

APPLICATION OF BIOINFORMATICS IN FRUIT PLANT BREEDING

Dimitar Vassilev, Asen Nenov, Atanas Atanassov,
George Dimov and Lubomir Getov

AgroBioInstitute, Sofia, BULGARIA

(Received September 15, 2005/Accepted October 30, 2005)

A B S T R A C T

The goal of fruit plant genomics is to understand the genetic and molecular basis of all biological processes in plants that are relevant to the species. This understanding is fundamental to allow efficient exploitation of fruit plants as biological resources in the development of new cultivars of improved quality and reduced economic and environmental costs. This knowledge is also vital for the development of new diagnostic tools. Traits considered of primary interest are, resistance to pathogens and abiotic stress, fruit quality, and yield.

A genome program can now be envisioned as a highly important tool for fruit plant breeding. Identifying key genes and understanding their function will result in a "quantum leap" in fruit quality improvement. Additionally, the ability to examine gene expression will allow us to understand how fruit 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 fruit 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 international research community. The data include not only sequence information, but also information on mutations, markers, maps and functional discoveries. The key objectives for fruit plant bioinformatics include:

- encouraging submission of all sequence data into the public domain, through repositories;
- providing rational annotation of genes, proteins and phenotypes, and
- elaborating relationships both within the data on individual fruits and between fruits and other organisms.

Keywords: fruit plants, bioinformatics, genomics, ESTs, QTL, MAS

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 EST, TIGR – The Institute of Genomic Research, QTL – Quantitative Trait Loci, MAS – Marker Assisted Selection.

INTRODUCTION

Over the past few decades, major advances in molecular biology and 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, generated a need for computerized databases to store, organize and index the data, and for specialized tools to view and analyze the data.

With the publication of the complete *Arabidopsis thaliana* genome sequence and the draft sequence for rice genome, plant research and industry have entered the age of genomics (AGI, 2000; Goff et al., 2002). Numerous applications of genomic information have created many opportunities for integrating the rich rewards from sub-systems biology, integrative biology and large scale systematic functional genomics projects. With this accumulation of various types of data, the universe of “genomic understanding” is wide open. With this understanding, it is possible to model and design the amount and sense of changes in the level of gene expression, or how to localize proteins and assess their interactions with other genes and proteins, and finally how they affect the metabolite pools within any given tissue. To reach these goals will require a huge scientific undertaking, many aspects of which will undoubtedly rely on bioinformatics (Rudd, 2004).

Bearing 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 logical to consider that bioinformatics has become a crucial part of modern genomics research. Bioinformatics is thoroughly involved with the completion and assessment of a multitude of different complete genome sequences. 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 few years. Bioinformatics is practiced worldwide to access various databases and to exchange information for comparison, confirmation, storage and analysis. To date, there have been several databases on proteins from humans, animals, plants, bacteria, and other life forms (Gibson and Muse, 2002).

In biology and medicine, these databases help in developing new inventions which are useful to mankind. Bioinformatics allows life scientists invent new drugs and drug delivery systems, which makes for greater progress in the field of biotechnology. For the future development of biotechnology, bioinformatics will have to take advantage of the internet and the World Wide Web (WWW). Future rDNA research should be guided largely by the databases available for generic or specific forms. Bioinformatics and biotechnology have to move hand in hand to advance. However, bioinformatics can now be considered as a *bona fide* discipline within information technology (Baxevanis and Oullette, 2001).

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of bioinformatics is to make possible new biological insights and create a global perspective on the unifying principles in biology. At the beginning of

the "genomic revolution", the task of bioinformatics was to create and maintain databases to store biological information, such as nucleotide and amino acid sequences. Database development involved not only design issues but the development of complex interfaces whereby researchers could both access existing data and submit new or revised data (Hack and Kendall, 2005).

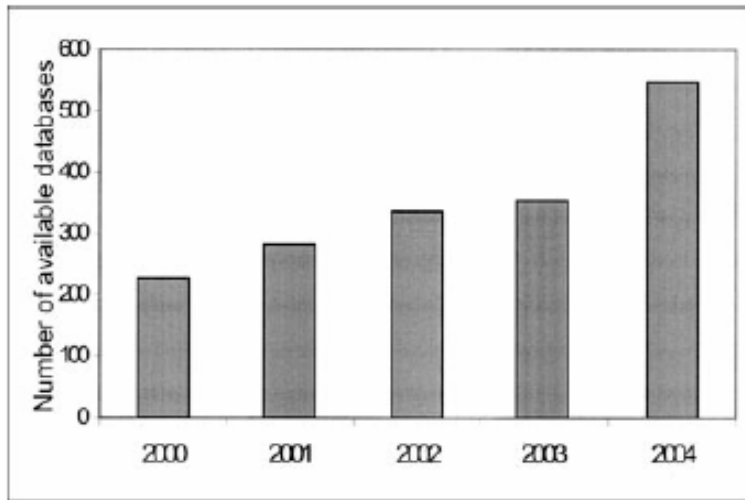


Figure 1. Growth in number of databases listed in the Molecular Biology Database Collection [2-6]

Table 1. Classification of databases in the 2004 edition of the Molecular Biology Database Collection (Hack and Kendall, 2005)

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

Therefore, the most urgent task for bioinformatics today has become the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analyzing and interpreting data applies to:

- the development and implementation of tools which enable efficient access to, and use and management of, various types of information, and
- 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 or function, and cluster protein sequences into families of related sequences (Baxevanis and Oullette, 2001; Davenport et al., 2004).

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 an environment for the vast biodiversity on earth. For centuries, humans have selected plant varieties that best fit their purposes and developed fruit plants that have many advantages compared to natural, wild plants in terms of quality, quantity and farming practices. However, multifactorial traits involved in resistance and quality have proven to be extremely difficult to improve, certainly in combination. The revolution in life sciences brought on by genomics dramatically increases the scale and scope of our experimental enquiry and applications in fruit plant breeding. The scale and high resolution power of genomics makes possible a broad and detailed genetic understanding of plant performance at multiple levels of aggregation. The complex biological processes that determine pathogen resistance and crop quality are now open for systematic functional analysis. In plant bioinformatics, these analyses are made with the help of special software on huge amounts of data in databases (Rudd, 2004; Meyer and Mewes, 2002).

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. It is often much easier to study a particular aspect in one organism than in others. Those organisms which are intensively studied are commonly called model organisms. Each has one or more characteristics that make it suitable for laboratory study. The most popular model organisms have great advantages for experimental research, such as rapid development with short life cycles, small adult size, ready availability, and tractability. They 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 gene regulation, genetic diseases, and evolutionary processes in humans and crop plants alike (Rudd, 2003)

Arabidopsis thaliana is a small flowering plant which belongs to the *Brassica* family, which includes species such as broccoli, cauliflower,

cabbage, and radishes. Because *Arabidopsis* has a smaller genome than other plants and is easily grown in the laboratory, it has become the organism of choice for basic studies on molecular genetics in flowering plants (AGI, 2000). Scientists expect systematic studies on *Arabidopsis* will facilitate basic research in genetics and molecular biology and will elucidate many questions in plant biology, including some of significant value to agriculture, energy, the environment, and human health.

In the 1980s, there was a growing awareness that investing heavily in studies on different plants such as corn, oilseed rape, and soybean were hampering efforts to fully understand basic properties in all plants. Scientists began to realize 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 can then study. Fortunately, because all flowering plants are closely related, the complete sequencing of all the genes in a single representative plant species will provide a lot of knowledge about all higher plants. Similarly, discovering the functions of the proteins produced by a model species will provide a lot of information on protein function in all higher plants.

Comparing genome sequences

The development of techniques for large-scale quantification and identification of biological molecules, together with combined advances in computer technology and the internet have made available large volumes of biological data that scientists can access from their desktops. 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 (Hesslop-Harrison, 2000.). The increase was achieved by automation, miniaturization, and integration of technologies. Applying this approach to other biological molecules including mRNA, proteins, and metabolites has massively accelerated the accumulation of biological data. This data has been made readily 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 a steady growth in the quality and size of the databases (Fig. 1). The 2004 edition contains 548 databases classified into eleven categories (Tab. 1).

A major aim of most genome projects is to determine the DNA sequence either of the genome or of a larger number of transcripts. This both leads to the identification of all or most genes and to the characterization of various structural features of the genome. Very often a common bioinformatics strategy 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: completeness, accuracy of the base sequence, and validity of the assembly.

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

Sequence alignment methods and applications

Sequence alignment is the arrangement of two or more amino acid or nucleotide sequences from one or more organisms so that the sequences sharing common properties are aligned. 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 an indicator of the genetic relatedness between the organisms (Baxevanis and Oullette, 2001).

Sequence Similarity Searching Algorithms. Smith-Waterman is an algorithm for local sequence alignment, using two sequences as input (Smith and Waterman, 1981). The difference between NCBI BLAST (also local alignment algorithm) and Smith-Waterman is that a) BLAST searches for a sequence throughout a database of sequences; and b) BLAST statistically calculates the most probable match, and Smith-Waterman is calculates the exact match.

Genome Comparison Tools. MegaBlast is an algorithm based on NCBI BLAST for large sequence similarity search (Hesslop-Harrison, 2000.). MegaBlast implements a greedy algorithm for the DNA sequence gapped alignment search. MegaBlast is used to compare 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 the draft-quality sequence or masked in the 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 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 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 or are being generated in several laboratories using different species and cultivars as well as various tissues at different developmental stages. This facilitates the identification of all of the genes expressed in a particular organism, such as the grape and some members of the *Rosaceae* family, including apples (*Malus*), strawberries (*Fragaria*), raspberries and blackberries (*Rubus*), peaches almonds and other stone fruits (*Prunus*).

The benefits arising from the rapid generation of large numbers of low-quality cDNA sequences were not universally recognized when the concept was originally proposed in the late 1980s (Baxevanis and Ouellette, 2001). Proponents of this approach argued that these cDNA sequences would allow for the quick discovery of hundreds or thousands of novel protein coding genes. Critics countered that cDNA sequencing would miss important regulatory elements that could be found only in genomic DNA. The cDNA sequencing advocates appear to have won. Since the original description of 609 Expressed Sequence Tags (ESTs) in 1991, the growth of ESTs in public databases has been dramatic. In mid-1995, the number of ESTs in GenBank surpassed the number of non-EST records in mid-1995. In June 2000, 4.6 million EST records comprised 62% of the sequences in GenBank. Although the original ESTs were of human origin, NCBI's EST database (dbEST) now contains ESTs from over 250 organisms. In addition, several commercial establishments maintain privately funded, in-house collections of ESTs. Throughout the genomics and molecular biology communities, ESTs are now widely used for gene discovery, mapping, polymorphism analysis, expression studies, and gene prediction.

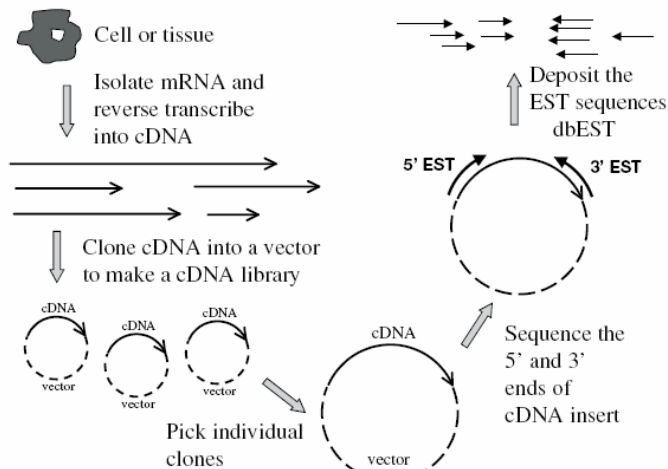


Figure 2. EST construction pipeline (*Isolation, Sequencing, Clustering, Assembly and Alignment*) (Baxevanis and Ouellette, 2001)

EST sequences are also an important resource for identifying single nucleotide polymorphisms, localizing and isolating gene sequences, and for producing cDNA microarrays for expression profile analyses. EST sequencing efforts will be greatly improved by sharing the information held by different laboratories and designing strategies to avoid duplication and extend the coverage of all expressed genes (Hide et al., 1999).

Expressed sequence tags (ESTs) can be used to discover new genes, map the genome, and identify 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. In a database like this, ESTs should be partitioned into clusters such that ESTs from each gene are put together in a distinct cluster. A further complication arises because DNA is a double stranded molecule and a gene could be part of either strand (Rudd, 2003).

DbEST 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) defines TC as Tentative Consensi (assemblies from ESTs) and ET as Expressed Transcripts (both non-human) when building TIGR Gene Indices (TGI).

dbEST release 040805

Number of public entries: 26,605,325

Malus x domestica	183 916
Vitis vinifera	147 300
Prunus armeniaca	15 081
Citrus x paradisi x Poncirus trifoliata	8 002
Vitis hybrid cultivar	6 533
Fragaria x ananassa	5 322
Prunus dulcis	3 864
Citrus reticulata	3 735
Citrus unshiu	2 561
Ananas comosus	1 547
Fragaria vesca	1 306
Citrullus lanatus	693
Citrus clementina x Citrus reticulata	74
Vitis cinerea x Vitis rupestris	61
Cucumis melo	60

Figure 3. Number of ESTs by fruit collected in dbEST (*release 040805*)

TIGR Gene Indices

The TIGR Gene Indices represent another effort to consolidate EST and other annotated gene sequences (Quackenbush, 2001). A significant difference between the Gene Indices and UniGene is that the Gene Indices are assemblies of ESTs and other gene sequences rather than clusters. The assemblies tend to represent

one transcript, so alternatively spliced products are grouped separately. Furthermore, the process generates a single consensus sequence per assembly. A Gene Index is maintained for fourteen organisms, including man, the mouse, the rat, *Drosophila*, the zebrafish, *Arabidopsis*, and several crop plants, including the grape. Gene Indices are created from publicly available GenBank and dbEST sequences by clustering ESTs with the DNA sequences encoding the coding sequences annotated on DNA and mRNA sequences.

ET sequences are extracted from appropriate divisions of GenBank and participate in the clustering and assembly process along with the cleaned ESTs. ESTs and ETs are compared and clustered together if they meet the following criteria: a minimum of forty base pairs match; identity in the overlap region is greater than 94%; and a maximum unmatched overhang of thirty base pairs. These clusters are then assembled into Tentative Consensus (TC) sequences. All sequences that do not belong to an EST cluster are called singletons, and they are used in analysis in rare cases.

UniGene is public domain transcriptome database that links ESTs in a cluster if the sequences have a fifty base pair overlap in the 3' untranslated region (3' UTR) with 100% identity. These clusters are not run through the more stringent assembly process and consensus sequences are not made. For this reason, several TIGR THCs are often contained within one UniGene cluster.

Integrated web resources for fruit plant genomes

TIGR Grape gene index




TIGR Grape Gene Index	
Search VvGI	
About	
Development and Goals	background information about VvGI
Release Summary	display a statistical summary of all VvGI releases
Category Comparison	display estimated number of genes among all plant releases
Sequence homology search	
BLAST	search TC sequences based on sequence similarity
VvGI sequence reports	
Identifiers or Keywords	search TC reports using TC identifiers, GB accessions or keywords
TC Annotator	list all TC annotation
EST Annotator	list all EST annotation
Libraries	search EST libraries by keywords or tissue origins
CAT# download	download EST and TC sequences originating from one library
Functional annotation and analyses	
EST Expression	compare EST expression between different libraries or tissues
Gene Ontology	classification of TCs by GO vocabularies
Metabolic Pathways	association of TCs with metabolic and signaling pathways
Oligomer Prediction	list all 70-mer oligo predictions
Attributions	
A significant number of ESTs used to construct this index were generated by:	
	
Department of Biochemistry University of Nevada	CAES Genome Facility

Figure 4. Grape ESTs/ETs collected in TIGR Grape Gene Index – *Factsheet*

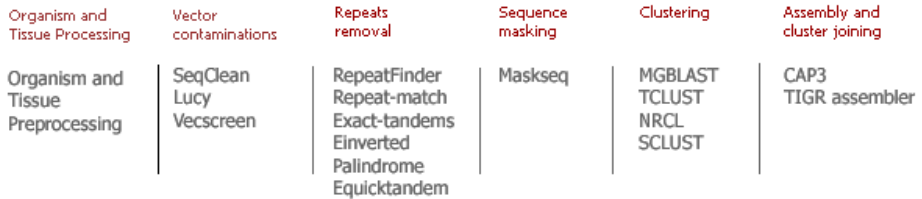








Figure 5. Sample EST Clustering and Assembly Pipeline


Summary of total unique sequences

	Number of unique sequences	TCs	Singleton ESTs	Singleton ETs	Total
Release 4.0 September 21, 2004		13,571	10,254	46	23,871
Release 3.1 November 13, 2003		13,218	9,837	54	23,109
Release 3.0 August 18, 2003		11,317	8,083	51	19,451
Release 2.0 May 7, 2003		9,571	7,469	51	17,091
Release 1.0 January 29, 2003		4,012	4,524	63	8,599

Scale:

 represents 10,000 sequences

Key:

 number of TCs



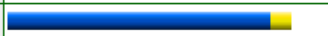



 number of singleton ESTs and ETs

Figure 6. TIGR Grape Index: Summary of total unique sequences


Distributions of input EST and ET sequences

		Distribution of input sequences	in TCs	in singletons	Total
Release 4.0 September 21, 2004	ESTs		129,126	10,254	139,380
	ETs		486	46	532
Release 3.1 November 13, 2003	ESTs		122,479	9,837	132,316
	ETs		194	54	248
Release 3.0 August 18, 2003	ESTs		100,823	8,083	108,906
	ETs		184	51	235
Release 2.0 May 7, 2003	ESTs		76,206	7,469	83,675
	ETs		169	51	220
Release 1.0 January 29, 2003	ESTs		25,843	4,524	30,367
	ETs		139	63	202

Scale:

 represents 100,000 sequences

Key:

 number of sequences found in TCs


 number of sequences found in singletons

Figure 7. TIGR Grape Index: Distributions of input EST/ET sequences

Distribution of TC Size of the current Release




	Number of TCs	Number of TCs
< 1KB		9,482
1KB <= size < 2KB		3,924
2KB <= size < 3KB		142
3KB <= size < 4KB		10
4KB <= size < 5KB		6
>= 5KB		7

Figure 8. TIGR Grape Index: Distributions of TC size of the current release (April 2005)

Fruit Transcriptome Based Clustering. Taxonomy ID:3750 in NCBI UniGene Database shows the known genes of *Malus x domestica* from GenBank, ESTs from dbEST, and alignments between all transcript sequences. UniGene clustering proceeds in several stages, with each stage adding less reliable data to the results of the preceding stage. This staged clustering affords greater control than a more egalitarian treatment of all links between sequences.

There is a range of contemporary genetic marker types and all have been exploited using attributes of EST data. Simple sequence repeats have been identified from the genome data and have applications in genotyping. Single nucleotide polymorphism (SNP) markers have been selected from various EST collections on the basis of available quality scores and, more recently, SNPs have been predicted and validated from various fruits by screening for conserved patterns of polymorphism within EST sequence clusters.

Genome Database for *Rosaceae* (GDR) is accurate and integrated web-based relational database. GDR contains comprehensive data of the genetically anchored physical map of the peach, an annotated peach EST database, *Rosaceae* maps and markers, and all publicly available *Rosaceae* sequences. Annotations of ESTs include contig assembly, putative function, simple sequence repeats, and anchored position to the peach physical map if applicable (Sook Jung et al., 2004).

The GDR has been initiated to meet the major deficiency in *Rosaceae* genomics and genetics research, namely a centralized web database and bioinformatics tools for data storage, analysis and exchange. GDR can be accessed at <http://www.genome.clemson.edu/gdr/>.

Molecular information and fruit 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 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, and 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 (Walsh, 2001; Dekkers and Hospital, 2002).

The vast 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 (Hospital et al., 2002). For instance, clearly visible phenotypic traits that are poorly understood at the biochemical level can be associated by high resolution mapping with candidate genes. Orthologous genes in a model species, such as *Arabidopsis* or rice, may not yet be associated 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 expected dramatic improvements in phenotypes of commercial interest 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:

1. abiotic stress tolerance (cold, drought and salt);
2. biotic stress tolerance (fungi, bacteria, viruses, chewing and sucking insects);
3. nutrient use efficiency;
4. manipulation of plant architecture and development (size, organ shape, number, and position, timing of development and senescence);
5. metabolite partitioning (redirecting of carbon flow among existing pathways, or shunting into new pathways).

Rational plant improvement

The implications of genomics for food, feed and fiber production can be envisioned on many levels. At the most fundamental level, advances in

genomics will greatly accelerate the acquisition of knowledge and that, in turn, will directly affect many aspects of 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 modeled from first principles.

Genotype building experiments

Biodiversity determined by the fruit plant genome analysis. In the last few years, an increasing amount of information on DNA polymorphism and sequencing has been accumulated for different plant varieties and cultivars. Most of this information was used for the recognition of different cultivars and for comparing the similarities and differences between them (Reif et al., 2005). These distances are measured by the polymorphism on a part of the chromosome whose function is unknown. This type of polymorphism is widely used in 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 an indirect marker (Kearsey, 1998).

Indirect markers are closely linked and sometimes overlap with the locus which determines the quantitative trait (QTL). QTLs are defined as genes or regions of chromosomes which affect a trait. QTLs by themselves are difficult to recognize. In both cases, these markers, can be used for further selection. This selection process is called MAS (Morgante and Salamini, 2003).

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 mostly on markers, a higher mapping density is important. An 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 (Sen and Churchill, 2001).

Some advantages of QTL/MAS selection approach come from:

1. measurement of the marker/QTL in early stages of development;
2. low heritability of the trait;
3. for animals – sex limited or measured after slaughtering – meat quality; for plants – malting quality, etc.

How QTL information could be of use?

4. it is assumed that some but not all loci are identified, so selection should be based on the combination of phenotypic and molecular information;

5. in the process of selection the link of markers and traits could decrease so this link should be observed throughout the generations;
6. in the selection process, QTLs prove the simultaneous existence of the desired genes in a line;
7. 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;
8. future prospective: with accumulation of molecular data, genotype building programs will be developed which will set homozygous desirable markers;
9. in introgression programs for combining the desirable traits from two lines in one;
10. finally, the real world of agriculture is at the stage of accumulation of molecular data.

Analytical approaches. One of the statistical tools for performing the QTL analyses is meta-analysis, which synthesizes 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. Further development and detailed discussion on QTLs includes statistical aspects of MAS, setting up the threshold of significance of marker effects, overestimation or bias in estimation of QTL effects, and 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 specific plant parts such as leaves, roots and fruits, as was proved for grapes (Morgante and Salamini, 2003).

Efficiency of QTLs

1. Traits of interest

Experimental results do not always confirm the efficiency of MAS over genotype building. The main reason is the insufficient precision of the initial assessment of a QTL, its location, and its effect. Some QTLs also could be lost in the GB process. For complex productivity traits, the epistatic lost would cause changes in the magnitude of the QTL effect in the parent and progeny generation. It is thus recommended that selection be based on allelic combinations rather on separate QTLs. This is in line with numerous GxE interactions and with selection within the environment of interest for disease or drought resistance.

Consequently, efficiency of MAS will depend on the complexity of the species/trait genetic architecture, on the development of the trait in the environment, and on the interactions between them For complex traits, QTLs

should be evaluated in different environments. Phenotypic evaluation over consecutive generations is also necessary. Drought resistance seemed to be a more complex trait than disease resistance.

2. Economics

From an economical point of view, the use of markers will be expensive in terms of DNA collection, genotyping, analyses, detection of QTLs, etc. This high price is paid for the genotype building (there is no other way of doing that) and for traits that are expensive to evaluate, such as disease resistance and traits with low heritability.

Species and traits of interest for MAS

Barley: disease resistance, malting quality;

Maize: drought tolerance, earliness, yield;

Rice: disease resistance;

Tomatoes: pest resistance, organoleptic qualities;

Apples (cultivar 'Galaxy'): clones resistant to fungal diseases; W100; W101

Peaches: results available in the Genome Database for *Rosaceae* (GDR), at <http://www.genome.clemson.edu/gdr/>

Sustainable fruit production and pomology: a knowledge-based approach

Sustainable production is related to obtaining optimum productivity in terms of yield and fruit quality. Two points are of interest:

1. knowledge of the factors which influence productivity;
2. management of the factors to obtain the necessary productivity.

The knowledge is based mostly on accumulation of the data from empirical observations and field experiments, proper planning, and analyses. The aim is to test as many of the factors which might influence important traits. The analyses would reveal the magnitude of different factors and their possible interaction. These factors can be manageable with agrotechnology in all its complexity, or unmanageable because they are random environmental factors, such as weather. Most variation in productivity is caused by these two types of factors.

Specific productivity depends mostly on the characteristics of the cultivar. In that sense, the sustainable production is related to the best fit of the manageable elements of agrotechnology to the specific requirements of the cultivars and to the creation of cultivars which are genetically less sensitive to weather conditions.

Studies on the sustainable production of orchard species include broad scale experiments, both on-going and already finished. They may be based on techniques such as *in vitro* culture, rooting, and grafting. Development and productivity of individual trees should be observed over their lifetime.

The information to be collected might include all possible observable traits, such as tree development, leaf morphology, branch morphology, growth, flowering, fruit quality, flavor, storability, transportability and resistance to disease and extreme environmental conditions. To complement the studies on DNA polymorphism and sequencing, further QTL analyses are going to be performed on different varieties and cultivars. The observed measurement data should be analyzed for each tree separately. Further analyses will be performed using different schemes in order to reveal possible important influences. These schemes will depend on the traits of interest because some of them cannot be measured individually.

The complex of results which is obtained for the influence of different factors on a given trait, on the similarity of influence across the traits, on the link between traits and on the cultivar specificity of these influences are the knowledge base on which proper management of the pomological production system is based. Results of the analyses of this information would give a possibility for qualifying and quantifying the magnitude elements of separate factors. The information obtained for the cultivars of interest could be transferred in a knowledge based system for future planning of the desired productivity.

Finally, who needs these results? On one hand, the farmers who are interested in choosing the best cultivars for their specific farming conditions: soil, climate, market conditions, production skills, etc. On the other hand, future selection should be based on information of interest: traits, factors, and influences. Finally, experience in testing could help in designing a proper system for comparing pomological species and cultivars in the future.

Proposed collaboration on the knowledge based systems in pomology – specialists on cultivars, molecular biologists and bioinformaticians should collaborate to build a knowledge based system to support decisions in pomological research.

REFERENCES

- AGI 2000. Analysis of genome sequence of the flowering plant *Arabidopsis thaliana*. NATURE 408: 796-815.
- Baxevanis A., Ouellette F. 2001. Bioinformatics: A practical Guide to the Analysis of Genes and Proteins. John Willey & Sons, Inc. NY., USA. 518p.
- Davenport G., Ellis N., Ambrose M., Dicks J. 2004. Using bioinformatics to analyse germplasm collections. EUPHYTICA 137: 39-54.
- Deckers J., Hospital F. 2002. The use of molecular genetics in the improvement of the agricultural populations. NATURE REV. GENET. 3: 22-32.
- Gibson G., Muse S. 2002. A primer in genome science. Sinauer Ass. Sunderland, USA, 347p.
- Goff S.A., Ricke D., Lan T.H., Presting G., Wang R., Dunn M. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*) SCIENCE 296: 92-100.

- Hack C., Kendall G. 2005. Bioinformatics: current practice and future challenges for life science education. *BIOCHEM. MOLECULAR BIOL. EDUC.* 33: 82-85.
- Hesslop-Harrison J.S. 2000. Comparative genome organization in plants: from sequence and markers to chromatin and chromosomes. *PLANT CELL*: 12: 617-636.
- Hide W., Miller R., Ptitsyn A., Kelso J., Gopallakrishnan C., Christoffels A. 1999. EST clustering tutorial. *ISMB*, 24 p.
- Hospital F., Bouchez A., Lecomete L., Causse M., Charcosset A. 2002. Use of markers in plant breeding: Lessons from genotype building experiments. 7th WCGALP, Montpellier, France, pp. 22-05.
- Kearsey M.J. 1998. The principles of QTL analysis (a minimal mathematics approach). *J. EXPER. BOT.* 49: 1619-1623.
- Meyer K., Mewes H.W. 2002. How we can deliver the large plant genomes ? Strategies and perspectives. *CURR.. OPIN. PLANT BIOL.* 5: 173-177.
- Morgante M., Salamini F. 2003. From plant genomics to breeding practice. *CURR. OPIN. BIOTECH.* 14: 214-219.
- Quackenbush J. 2001. The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *NUCLEIC ACIDS RES.* 29: 159-164.
- Reif J.C., Melchinger A.A., Frisch M. 2005. Genetics and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *CROP SCI.* 45: 1-7.
- Rudd S. 2003. Expressed sequence tags: alternative or complement to whole genome sequences ? *TRENDA IN PLANT SCI.* 7: 321-329.
- Rudd S. 2004. Bioinformatics, plant genomes and biosafety: can genomics help. In: J.P.H.Nap, Atanassov A., Stiekema W.J. (eds), *Genomics for Biosafety and Plant Biotechnology*. IOS Press, pp. 61-76.
- Sen S., Churchill G. 2001. A statistical framework for quantitative trait mapping. *GENETICS* 159: 371-387.
- Smith T.F., Waterman M.S. 1981. Identification of common molecular subsequences. *J. MOLECULAR BIOL.* 147: 195-197.
- Sook Jung C., Jesudurai M. Staton, Zhidian Du, Ficklin, Ilhyung Cho S., Abbott A., Tomkins J., Main D. 2004. GDR (Genome Database for Rosaceae): integrated web resources for Rosaceae genomics and genetics research. *BMC BIOINFORMATICS* 5: 130.
- Walsh B. 2001. Quantitative genetics in the age of genomics. *THEOR. POP. BIOLOGY* 59: 175-184.

ZASTOSOWANIE BIOINFORMATYKI W HODOWLI ROŚLIN SADOWNICZYCH

Dimitar Vassilev, Asen Nenov, Atanas Atanassov,
George Dimov i Lubomir Getov

S T R E S Z C Z E N I E

Współczesna hodowla wytwarza ogromną liczbę danych zarówno fenotypowych, obejmujących wyniki hodowli klasycznej jak i tych związanych z analizą genomu, które pozwalają na poznanie genetycznych i molekularnych podstaw procesów biologicznych i znajdują bezpośrednie zastosowanie w praktyce ogrodniczej. Bioinformatyka umożliwia skorelowanie tych danych i opracowanie zależności między poszczególnymi informacjami a rzeczywistymi cechami roślin i zachodzącymi w nich procesami. Znaczącą rolę mają w tym dane charakteryzujące rośliny modelowe, metody porównywanie sekwencji pochodzących ze sztucznych chromosomów bakterii i drożdży czy sekwencji ekspresyjnych EST, ustalanie homologii sekwencji, metody gromadzenia wyników do oceny cech ilościowych (QTL) i mapowania genomów. Autorzy prezentują także odnośniki www. do baz danych dotyczących tych zagadnień. Zbudowanie pełnego systemu bazy danych dla potrzeb nauki i praktyki ogrodniczej wymaga współpracy specjalistów z dziedziny pomologii, biologii molekularnej i bioinformatyki.

Słowa kluczowe: rośliny sadownicze, bioinformatyka, genom, sekwencje, EST, MAS, QTL