutrition and health and they provide the environment for the vast biodiversity on earth. For centuries, humans have selected plant varieties that best fit their purposes and developed crop plants that have many advantages compared to natural (wild) plants in quality, quantity and farming practises. However, multifactorial traits involved in resistance and quality have proven to be extremely difficult to improve, certainly in combination. The revolution in life sciences signalled by genomics dramatically changes the scale and scope of our experimental enquiry and application in plant breeding. The scale and high resolution power of genomics enables to achieve a broad as well as detailed genetic understanding of plant performance at multiple levels of aggregation. The complex biological processes that make up the mechanisms of pathogen resistance and provide quality to our crops are now open for a systematic functional analysis. These analysis are made with specific software on the high amounts of data generated in databases and is the field of plant bioinformatics. (15, 16).

Genome initiatives are under way for more than 60 different plant species. From the point of view of economics, the most important of these are those of the major feed crops – the grasses maize, rice, wheat, sorghum and barley; and the forage legumes soybean and alfalfa. Several of these genomes are so large (as result of autopolyploidization and the dramatic expansion of repetitive DNA) that whole genome sequencing is impractical, and efforts have instead been focused on comparative genome methods. Both rice and maize, however, have relatively small genomes and are such key elements of the agricultural economies of the developed world that complete genome sequences have been prioritazed.

The role of model organism. Over the last century, research on a small number of organisms has played a pivotal role in advancing our understanding of numerous biological processes. This is because many aspects of biology are similar in most or all organisms, but it is frequently much easier to study a particular aspect in one organism

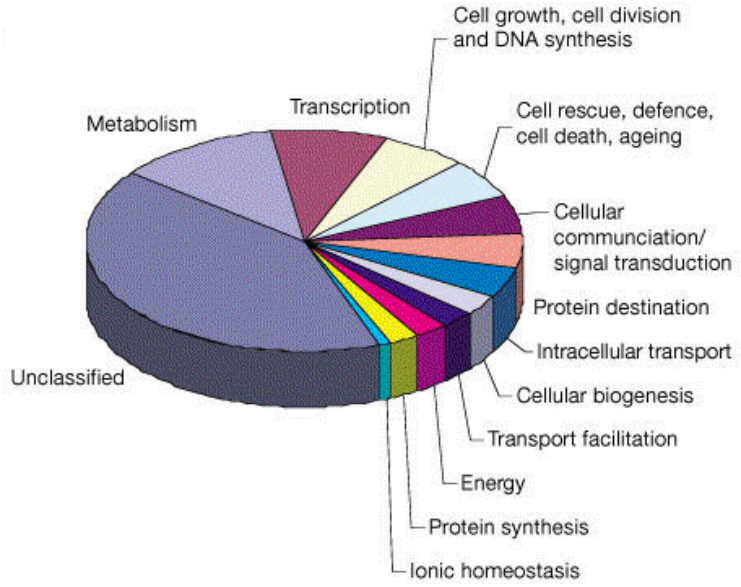
than in others. These much-studied organisms are commonly referred to as model organisms, because each has one or more characteristics that make it suitable for laboratory study. The most popular model organisms have strong advantages for experimental research, such as rapid development with short life cycles, small adult size, ready availability, and tractability, and become even more useful when many other scientists work on them. A large amount of information can then be derived from these organisms, providing valuable data for the analysis of normal human or crop development; gene regulation, genetic diseases, and evolutionary processes [http://www.bioinformatics.nl].

Comparison of genome sequences of rice and Arabidopsis suggests that extensive but complex patterns of synteny will be a useful feature of plant genomics. Medicago (alfalfa) is a true diploid that, along with its crucial role in the fixing soil nitrogen, constitutes a major part of forage diets. Other grasses and legumes are the subjects of extensive EST sequencing and high resolution genetic map construction, in some cases involving radiation hybrid mapping, in hopes of taking advantage of the expected pervasive synteny within these families. Web sites established by individual research groups integrate research efforts from around the globe. Some useful web sites include the UK CropNet (http://uk-crop.net), the U.S. Agricultural Research Service's http://ars-genome.cornell.edu and organism-specific resources such as MaizeDB (http://www.agron.missouri.edu). A common objective of these sites is to link seed stock and real genetic resources to virtual data on linkage and mapping data, for which purpose search engines and ever more sophisticated relational databases are under development.

In the 1980s, there was a growing awareness that significant investments in studies of many different plants, such as corn, oilseed rape, and soybean, were diluting efforts to fully understand the basic properties of all plants. Scientists began to realise that the goal of completely understanding plant physiology and development is so ambitious that it can best be accomplished by turning to a model plant species that many scientists then study. Fortunately, because all flowering plants are closely related, the complete sequencing of all the genes of a single, representative, plant species will yield much knowledge about all higher plants. Similarly, discovery of the functions of the proteins produced by a model species will offer much information about the roles of proteins in all higher plants.

Arabidopsis thaliana has become universally recognised as a model plant for study. It is a small flowering plant that belongs to the *Brassica* family, which includes species such as broccoli, cauliflower, cabbage, and radish. Although it is a non-commercial plant, it is favoured among basic scientists because it develops, reproduces, and responds to stress and disease in much the same way as many crop plants. Scientists expect that systematic studies of Arabidopsis will offer important advantages for basic research in genetics and molecular biology and will illuminate numerous features of plant biology, including those of significant value to agriculture, energy, environment, and human health. Because of several reasons Arabidopsis has become the organism of choice for basic studies of the molecular genetics of flowering plants.

Arabidopsis thaliana has a small genome (125 Mb total), which already has been sequenced in the year 2000, [www.bioinformatics.nl], and it lacks the repeated, less-informative DNA sequences that complicate genome analysis. It has extensive genetic and physical maps of all 5 chromosomes (MapViewer); a rapid life cycle (about 6 weeks from germination to mature seed); prolific seed production and easy cultivation in restricted space; effi-



cient transformation merits utilising *Agrobacterium tumefaciens*; a large number of mutant lines and genomic resources (Stock Centers) and multinational research community of academic, government and industry laboratories.

During the *Arabidopsis* evolution the whole genome has duplicated once, followed by subsequent gene loss and extensive local gene duplications. The genome contains 25.498 genes encoding proteins from 11.000 families. The genome is compared to previous sequenced genomes and ESTs, and as much functions of genes are predicted. This functional analysis of the *Arabidopsis* genome showed the following proportion of predicted function. As with other model organisms there is much more to the Arabidopsis genome project than the complete genome sequence. The Web site for the *Arabidopsis* Information Resource, TAIR (http://www.arabidopsis.org), allows researchers to integrate the genome sequence with an extensive EST data base and with the genetic and physical maps; it provides links to functional and molecular genetic data and the literature for specific genes; and it shows an ever expanding list of mutant stocks. An alternative resource for *Arbabidopsis* and many other plants UK CropNet, uses common AceDB or WebAce platforms to coordinate genetic and molecular data.

As plant model organisms also could be considered**:**

- **Medicago truncatula** (commonly known as "barrel medic" because of the shape of its seedpods) is a forage legume commonly grown in Australia.
- **Tomato.** The overall goal of the **Tomato Genomics Project** includes the development of an integrated set of experimental tools for use in tomato functional genomics. The resources developed will be used to further expand our understanding of the molecular genetic events underlying fruit development and responses to pathogen infection, and will be made available to the research community for analysis of diverse plant biological phenomena. http://www.sgn.cornell.edu/about/tomato_project/project.html

- **Rice.** Some desired features of future improved rice varieties are superior grain quality, higher yield potential, enhanced resistance to insect pests and diseases, and greater tolerance to stresses such as drought, cold, and nutrient deficiencies. http://www.riceweb.org/research/Res_ntbio.htm
- **Maize.** Maize products produce about $30 billion every year, and is used for food, rubber, plastic, fuel, and clothing. The maize genome is about 20 times larger than the one from Arabidopsis. This means that it is as big as the human genome. However, it's organisation is more complex to sequence than that of all organisms that are sequenced today. The genes are situated in clusters through the genome, with high amounts of **repetitive sequences** in between. The genes containing regions make up to 15% of the total genome. Other significant characteristics of the maize genome are that it contains **multiple copies** of most genes and the existance of jumping genes or transposons that make up a large portion of the genome.
  http://www.agron.missouri.edu
- **Wheat.** Recent advances in plant genetics and genomics offer unprecedented opportunities for discovering the function of genes and potential for their manipulation for crop improvement. Because of the large size of the wheat genome, it is unlikely that the actual base pair sequences of the DNA molecules will be learned completely in the near future. This project takes an alternative strategy to realise the benefits of new techniques for discovering genes and learning their function (functional genomics). Following the identification of 10,000 wheat ESTs, they will be mapped to their **physical location** on the chromosomes of wheat. This process utilizes a unique feature of the wheat chromosomes, their ability to tolerate deletions of portions of the chromosomes and still produce a viable plant. The mapping logic is direct: if an EST is present in a plant with complete chromosomes, but absent in a plant missing a known part of a single chromosome, then it can be inferred that the DNA sequence that corresponds to that EST is located in that segment of the chromosome. By the end of the mapping component of this project, a most valuable tool will have been produced: 10,000 unique DNA sequences, likely corresponding to genes, whose physical location in the chromosomes of wheat are known. This sets the stage for the next phase of the project, the analysis of this array of mapped ESTs to **determine function**.
  http://wheat.pw.usda.gov/NSF/images.html
- **Other flowering plants.** Over 90 different angiosperm genome projects around the world are listed on the United States Department of Agriculture Web site (http://www.usda.go/pgdic/Map.proj)
  The list includes African projects on beans, corn and fungal pathogens; Australian projects on cotton, wheat, pine, sugarcane, and nine others; at least 24 European projects that include vegetables such as cabbage, cucumber, and pea, and fruits such as apple, peach and plum; and over 50 North American projects as diverse as turf grass, chrysanthemum, almond, papaya grape and poplar. The common denominator among all of these projects is the assembly of genetic maps (and in some cases physical maps) and the placement of a common set of plant genes on them. For some species large EST sequencing projects are also in place with the twin objectives of enabling comparative genomic analysis (particularly in the regions of synteny) and QTL mapping.

Managing and distributing plant genome data

As with many areas of science and technology genome science has benefited

greatly from advances in computing capabilities and bioinformatics. Improved computational speed has been important but a strong argument can be made that the growth of the Internet has been even more crucial for genome scientists. In conjunction with the maturation of modern database technology the World Wide Web has become the natural medium for managing and distributing genomic resources.

The emergence of the Internet allowed the creation of centralized data warehouses. Just as important, it led to the creation shared public resources for searching and analyzing the contents of genomic databases. Full-featured Web sites such as those at NCBI (http://www.ncbi.nlm.nih.gov) and EMBL (http://www.embl-hedelberg.de) provide immediate access to enormous amounts of data and analysis tools, free of charge, from anywhere of the globe. This is dramatic change from the situation about a decade go, when GenBank database was distributed by paid subscription in a small notebook, full of 5.25" floppy disks.

Networking advances have also been important for within laboratory data management and with little or no human intervention. Centralized laboratory information management systems or LIMS, then allow users at multiple workstations or even multiple geographic locations to browse, edit, analyze and annotate the data.

The core item of genomic data is a database system. Most databases can be classified as either relational databases (RDB) or object-oriented data bases (OODB).

There are three primary sequence databases: GenBank (NCBI), the Nucleotide Sequence Database (EMBL) and the DNA Databank of Japan (DDBJ). These are repositories for raw sequence data, but each entry is extensively annotated and has features table to highlight the important properties of each sequence. The three databases exchange data on a daily basis. Similarly , SWISS-PROT and TrEMBL are the major primary databases for the storage of protein sequences. There are also secondary databases of protein families and sequence patterns such as PROSITE, PRINTS and BLOCKS. These are called secondary databases because the sequences they contain are not raw data, but they have been derived from the data in the primary databases. The journal Nucleic Acid Research devotes its first issue every year to articles describing new databases and updates in existing ones. It is called Molecular Biology Database Collection and is freely available immediately upon publication. (www.nar.oupjournals.org) (15).

The early bioinformatics databases emphasized primary data capture. GenBank, established in the late 1980's began with staff at Los Alamos National Laboratoty (and later the National Center for Biotechnology Information) manually keying DNA sequences from published papers into the computer. It neither added nor removed any information from these sequences, nor did it perform any integration of multiple overlapping sequences. Other databases founded around this time also focused on a single data type: SwissProt for protein sequences and PDB for X-ray crystallographic structures. From the mid-1990s to the early part of this decade the emphasis shifted from data capture to data aggregation and integration. This was largely due to the limitations of primary data archives: what if one wanted to correlate DNA sequencing information with data from biochemical or genetic studies? Model Organism Databases (MODs), integrated repositories of all the electronic information resources pertaining to a particular experimental plant or animal species, became the darlings of the bioinformatics world (29).

Now the MOD paradigm is itself giving way to new, higher level concepts, such as clade-specific and pathway databases. Integrating multiple types of biological data across several species, these resources enable researchers to make discoveries that wouldn't be possible by examining a single

species alone. It could be predicted that in another decade the idea of biological database devoted to one species will seem as quaint as the idea of a database devoted to a single type of laboratory data seems today.

Though MOD are still going strong, their preeminence is now being challenged by multispecies, comparative-genomics databases, sometimes called clade-specific databases. These systems integrate information on multiple organisms and use comparative analysis to discover patterns in genome that might otherwise be missed. Well known clade specific databases include: EnsEMBL at the European Bioinformatics Institute (EBI; [http://www.ensembl.org]), Entrez at the NCBI [http://www.ncbi.nlm.nih.gov], and the Genome Browser at the University of California, Santa Cruz [http://www.genome.ucsc.edu], all of which relate information on the human genome to data gathered over plants, vertebrates, invertebrates and prokaryotes.

Some years ago the Cold Spring Harbor Laboratory, New York established Gramene [http://www.gramene.org], a comparative genomics resource for crop grasses. This database integrates genome sequences, genetic maps, mutation and trait data across rice, maize wheat and a large number of other cereals. Gramene gives researchers the benefit of genome sequencing even before their favorite organism actually has been sequenced.

The maize genome, for instance, is about the same length as the human genome, and won't be fully sequenced for another several years, but rice, with a compact genome one-tenth the size of human's, already is. Because the two grains are closely related evolutionarily, we have been able to create maps that relate maize's genetic map to the rice genome sequence. This allows to researchers to follow a genetically mapped trait in maize, such as tolerance to high salt levels in the soil, and move into the corresponding region in the rice genome, thereby identifying candidate genes for salt tolerance. Similar techniques helped cattle researchers identify in 1997 a gene responsible for muscle growth based on the existence of a genetic mutation in a corresponding region of the mouse genome.

The next very important class of databases in near future could be considered as pathway databases. Traditional databases are linear catalogues of sequences, genes, proteins, genomes, and genome-to-genome alignments. Such databases have one or a small number of central data objects, such as gene record, and all the other information hangs off that object. A typical research project describes the model of the series of experiments. The model usually describes the series of molecular events (the pathway) that is responsible for whatever phenomenon is studied, whether it be embryonic development, neuronal signaling in the brain, or the transformation of healthy cells into cancerous ones. These pathways are the ultimate output of biological research. The initial project in pathway databases is the Reactome [http://www.reactome.org]. Current entries describe energy metabolism, DNA replication, RNA transformation and splicing, protein translation and cell cycle regulation. Each pathway is linked to the literature references that provide experimental support for it, and to the database records for genes, sequences and proteins that participate in the pathways.

A new standard in transferring metabolomics data has been developed since 2002 by several BioPAX project, and financed by Department of Energy of United States. The goal of the BioPAX group is to develop a common exchange format for biological pathways data.

A sample pathway entry in the new standard, implemented in OWL

(http://www.w3.org/TR/owl-features)

will look like this:

```
<owl:Ontology rdf:about="">
<owl:imports
rdf:resource="http://www.biopax.org/release/biopax-
level1.owl"/>
</owl:Ontology>
<bp:bioSource rdf:ID="bioSource33">
<bp:NAME
rdf:datatype="http://www.w3.org/2001/XMLSchema
#string">Escherichia coli</bp:NAME>
   <bp:TAXON-XREF>
    <bp:unificationXref rdf:ID="unificationXref34">
     <bp:ID>562</bp:ID>
     <bp:COMMENT
rdf:datatype="http://www.w3.org/2001/XMLSchema
#string">This is an example of BioPAX
project</bp:COMMENT>
      <bp:DB
rdf:datatype="http://www.w3.org/2001/XMLSchema
#string">taxon</bp:DB>
    </bp:unificationXref>
   </bp:TAXON-XREF>
</bp:NAME>
</bp:bioSource>
</rdf:RDF>
```

Sequence alignment methods and applications

Comparing genome sequences. The development of technologies for the large-scale quantification and identification of biological molecules combined with advances in computing technologies and the internet has served to facilitate the delivery of large volumes of biological data to the scientists' desktop. By the time the human genome sequence was published in 2001, the rate of DNA sequencing had increased 2,000-fold since the inception of the technology in 1986 (12). The increased productivity was gained through automation, miniaturization, and integration of technologies; applying this approach to the analyses of other biological molecules including mRNA, proteins, and metabolites has resulted in a massive increase in the generation of biological data. This data has been made easily accessible, in part due to publications such as the Molecular Biology Database Collection, an annual listing of the best databases publicly available to the biological community. Analysis of the collection reveals the steady growth in the quality and size of the databases (**Fig. 1**), with the 2004 edition containing 548 databases classified into 11 categories (**Table 1**).

DNA sequencing is performed using and automated version of the chain termination reaction, in which limiting amounts of dideoxyribonucleotides generate sets of DNA fragments with specific terminal bases. Four reactions are set up, one for each of the four bases in DNA, each incorporating different fluorescent label. The DNA fragments are separated by the PAGE and the sequence is read by a scanner as each fragment moves to the bottom of the gel.

DNA sequences come in three major forms. Genomic DNA comes directly from the genome and includes extragenic material, as well as genes. In eukaryots, genomic DNA contains introns. cDNA is reverse transcribed from mRNA and corresponds only to expressed parts of the genome. It does not contain introns. Finally, recombinant DNA comes from the laboratory and comprises artificial DNA molecules such as cloning vectors.

Major aim of most genome projects is to determine the DNA sequence either of the genome or of a larger number of transcripts. This endeavor both leads to the identification of all or most genes and to the characterization of various structural features of the genome. Very often the major essence of the bioinformatics strategies for sequence alignment is the comparison of cDNA/EST and genomic sequences and annotation. The veracity of any whole genome sequence must be assessed at three levels: its completeness, the accuracy of the base sequence and the validity of its assembly.

In addition to whole genome sequencing, plant sequence data have been accumulating from three major sources: sample sequencing of bacterial artifcial chromosomes (BACs), genome survey sequencing (GSS) and sequencing of expressed se-

quence tags (ESTs).

Sequence alignment is the arrangement of two or more amino acid or nucleotide sequences from an organism or organisms in such a way as to align areas of the sequences sharing common properties. The degree of relatedness or homology between the sequences is predicted computationally or statistically based on weights assigned to the elements aligned between the sequences. This in turn can serve as a potential indicator of the genetic relatedness between the organisms (3).

Alignment is a computational problem. There is a certain degree of convinction, that two similar sequences can be lined up in such a way that identical bases (or amino acids) are all matched. However from a computers point of view the alignment process is far from trivial. If gaps are allowed there are a tremendous number of of different alignments possible for any two sequences.

Dynamic programming algorithms can calculate the best alignment of two sequences. Well known variants for pairwise alignment are the Smith-Waterman algorithm for local alignment and the Needelman-Wunsch (19) algorithm for global alignment. Local alignments are useful when sequences are not related over their full lengths, for example proteins sharing only certain domains, or DNA sequences related only in exons.

Multiple sequence alignment. Multiple alignment illustrates relationships between two or more sequences. When the sequences involved are diverse, the conserved residues are often key residues associated with maintenance of structural stability or biological function. Multiple alignments can reveal many clues about protein structure and function. The most commonly used alignment software is the ClustalW package. A desktop version of ClustalW software is freely available by ftp from:

ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX

Alignment scores and gap penalties. A simple alignment score measures the number of proportion of identically matching residues. Gap penalties are subtracted from such scores to ensure that alignment algorithms produce biologically sensible alignments without too many gaps. Gap penalties may be constant (independent of the length of the gap), proportional (proportional to the length of the gap) or affine (containing gap opening and gap extension contributions). Gap penalties can be varied according to the desired application. Sequence similarity can be quantified using score from the alignment algorithm, percentage sequence identities or more complex measures.

Sequence Similarity Searching Algorithms

Dynamic programming algorithms are guaranteed to find the best alignment of two sequences for given substitution matrices and gap penalties. This is impressive, but the process is often quite slow, perhaps taking hours for a search of a large database. For these reason alternative methods have been developed. Perhaps the most used of these are FASTA [http://www.ebi.ac.uk/fasta] and BLAST http://www.ncbi.nlm.nih.gov/BLAST/ (2). Both tools BLAST and FASTA provide very fast searches of sequence databases. Unlike dynamic programming, they do not guarantee to find the best possible alignment to each database sequence, but in practice the effect of performance is usually minimal. Each operates by first locating short stretches of identically or near identically matching letters (words) that are eventually extended into longer alignments. Best BLAST server runs the NCBI and can be used to search many general-purpose sequence databases. A similar FASTA implementation is available at the EBI.

Smith-Waterman is an algorithm for local sequence alignment, using as input two sequences (26). The difference between

NCBI BLAST (also considered as local alignment algorithm) and Smith-Waterman is that a) BLAST is searching query sequence over a database of sequences; and b) BLAST is calculating statistically most probable alignment match, since Smith-Waterman is calculating the exact match.

Genome Comparison Tools**.** MegaBlast is NCBI BLAST based algorithm for large sequence similarity search (12). MegaBlast implements a greedy algorithm for the DNA sequence gapped alignment search. MegaBlast is used to compare the raw genomic sequences to a database of contaminant sequences (including the UniVec database of vector sequences, the Escherichia coli genome, bacterial insertion sequences, and bacteriophage databases). Any foreign segments are removed from draft-quality sequence or masked in finished sequence to prevent them from participating in alignments.

Jim Kent's BLAT (BLAST-Like Alignment Tool) is a tool which performs rapid mRNA/DNA and cross-species protein alignments. BLAT is more accurate, 500 times faster than popular existing algorithms for mRNA/DNA alignments, and 50 times faster for protein alignments at sensitivity settings typically used when comparing vertebrate sequences.

Genome based multiple alignment using BlastZ**.** BLASTZ is a multiple sequence alignment program basically used for the whole-genome human-mouse alignments. Blastz output can be viewed with the LAJ interactive alignment viewer, converted to traditional text alignments. LAJ is a tool for viewing and manipulating the output from pairwise alignment programs such as BLASTZ. It can display interactive dotplot, pip, and text representations of the alignments, a diagram showing the locations of exons and repeats, and annotation links to other web sites containing additional information about particular regions.

EST sequencing ESTs are partial gene sequences which have been generated or are in the process of being produced in several laboratories using different species and cultivars as well as varied tissues and developmental stages (13). This represents an important step towards the identification of all expressed genes for instance in grapevine, and some members of the *Rosaceae* family: a large part of the *Malus* (apple) genome, raspberries/blackberries *Rubus*, stone fruits *(Prunus)*, strawberry *(Fragaria)* peach, almond (21).

ESTs are now widely used throughout the genomics and molecular biology communities for gene discovery, mapping, polymorphism analysis, expression studies, and gene prediction

Expressed sequence tags (ESTs) have applications in the discovery of new genes, mapping of the genome, and identification of coding regions in genomic sequences. An EST database consists of ESTs drawn from multiple cDNAs, and there could be potentially many ESTs drawn from each cDNA. Given such a database, the EST clustering problem is defined as follows: The ESTs should be partitioned into clusters such that ESTs from each gene are put together in a distinct cluster. A further complication arises due to the fact that DNA is a double stranded molecule and a gene could be part of either strand (23).

dbEST (30) is a division of GenBank that contains sequence data and other information on "single-pass" cDNA sequences, or Expressed Sequence Tags, from a number of organisms. The Institute for Genomic Research (TIGR) is defining also TC as Tentative Consensi (assemblies from ESTs) and ET as Expressed Transcripts (both non-human) when building TIGR Gene Indices (TGI) (**Table 2**).

Molecular information and plant breeding – a bioinformatics approach

Molecular plant breeding. As the resolution of genetic maps in the major crops increases, and as the molecular basis for

TABLE 2
Number of ESTs by fruit collected in dbEST (*release 040805*)

dbEST release 040805
Number of public entries: 26,605,325

Malus x domestica	183916
Vitis vinifera	147300
Prunus armeniaca	15181
Citrus x paradisi x Poncirus trifoliata	8002
Vitis hybrid cultivar	6533
Fragaria x ananassa	5322
Prunus dulcis	3864
Citrus reticulata	3735
Citrus unshiu	2561
Ananas comosus	1547
Fragaria vesca	1306
Citrullus lanatus	693
Citrus clementina x Citrus reticulata	74
Vitis cinerea x Vitis rupestris	61
Cucumis melo	60

specific traits or physiological responses becomes better elucidated, it will be increasingly possible to associate candidate genes, discovered in model species, with corresponding loci in crop plants. Appropriate relational databases will make it possible to freely associate across genomes with respect to gene sequence, putative function, or genetic map position. Once such tools have been implemented, the distinction between **breeding** and **molecular genetics** will fade away. Breeders will routinely use computer models to formulate predictive hypotheses to create phenotypes of interest from complex allele combinations, and then construct those combinations by scoring large populations for very large numbers of genetic markers (27) , (7).

The vast resource comprising breeding knowledge gathered over the last several decades will become directly linked to basic plant biology, and enhance the ability to elucidate gene function in model organisms (8). For instance, traits that are poorly defined at the biochemical level but well established as a visible phenotype can be associated by high resolution mapping with candidate genes. Orthologous genes in a model species, such as *Arabidopsis* or rice, may not have a known association with a quantitative trait like that seen in the crop, but might have been implicated in a particular pathway or signaling chain by genetic or biochemical experiments. This kind of cross-genome referencing will lead to a convergence of economically relevant breeding information with basic molecular genetic information. The specific phenotypes of commercial interest that are expected to be dramatically improved by these advances include both the improvement of factors that traditionally limit agronomic performance (input traits) and the alteration of the amount and kinds of materials that crops produce (output traits). Examples include:

- abiotic stress tolerance (cold, drought, salt)
- biotic stress tolerance (fungal, bacterial, viral, chewing and sucking insect attack (feeding))
- nutrient use efficiency
- manipulation of plant architecture and development (size, organ shape, number, and position, timing of development, senescence)
- metabolite partitioning (redirecting of carbon flow among existing pathways, or shunting into new pathways)

Rational plant improvement. The implications of genomics with respect to food, feed and fibre production can be envisioned on many fronts. At the most fundamental level, the advances in genomics will greatly accelerate the acquisition of knowledge and that, in turn, will directly impact many aspects of the processes associated with plant improvement. Knowledge of the function of all plant genes, in conjunction with the further development of tools for modifying and interrogating genomes, will lead to the development of a genuine genetic engineering paradigm in which rational changes can be designed and mo-

deled from first principles.

Genotype building experiments

Biodiversity determined by the plant genome analysis. In the recent years an increasing amount of information for the DNA polymorphism and sequencing was accumulated in different plant varieties and cultivars. Most of this information was used for the purpose of recognition of different cultivars as well as for their comparison – distances and similarities (22). These distances are measured by the polymorphism on a part of the chromosome with unknown function. This type of polymorphism is widely used in the genomic studies across the species. The data for the polymorphism are analyzed for a possible link with a quantitative trait of interest of the individual phenotypes. Once such a link is detected it is called indirect marker (14).

Indirect markers are closely linked, sometimes they may overlap, with a locus which determine this quantitative trait – QTL. QTLs are defined as genes or regions of chromosomes which affect a trait. QTLs by themselves are difficult to be recognized. In both cases this information, or as it is called – markers, can be used in further selection purposes. This selection process is named as MAS (18).

QTLs and mapping. The major problem is to define which populations are suitable for QTL-analyses – unstructured and F2 crosses and in plant - large scale populations in order to screen for possible QTLs.

As selection is based most on markers, higher density of mapping is important. The interval between marker and QTL of about 5 centiMorgans (cM) seemed sufficient for effective selection. The simulation studies however showed that selection accuracy dropped down to 81% and 74% with 2 cM and 4 cM distance compared to 1cM (25).

How QTL information could be of use?

- it is assumed that some but not all loci are identified, so selection should be based on the combination of phenotypic and molecular information;
- in the process of selection the link of markers and traits could decrease so this link should be observed throughout the generations;
- in the process of selection QTLs prove simultaneous existence of the desired genes in a line;
- in crossbred programs QTLs could predict the productivity of untested crosses, including their non-additive effect on the information of the parent lines and limited number of crosses;
- future prospective – with accumulation of molecular data genotype building programs will be developed which will set homozygous desirable markers;
- in intrigression programs for combining the desirable traits from two lines in one;
- finally – the real world of agriculture is on the stage of accumulation of molecular data.

Analytical approaches. One of the statistical tools for performing the QTL analyses such is the meta-analysis, which synthesize dense QTL information and refines the QTL position. A program of this class is the French BioMercator. An environment with complex research opportunities is also PlaNet – the European plant genome database network, which is available at [http://www.eu-plant-genome.net].

Further development and efficiency of QTLs and MAS. Further development and detailed discussion on QTLs includes the statistical aspects of MAS, setting up the threshold of significance of marker effects, overestimation or bias in estimation of QTL effects, optimization of selection programs for several generations with simultaneous utilization of MAS and phenotypic data. A specific feature is that detection should be made on plant specific parts – leaves, roots, fruits etc., as it was proved for the grapes (18).

Experimental results not always confirm the efficiency of MAS over the genotype

building. The main reason is in unsufficient precision of the initial assessment of a QTL, its place and effect. Some QTLs also could be lost in the genotype building process. For complex productivity traits the epistatic lost would be a reason for changes in the magnitude of QTL effect in the parent and progeny generation. Then it is recommended that selection is based on the allelic combinations rather on the separate QTLs. It is in line with the numerous GxE interactions and with the selection within the environment of interest in the case of disease/drought resistance (17).

Consequently, efficiency of MAS will depend on the complexity of species/trait genetic architecture, on the development of the trait in the environment and on their interaction. For complex traits the evaluation of QTLs should be in different environments. Phenotypic evaluation/check throughout the consecutive generations is also necessary (20). For instance: drought resistance seemed to be more complex trait vs. disease resistance.

From the economics point of view the use of markers will cost collection of DNA, genotyping, analyses, detection of QTLs etc. This high price is paid for the genotype building (there is no other way of doing that), for traits that are expensive for evaluation – disease resistance, or traits with low heritability.

REFERENCES

1. **AGI** (2000) Nature, **408**, 796-815.
2. **Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J.** (1990) Journal of Molecular Biology, **215**, 403–410.
3. **Baxevanis A., Ouellette F.** (2001) Bioinformatics: A practical guide to the analysis of genes and proteins. John Willey & Sons, Inc. N.Y., USA. p. 518.
4. **Carlborg O., Haley C.** (2004) Nature Reviews Genetics, **5**, 618-625.
5. **Claverie J-M., Notredame C.** (2003) Bioinformatics for Dummies. Willey Publ. Inc. N.Y., USA, p. 452.
6. **Davenport G., Ellis N., Ambrose M., Dicks J.** (2004) Euphytica, **137(1)**, 39-54.
7. **Deckers J., Hospital F.** (2002) Nature Reviews Genet., **3**, 22-32.
8. **Hospital F., Bouchez A., Lecomete L., Causse M., Charcosset A.** (2002) 7th WCGALP, Montpellier, France, 22-05.
9. **Gibson G., Muse S.** (2002) A primer in genome science. Sinauer Ass. Sunderland, USA, p. 347.
10. **Goff S.A., Ricke D., Lan, Presting T.H., Wang G., Dunn R.M. et al.** (2002) Science, **296,** 92-100.
11. **Hack C., Kendall G.** (2005) Biochemistry and Molecular Biology Education, **33(2)**, 82-85.
12. **Hesslop-Harrison J.S.** (2000) Plant Cell, **12,** 617-636.
13. **Hide W., Miller R., Ptitsyn A., Kelso J., Gopallakrishnan C., Christoffels A.** (1999) EST Clustering Tutorial. ISMB, p. 24.
14. **Kearsey M.J.** (1998) Journal of Experimental Botany, **49(327)**, 1619-1623.
15. **Neerincx P., Leunissen J.** (2005) Briefings in Bioinformatics, **6(2)**, 178-188.
16. **Meyer K., Mewes H.W.** (2002) Curr. Opin. Plant Biol., **5**, 173-177.
17. **Mohammadi S.A., Prasanna B.M.** (2003) Crop Sci., **43**, 1235-1248.
18. **Morgante M., Salamini F.** (2003) Current Opinion in Biotechnology, **14**, 214-219.
19. **Needleman S.B., Wunsch C.D.** (1970) Journal of Molecular Biology, **48,** 443–453.
20. **Orr H.A.** (2005) Nature Review Genetics, **6,** 119-127.
21. **Quackenbush J.** (2001) Nucleic Acids Res., **29,** 159-164.
22. **Reif J.C., Melchinger A.A., Frisch M.** (2005) Crop Sci., **45**, 1-7.
23. **Rudd S.** (2003) Trends in Plant Science, **7,** 321-329.
24. **Rudd S.** (2004) Bioinformatics, Plant Genomes and Biosafety: can genomics help. In: Genomics for Biosafety and Plant Biotechnology, (J.P.H. Nap, A. Atanassov, W.J. Stiekema, Eds.), IOS Press, 61-76.
25. **Sen S., G. Churchill** (2001) Genetics, **159,** 371-387.
26. **Smith T.F., Waterman M.S.** (1981) Journal of Molecular Biology, **147**, 195–197.
27. **Walsh B**. (2001) Theor. Pop. Biology, **59**, 175-184.
28. **Ben Fry** (2004) MIT Media Laboratory.
29. **Stein L.** (2005) The Scientist, **19(10)**, 31-33.
30. **Nature Genetics** (2004) **4,** 332-333.