Genome Analysis of Species of Agricultural Interest

Maria Luisa Chiusano, Nunzio D'Agostino, Amalia Barone, Domenico Carputo, and Luigi Frusciante

Abstract In recent years, the role of bioinformatics in supporting structural and functional genomics and the analysis of the molecules that are expressed in a cell has become fundamental for data management, interpretation, and modeling. This interdisciplinary research area provides methods that aim not only to detect and to extract information from a massive quantity of data but also to predict the structure and function of biomolecules and to model biological systems of small and medium complexity. Although bioinformatics provides a major support for experimental practice, it mainly plays a complementary role in scientific research. Indeed, bioinformatics methods are typically appropriate for large-scale analyses and cannot be replaced with experimental approaches. Specialized databases, semiautomated analyses, and data mining methods are powerful tools in performing large-scale analyses aiming to (i) obtain comprehensive collections; (ii) manage, classify, and explore the data as a whole; and (iii) derive novel features, properties, and relationships. Such methods are thus suitable for providing novel views and supporting in-depth understanding of biological system behavior and designing reliable models.

The success of bioinformatics approaches is directly dependent on the efficiency of data integration and on the value-added information that it produces. This is, in turn, determined by the diversity of data sources and by the quality of the annotation they are endowed with. To fulfill these requirements, we designed the computational platform ISOLA, in the framework of the International Solanaceae Genomics Project. ISOLA is an Italian genomics resource dedicated to the Solanaceae family and was conceived to collect data produced by 'omics' technologies. Its main features and tools are presented and discussed as an example of how to convert experimental data into biological information that in turn is the basis for modeling biological systems.

M.L. Chiusano
Department of Soil, Plant, Environmental and Animal Production Sciences,
University of Naples Federico II, Naples, Italy
e-mail: chiusano@unina.it

P.J. Papajorgji, P.M. Pardalos (eds.), *Advances in Modeling Agricultural Systems*,
DOI 10.1007/978-0-387-75181-8_18,

1 Introduction

Bioinformatics is a discipline in which biology, computer science, statistics, applied mathematics, and information technology merge. The use of computers to study biological systems encompasses the design of methods and the implementation of algorithmic tools to facilitate the collection, organization, and analysis of large amounts of data. Most molecular data are in the form of a succession of letters that summarizes a basic level of the structural organization of a DNA, mRNA, or a protein molecule, commonly defined *primary structure*. This symbolic linear depiction is referred to as a *biomolecular sequence*. In this mold, molecular data can be easily edited, manipulated, compared, and analyzed. Computational methods proved early on to be especially suitable for many of these basic approaches. The growing availability of 'sequence data' combined with the 'high throughput' technologies makes bioinformatics essential in supporting the analysis of the structure and function of biological molecules. To this end, bioinformatics plays a key role in data mining and has broad applications in the molecular characterization of an organism's gene and protein space (i.e., genomics and proteomics), in the genome-wide study of mRNA expression (transcriptomics), the systematic study of the chemical fingerprints that specific cellular processes leave behind (metabolomics), drug discovery, and in the identification of biomarkers as biological indicators of disease, toxicity of pathogens, or effectiveness of healing. Nevertheless, the substantial improvement in technologies and data processing tools also drives the generation of 'digital data' and metadata that can be investigated by large-scale analytical strategies.

Despite the advances in experimental strategies and computational approaches that support the characterization of genomes and their products, only 10% of genome organization and functionality is today understood. This means that we are still far from achieving the ambitious goal of the *in silico* simulation of complex living systems. Indeed, this would require deeper knowledge of the structure and functions of an organism's genome, transcriptome, and proteome and of their links with cellular physiology and pathophysiology. The real challenge of bioinformatics is to design suitable computational methods for revealing the information that biological data still hide and to integrate the large amount of the so-called 'omics' data to approach a systems biology view. The long-term goal is the creation of models to simulate biological system behaviors and their exploitation for applications in medicine, biotechnologies, and agricultural sciences.

This chapter presents an overview of the importance of 'omics' approaches in agriculture and plant science. Bioinformatics strategies for exploiting 'omics' data follows. The discussion is mainly focused on the state of the art of genome data analysis and on the methods employed for gene discovery and for comprehension of gene functionalities. The purpose is to highlight the strong connection between bioinformatics and the need for models that describe genome functionalities of living organisms.

Finally, the chapter presents our efforts to organize and integrate 'omics' data to model genetic expression maps and study regulatory networks and mechanisms of genetic control. As partners of the consortium established by the International Solanaceae (SOL) Genomics Network (http://www.sgn.cornell.edu/solanaceae-project/), the authors designed and currently maintain an **I**talian **Sola**naceae genomics resource (ISOLA), which was conceived as a multilevel computational environment based on effective data integration using comparative and evolutionary approaches. The computational platform described as a sample application, together with the multiple aspects of data modeling it covers, may represent a reference for analysis and modeling of molecular mechanisms in species of agricultural interest. The strategy of *moving from the 'omics' to systems biology* is one of the possible pathways for tackling the incipient challenges in bioinformatics.

2 Genome Analysis and Applications in Agriculture

Genome analysis involves a wide variety of studies and is aimed at understanding the structure, function, and evolution of all the hereditary information of a given organism. Knowledge of the genetic make-up of living organisms could dramatically and effectively change our approach to agriculture. Species with large amounts of DNA make genetic studies and manipulations difficult. By contrast, species with a small DNA content can be more easily analyzed. Unsurprisingly, the latter species were the first candidates for genome sequencing projects. Because of major differences among living organisms in terms of fundamental characteristics, their genes must encode different enzymes that are necessary for their metabolic processes. Therefore, a wide range of variation among living organisms is found in terms of DNA quantity, karyotype (i.e., number, type, shape, etc., of chromosomes), and DNA sequence. As reported by Cullis [1], genomes show amazing diversity in terms of shapes (chromosomes) and sizes. Nuclear DNA content has been generally referred to as the amount of DNA contained in the nucleus of a gamete (the 1C value), without any regard to the actual ploidy of the gamete-producing individuals. The largest range of nuclear DNA content occurs in the plant kingdom, even between closely related species. In plants, one of the smallest genomes belongs to *Arabidopsis thaliana* (0.125 picogram), which is about 200 times smaller than other plant genomes (www.rbgkew.org.uk/cval/homepage.html). The largest genome reported to date belongs to *Fritillaria assyriaca*, with roughly 127 pg. This represents a 1000-fold difference between the largest and smallest genomes characterized to date in the plant kingdom. Living organisms show high variability also in terms of somatic (2n) chromosome numbers (e.g., among animals, *Homo sapiens* 2n = 46, *Felix domesticus* 2n = 38, *Drosophila melanogaster* 2n = 8; among plants, *Solanum lycopersicum* 2n = 24, *Arabidopsis thaliana* 2n = 10, *Prunus persica* 2n = 16). A relatively common phenomenon in the plant kingdom is the occurrence of polyploids. These have more than

two sets of the basic chromosome number x, and species in which x is a multiple of 6 or 7 are frequent [2]. Additional differences among living organisms are given by their karyotypes. Most of the species possess a single centromere for each chromosome. However, some organisms have polycentric or holocentric chromosomes, with multiple or diffused centromeres.

At the molecular level, the large-scale organization of the chromosomes relies on repetitive DNA sequences. Early studies on the kinetics of the DNA reassociation [3] already indicated that the genome of higher eukaryotes consisted of large amounts of highly repeated DNA, whereas a small portion of the genomic sequence was present as unique or very low copy. Nowadays, most genes found in species with much larger genomes are also present in species with smaller genomes. The smallness of some species genomes is due to the lower abundance of highly repeated sequences compared with that of other species [2]. The proportion of the genome taken up by repetitive DNA varies widely and may represent more than 90% of the whole genome [4]. The sequences in a genome are generally classified with respect to the number of times they are represented. Three main classes can be identified: unique sequences (which probably represent the genes), moderately repetitive sequences, and highly repetitive sequences.

Technologies for acquiring data on genomes have progressed greatly in the past few years, and a large amount of cytological and molecular information has been generated. Particularly attractive is the availability of the complete DNA sequence of hundreds of viruses and bacteria and a few eukaryote organisms, including mammals and flowering plants. Also important are advances that are being made in understanding gene and protein structure, expression, and function. These developments are strongly accelerating the speed at which scientists can exploit the knowledge gained in understanding the biochemical and molecular basis of metabolic, physiological, developmental, and reproductive traits of interest. This can result in several practical applications to challenging problems in agriculture [5]. The generation of high-quality molecular maps, for example, can allow better characterization and manipulation of genes underlying complex noteworthy traits, such as those associated with stress tolerance and yield. In addition, it will make marker-assisted selection possible to a greater extent both for animal and plant breeding efforts [6–9]. Advances in knowledge may also provide novel strategies to stabilize agricultural yields in concert with the environment (e.g., through the development of plants that tolerate drought and pathogens). They will also contribute to the introduction of improved and novel crops/animals that may create new economies based on agricultural products and may increase the diversity of product packaging available to consumers. On this point, the creation of animals that produce biomedically useful proteins in their blood or milk is an attractive perspective. Of equal interest is the development of plants containing more essential macro- and micronutrients and the exploitation of the thousands of secondary metabolites that higher plants synthesize.

In addition, solutions to challenges related to environmental management and energy can be met through the exploitation of genomic knowledge and the application of novel technologies. Reduction in greenhouse gases achievable by the production of plant biofuels and rehabilitation of chemically contaminated sites by phytoremediation are two examples of these synergic effects.

As a whole, the major challenges facing mankind in the 21st century are the need for increased food and fiber production, a cleaner environment, and renewable chemical and energy resources. Plant-based technologies and a deeper understanding of numerous fundamental aspects of plant biology can play a major role in meeting each of these challenges, allowing plants to be used for biomass, chemical feedstocks, and production of biodegradable materials.

Indeed, information gained from genomics (mapping, sequencing, and understanding gene function) is being used to improve traits through genetic engineering and new breeding strategies. Genomics approaches are revolutionizing biology as they affect our ability to answer questions concerning the whole genome, which cannot be answered using a gene-by-gene approach. This could be the case of complex traits such as fruit development in agronomically important species like tomato and pepper, or tuber development in potato, which are controlled by many genes, each exhibiting a function not easily determined. In higher plants, sugars are known to be produced through photosynthesis in leaves and are then transported to other organs (fruits, tubers, seeds, roots) where they are metabolized or stored. This source-sink balance is of basic importance for human food and nutrition. Therefore, the understanding of its genetic determination is fundamental for crop yield and quality.

Besides the comprehension of basic biological processes, the genomics revolution allowed marker-assisted breeding to evolve into genomics-assisted breeding for crop improvement [10]. Indeed, the increasing amount of sequencing information today available for many agriculturally important species and the research that successfully unraveled metabolic pathways have allowed a huge number of new molecular markers for agronomic traits to be discovered and used by breeders. Genomic characterization and germplasm phenotyping are becoming fundamental tools for increasing crop genetic diversity, thus allowing a broad range of genes and genotypes for important traits to be identified. An extremely important advantage deriving from available molecular markers clearly distinguishing genotypes is the realization of genetic traceability of food products. Public questions over food safety are creating demand for food tracking and traceability systems to ensure it, where traceability is defined as the ability to follow and document the origin and history of a food or feed product. DNA-based traceability using Single Nucleotide Polymorphism (SNP) markers is applicable to every organism for which genetic variation is known, thus leading DNA to become nature's bar code to trace products from the consumer to the farm of origin. Moreover, genomics tools could increase the ability to guarantee and protect all agricultural products and their derivatives (milk, meat, vegetable, wine, cheese) that can be better distinguishable and characterized by PDO (Protected Designation of Origin), PGI (Protected Geographical Indication),

TSG (Traditional Speciality Guaranteed) certifications. A wide discussion is still open on the development of breeding and molecular strategies that can efficiently exploit the available genomic resources and genomics research for crop improvement [11], which should combine high-throughput approaches with automation and enhanced bioinformatics techniques.

3 Biological Data Banks and Data Integration

Biological data are collected in molecular databases that may be accessed using the Internet (see Database issue 35 of *Nucleic Acids Research* [2007] for a description of the currently available biological databases).

Biological knowledge is today distributed among many different data banks, which are increasingly indispensable and important tools in assisting scientists to understand and explain biological phenomena despite the trouble to ensure the consistency of information delivered. The spread of these collections was mainly due to the need to organize and distribute the amount of molecular data from 'omics' approaches. Different types of biological databases are available today: large-scale public repositories, meta-databases, genome browser databases, community-specific and project-specific databases. Large-scale public repositories are places for long-term storage and represent the current knowledge on the nucleotide or protein primary sequences of all organisms as well as the three-dimensional structure of biological molecules. They interchange the deposited information and are the source for many other specialized databases. Examples include GenBank for nucleotide sequences [12] and UniProt [13] for protein sequence information and the Protein Data Bank [14] for structure information. A meta-database can be considered a database of databases that collects information from different sources and usually makes them available in new, more convenient and user-friendly forms. Examples include Entrez [15] and GeneCards [16]. Another category of databases are the genome browsers, which enable scientists to visualize and browse entire genomes with annotated data that come usually from multiple diverse sources. Examples include the Ensembl Genome Browser [17], the GMOD Gbrowse Project [18], and the UCSC Genome Browser [19]. There are a number of community-specific databases that address the needs of a particular community of researchers. Examples of community-specific databases are those focused on studying particular organisms [20] or on specific types of data [21]. Project-specific databases are often short-lived databases that are developed for project data management during the funding period. Usually these databases and the corresponding Web resources are not updated beyond the funding period of the project. Since 1996, the journal *Nucleic Acids Research* has provided an annual compilation of the most important databases, noting their growing relevance to working scientists.

Because of the relevance of the data collected for the whole scientific community, data sharing has been the most interesting achievement in science in

recent years. Under the policy of the Bermuda principle (http://www.gene.ucl.ac.uk/hugo/bermuda.htm) and its extension to large-scale results [22], data are required to be publicly available and submitted to public repositories (e.g., sequence data to GenBank). However, there are no standards established for many data types (e.g., proteomics data, metabolomics data, protein localization, *in situ* hybridization, phenotype description, protein function information). Standards, specifications, and requirements for publication of data in repositories should be established in general agreement and made accessible to researchers early on in their data-generation and research activity processes.

Data integration is one of the major challenges in bioinformatics. By means of the analytical methods available, data with the same content type, but which originate from different experimental approaches, must be organized, integrated, and analyzed with distinct yet related cognitive goals. Furthermore, bioinformatics approaches must be successfully applied to integrate different data types (such as those from genomics, proteomics, and metabolomics approaches as well as experimental or clinical measurements) to provide an overview of the biological system under investigation. The long-term goal of this level of integration is to drive the research community closer to understanding the biological system physiology at a more holistic level. This strategy is useful for deeper comprehension of the response to stimulation mechanisms. It can also be exploited for the development of novel diagnostic approaches and therapies and for the identification of effective biomarkers.

4 Analysis of Biological Sequences: Sequence Comparison and Gene Discovery

Improvement of automated sequencing technology and proliferation of large-scale sequencing projects have supported the building of bioinformatics infrastructures for the management, processing, and analysis of genome sequence data. Sequence comparison, which focuses on finding all significant regions of similarities among two or more sequences, is the rationale that underpins understanding of the function and evolutionary processes that act on genomes. It is a widely used approach in bioinformatics whereby both similarities and differences in molecular sequences, genomes, RNAs, and proteins of different organisms are investigated to infer how selection has acted upon these elements. Furthermore, sequence comparison is fundamental in molecule annotation, that is, in defining the structural and functional role of molecules, exploiting the paradigm *structure determines function*. However, whereas this paradigm holds biologically true generally, caution must be exercised while assigning function to orphan sequences [23]. Therefore, a suitable method to avoid wrong or misleading (false positive) deductions is to cross-link structural and functional information so as to obtain predictions that are as reliable as possible.

As the value of genome sequences lies in the definition of their structural and functional role, the tasks of genome annotation is the first to be undertaken when facing a genome sequencing project. The primary task of genome annotation involves the identification of gene locations and definition of gene structures into raw genomic sequences. Several *in silico* methods can be used for gene annotation in a genome [24–26]. These methods can be based on *ab initio* predictions or on similarity based methods that identify sequences sharing high level of similarity in other genomes and/or in public sequence databases using specific algorithms (e.g., Smith-Waterman and BLAST [27]). The *ab initio* gene finder tools [28] attempt to recognize coding regions in genomic sequences, exploiting compositional properties in order to discriminate between coding and noncoding segments of the sequence and to detect signals involved in gene specification such as initiation and termination signals, exon and intron boundaries, and so forth. *Ab initio* methods need to be trained on a set of known genes, assuming that all coding regions within a particular genome will share similar statistical properties. Therefore, gene predictors require real models of genes that encode for the different possible classes of molecules such as mRNAs or other RNA types (tRNA, rRNAs, small RNAs) that a genome usually expresses.

Identification of a gene and definition of its structure (exon–intron regions) is the first step toward characterizing genes and hence genome functionality. However, recognizing genes in the DNA sequences remains a problem in genome analysis because the features that define gene structures are not clearly and univocally defined. Although much software for gene finding has been developed to detect both plant and animal genes [29, 30], it is still impossible to overcome limits concerning gene discovery and gene structure prediction because (i) the definition of the exact gene boundaries in a genomic region, such as the detection of the transcription start sites (TSS), is still not accurate; (ii) the prediction of small genes (300 nucleotides or less) or genes without introns is missed by many methods; (iii) partial knowledge about non–protein-coding RNAs (ncRNAs) such as microRNA or small nucleolar RNA (snoRNA); and (iv) the ambiguities concerning the prediction of alternative gene structures, such as alternative transcripts from the same gene locus or alternative splicing from the same gene product. All the issues mentioned above are research topics of the same area as bioinformatics, which is dedicated to designing methods for gene discovery. However, persistent efforts to improve the quality of gene predictions have not yet solved problems related to false-positive and false-negative predictions.

Currently, the most direct and valuable method for protein-coding gene identification and for gene structure definition relies on full-length sequencing cDNAs (i.e., more stable DNA molecules synthesized from mRNA templates), with subsequent alignments of the cDNA sequences to the genomic DNA [31, 32]. However, experimental determination of full-length cDNA sequences is an expensive and time-consuming approach when compared with high-throughput Expressed Sequence Tag (EST) sequencing. An EST represents a tiny portion

of an entire mRNA and thus provides a 'tag level' association with an expressed gene sequence. Thus, despite EST intrinsic shortcomings due to limited sequence quality, these data represent a valuable source of information to accomplish the task of gene identification and gene model building [33, 34, 35].

A further aspect to consider while annotating a genome is the analysis of inter-genic regions (i.e., stretches of DNA sequences located between the genes) of 'repetitive DNAs' [36] (which are especially copious in plant genomes [37]) and of the so-called 'junk DNA' (i.e., DNA that has no apparent gene function). Such studies hold great promise for additional insights into the evolution and organization of genomes [38, 39].

5 Transcriptome Analysis

Gene expression is the process by which the information in a gene is made manifest as a biologically functional gene product, such as an RNA molecule or a protein. A genome expresses at any moment only a tiny portion of its genes, and none of the available modeling methods are yet able to suggest which genes are expressed under specific conditions.

Because among the RNA molecules expressed by a genome the mRNA is commonly accepted to be the one that most determines the specific functionality of a cell, the sequencing of mRNA molecules is the goal of many transcriptome projects. However, full-length mRNA sequencing is a more expensive and time-consuming approach than is the high-throughput sequencing of ESTs.

An EST is produced by one-shot sequencing of a cDNA. The resulting sequence is a relatively low-quality fragment whose length is limited by current technology to approximately 500 to 800 nucleotides. ESTs represent the first truly high-throughput technology to have populated the databases and have made the rapid growth of advanced computational studies in biology inevitable. Expressed sequence tags are generated and deposited in the public repository as redundant and unannotated sequences, with negligible biological information content. The weak signal associated with an individual raw EST increases when many ESTs are analyzed together so as to provide a snapshot of the transcriptome of a species. EST sequences are a versatile data source and have multiple applications: due to exponential growth in genome sequencing projects, they are widely used for gene location discovery and for gene structure prediction. As already discussed in the previous section, predictions are usually based on spliced-alignment of source-native ESTs onto the genomic sequences [33]. EST sequences can provide useful information on the putative functionality of the tissue that the mRNA they represent were collected from. They may provide digital gene expression profiles (digital Northern) to infer the expression levels of different genes. The strategy is based on the fact that the number

of ESTs is reported to be proportional to the abundance of cognate transcripts in the tissue or cell type used to make the cDNA library [40]. Large-scale computer analyses of EST sequences can be used in the identification and analysis of coexpressed genes [41, 42]. It is important to find genes with similar expression patterns (coexpressed genes) because there is evidence that many functionally related genes are coexpressed and because this coexpression may reveal much about the genes' regulatory systems. Similar analyses can be carried out in order to detect genes exhibiting tissue- or stimuli-specific expression [43]. For this purpose, it is necessary to (i) reconstruct the full-length mRNA from EST fragments exploiting a clustering/assembling procedure [44, 45, 46]; (ii) assign a putative function to the mRNA [47, 48]; (iii) implement classification tools and data-mining techniques for reconstructing expression patterns. Transcript levels for many genes at once can be measured with DNA microarray technology, providing a systematic study of gene expression levels. Microarray experiments offer an efficient way to rapidly analyze the activity of thousands of genes simultaneously aiding gene functional assignment. Such experiments produce large amounts of data that are analyzed using data mining techniques. The key impasse is the interpretation of the results and the selection of those transcripts on which to focus attention.

The discussion on the *minimum* information required to interpret unambiguously and potentially reproduce and verify an array-based gene expression experiment is still open, as is the debate on how results from different gene expression experiments have to be interpreted for the definition of genes showing the same trend (i.e., expression profiles) or genes showing the same regulation pattern [49].

6 Systems Biology: The Major Challenge

This section describes emerging areas in bioinformatics. Indeed, *systems biology* may provide a relevant contribution to modeling and the comprehension of living organism's functionalities in the 'omics' era.

As discussed above, bioinformatics is still facing both computational and biological challenges due to the need to collect and organize large biological data collections and to manage the biological complexity revealed by increasing scientific knowledge. This is only the starting point for data analysis and, all the more so, for modeling biological processes. Computational tools and methods for data analysis are available, and the demand for more sophisticated bioinformatics solutions is expanding. Data need to be investigated to understand the structure, organization, and function of molecules for deriving peculiarities and similarities from specific cellular systems under specific conditions. Data are analyzed also to provide information on the building blocks of living organisms also in light of their evolution. This is the key to enhance our knowledge of what

has determined the organization and physiology of living organisms. However, the study of biological systems cannot be limited to simply provide an exhaustive list of their components (chemical components, reactions, proteins, genes, cells, etc.) but has to point out how and when such components are assembled and how they interact with one another. In other words, structural and dynamic aspects must be considered simultaneously in order to understand the inner make-up and workings of the system to provide information on the mechanisms that control the expression of a genome in time (i.e., during development) and space (i.e., in the single compartments of a multicellular life form) from the cellular locations to the tissue and to more complex anatomic organizations.

Interest in employing methods of knowledge discovery and text and data mining is strong and consistent with pursuing this goal and generating models of biological systems. As scientific publications are the major tool for the exchange of new scientific facts, automatic methods for extracting interesting and non-trivial information from unstructured text become very useful. Indeed, the understanding and modeling of biological systems rely on the availability of numerical values concerning physical and chemical properties of biological macromolecules, and their behavior in a cell (e.g., kinetic parameters) is not yet widely reported in public standardized repositories like molecular databases. The gathering and convergence of these data, supported by automated processes and summarized by integrative approaches, represent the founding elements in systems biology. This can be considered the new challenge in biological research. As Denis Noble wrote: *"Systems biology . . . is about putting together rather than taking apart, integration rather than reduction"* [50]. Systems biology strategies can therefore be viewed as a combination of 'omics' approaches, data integration, and modeling [51]. They require not only high-throughput biological results such as those from DNA sequencing, DNA arrays, genotyping, proteomics, and so forth, but also computational power and space for model generation and for integration of different levels of biological information. This provides a global view on the data collected and yields models that describe system behavior as a whole.

The need of a 'whole-istic' biological view [52] is by now an accepted fact. This is manifested by coordinated efforts from a wide variety of scientists worldwide (computational biologists, statisticians, mathematicians, computer scientists, engineers, and physicists) in modeling metabolic networks, cell-signaling pathways, regulatory networks, and in developing syntactically and semantically sound ways to describe these models so as to construct virtual cells on computers [http://www.e-cell.org/;. http://www.nrcam.uchc.edu/]. The results of these efforts are still far from exhaustive. Nonetheless, the interdisciplinary field of systems biology is continuing to evolve. Therefore, the determination of standards and ontologies within systems biology [53], improvements in sampling methods for molecular simulation [54], and the expansion of computational toolboxes give cause for confidence in the impact of future applications.

7 An Italian Resource for Solanaceae Genomics

We present here our effort as partners of the consortium established by the International Solanaceae (SOL) Genomics Network. The long-term goal of the consortium is to build a network of resources and information dedicated to the biology of the Solanaceae family, which includes many species of major agricultural interest such as tomato and potato. The Solanaceae comprises about 95 genera and at least 2400 species. Many of these species have considerable economic importance as food (tomato, potato, eggplant, garden pepper), ornamental (petunia), and drug plants (tobacco). Solanaceae species show a wide morphologic variability and occupy various ecological niches though they share high genome conservation. The need to enhance our knowledge of the genetic mechanisms that determine Solanaceae diversification and adaptation has led scientific efforts to be gathered under the International Solanaceae (SOL) Genome Project. The cultivated tomato, *Solanum lycopersicum*, has been chosen by the SOL initiative for BAC-based [55] genome sequencing. The long-term goal is to exploit the information generated from the tomato genome sequencing project to analyze the genome organization, its functionality, and the evolution of the entire Solanaceae family. To address key questions raised by the rationale of the SOL project, large amounts of data from different 'omics' approaches are being generated. The raw data are of little use as they stand and need to be converted into biologically meaningful information. Therefore, bioinformatics approaches become preeminent, though their results may be far from exhaustive and complete.

We present here the design and organization of ISOLA (http://biosrv.cab.unina.it/isola/), an **I**talian **Sola**naceae genomics resource. It was conceived to collect, integrate, and converge results generated from different 'omics' approaches within the consortium and to manage the overwhelming amounts of data already available and under production for investigating multiple aspects of the Solanaceae genomics. ISOLA is a multilevel computational environment and consists of different data sources and tools that are necessary to enhance data quality, to extract their information content, and to exploit their integration efficiently. The multilevel structure of ISOLA summarizes the semantics of biological data entities (Fig. 1). ISOLA currently consists of two main levels: the genome and the expression level. The cornerstone of the genome level is represented by the *Solanum lycopersicum* genome draft sequences produced by the International Tomato Genome Sequencing Consortium and by the data generated during the genome annotation procedure. Instead, the basic element of the expression level is the transcriptome information from different Solanaceae species, in the form of species-specific comprehensive collections of ESTs.

Cross-talk between the genome and expression levels ensures the sharing of data sources. This is achieved by tools that extract information content from the levels' subparts and produce value-added biological knowledge. The multilevel

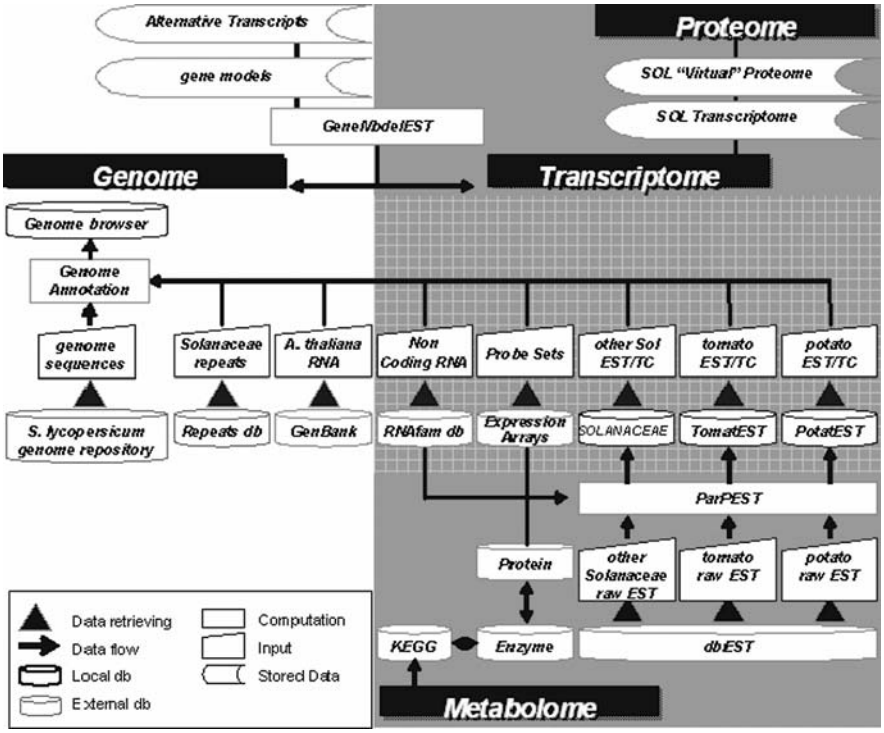

Fig. 1 ISOLA multilevel environment. Data collection and tools of the platform: The subparts of the Genome and the Transcriptome levels are included in the light area and the dark gray area, respectively. The data that are shared are located in the grid area. Subsidiary tools lay on the interface of the two levels (GeneModelEST). Value-added data gathered from the platform are listed and enclosed in the level to which they contribute more. Entry points for proteome and metabolome approaches are indicated

environment includes (i) 'basic' tools for enhancing data quality and increasing data information content and (ii) 'subsidiary' tools, which lie over the existing multilevel environment, exploiting the synergy between the levels. Each level can be independently accessed: the genome browser gateway was created for exploring the annotation of the draft tomato genome sequences, whereas the EST database gateway was created for browsing the EST-based transcriptome resources. Both the access points allow user-driven data investigation and are cross-linked to support Web-based navigation.

The genome level is enriched with reliable annotations of all the gene classes (mRNA and other non-protein-coding RNAs) and of the repeats (simple and complex ones). As already discussed, integration with the rich collections of preannotated Solanaceae ESTs represents a valuable resource for annotating tomato draft genomic sequences efficiently and effectively. In addition, annotation based on data from different species can be exploited for genome-based

comparative analyses of the Solanaceae transcriptomes. The genome annotation process is made as reliable as possible: identification of genes is performed by exploiting the convergence of the collective evidence from ESTs, mRNA, and proteins. Results from predictions are neglected because gene predictors still need to be trained on a consistent set of tomato gene models.

Definition of good quality and a representative data set of gene models is one of the tasks of the International Tomato Genome Sequencing Project and a preliminary requirement for the training of *ab initio* gene predictors. The building of gene models represents value-added information that can be elicited from effective integration of the main levels of ISOLA. The subsidiary tool GeneModelEST [35] selects candidate gene models by evaluating the tentative consensus sequences that are generated from EST-based clustering/assembling procedures and are then aligned to the tomato genome sequences. EST-to-genome alignment supports the evaluation of the exons that the tentative consensus sequences describe along the genome sequence. Furthermore, alternative transcripts are catalogued because alternative gene structures must be avoided in the definition of candidate gene models.

The expression level is mainly represented by the EST data from cDNA libraries of 16 different Solanaceae species. Each source-specific EST collection is processed and annotated using the ParPEST pipeline [48], a *basic* tool in ISOLA. The set of information the ParPEST pipeline generates is deposited in relational databases, and user-friendly Web interfaces are developed to easily manage and investigate the EST collections. The data/tool design at this level permits the study of species-specific expression patterns and their time course, in normal or pathologic conditions and/or under specific biotic or abiotic stimuli. The protein-based functional annotation of EST sequences and the detection of the putative protein-coding region (Open Reading Frames) can undoubtedly enrich our knowledge of Solanaceae proteomes. This gives the opportunity to plug into ISOLA data generated from proteome efforts. In addition, the association of ESTs to metabolic pathways as described in the KEGG database [56] provides an entry point to integrate into ISOLA metabolome data, which can further support the definition of gene expression patterns. On the other hand, the association between EST sequences and the oligonucleotide probe-set from tomato expression arrays [57] [GeneChip Tomato Genome Array; http://www.affymetrix.com/products/arrays/specific/tomato.affx] leaves data from expression profiling arrays to be part and parcel of ISOLA.

Organization, classification, and annotation of source-specific EST data provide an estimation of the transcriptome space for each Solanaceae species. They also permit gene expression maps, regulatory networks, and metabolic processes to be modeled according to the paradigm 'from the omics to systems biology.' Annotation of the tomato draft genome sequences, on the other hand, provides assessment of the tomato gene space and shows how fundamental it is to achieve in-depth knowledge of a reference genome within a flowering plant family of such considerable economic and agronomic importance. All the information generated in annotating the tomato genome can be useful for the

comprehension of the genome organization, its functionality, and the evolution of the entire Solanaceae family.

ISOLA is not a static 'platform.' It is under continuous evolution in that it considers the ongoing growth of data sources as well as the production of novel computational and experimental methods. We are well aware that ISOLA is still far from fulfilling the challenge of systems biology approaches. However, it lays the foundations for providing reliable molecular information on what is acting in a biological system (cell, tissue, organism), as well as where and when. Of course, methods must be improved to efficiently investigate the basis of comparative approaches, common features that could highlight how biological phenomena take place. We also believe that studying the structure, its function, and the evolution of the Solanaceae genomes is a suitable test-bench to challenge and expand this effort.

8 Conclusions

There are many bioinformatics applications in support of high-throughput experimental technologies, and high-level computational requirements are necessary. *Ad hoc* methods for data storage, data warehousing, data integration, data visualization, and data modeling are fundamental. The main target is to appropriately model structures and functions of molecules and the biological phenomena they give rise to.

The demand for bioinformatics tools and data banks able to fulfill such requirements also shows that bioinformatics must be based on multidisciplinary competences covering many different scientific aspects and challenging many different fields of research in the context of rapidly evolving scientific research and technologies. An additional requirement is therefore the need to train researchers for the integrated research environment, which also includes competences from those who plan and conduct experimental analyses and have in-depth knowledge of the biological systems to be modeled.

Acknowledgments We wish to thank Prof. Gerardo Toraldo for useful discussions and constant support. This is the contribution DISSPAPA book 3. Part of the presented work is supported by the Agronanotech Project (Ministry of Agriculture, Italy) and by the PRIN 2006 (Ministry of Scientific Research, Italy) and is in the frame of the EU-SOL Project (European Community).

References

1. Cullis, C.A.: Plant genomics and proteomics. Hoboken, NJ: John Wiley and Sons, pp. 214 (2004).
2. Heslop-Harrison, J.S., Murata, M., Ogura, Y., Schwarzacher, T., Motoyoshi, F.: Polymorphisms and genomic organization of repetitive DNA from centromeric regions of *Arabidopsis thaliana* chromosomes. Plant Cell 11: 31–42 (2000).

3. Britten, R.J., Kohne, D.E.: Repeated sequences in DNA. Science 161: 529–540 (1968).
4. Zwick, M.S., Hanson, R.E., McKnight, T.D., Islam-Faridi, M.N., Stelly, D.M., et al.: A rapid procedure for the isolation of Cot1 DNA from plants. Genome 40: 138–142 (1997).
5. EPSO: European plant science: A field of opportunities. Journal of Experimental Botany 56: 1699–1709 (2005).
6. Iovene, M., Barone, A.., Frusciante, L.., Monti, L.., Carputo, D.: Selection for aneuploid *Solanum commersonii-S. tuberosum* hybrids combining low wild genome content and resistance traits. Theoretical and Applied Genetics 119: 1139–1146 (2004).
7. Barone, A., Frusciante L.: Molecular marker-assisted selection for resistance to pathogens in tomato. In: Marker-assisted selection: Current status and future perspectives in crops, livestock, forestry and fish. E.P. Guimaraes, J. Ruane, B.D. Scherf, A. Sonnino, J.D. Dargie (eds), FAO, Rome, Italy pp. 151–164 (2007).
8. Barone, A.: Molecular marker-assisted selection for potato breeding. American Journal of Potato Research 81: 111–117 (2004).
9. Ruane, J., Sonnino, A.: Marker-assisted selection as a tool for genetic improvement of crops, livestock, forestry and fish in developing countries: On overview of the issues. In: Marker-assisted selection: Current status and future perspectives in crops, livestock, forestry and fish. E.P. Guimaraes, J. Ruane, B.D. Scherf, A. Sonnino, J.D. Dargie (eds), FAO, Rome, Italy pp. 4–13 (2007).
10. Varsheney, R.K, Graner, A., Sorrells, M.E: Genomics-assisted breeding for crop improvement. Trends in Plant Science 10(12): 621–630 (2005).
11. Peleman, J.D., Rouppe van der Voort, J.: Breeding by design. Trends in Plant Science 8: 330–334 (2003).
12. Wheeler, D.L., Smith-White, B., Chetvernin, V., Resenchuk, S., Dombrowski, S.M., et al.: Plant genome resources at the national center for biotechnology information. Plant Physiology 138: 1280–1288 (2005).
13. Schneider, M., Bairoch, A., Wu, C.H., Apweiler, R.: Plant protein annotation in the UniProt Knowledgebase. Plant Physiology 138: 59–66 (2005).
14. Deshpande, N., Addess, K.J., Bluhm, W.F., Merino-Ott, J.C., Townsend-Merino, W., et al.: The RCSB Protein Data Bank: A redesigned query system and relational database based on the mmCIF schema. Nucleic Acids Research 33: D233–D237 (2005).
15. Ostell, J.: The NCBI Handbook: The Entrez Search and Retrieval System. National Library of Medicine (NLM), Washington, DC, USA Part 3, section 15, Rome, Italy (2003).
16. Safran, M., Chalifa-Caspi, V., Shmueli, O., Olender, T., Lapidot, M., et al.: Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE . Nucleic Acids Research 31: 142–146 (2003).
17. Hubbard, T. J. P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., et al.: Ensembl 2007. Nucleic Acids Research 35: D610–D617 (2007).
18. Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., et al.: The generic genome browser: A building block for a model organism system database. Genome Research 12(10): 1599–1610 (2002).
19. Kuhn, R.M., Karolchik, D., Zweig, A.S., Trumbower, H., Thomas, D.J., et al.: The UCSC Genome Browser database: Update 2007 Nucleic Acids Research. 35: D668–D673 (2007).
20. Yamazaki, Y., Jaiswal, P.: Biological ontologies in rice databases. An introduction to the activities in Gramene and Oryzabase. Plant Cell Physiology 46: 63–68 (2005).
21. D 'Agostino, N., Aversano, M., Frusciante, L., Chiusano, M.L.: TomatEST database: In silico exploitation of EST data to explore expression patterns in tomato species. Nucleic Acids Research 35: D901–D905 (2007).
22. The Wellcome Trust. Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility. Fort Lauderdale, FL: Wellcome Trust (2003).

23. Noel, J.P., Austin, M.B., Bomati, E.K.: Structure-function relationships in plant phenylpropanoid biosynthesis. Current Opinion in Plant Biology 8: 249–253 (2005).
24. Claverie, J.M.: Computational methods for the identification of genes in vertebrate genomic sequences. Human Molecular Genetics 6**:** 1735–1744 (1997).
25. Stormo, G.D.: Gene-finding approaches for eukaryotes. Genome Research 10(4): 394–397 (2000).
26. Davuluri, R.V., Zhang, M.Q.: Computer software to find genes in plant genomic DNA. Methods in Molecular Biology 236: 87–108 (2003).
27. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology 215(3): 403–410 (1990).
28. Yao, H., Guo, L., Fu, Y., Borsuk, L.A., Wen, T.J., et al.: Evaluation of five ab initio gene prediction programs for the discovery of maize genes. Plant Molecular Biology 57(3): 445–460 (2005).
29. Zhang, M.Q.: Computational prediction of eukaryotic protein-coding genes. Nature Reviews Genetics. 3: 698–709 (2002).
30. Schlueter, S.D., Dong, Q., Brendel, V.: GeneSeqer@PlantGDB: Gene structure prediction in plant genomes. Nucleic Acids Research. 31: 3597–3600 (2003).
31. Seki, M., Naruska, M., Kamiya, A., Ishida, J., Satou, M., et al.: Functional annotation of a full-length *Arabidopsis* cDNA collection. Science 296: 141–145 (2002).
32. Alexandrov, N.N., Troukhan, M.E., Brover, V.V., Tatarinova, T., Flavell, R.B.,et al.: Features of *Arabidopsis* genes and genome discovered using full-length cDNAs. Plant Molecular Biology 60(1): 69–85 (2006).
33. Adams M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., et al.: Complementary DNA sequencing: Expressed sequence tags and human genome project. Science 252: 1651–1656 (1991).
34. Zhu, W., Schlueter, S.D., Brendel, V.: Refined annotation of the *Arabidopsis* genome by complete expressed sequence tag mapping. Plant Physiology. 132: 469–484 (2003).
35. D 'Agostino, N., Traini, A., Frusciante, L., Chiusano, M.L.: Gene models from ESTs (GeneModelEST): An application on the *Solanum lycopersicum* genome. BMC Bioinformatics 8(Suppl 1): S9 (2007).
36. Lewin, B.: Genes VIII. Upper Saddle River, NJ: Prentice Hall (2003).
37. Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., Wessler, S.R.: Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569–573 (2004).
38. Morgante, M.: Plant genome organisation and diversity: The year of the junk! Current Opinion in Biotechnology 17(2): 168–173 (2006).
39. Morgante, M., De Paoli, E., Radovic, S.: Transposable elements and the plant pangenomes. Current Opinion in Plant Biology 10(2): 149–155 (2007).
40. Audic, S., Claverie, J.M.: The significance of digital gene expression profiles. Genome Research 7(10): 986–995 (1997).
41. Ewing, R.M., Ben Kahla, A., Poirot, O., Lopez, F., Audic, S., et al.: Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Research 9: 950–959 (1999).
42. Wu, X., Walker, M.G., Luo, J., Wei, L.: GBA server: EST-based digital gene expression profiling. Nucleic Acids Research 33 (Web Server issue): W673–W676 (2005).
43. M égy, K., Audic, S., Claverie, J.M.: Heart-specific genes revealed by expressed sequence tag (EST) sampling. Genome Biology 3(12): RESEARCH0074, Rome, Italy, 11 (2003).
44. Burke, J., Davison, D., Hide, W.: d2_cluster: A validated method for clustering EST and full-length cDNA sequences. Genome Research 9: 1135–1142 (1999).
45. Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., et al.: TIGR Gene Indices clustering tools (TGICL): A software system for fast clustering of large EST datasets. Bioinformatics 19: 651–652 (2003).
46. Kalyanaraman, A., Aluru, S., Kothari, S., Brendel, V.: Efficient clustering of large EST data sets on parallel computers. Nucleic Acids Research 31: 2963–2974 (2003).

47. Hotz-Wagenblatt, A., Hankeln, T., Ernst, P., Glatting, K.H., Schmidt, E.R., et al.: ESTAnnotator: A tool for high throughput EST annotation. Nucleic Acids Research 31: 3716–3719 (2003).
48. D 'Agostino, N., Aversano, M., Chiusano, M.L.: ParPEST: A pipeline for EST data analysis based on parallel computing. BMC Bioinformatics 6 (Suppl 4): S9 (2005).
49. Van Helden, J.: Regulatory sequence analysis tools. Nucleic Acids Research 31: 3593–3596 (2003).
50. Noble, D.: The music of life. Oxford: Oxford University Press (2006).
51. Ge, H., Walhout, A.J., Vidal, M.: Integrating 'omic' information: A bridge between genomics and systems biology. Trends in Genetics. 19(10): 551–560 (2003).
52. Chong, L., Ray, L.B.: Whole-istic Biology. Science 295(1): 1661 (2002).
53. Strömbäck, L., Hall, D., Lambrix, P.: A review of standards for data exchange within systems biology. Proteomics 7(6): 857–867 (2007).
54. Lei, H., Duan, Y.: Improved sampling methods for molecular simulation. Current Opinion in Structural Biology 17(2): 187–191 (2007).
55. Mueller, L.A., Tanksley, S.D., Giovannoni, J.J., van Eck, J., Stack, S., et al.: The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL). Comparative and Functional Genomics 6: 153–158 (2005).
56. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., et al.: From genomics to chemical genomics: New developments in KEGG. Nucleic Acids Research 34: D354–D357 (2006).
57. Fei, Z., Tang, X., Alba R., Giovannoni, J.: Tomato Expression Database (TED): A suite of data presentation and analysis tools. Nucleic Acids Research 34: D766–D770 (2006).