

Chapter 2

Bioinformatics

Abstract Bioinformatics is a major research area in its own right, as well as a source of tools, databases and services. This research aspect is highlighted in the area of genome annotation, in its broadest sense of defining the biological role of a molecule in all its complexity. This complexity is explored in this chapter, and involves gene definition, alternative transcripts and splicing, gene regulation and expression, the functional annotation of proteins, post-translation modification, membrane and localisation prediction, protein complexes, networks and pathways. Annotation is further unified in an international collaborative effort on compiling an encyclopaedia of DNA elements.

Introduction

Genome Sequences

The genome projects have revealed and codified the entire DNA sequence of humans and other organisms, which if not entirely providing a “blueprint” of life describes many key elements. The first draft of the human sequence was published in 2001 (The International Human Genome Mapping Consortium 2001), and there are now over 53 eukaryota, 46 archaea and 517 bacteria genome sequences in the public domain (EBI 2007). This explosion in genomic information has been achieved in a remarkably short period of time, and the flood of new sequence data is likely to continue for the foreseeable future. However, a DNA sequence is a string of letters; it must be interpreted in terms of the RNA and proteins that it encodes and the promoter and regulatory regions that control transcription and translation.

Annotation

Annotation can be described as the process of “defining the biological role of a molecule in all its complexity” and mapping this knowledge onto the relevant gene

products encoded by genomes. This involves both experimental and computational approaches and, indeed, absolutely requires their integration.

European Contributions

European scientists have been very active in the field of genome and protein annotation, with Ensembl (2007) and Swiss-Prot (2007), now integrated in UniProt (2007), being among the primary resources in use worldwide. Many of the tools used in genome and protein sequence and structure annotation, prediction and validation, as well as in pathway analysis and secondary resources derived from protein sequences and structures were developed in Europe. The fragmentation of resources for genome annotation meant that only a few bioinformatics experts knew where to look for them. Consequently, most experimentalists had been unable to access all the best information about a genome. In what follows, key recent contributions from several projects are discussed, including ATD (2007), BaSysBio (2007), BioSapiens (2007), BioBabel (2007), ENCODE (2007), EuTRACC (2007), GO (2007), IIMS (2007), SPINE (2007) and TEBLOR (2007),

Key Areas

Some of the key research areas in the bioinformatics of genome annotation are systematically discussed in this chapter, and include:

- Gene definition/alternative splicing
- Regulators and promoters
- Expression
- Genetic variation (haplotypes, single-nucleotide polymorphisms, etc.)
- Protein families, orthologues
- Membrane proteins and ligands
- Three-dimensional protein structure
- Post-translation modification and localisation
- Sequence and structure to function
- Protein–protein complexes
- Pathways and networks

Genome Annotation

A European Virtual Institute for Annotation

In response to the topic published in support of genome annotation (FP6-2002-LIFESCIHEALTH, 2002), the BioSapiens (2007) project was successfully proposed, and has created a fully functioning European Virtual Institute for Genome

Annotation BioSapiens (2005). This virtual institute has established tools and work flows that allow annotation over a large part of the range of biological knowledge, and it addresses the full range of research topics listed above. The institute is improving bioinformatics research in Europe, by providing a focus for annotation and through the organisation of European meetings and workshops to encourage cooperation, rather than duplication of effort.

Distributed Annotation

An important aspect of the network activities is to achieve closer integration between experimentalists and bioinformaticians, through a directed programme of genome analysis, focused on specific biological problems. The annotations generated by the Institute and external participants are available in the public domain and are easily accessible on the Web. This has been achieved through a distributed annotation system (DAS 2007), which has evolved to take advantage of the GEANT2 (2007) pan-European research and education network supported by Enabling Grids for E-Science in Europe (EGEE 2007). The groups also focus on the development of improved computational methods for annotation through new methods available via the Web. Annotations from these new methods are available via DAS (2007), which is available via the website as a DAS portal and has a DAS server information service (DAS-Information 2007). There are over 23 distinct DAS servers providing 69 different data sources.

An Integrated Approach

Many of the tools used in genome and protein sequence and structure annotation, prediction and validation, and pathway analysis have been developed in Europe. BioSapiens (2007) has been instrumental in creating increased integration, expert training and improved tools and services, and an enhanced European role in the academic and industrial exploitation of genomics. Some of the main results being produced by the project include the development of an integrated approach to genome annotation from gene to function, and ultimately the establishment of an integrated and distributed website for genome annotation. A description of the individual research areas is available via the work packages as shown in Table 2.1.

Annotation Deliverables

References to the bioinformatics methods used in each particular area are available via deliverables within the project work package descriptions. Work packages in the range of more than 100 are for work in progress. For example, in the area

Table 2.1 BioSapiens scientific areas for genome annotation

Scientific area	Reference
Gene definition/alternative splicing	BioSapiens-WP1 (2007) BioSapiens-WP101 (2007)
Gene regulation and expression	BioSapiens-WP2 (2007) BioSapiens-WP3 (2007) BioSapiens-WP102 (2007)
Variation (haplotypes and single-nucleotide polymorphisms)	BioSapiens-WP4 (2007) BioSapiens-WP103/110 (2007)
Functional annotation of proteins	BioSapiens-WP5 (2007) BioSapiens-WP7 (2007) BioSapiens-WP9 (2007) BioSapiens-WP104 (2007)
Post-translational modification, membrane and localisation prediction	BioSapiens-WP6 (2007) BioSapiens-WP8 (2007) BioSapiens-WP105 (2007)
Protein complexes, networks and pathways	BioSapiens-WP10 (2007) BioSapiens-WP11 (2007) BioSapiens-WP106 (2007)

of gene definition/alternative splicing (BioSapiens-WP101 2007), deliverable “Del 1.1: A list of experimentally validated gene structures for the human genome and other mammalian genomes” contains a full list of references at the end of the document, including a reference to the major tool used, Ensembl (Hubbard et al. 2005), supplementing the overall project list of more than 75 major publications at BioSapiens-Publications (2007)

Genome Browser and Distributed Annotation Viewer

Ensembl (2007) is a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl provides accurate and automatic analysis and annotation of genome data, concentrating on vertebrate genomes, but includes a wide range of other model organisms. There are over 33 genomes available in Ensembl; these include those of human, several other mammals, chicken, four species of fish, several insect species and a nematode. Prereleases of three further organisms, including the start of a large number of low-coverage mammalian genomes such as that of elephant, are available through Pre-Ensembl. Ensembl automatically annotates genome sequence and predicts the positions of genes, to provide a comprehensive range of sequence features and genome-wide gene and protein sets. The system is applied in a consistent way to different species, and incorporates between-species comparisons of genome sequence and homologous genes. A rich variety of links to external databases helps to make Ensembl a key starting and central reference point for studies in genetics and molecular biology. Ensembl continues to improve annotation to both the human and the mouse genomes. It also provides timely annotation, both for newly sequenced genomes (such as that of platypus) and

for previously sequenced genomes for which information continues to be refined (such as that of chicken). In particular, Ensembl has an effective pipeline for calculating non-coding RNA gene models. These include structural RNAs, such as U6 RNA, and regulatory RNAs, such as micro-RNAs. In the human genome there are also results on *cis*-regulatory networks. This is an actively growing area of research with many groups developing methods

As a distributed annotation viewer, the Ensembl (2007) genome browser is used with its DAS (2007) display to look at the p53 gene, an apoptosis-regulating gene with nearly 44,000 publications, shown in Fig. 2.1, providing the following information, and much more:

Gene, TP53 (HUGO Gene Nomenclature Committee, HGNC, symbol); synonyms, p53.

This gene is a member of the human CCDS set: CCDS11118.

Ensembl gene ID: ENSG00000141510

Genome location: This gene can be found on chromosome 17 at location 7,512,464–7,531,642.

The start of this gene is located in contig AC087388.9.1.121017.

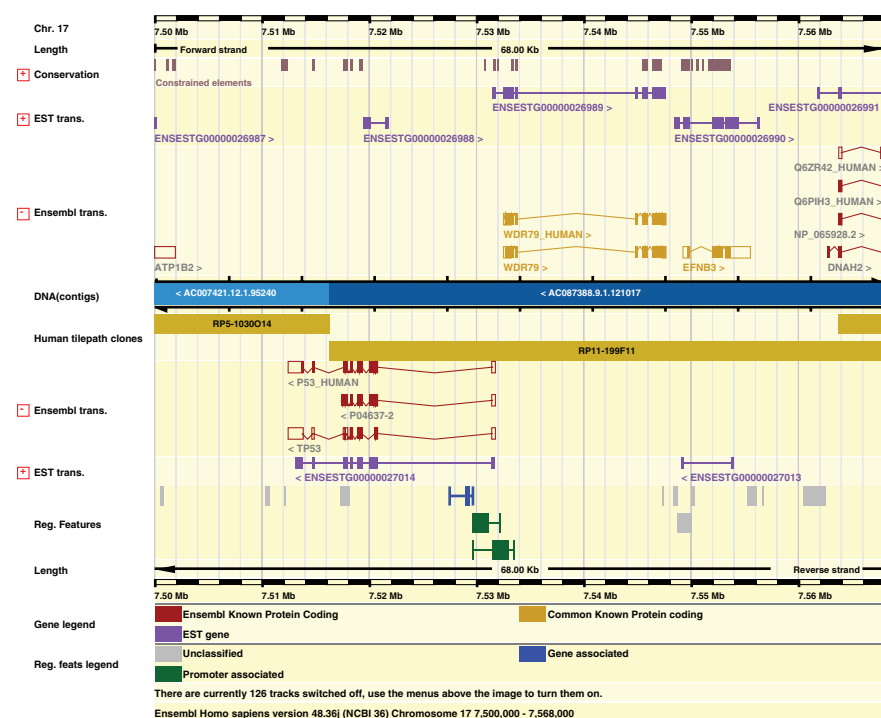


Fig. 2.1 Ensembl (2007) browser with DAS display to look at p53, Ensembl gene ID ENSG00000141510

Description: Cellular tumour antigen p53 (tumour suppressor p53) (phosphoprotein p53) (antigen NY-CO-13). Source, UniProt/Swiss-Prot P04637

Prediction method: Genes were annotated by the Ensembl automatic analysis pipeline using either a GeneWise/Exonerate model from a database protein or a set of aligned complementary DNAs (cDNAs) followed by an open reading frame prediction. GeneWise/Exonerate models are further combined with available aligned cDNAs to annotate untranslated regions (for more information see Curwen et al. 2004).

Gene DAS report: DAS sources

AltSplice (alternative Splice database)

AltTrans (alternative transcript diversity database)

ArrayExpress (gene expression database)

GAD (genetic association database)

HGNC (HUGO Gene Nomenclature Committee)

HUGO_text (PubMed text mining via HGNC symbol)

Phenotypes (associated directly or via orthologues or protein families)

Protonet (global classification of proteins into hierarchical clusters)

RZPD verif. cDNA (RZPD sequence verified non-redundant cDNA clone sets)

RZPD esiRNA (RZPD gene silencing (RNA interference) resources)

RZPD Prot Exp (RZPD clones ready for protein expression)

Reactome (knowledgebase of biological processes)

UniProt (protein knowledgebase)

Reaction Pathways

All of the above DAS reports could be selected, in which case much more information appears. As one example, Fig. 2.2 shows the Reactome (2007) path provided (Reactome-1756 2007), which consists of fully curated pathway data, and is described as phosphorylation of p53 at ser-15 by ataxia-telangiectasia mutated

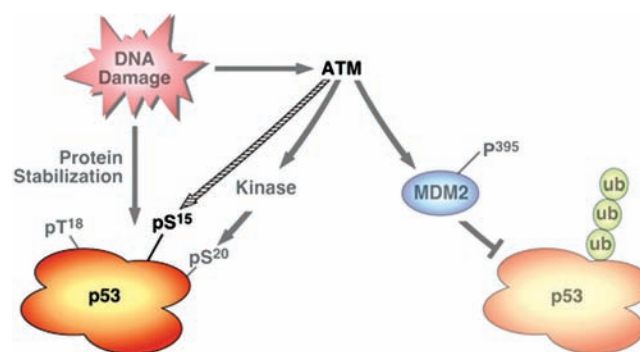


Fig. 2.2 Reactome-1756 (2007), A curated knowledgebase of biological pathways – path 1756 – phosphorylation of p53 at ser-15 by ataxia-telangiectasia mutated (*ATM*) kinase (*Homo sapiens*)

(ATM) kinase stable identifier REACT_1756.1. In response to DNA damage due to ionising radiation, the serine at position 15 of the p53 tumour suppressor protein is rapidly phosphorylated by the ATM kinase. This serves to stabilise the p53 protein. A rise in the levels of the p53 protein induces the expression of the p21 cyclin-dependent kinase inhibitor. This prevents the normal progression from G1 to S phase, thus providing a check on replication of damaged DNA.

Experimental–Computational Collaboration

BioSapiens (2007) also stimulates cooperation between experimental scientists and computational biologists for genome annotation, in the form of meetings and joint collaborations. Experimental validation of predictions made *in silico* forms part of these collaborations. The tools are also validated and applied via thematic work packages.

Thematic Collaborations

BioSapiens (2007) consciously chooses particular thematic areas where the full power of the Virtual Institute can be directed towards particular scientific problems. These areas are summarised in Table 2.2. These thematic areas show the power of these large networks, since they can apply the tools they develop to a wide range of problems, including relevant disease research. The disease themes are discussed later in the appropriate book sections. The exploitation of the biological information enabled by BioSapiens (2007) will in some cases be relatively direct, e.g. improved health-care through better drugs, new vaccines and personalised medicines for individuals and subpopulations, and improved understanding of diet and health.

Critical Mass of Resources

BioSapiens (2007) has had an important impact on the establishment of a European research structure that supports the coordination of bioinformatics research activities across different subareas, and across different areas of medical and biotechnological application. It has developed the required level of critical mass so that Europe, with primarily nationally based funding schemes, can compete with the major investments

Table 2.2 BioSapiens (2007) thematic areas of scientific collaboration

Infectious diseases	BioSapiens-WP15 (2007)
Down syndrome	BioSapiens-WP16 (2007)
ENCODE project	BioSapiens-WP20 (2007)
	BioSapiens-WP108 (2007)
Cancer	BioSapiens-WP109 (2007)

made in the USA and Japan. The integration between the groups has already had a lasting impact on the European bioinformatics infrastructure, and on the sharing of human resources, infrastructure databases and tools. Through cutting-edge research, high-level training and vigorous European-level interaction, BioSapiens has made a substantial contribution to improving Europe's knowledge base.

Bioinformatics Tools For Annotation

Integrated Tool Development

The TEMBLOR (2007) project, the European molecular biology linked original resources, received almost €20 million over 3 years. The project concentrated on research and development to build major bioinformatics resources. These resources were embedded in an integrated layer known as Integr8 (2007), allowing biomedical researchers to fully exploit genomic and proteomic data. Integr8 draws on databases that are maintained at major bioinformatics centres in Europe, and also on important new resources. The main aim of TEMBLOR (2007) is to allow users to carry out complex queries across databases in a much simpler way than has previously been possible, by accessing all of these databases through Integr8.

A summary of the projects within TEMBLOR includes:

- Integr8 (2007) – an integrated layer for the exploitation of genomic and proteomic data
- EMSD (2007) – storing and analysing the structures of large molecules
- DESPRAD (2007) – standards and repositories for gene expression experiments
- IntAct (2007) – standards and resources for protein–protein interaction data

Integrated Layer for Genomic and Proteomic Data

Integr8 (2007), described by Kersey et al. (2005) is a Web portal for exploring the biology of organisms with completely deciphered genomes. For 53 eukaryota, 46 archaea and 517 bacteria, Integr8 provides access to general information, recent publications, and a detailed statistical overview of the genome and proteome of the organism. Integr8 (2007) also provides access to complete genomes and proteomes, as part of developing integrated search capabilities, resulting in a major strengthening of individual database capabilities for protein sequence work, taxonomy and ontologies, via support given to projects such as UniProt (2007), InterPro (2007), NEWT (2007), GO (2007) and GOA (2007), which had already been partially developed by BioBabel (2007). Although these databases are centralised at the EBI, their establishment and continuing development are based on multilaboratory collaboration in the development of Integr8 (2007) and Genome Reviews (2007). The Integr8 (2007) Web portal provides easy access to inte-

grated information about deciphered genomes and their corresponding proteomes. Available data include:

- DNA sequences from databases including the EMBL nucleotide sequence database, Genome Reviews, and Ensembl
- Taxonomy of the organism via NEWT (2007)
- Protein sequences from databases including the UniProt knowledgebase and IPI (2007)
- Statistical genome and proteome analysis performed using InterPro (2007), CluSTr (2007), and GOA (2007)
- Information about orthology, paralogy, and synteny

Protein Structure

The Macromolecular Structure Database (MSD 2007) group is one of the three partners in the worldwide Protein Data Bank (PDB), the consortium entrusted with the collation, maintenance and distribution of the global repository of macromolecular structure data, especially protein structure data. The PDB is the international repository for three-dimensional structures of macromolecular complexes of proteins, nucleic acids and other biological molecules. The data range from those of small protein fragments to those of large macromolecular assemblies such as viruses and ribosomes, whose structures have been determined by experimental methods such as X-ray crystallography, nuclear magnetic resonance spectroscopy or electron microscopy. Many of the electron microscopy analysis capabilities were developed in the IIMS (2007) project. These data are publicly accessible, and are used by scientists, researchers, bioinformaticians, educators, students and lay audiences. By annotating and archiving the data in an efficient and consistent way, the PDB supports the understanding of biological phenomena at a structural level and facilitates new discoveries in science. The MSD (2007) tools available include:

- MSDlite (simple search of relational PDB)
- MSDpro (advanced search system)
- MSDmotif (small three-dimensional motif statistics with extensive Φ , Ψ , χ search options)
- MSDtemplate (local residue interactions in the PDB)
- MSDpisa (search and analysis of protein interfaces, surfaces and assemblies)
- MSDchem (ligand search)
- MSDmine (ad hoc queries and data analysis)
- MSDsite (ligand-environment search)
- MSDfold (secondary structure matching)
- MSDanalysis (validation and analysis of MSD data)
- MSDtarget (sequence target search)
- EMsearch (search the electron microscopy database)
- MSDbar (search system using toolbar application)

- PQS (protein quaternary structure server)
- PQS-Quick (simple PQS search)
- NMR Representatives (representative model from NMR ensemble)
- Reference Server (search by author/ID for PDB structures without final reference)
- Relibase (a program for searching protein-ligand databases)
- Biotech (validation suite for protein structures)
- Search OCA (enter OCA search system)
- PDB Pending (search pending and waiting list for status of file under processing)
- PDB New Entries (PDB latest releases)
- SPINE @ EBI (direct to spine targets)

The tools of MSD (2007) provide a wide range of options, for example visualising protein structures that were analysed in the FP5 (2007) collaborative project SPINE (2007) – Structural Proteomics in Europe. Figure 2.3 shows an example of a surface protein of a cancerous cell:

Expression Data

The DESPRAD (2007) subproject was aimed at developing ArrayExpress (2007), a public repository for microarray data, and the standards and ontologies needed to describe, exchange and store microarray data (experiments, protocols and array designs). Also, software tools were developed for querying the database, and for

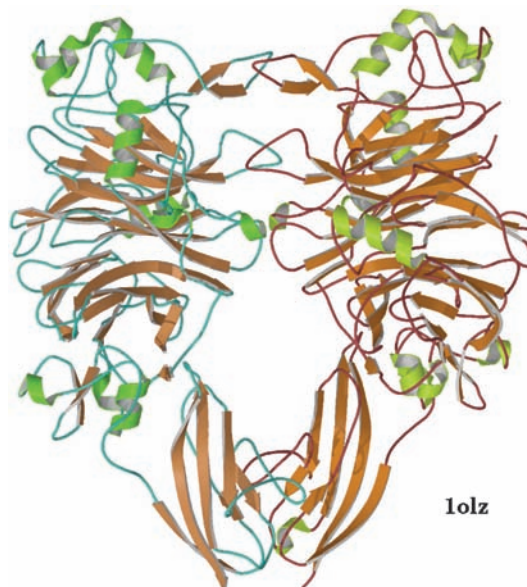


Fig. 2.3 Semaphorin 4D precursor protein structure, *Homo sapiens*, Protein Data Bank accession 1OLZ, Swiss-Prot accession Q92854 (see also Love et al. 2003)

curation and submission of data. Analysis tools, standalone or integrated with the database, were also goals for this project.

Elements of this project include:

- Minimum Information About a Microarray Experiment (MIAME 2007). This has become an accepted worldwide standard.
- Microarray and Gene Expression (MAGE 2007). These standards have been adopted by The Object Management Group (OMG 2007).
- MGED (2007) is an ontology for describing microarray experiments.
- ArrayExpress (2007) is a public repository which is online and accepting submissions.
- MIAMExpress (2007) is a MIAME (2007) compliant microarray data submission tool.
- Expression-Profiler (2007) is an open, extensible Web-based collaborative platform for microarray gene expression, sequence and protein–protein interaction data analysis.

Storing and Interpreting Microarray Data

ArrayExpress (2007) is now one of the major tools of modern biology, essential for storing and interpreting microarray data. ArrayExpress is a public repository for microarray data, which is aimed at storing MIAME-compliant data in accordance with MGED (2007) recommendations. The ArrayExpress (2007) data warehouse stores gene-indexed expression profiles from a curated subset of experiments in the repository. Public data are made available for browsing and querying on experiment properties, submitter, species, etc. Queries return summaries of experiments and complete data, or subsets can be retrieved. A subset of the public data are reannotated to update the array design annotation and curated for consistency. These data are stored in the data warehouse and can be queried on gene, sample, and experiment attributes. The results return graphed gene expression profiles, one graph per experiment.

Microarray Expression, Sequence and Protein–Protein Data Analysis

Coupled to ArrayExpress is Expression-Profiler (2007). Expression Profiler: Next Generation (Kapushesky et al. 2004) is an open, extensible Web-based collaborative platform for microarray gene expression, sequence and protein–protein interaction data analysis, exposing distinct chainable components for clustering, pattern discovery, statistics (via the R programming language), machine-learning algorithms and visualisation. The architecture modularises the original design and allows individual analysis-task-related components to be developed by different groups and yet still seamlessly to work together and share the same user-interface look and feel. Data analysis components for gene expression

data before processing, missing value imputation, filtering, clustering methods, visualisation, significant gene findings, between-group analysis and other statistical components are available from the EBI (2007) website. The Web-based design of Expression-Profiler (2007) supports data sharing and collaborative analysis in a secure environment. Developed tools are integrated with the microarray gene expression database ArrayExpress (2007) and form the exploratory analytical front end to those data.

Protein-Protein Interactions

Major capabilities for studying protein-protein interactions were established by the TEMBLOR (2007) subproject IntAct (2007). IntAct (2007) provides a freely available, open source database system and analysis tools for protein interaction data. All interactions are derived from literature curation or direct user submissions and are freely available. An experiment consists of many interactions which contain two or more interactors. An interactor can be either an individual protein or a protein complex (i.e. the result of a previous interaction). Therefore, an interaction can consist of two or more proteins or complexes. Each object (experiment, interaction, interactor) has attributes assigned which provide a detailed description. This is important as specific features of an experiment can have a profound impact on the type of interaction. Every IntAct (2007) object has a unique accession code which starts with “EBI-”, followed by a number. It is these accession codes that enable the hierarchical data structure.

Protein Sequence and Function Database

UniProt (2007) is the world’s most comprehensive catalogue of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in UniProt Knowledgebase (UniProtKB)/Swiss-Prot, UniProtKB/TrEMBL, and PIR. UniProt is composed of three components, each optimised for different uses. UniProtKB is the central access point for extensive curated protein information, including function, classification, and cross-reference. The UniProt Reference Clusters (UniRef) databases combine closely related sequences into a single record to speed searches. The UniProt Archive (UniParc) is a comprehensive repository, reflecting the history of all protein sequences.

Protein Sequence Grouping

InterPro (2007) is a searchable database providing information on sequence, function and annotation. Sequences are grouped on the basis of protein signatures or “methods”. These groups represent superfamilies, families or subfamilies of sequences. The groups may be defined as families, domains, repeats or sites. The function of sequences within any group may be confined to a single biological

process or it may be a diverse range of functions (as in a superfamily) or the group may be functionally uncharacterised, but without exception every entry has an abstract and references are provided where possible. It is well worth browsing the database and going through the InterPro frequently asked questions.

Gene Definition/Alternative Transcripts and Splicing

Gene Definition

Although the human genome sequence has been available in at least draft form for several years, the complete list of all of its functional regions is far from complete (BioSapiens-WP1 2007). Genome annotation relies on computational methods to integrate information from both de novo gene prediction algorithms and protein databases and other sources of expressed sequences such as expressed sequence tags (cDNA) and high-quality reference sequence messenger RNAs (mRNAs). Each of the sources of expressed sequence must be accurately mapped to the exact genome locations corresponding to the sequence to discover the gene responsible for the given sequence. These processes require significant computational resources. There is still considerable question about the total number of protein-coding genes contained within the human genome. The currently accepted estimate is approximately 25,000 genes, and many may include multiple transcribed forms. This estimate is based on a number of independent methods for annotating the human genome (e.g. Curwen et al. 2004). Essentially the same number of genes are thought to be present in the mouse genome.

Alternative Transcripts and Splicing

A single human gene can produce a variety of alternative transcripts (or mRNA isoforms) (pp. 436–437 in Alberts et al. 2002), which differ in terms of their transcription initiation, splicing or polyadenylation patterns. Expression of alternative transcripts has been observed to be specific to tissue type or developmental stage. Disruptions in alternative transcript expression have serious consequences for an organism and are associated with numerous diseases, including cancer, multiple sclerosis, heart failure and neurodegenerative disorders.

Gene Definition and Alternative Splicing Methods

In BioSapiens (2007), the goals in the areas of gene definition and alternative splicing (BioSapiens-WP1 2007; BioSapiens-WP101 2007) are to study functional regions of genomes, in particular the genes, focusing on four main areas of investigation:

1. The basic gene structure (intron–exon structure)
2. The presence of differential gene structure (alternative splicing)
3. The evolution of gene structure
4. The alternative splicing process

Methods involve combining classical machine learning algorithms, theoretical studies of evolution and experimental techniques, and using genomes across all eukaryota where appropriate, but with a focus on mammalian and in particular human, mouse and rat genomes. As well as multigroup interactions, providing the crucial feedback loop between experiments and predictions, there is networking with outside groups; in particular, those with prediction algorithms that influence the understanding of functional gene content (e.g. signal peptide prediction and structural modelling of protein sequences), and with the thematic disease foci by providing in-depth analysis of genomic regions and genes of interest.

Alternative Transcription Goals

The Alternate Transcript Diversity (ATD 2007) project has investigated the mechanisms responsible for the formation of different alternative transcripts. These mechanisms are discussed extensively in the ATD (2007) project “literature” section, containing general references and project publications. All public deliverables are available on the website under “ATD data releases”. It is also anticipated that studies in the field of alternative transcripts will have direct applications for pharmaceutical industries. These applications include disease diagnosis or prognosis of risk patients, as well as identification of new drug targets. ATD (2007) is a follow-on project from the Alternate Splicing Diversity project (Stamm et al. 2006; Thanaraj et al. 2004).

Alternative Transcription Methods

ATD (2007) is a collaborative multidisciplinary project. It has comprehensively characterised alternative transcript forms throughout the human genome, and has assessed the differential expression of these forms in time and space, in normal and disease-related tissues. This was accompanied by quality control procedures, such as research for evolutionary proof through comparative sequence data analysis, between human and mouse. Further characterisation of alternative transcripts was implemented through activities such as identification of regulatory patterns, and derivation of expression states (i.e. expression specificity in terms of association with diseases, developmental stages, or tissue specificity). The project developed standard vocabularies and models that represent gene structures and their expression patterns. The validity of the bioinformatics prediction of disease-specific alternative transcripts has been examined through the execution of reverse

transcription polymerase chain reaction experiments on selected tissues. The discovery effort was accompanied by database integration, and also by dissemination to the scientific community.

Alternative Transcription Results

Some major results of ATD (2007) are summarised in two publications (Le Texier et al. 2006; Stamm et al. 2006), as follows:

- The creation of a unified ATD (2007) database integrating various information levels such as gene, feature variants, transcript variants, annotations, derived expression states, protein functionalities, results of experimental validations and associations with diseases. Fully developed query interfaces and toolboxes were available in the databases created by ASD (2007) and ATD (2007), which have been combined and upgraded to create the ASTD (2007) database.
- The definition of standards to represent gene structures and variants, and the creation of vocabularies for the representation of annotations.
- The confirmation of differentially expressed alternative transcripts in healthy and diseased tissues from human and mouse.
- The prediction of the regulatory motifs involved in alternative transcript formation.

Future Research

Traditional molecular biology approaches founded on a “one gene at a time” basis are no longer practical when detecting new disease-specific alternative transcripts. There is currently a need for the execution of genome-wide alternative transcript detection, followed by high-throughput analysis of transcript expression.

Gene Regulation and Expression

Gene Regulation and Expression Processes

Gene expression is the process by which the DNA sequence is transcribed into a gene product such as a protein or RNA. Several steps in the gene expression process may be modulated, including the transcription step and the post-translational modification of a protein. Gene transcription regulation gives the cell control over structure and function, and is the basis for cellular differentiation, morphogenesis and the versatility and adaptability of any organism.

DNA Microarray Data

A DNA microarray is a collection of microscopic DNA spots, commonly representing single genes, arrayed on a solid surface. DNA arrays are commonly used for expression profiling, i.e. monitoring expression levels of thousands of genes simultaneously. Software to store and analyse this type of data was developed in a sub-project within a major European Commission infrastructures collaborative project (TEMBLOR 2007) with four major components, one of which was DESPRAD (2007), which resulted in the development of ArrayExpress (2007).

Expression Research Goals

The BioSapiens (2007) goals for gene expression (BioSapiens-WP3 2007) have been:

- Development of methods and tools enabling the building of a human gene expression compendium characterising expression patterns of all genes in different tissues and cell types in different states by integration and analysis of data from a variety of sources, including testing and application of these methods
- Development and testing of methods for using gene expression and comparative genomics data for promoter prediction and analysis
- Development and evaluation of statistical and algorithmic methods for the analysis of gene expression data in the context of biological networks

This work addresses some of the major questions in modern biology. BioSapiens-WP3 (2007) deliverable DE3.4, “Documentation on DAS links to ArrayExpress (2007) [Au3] Data”, describes how a human gene expression compendium has been constructed by developing a protocol that links human gene sequences and their annotations with expression profiles. A major part of the implementation of this protocol has been the creation of a method that parses files describing the design elements comprising microarrays, with the presence of links to DNA, protein or model organism databases.

Expression Results

Several types of analysis have been performed, an example being given in BioSapiens-WP3 (2007) deliverable 3.2, “Documentation on comparative analysis of mammalian gene expression”. DNA microarrays were used to characterise gene expression patterns in skin biopsies from individuals with a diagnosis of systemic sclerosis with diffuse scleroderma, and these patterns were compared with those of gene expression seen in biopsies from normal unaffected individuals. The expression profiles of the transcription factor genes, and genes exhibiting correlated expression, were obtained from these microarray experiments (stored in the data

warehouse) that compared transcription patterns in organism parts and disease states. These analyses were designed to determine whether orthologous transcription factors control the expression of the same sets of genes, and in the same tissues, in both humans and mice. Additionally, the determination of the patterns of expression of human-specific transcription factors highlights divergences in basic biological processes between the two species. Furthermore, the identification of the misregulation of transcription factors in the various disease states may give a greater understanding of the molecular mechanisms underlying a disease.

Gene Regulation Research Goals

In BioSapiens (2007) gene regulation and expression studies (BioSapiens-WP2 2007; BioSapiens-WP3 2007; BioSapiens-WP102 2007), the main goals for regulation research are:

- Implementation and further development of novel sequence annotation tools for promoter analysis in the human genome
- Development of statistical methods and tools enabling the prediction of likely regulators for given genes or groups of genes
- Discovery of *cis*-regulatory modules in mammals, and the development and analysis of gene regulatory network models utilising gene expression data
- Development and testing of a similarity search engine for expression data repositories
- Improvement of predictive methods for higher organisms and interoperability of tools

Regulation Results

This work has already led to a number of tools and results, for example as reported in BioSapiens-WP2 (2007) deliverable De2.7, “Report on the utility of Web services for *cis*-regulation”. They are in the process of linking worldwide databases from the following sources:

- RSAT (2007)
- Ensembl (2007)
- T-Reg (2007)
- ArrayExpress (2007)
- STRING (2007)
- RegulonDB (2007)
- ORegAnno (2007)

Several important results have been obtained, for example those reported in BioSapiens-WP2 (2007), in De2.6, report on the analysis of multiple ChIP-chip

(chromatin immunoprecipitation on a DNA microarray chip) datasets in human. Results from the multiple ChIP-chip datasets produced within the ENCODE (2007) consortium are leading to an expanded understanding of the complexity of mammalian transcription. These experiments have defined approximately 4,500 transcription start sites within the ENCODE regions of the human genome. This is approximately 10 times the number of known genes in these regions and highlights the increasing complexity of mammalian transcription as described within the ENCODE project. This analysis of ChIP-chip data from multiple experiments has also demonstrated that the binding of transcription factors is symmetric around the transcription start site. This result will affect other assays for promoter regions that have traditionally concentrated only on the regions immediately upstream of the transcription start site. A key publication discussing the ENCODE (2007) and GENCODE (2007) work is Koch et al. (2007).

Systems Biology of Transcription and Regulation

Two new projects have begun, (see BaSysBio (2007) described in Chap. 4), which are taking a full systems biology approach and which will be making major contributions to data generation concerning transcription regulation and expression. BaSysBio (2007) has the goal of understanding of dynamic transcriptional regulation at global scale in bacteria. The European Transcriptome, Regulome and Cellular Commitment Consortium (EuTRACC 2007), will participate in the International Regulome Consortium (IRC 2007).

Functional Annotation of Proteins

Protein Sequence, Structure and Function Integration

Much of the bioinformatics structure for analysis of protein sequence, structure and annotation was established, extended or improved in the TEMBLOR (2007) project, via the European Molecular Structure Database (EMSD 2007), for protein structure, and Integr8 (2007) for integration of protein sequence and structure related data. EMSD (2007) provided the basis for the current MSD (2007), the EBI Macromolecular Structure Database, which is the European project for the collection, management and distribution of data about macromolecular structures, derived in part from the PDB (2007). A wide range of structure and sequence databases have been developed, along with many tools to infer protein sequence from gene sequence, and protein structure from protein sequence. A major effort has also been under way to comprehensively link these resources. Much of this effort has occurred in European collaborative programmes, as follows.

Sequence to Structure to Function Results

BioSapiens (2007) has built on this major infrastructure of databases and tools to mount a major annotation programme for proteins (BioSapiens-WP5 2007; BioSapiens-WP7 2007; BioSapiens-WP9 2007; BioSapiens-WP104 2007). This work has been focused on the integration of methods for the construction and the validation of three-dimensional models of protein structures. The models incorporate confidence values both at the protein and at the residue level and internal quality checks. These predictions, based on the family analysis and integrated with the predictions of integral membrane proteins, were channelled to the other participants for structure-based functional annotation. Functional information gained from structure analysis is highly complementary to that obtained by high-throughput experimental, sequence-based, or genomic context methods. Despite the fact that the relationship between structure and function is a central problem in molecular biology, and thus is critical for protein engineering and drug design, there are only a few methods able to generate function predictions from the analysis of protein structures. As a consequence, many proteins with known structure are not yet functionally characterised. Methods are being established for fully automatic inference of structure from function, aiming at the identification and characterisation of functional regions in proteins. An integrated Web resource was implemented at EBI (2007). Contributions were provided by the connections to the Web server implementations of their corresponding methods by the participating groups. The results involve:

- Combining very different methods into a working pipeline
- Comparing and benchmarking the predictions to obtain a combined approach
- Contributing to the annotation of binding (and specificity) sites in protein models

Functional Sites Results

An example result for determining functional sites is found in BioSapiens-WP9 (2007) deliverable De9.14, “A Web tool for the prediction of residues of functional specificity from multiple alignments”. TreeDet (2007) (Carro et al. 2006) predicts evolutionary importance and functional sites in protein families. The server integrates the results of three separate methods for the prediction of residues of functional interest in protein families. These tree-determinant methods are based on the relation between sequence conservation and evolutionary importance and include a tree-based method, a correlation-based method and a method that employs a principal component analyses coupled to a cluster algorithm. Accurate alignments are crucial to the prediction of tree-determinant residues and for that reason a tool for the evaluation of alignment reliability (SQUARE) has been included in the package.

Small-Ligand Binding

Some of the major problems related to metabolism and drug development are related to the binding of small ligands to proteins. An example of the important results being obtained in this area is found in BioSapiens-WP9 (2007) deliverable 9.9, “New methods for characterising ligand binding sites”. Stockwell and Thornton (2006) observed that the phenomenon of molecular recognition, which underpins almost all biological processes, is dynamic, complex and subtle. They presented an analysis of the conformational variability exhibited by three of the most ubiquitous biological ligands in nature, ATP, NAD and FAD, and demonstrated qualitatively that these ligands bind to proteins in widely varying conformations, including several cases in which parts of the molecule assume energetically unfavourable orientations. Several other results are presented that are fundamental to structure to function interpretations concerning proteins and ligands.

Future Plans

BioSapiens’s (2007) plans for the future aim at establishing methods for fully automatic inference of protein function. The main goal will be the identification and characterisation of functional regions in proteins. An integrated Web resource will be provided through the BioSapiens (2007) portal, through the DAS (2007) protocol or by Web services. The objectives are:

- To develop tools to improve the classification of proteins, using sequence and structure information, into protein domain families
- To improve methods for modelling protein structures from sequence and to develop quality indicators for different structures.
- To develop new methods for functional annotation from sequence and structure
- To make all the knowledge generated available through the BioSapiens (2007) portal utilising DAS (2007) where appropriate or Web services

Post-translation Modification, Membrane and Localisation Prediction

Membrane Proteins and Results

An important basis for the work of BioSapiens-WP6 (2007) on membrane proteins was established by annotation of integral membrane proteins, in terms of function, subcellular localisation and topology/structure. New methods were developed that used experimental and theoretical results from widely different studies to enhance

the transmembrane topology prediction. New molecular-class specific information systems were created to better present the results to bioscientists. An example of the results is found in BioSapiens-WP6 (2007) deliverable De6.9, "Integration of the Ensembl 2.0 predictor into the TRAMPLE (2007) transmembrane protein labelling environment", which describes the results of joint work on the task of annotating *in silico* all the available sequences of the human genome according to the UniProt (2007) database previously selected. A new integrated and browsable database was developed that improves the previously developed TRAMPLE (2007). The new environment/Web server/DAS (2007) server is called PONGO (2007), a Web server for multiple predictions of all- α transmembrane proteins (Amico et al. 2006). It is based on a relational database containing all the data generated. The local DAS (2007) annotation server is resident on the same machine. The results of the effort involve the inclusion of new methods for the *in silico* annotations of transmembrane predicted regions. The user is able to trace for each UniProt (2007) sequence whether the protein is or is not endowed with a signal peptide, whether the sequence is or is not a membrane protein, and its putative topology, as computed by six different predictors, two of which are newly included in the latest version of the Web server. This allows users to compare among different predictors at the same time and assess whether the expected results are in agreement with their own experimental findings. Alternatively, different predictions, especially when in agreement, may enforce the expectation that a given chain is a membrane protein and in this case the putative topology may help in designing experiments in order to validate (or not) the number of transmembrane helices and the location of the N and C termini of the protein with respect to the plane of the membrane. This may be particularly useful when the chain has no homologous counterpart in the database of sequences and may help in highlighting also its function.

Post-translation Modification and Localisation and Results

In BioSapiens-WP8 (2007) work on post-translation modification and localisation, the objectives were to predict protein features, in particular post-translational modifications, localisation signals, and to use combinations of such features to predict cellular role and molecular function for proteins without sequence similarity to proteins of known function. Datasets were constructed and verified for data-driven prediction algorithms, and were made publicly available. One of the tools developed is described in BioSapiens-WP8 (2007) deliverable De8.8, "A neural network based method for prediction of protein localisation to the nucleolar proteome". The nucleolus is the most prominent substructure of the nucleus. To predict nucleolar protein localisation, different data sources were integrated using a semiautomated neural network scheme which was later used to assign and rank nucleolus proteins to highly connected protein complexes. The procedure has an exploratory part and an evaluation part. The exploratory part consists of protein interaction database mining—building interaction complexes with a compiled list of the human nucleolus

proteome as “seed.” Subsequently, a “reverse” proteomics step is implemented in order to gain experimental evidence of nucleolar localisation of proteins that were not in the seed list but that are predicted to be nucleolar on the basis of their presence in the *in silico* generated high-confidence protein complexes. As the evaluation part of the procedure, a machine-learning method was constructed to produce a score indicating the likelihood that a given *in silico* generated complex is nucleolus-localised. The full list of nucleolus proteins from the top 15 complexes can be found at NUCLEOLUS (2007), and for details, see Hinsby et al. (2006) A picture of the human nucleolus is shown in Fig. 2.4.

In further development of this work, Lage et al. (2007) combined protein–protein interaction data and text mining for disease gene finding in novel ways. They performed a systematic, large-scale analysis of human protein complexes comprising gene products implicated in many different categories of human disease to create a phenome–interactome network. This was done by integrating quality controlled interactions of human proteins with a validated, computationally derived phenotype similarity score, permitting identification of previously unknown complexes likely to be associated with disease. Novel candidates were proposed as being implicated in disorders such as retinitis pigmentosa, epithelial ovarian cancer, inflammatory bowel disease, amyotrophic lateral sclerosis, Alzheimer disease, type 2 diabetes and coronary heart disease.

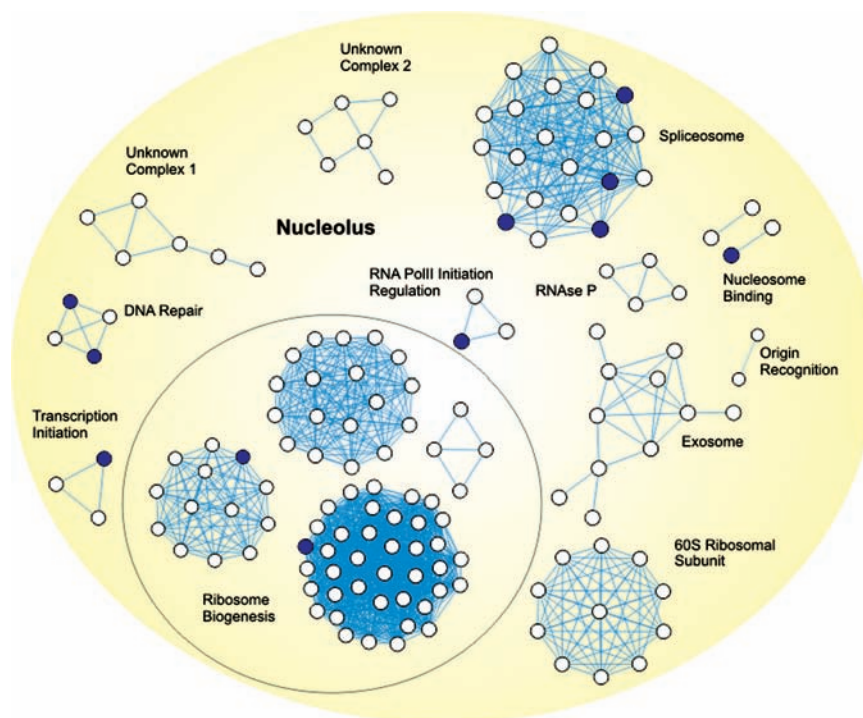


Fig. 2.4 A wiring of the human nucleolus. (NUCLEOLUS 2007); Hinsby et al. 2006)

Future Post-translation Modification and Localisation

Future BioSapiens-WP105 (2007) plans are to predict protein features, in particular post-translational modifications and localisation signals, and to use combinations of such features to predict cellular role and molecular function for proteins without sequence similarity to proteins of known function. It is planned to construct and verify datasets for data-driven prediction algorithms, and to make these publicly available. The work package also includes close link with experiments. A new computational model for transmembrane helices in mammalian proteins will be developed. A Web server that predicts the membrane insertion free energy of peptide segments will be constructed.

Protein Complexes, Networks and Pathways

Protein-Protein Complexes

In BioSapiens-WP10 (2007), the objective is the automatic identification, prediction and analysis of protein interaction partners. The availability of genome sequences and high-throughput biology enables fundamentally different approaches for function prediction. A major resource developed is described in BioSapiens-WP10 (2007) deliverable 10.5, “Integration of experimental data sources into STRING (2007)”. The database provides a cross-species protein-protein network of functional interactions. Until recently, however, it only included predicted interactions based on so-called genomic context methods, which greatly limited its usefulness for the analysis of eukaryotic proteomes. BioSapiens deliverable 10.3 describes a unified scoring scheme for experimental and computational protein-protein interactions. STRING (2007) now integrates and scores physical protein-protein interaction evidence from five different databases. The scoring scheme has now been extended to also cover microarray expression data. The latter have been implemented in the form of a Web server called ArrayProspector, which provides microarray-based evidence for STRING (2007). The resulting functional interactions have been benchmarked against the same reference set used for all other evidence types in STRING (2007) to make evidence of different types directly comparable.

Network Prediction

In the area of networks where bioinformatics starts to merge with systems biology, BioSapiens-WP11 (2007) describes moving the field of network prediction and analysis to a status that allows everybody in the community without much effort to

exploit the knowledge in the field, with the specific focus of establishing the technology for robust annotation based on network and pathway information for complete genomes. Protocols are described for the combination of the information stored in existing databases on pathways. Procedures are explored for the integration with the pathway information of the network predictions (and the corresponding quality controls). Methods for detailed analysis of the networks, leading to biological discoveries and final functional annotations, are developed.

Metabolic Pathway Net

An important application of this method is shown in BioSapiens-WP11 (2007) deliverable 11.3, “Prediction and annotation of a metabolic net in a model organism”. Some bacteria, yeasts, plants, mice, rats and humans utilise the methionine salvage pathway. In this pathway, organic sulphur is salvaged from methylthioribose, which is derived from the methylthioadenosine that is a by-product of the synthesis of spermidine and spermine. This pathway regenerates reduced sulphur and metabolically links it to polyamine biosynthesis, but details of the physiological roles of this pathway remain obscure. The STRING (2007) database employs two different strategies for transferring known and predicted associations between organisms: the first (“COGmode”) relies on externally provided orthology assignments and transfers interactions in an all-or-none fashion, whereas the second (“protein mode”) uses quantitative sequence similarity searches and often distributes a given interaction fractionally among several proteins of the target organism. With use of protein mode, a network of functional associations was derived for the *Bacillus subtilis* methionine salvage pathway, and for other organisms, based on three types of genomic context evidence. Multiple types of evidence support several of the relations, and they include additional proteins, which are likely to play a role in methionine salvage.

Future Pathway Work

BioSapiens-WP106 (2007) integrates known and predicted protein–protein interactions from a number of databases and prediction methods from the various partners. To be able to do this at a genome-wide scale with the highest possible quality, automation and critical evaluation and scoring of the individual sources are the key principles in this work. While previous work focused on the development of common platforms and capturing of knowledge from existing databases, future activities make use of these resources for predictions and the annotation of more probabilistic features, e.g. the prediction of a missing enzyme or the regulator or transporter with which a pathway is associated. Exploratory work is carried out to capture other cellular processes.

Encyclopaedia of DNA Elements

Functional Elements in the Human Genome

The Encyclopaedia of DNA Elements (ENCODE 2007) was launched in September 2003 by the National Human Genome Research Institute (NHGRI 2007) of the National Institutes of Health (NIH 2007). The goal is to identify all functional elements in the human genome sequence. The pilot phase aims at analysing defined regions of the human genome sequence using existing testing methods and close interactions between computational and experimental scientists. Regions representing approximately 1% (30 Mb) of the human genome have been chosen and were analysed by ENCODE consortium researchers. Fourteen regions were chosen because they were regions of special interest and 30 more regions were chosen randomly from cluster regions that were grouped according to non-exonic conservation and gene density. Whereas this book concentrates mostly on European collaborative research, ENCODE (2007) is an excellent example of a broad collaborative effort based in the USA, with a wide range of international partners (ENCODE-Participants 2007) who are funded, and others with whom major informal collaborations occur at the project level.

International Collaboration

The contribution of BioSapiens (2007) to ENCODE (2007) is vital, especially with the wide range of tools available within the BioSapiens (2007) European Virtual Institute for Annotation. GENCODE (2007) is a BioSapiens-WP20 (2007) and BioSapiens-WP108 (2007) subproject which is associated with ENCODE, which seeks to identify all protein-coding genes in the ENCODE (2007) selected regions.

Functional Identification Methods

Important advances have been made in functional identification by using the full power of the BioSapiens (2007) network for ENCODE (2007), as shown in BioSapiens-WP20 (2007) deliverable De20.1, “BioSapiens–ENCODE collaboration”, containing a report describing the status of the work performed. Data are generated by the means shown in Table 2.3, where functional genomic elements are identified. The methods indicated are being used to identify different types of functional elements in the human genome. For each protein-coding gene, the delineation of a complete mRNA () sequence is performed for at least one splice isoform, and often for a number of additional alternative splice forms. Coding sequences for the 44 regions in the study

have been ascertained by the Human And Vertebrate Analysis and Annotation (HAVANA 2007) group. In total there are 1,097 coding sequences from the 44 selected regions of the human chromosome. The contributions from the BioSapiens (2007) partners were focused on information from a protein annotation perspective so that, where possible, annotations can be viewed from all groups simultaneously through DAS (2007) servers. Special attention is given to the potential aspect of alternative splicing and the putative effect it has on function by altering domain, structure, localisation and post-translational modification.

Genome Analysis Future

The groups participating in ENCODE (2007) plan to cover 100% of the human genome. BioSapiens (2007) has been participating in the process and the deliverables from this work package will be tailored to the final plan adopted by the ENCODE (2007) partners. A major task is to enable scaling of the protein analysis for full coverage of the human genome, including all the isoforms. The BioSapiens (2007) consortium is particularly interested in the experimental verification of the translation of these genes into proteins.

Future work includes:

- Gene mapping of the 434 gene loci in the set
- Assignment of UniProt sequences, PDB templates, Gene Ontology terms and Pfam domains to the 1,097 sequences in the set
- Comparison of the gene/variants from the randomly and manually selected regions
- Detailed study of a large number of examples from the set in respect of their function and relationship to disease
- Comparison of the supporting evidence for the most interesting splice variants
- Study of the TRANSFAC (2007) sequences from the set

Table 2.3 Indicated methods being used in ENCODE to identify functional elements in the human genome. (From ENCODE-Project-Consortium 2007)

Feature class	Experimental techniques
Transcription	Tiling array, integrated annotation
5' ends of transcripts	Tag sequencing
Histone modifications	Tiling array
Chromatin structure	Quantitative PCR, tiling array
Sequence-specific features	Tiling array, tag sequencing, promoter assays
Replication	Tiling array
Computational analysis	Computational methods
Comparative sequence analysis	Genomic sequencing, multisequence alignments, computational analysis
Polymorphisms	Resequencing, copy number variation

Major Result: Most DNA Is Transcribed to RNA

A major paper describing the work of the NIH (2007) funded ENCODE (2007) consortium and a number of accompanying papers have been published by the ENCODE-Project-Consortium (2007). The BioSapiens (2007) results are available at GENCODE (2007), see Tress et al. (2007). The findings of ENCODE-Project-Consortium (2007) promise to reshape our understanding of the functioning of the human genome. They challenge the traditional view of our genetic blueprint as a tidy collection of independent genes, pointing instead to a network in which genes, regulatory elements and other types of DNA sequences interact in complex, overlapping ways. In an analysis effort led by the European partners, the ENCODE (2007) consortium's major findings include the discovery that the majority of human DNA is transcribed into RNA and that these transcripts extensively overlap one another. This broad pattern of transcription challenges the long-standing view that the human genome consists of a small set of discrete genes, along with a vast amount of "junk" DNA that is not biologically active. The new data indicate that the genome contains few unused sequences; genes are just one of many types of DNA sequences that have a functional impact. These discoveries are fundamental to the future course of biomedical research.



<http://www.springer.com/978-3-540-78352-7>

Bioinformatics and Systems Biology
Collaborative Research and Resources

Marcus, F.

2008, XXVI, 287 p., Hardcover

ISBN: 978-3-540-78352-7