Protein Annotation at Genomic Scale: The Current Status

Dmitrij Frishman*

Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenchaftszentrum Weihenstephan, 85350 Freising, Germany, and Institute for Bioinformatics, GSF-National Research Center for Environment and Health, Ingolstaedter Landstrasse 1, 85764 Neuherberg, Germany

Received December 22, 2006

Contents

Dmitrij Frishman received a M.S. in Biomedical Electronics from the Saint Petersburg Electrotechnical University in 1984 and a Ph.D. in Biochemistry from the Russian Academy of Sciences in 1991. An Alexander von Humboldt Research Fellowship he received at the end of 1991 allowed him to join the Pat Argos group at the Biocomputing Department of EMBL in Heidelberg, where he pursued postdoctoral training in structural bioinformatics until 1996. He subsequently joined the Munich Information Center for Protein Sequences as a senior scientist and later became Deputy Director of the Institute for Bioinformatics at the German Research Center for Health and Environment. In 1997, he cofounded a bioinformatics company called Biomax Informatics AG, which provides computational solutions for better decision making and knowledge management in the life science industry. Since 2003, Dmitrij Frishman has been Professor for Bioinformatics at the Technical University of Munich. His current research interests focus on genome annotation, prediction and analysis of protein interactions, and structural genomics.

1. Introduction: From Data Deluge to Information Desert

Genomic sequences are being deciphered at an unprecedented pace, and the demand for sequence data is also continuously growing, fuelled by thrilling potential applications which range from personalized genome-based medicine and targeted cancer therapies to microbial strain optimization and bioterrorism prevention. Sophisticated sequencing procedures ultimately result in plain text files in which the DNA molecules, the carriers of the precious code of life, are represented by endless strings of the characters A, C, G, and T. These sequences are entirely incomprehensible and all but useless unless meaningful biological facts are associated with them in the course of genome annotation. The Webster dictionary defines annotation as "a note added by way of comment or explanation". In molecular biology databases, such notes typically contain information about the cellular role and mechanism of action of genes and their products. Throughout the 1980s and most of the 1990s, the biological community critically relied on high-quality protein annotation produced by relatively small groups of enthusiasts at the PIR,[1] Swiss-Prot,[2] and DDBJ[3] databanks in a process involving careful analysis of experimental facts published in the literature as well as bioinformatics analyses by highly

* Address correspondence to the author at Technische Universität München.

10.1021/cr068303k CCC: $65.00
Published on Web 07/21/2007

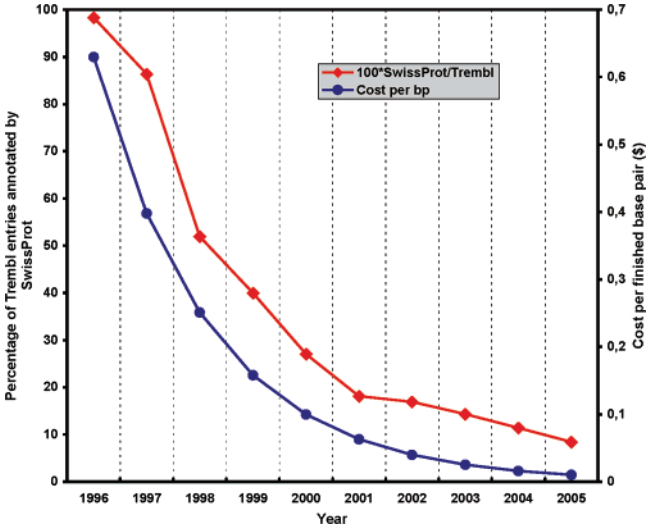

Figure 1. Cost per finished base pair[239] and percentage of manually annotated sequences. The latter is estimated as the ratio of the number sequences in the SwissProt database and all known sequences available in the Trembl database.

experienced staff. This time-consuming process resulted in invaluable datasets which represent the core of today's protein knowledge base.

At the onset of the genomic era, essentially every new protein sequence determined received the attention of human experts and was annotated to the maximally possible extent. While planning early genome sequencing projects, typically 10% of the budget was allocated for bioinformatics, including careful curation of data. In the past 10 years, the speed of sequencing has continuously grown while the total number of computational biologists directly involved in manual curation of molecular data has increased only insignificantly. As illustrated in Figure 1, the dramatic fall of sequencing prices is accompanied by the decrease in the percentage of annotated proteins from nearly 100% only a decade ago to less than 5% now. Assuming that one needs on average roughly 30 min to assess published facts and bioinformatics evidence for one protein, one thousand annotators would have to work 1 year long, 8 h a day to annotate all 5 million sequences that are currently known. However, since the size of the protein database has been consistently doubling every 18 months, the moving target of annotating all proteins will never be achieved. On a more anecdotal vein, according to some conservative estimates,[4] the total number of proteins on Earth is in excess of 10^{10}. I do not want to speculate what kind of human resources one would need to analyze all these proteins by hand.

A new generation of superfast and ultralow-cost DNA sequencing technologies (reviewed by Metzker)[5] is expected to have the throughput of hundreds or even thousands of megabases per day.[6] The National Human Genome Research Institute issued a request for applications, seeking to further reduce the cost of sequencing mammalian genomes by 4 orders of magnitude: from tens of millions of dollars today down to merely a thousand dollars.[7] These new technologies also require much more modest in-house resources than current state-of-the art techniques, and they will render sequencing entire genomes accessible to relatively small institutions or even individual research groups which will not be able to invest significant time and resources into manual curation of the resulting data. These future advances will convert genome sequencing into an easily accessible routine lab technique and will often result in a situation where sequence data do not even get submitted to the central repositories.

In addition to the quantitative challenge outlined above, we are increasingly facing a qualitative challenge: for most of the newly determined proteins, there is simply no experimental evidence to annotate, as they have never seen a test tube. No primary experimental information is available, nor are these proteins described in journal articles, and any functional inference for them can only be made by homology-based transfer of annotation. Even for the best studied organism, *Escherichia coli*, experimental information is available for only 54% of the gene products.[8] Instead, we are observing a surge in the quantity of high-throughput experimental data available for entire proteomes. For example, the number of proteins for which some information on protein−protein interactions is available is approaching 100 000. Individual biochemical experiments are phased out by mass measurements often plagued by artifacts and noise. In contrast to the detailed and information-rich results of a classical biochemical study, high-throughput experiments typically deliver a single measured value for each protein under study which cannot be easily interpreted in isolation.

To make things worse, there is even no agreement on the actual number of genes in a completely sequenced genome. Especially for complex genomes, the set of genes predicted for a particular chromosome is still substantially dynamic and subject to constant change as cDNA and EST databases grow, more experimental data becomes available, and automatic gene predictors get better. For example, in 2003, Collins et al.[9] revised their own annotation of 198 out of 936 genes (including noncoding and pseudogenes) of the human chromosome 22 produced 4 years earlier.

The main difficulty of writing a review about genome annotation lies in finding an appropriate scope. Today's bioinformatics is essentially genome informatics, and virtually any new method, database, or biological discovery is relevant for analyzing genomic data. Here, I mainly focus on modern procedures used for manual genome annotation and its current status as well as computational methods and software infrastructure for analyzing protein function at the genomic scale (Figure 2). I also discuss possible ways to reduce the error level of automatically generated annotation both by experimental means and by using machine intelligence. There are many important aspects of genome analysis that could not be addressed here but have been extensively reviewed elsewhere. Problems in gene prediction are summarized only briefly, with the main focus on improving the quality of functional inference; for in-depth coverage of computational gene finding, the reader is referred to specialist publications (see below). Excellent accounts on text mining,[10] pseudogenes,[11] alternative slicing (Artamonova and Gelfand, this volume), finding microRNAs and their targets,[12,13] gene regulation,[14,15] chemical approaches in protein function annotation,[16] genome comparison,[17] and biological data integration[18] exist.

2. Gene Modeling: Still an Open Problem

Gene prediction is a crucially important, quality determining stage in genome annotation. Missed genes translate into gaps on reconstructed metabolic maps as well as misinterpreted microarray experiments, while missed or overpredicted exons and fused genes complicate interpretation of sequence

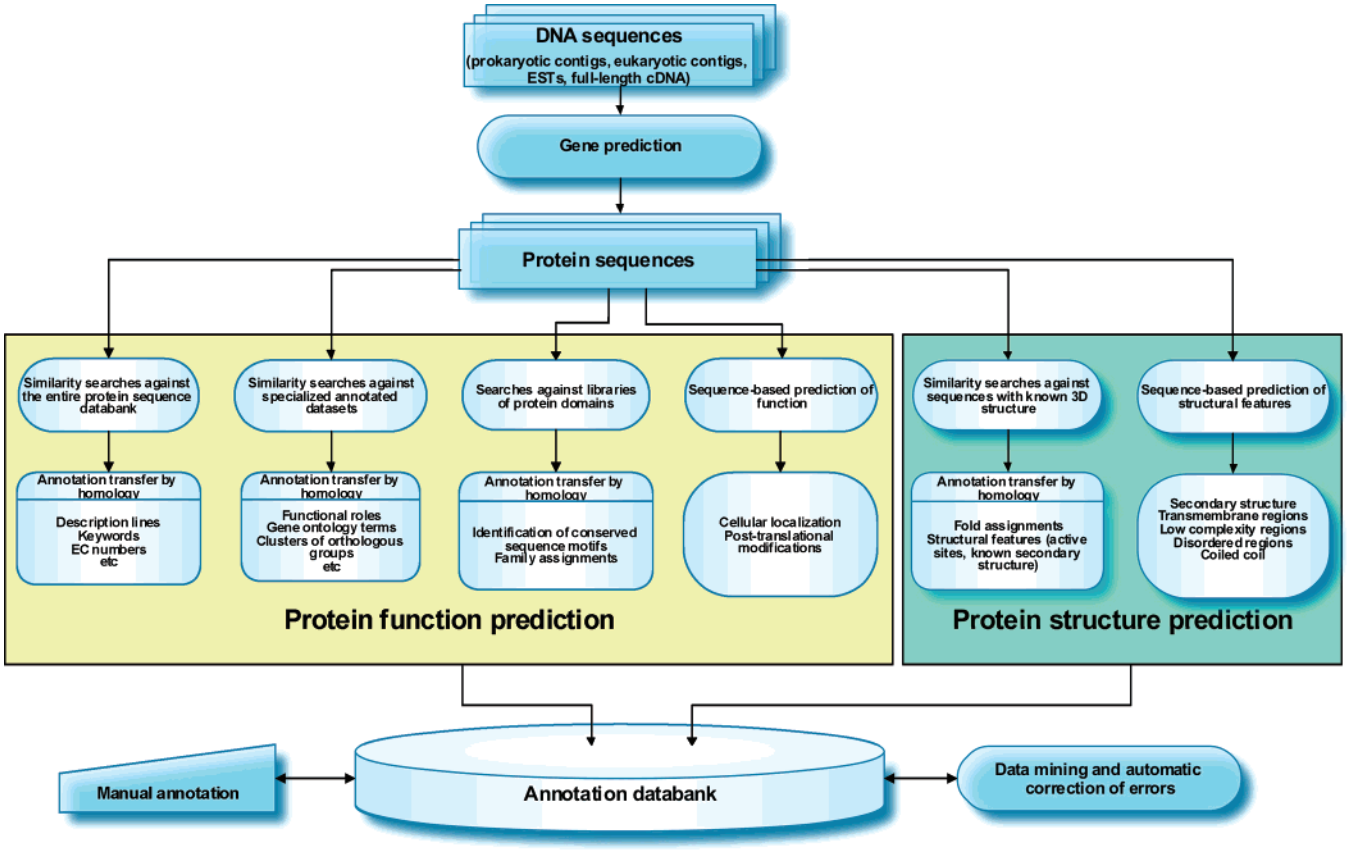

Figure 2. General overview of the genome annotation process.

alignments, delineating functional specificity of proteins, and building structural models.

Remarkable advances in eukaryotic gene prediction accuracy achieved in the past decade can be easily seen by comparing evaluations of state-of-the art methods made 10 years ago[19] and just recently.[20] Specificity and sensitivity at both exon and nucleotide levels have been substantially improved, especially due to the application of novel strategies based on alignment of closely related genomes.[21] Besides the use of dual genomes, important sources of information for eukaryotic gene prediction are expressed sequences (EST and cDNA). However, even the best currently available automatic tools are still able to accurately define complete structures for only 40% of human genes. Gene prediction in prokaryotes is an easier, but in no way completely solved, problem. Nielsen and Krogh[22] estimated that 60% of prokaryotic genes in public databases have wrong predicted starts (another estimate − 40% (Mark Borodovsky, private communication)) and that 30% of all genomes are overannotated by more than 5%, while nearly 8% are underannotated, particularly due to difficulties in correctly detecting short genes. A detailed review of modern computational methods to predict genes from DNA sequences is beyond the scope of this article and has been done effectively by specialists in the field.[23−25] Here, we focus on selected aspects of gene prediction in the context of functional annotation of genomes (see Figure 3).

Given that six out of every ten predicted genes in complex eukaryotic organisms contain errors, high-quality gene models can only be produced by hand curation and improvement of results generated by a host of carefully selected automatic algorithms. This tedious and time-consuming task involves resolving conflicts between alternative gene models by referring to literature sources, identifying sequence motifs and domains, and incorporating supporting evidence derived from alignments with EST, cDNA, and reference gene structures, if available. The first experience in manual improvement of predicted gene structures was accumulated while annotating the *Saccharomices cerevisiae* genome, in which only slightly more than 200 genes have introns. The task is incomparably more complicated for complex multiexon metazoan genes. Human genes models are being actively annotated by a powerful consortium including the Sanger Center, RIKEN, the Joint Genome Institute, Genoscope, and the Washington University Genome Sequencing Center, which makes them available via the Vertebrate Genome Annotation Database (VEGA).[26] A growing body of manually annotated gene models for mouse, dog, pig, and zebrafish is also available via VEGA. Among other model genomes with comprehensively curated gene structures are *Caenorhabditis elegans*,[27] *Arabidopsis thaliana*,[28] and *Drosophila melanogaster*.[29] In general, systematic maintenance and improvement of gene models for eukaryotic genomes can only be implemented within large-scale annotation efforts such as Ensembl.[30]

There are two ways to cope with the gene modeling challenge at large scale: either by intelligent software-based decision support or through increasing the number of contributing scientists by involving the broad biological community into genome annotation. Software tools have been developed that attempt to automate the process of combining gene calls produced by different algorithms. GeneComber[31] selects the most reliable coding regions predicted by Genscan[32] and HMMgene[33] by applying simple logical rules to exon probabilities calculated by these methods. A more sophisticated approach, JIGSAW,[34] requires a set of trusted

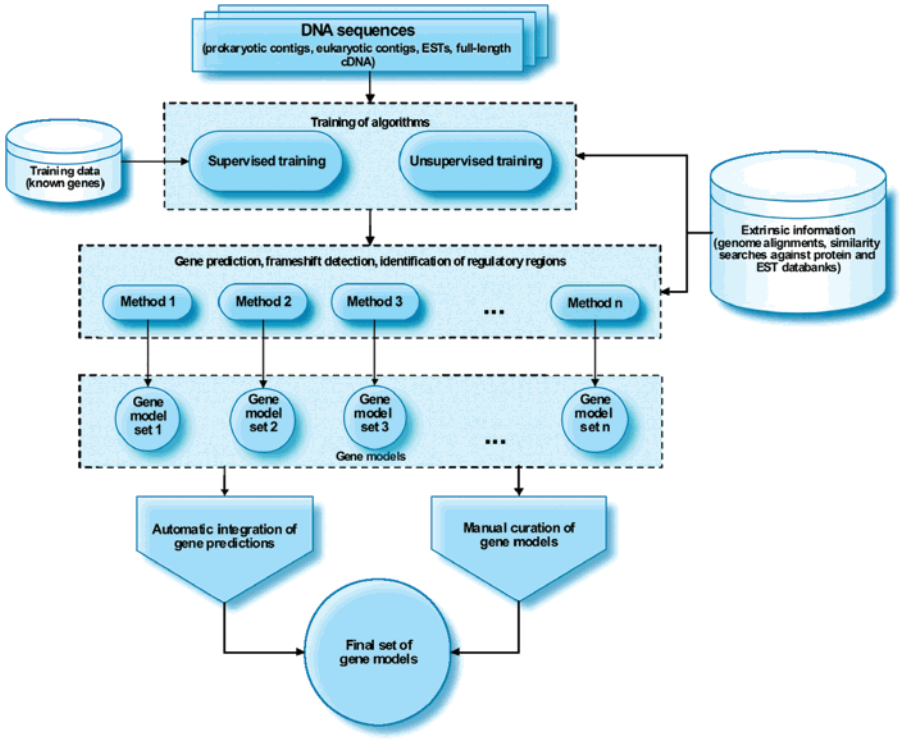

Figure 3. Typical gene prediction pipeline.

genes for a given organism to be available. Based on these training data it assigns relative weights to each genomic feature and uses dynamic programming to integrate evidence from any given number of sources, including *ab initio* and similarity-based prediction techniques, known gene indices, cDNA alignments, and predicted splice sites.

On the other hand, in view of the current rate of data generation, some researchers advocate wide involvement of the biological community as a promising way of augmenting the depth and quality of genome annotation.[35] Although this approach has its intrinsic problems, such as potential lack of consistency, it also has the advantage of attracting highly motivated biologists willing to share their specific problem domain expertise not easily accessible to professional curators. While general purpose software solutions to support community-wide genome annotation have been available for quite some time (see below), specialized resources for collective and distributed modeling of gene structure are now beginning to emerge. For example, yrGATE (Your Gene structure Annotation Tool for Eukaryotes)[36] allows annotators to access over the Internet precalculated exons, evaluate supporting evidence (such as EST alignments), introduce custom evidence, edit the DNA sequence to correct errors, and submit their annotation, including user-defined exons, for community review.

Another major bottleneck in high-throughput gene finding lies in the fact that most of the individual gene recognition algorithms require a training set of known genes. Compiling such experimentally validated sets is a nontrivial task even for highly annotated model genomes, and it may not be possible at all for less studied genomes. In order to circumvent this problem, self-training algorithms for prokaryotes attempt to extract reliable open reading frames that either are confirmed by alignments with known proteins[37] or are long enough to make the probability or their occurrence by chance very low.[38,22] GeneMarkS[39] finds bacterial gene starts using an unsupervised training procedure which involves iterative cycles of gene prediction and detection of ribosome binding sites in gene upstream regions by Gibbs sampling. In a significant recent advance, a self-training *ab initio* gene prediction technique for eukaryotes has been developed.[40] After initially estimating parameters of Markov chains based on compositional features of the DNA sequence or sufficiently long Open Reading Frames (ORFs), the algorithm proceeds by systematically refining the set of trusted genes and updating the HMM architecture until convergence.

In many practical applications of genome sequencing, it is desired to quickly obtain low coverage genome assembly. Errors in DNA sequences can be efficiently identified by evidence-based gene predictors that detect frameshifts and in-frame stop codons based on spliced alignment with EST and cDNA.[41,42] *Ab initio* framshift finding, especially in eukaryotic genomes, is a much more difficult task, and most of the gene predictors cannot handle sequence errors. The FrameD algorithm[43] allows for detecting and correcting frameshifts in prokaryoric genomes and EST sequences by explicitly modeling deletions and insertions as additional edges on a directed acyclic graph on which every path represents a possible gene prediction. In ESTscan,[44] expected probabilities of insertions and deletions are utilized to build an error-tolerant hidden Markov model of coding regions in EST. Overall, accurate gene finding in sequence data obtained by low coverage sequencing as well as in short contigs and ESTs of poor quality still represents a significant challenge.

In summary, in addition to the obvious need for more accurate prediction algorithms which, in particular, should be capable of handling incomplete and low-quality sequence data, further progress in obtaining the complement of protein-coding genes of acceptable quality will require better models of human logic created by intelligently combining various computational methods, an increase in the degree of automa-

Table 1. Selected Manually Curated Model Genomes

organism	URL	ref
Escherichia coli	ecocyc.org	49
Bacillis subtilis	genolist.pasteur.fr/SubtiList/	231
Saccharomices cerevisiae	mips.gsf.de/genre/proj/yeast	50
	yeastgenome.org	45
Dictyostelium discoideum	dictybase.org	232
Caenorhabditis elegans	wormbase.org	233
Drosophila melanogaster	flybase.org	234
Mus musculus	www.informatics.jax.org	235
Rattus norvegicus	rgd.mcw.edu	236
Homo sapiens	vega.sanger.ac.uk/Homo_sapiens/index.html	26
Arabidopsis thaliana	mips.gsf.de/proj/thal/db	28
	www.arabidopsis.org	237

tion by employing self-training algorithms, as well as stronger reliance on the broad support of the biological community.

3. Manual Annotation of Genomes

3.1. Islands in The Sea: Manual Annotation of Model Genomes

In spite of the growing maturity and usefulness of automated bioinformatics and linguistic methods, reliable high-quality annotation of sequence data can only be produced by careful manual curation. In general, only a human expert can draw from the biomedical literature and appropriately represent experimentally determined functional information. Manual annotation remains the only way to make the crucial step from merely displaying the description line of the best database hit for a given protein, as is done by automated methods, to formulating its precise functional role, carefully documenting data origin, validating computer-generated information, and assigning confidence levels to various pieces of evidence.

Scientific curators of genomes have been compared to museum curators who systematically collect and display information on art objects.[45] In addition to encyclopedic knowledge of biology combined with bioinformatics expertise, they are also required to have strong communication and even diplomatic skills, allowing them to preserve neutrality while resolving conflicts that may arise between members of the scientific community centered around a certain model organism. Full-time genome annotators in large organizations, such as TIGR, EBI, or MIPS, acquire a special professional status which, in contrast to standard scientific careers, stresses teamwork toward achieving the highest possible data quality over individual ambitions. Annotation teams strive to improve data quality and consistency by developing standard operational procedures and rules for capturing and describing data in a particular knowledge domain, often summarized in publicly available manuals (see, e.g., ref 46).

Most of the past and current manual annotation efforts focus on a limited number of selected model genomes (Table 1). Beyond sequence data, model organism databases cover a broad range of biological aspects, including metabolic and signaling pathways, transcription units, known protein structures, computational models, and disease and phenotype information. With the advent of high-throughput omics technologies, annotation groups are actively incorporating datasets from microarray, RNAi, two-hybrid, SAGE, and other experiments and linking them to functional annotation.

Model genome projects represent community efforts and are usually characterized by close collaboration between bioinformaticians and biologists.[47] For example, over 100 researchers continually provide updates to the *Pseudomonas* Genome Database,[48] typically without solicitation. The EcoCyc database[49] formed an advisory board consisting of leading scientists to guide its work and partners with outside experts who help annotate individual cellular systems. The Saccharomices Genome Database (SGD) is getting 50 user mails a week reporting new data and inaccuracies in the existing annotation.[45] It also cooperates with remote curators who spend two weeks a year at SGD, where they take part in hands-on practical work, meetings, and discussions with the core staff.

Collins et al.[9] noted that a detailed annotation of 1% of the human genome, including bioinformatics analyses, manual curation, and experimental verification, took 6 person years to complete. However, there is little doubt that in a reasonable time frame the entire human genome will be annotated to the highest possible standard of quality due to its crucial importance. Significant progress has been made in manual annotation of smaller microbial genomes sequenced a decade ago. Just recently (September 2006), the EcoCyc team reported that it has manually curated all *Escherichia coli* gene products. For 3557 out of 4449 *E. coli* genes, summaries written by human experts are available based on over 14 000 literature citations, while for the remaining genes no literature references are available. The SGD database has sufficient staff to incorporate all yeast-related references as they appear.[45]

Starting from an initial seed annotation of the most established model genome within a given community, quick expansion is taking place by adding further closely related species and thus leveraging on already available annotation and expertise of highly specialized scientific curators. One of the prominent examples of this approach is the MIPS fungal genome database (mips.gsf.de/projects/fungi), which started more than a decade ago as the central resource for *Saccharomices cerevisiae*[50] and then added careful manual annotation of three further completely sequenced fungal genomes, *Neurospora crassa*, *Fusarium graminearum*, and *Ustilago maydis*, which, combined with automatic annotation of all 38 sequenced fungal genomes, represents a comprehensive comparative fungal genome resource. In a similar fashion, both MIPS and TIGR are expanding their plant databases from the core *Arabidopsis thaliana* dataset to multiple species including maize, *Medicago truncatula*, *Lotus japonicus*, rice, tomato, wheat, potato, and others.

Overall, the total number of model genomes that are being actively curated by hand is probably on the order of 30, while the Genomes Online Database[51] currently lists over 2000 genome sequencing projects. The smaller percentage of sequences annotated manually, the more precious these datasets become for the biological community, as they represent the only available source of trustworthy information and serve as a gold standard for benchmarking computational methods.

3.2. Protein–Protein Interactions

While the amount of mass data on protein−protein interactions obtained by high-throughput techniques, such as the two-hybrid system, is quickly growing, manually annotated protein interactions are a relatively rare commodity. Experimental investigation and annotation of protein

interactions is being done nearly exclusively for several well-studied model organisms. For example, the DIP database[52] contains interactions from 110 organisms, but 96% of them are from just eight organisms: *D.melanogaster*, yeast, *E. coli*, *C. elegans*, *H. pylori*, human, mouse, and rat. *S. cerevisiae* is the undisputed champion in terms of quantity and quality of reliable information per each gene. The MIPS yeast interaction database MPact,[53] initiated more than a decade ago, is the most complete source of interactions for this organism and often serves as a gold standard against which the quality of high-throughput data[54,55] and bioinformatics methods is being assessed. On average, MPact contains 2.6 manually annotated interactions per protein for 1500 proteins, and each interaction is supported by 1.2 evidences, with 2.5 interactions described in each literature reference. Another carefully annotated MIPS dataset contains interactions of over 900 proteins from mammalian species, primarily in human, mouse, and rat.[56] A much larger literature-derived human dataset involving more than 30 000 interactions for over 20 000 proteins is available via the Human Protein Reference Database.[57]

3.3. Ontologies

Biological ontologies provide an extremely efficient framework for structuring and organizing functional information about proteins. They constitute a common language for formalizing knowledge about cellular roles of gene products based on a controlled vocabulary and play a crucial role in streamlining and standardizing annotation work. Ontologies are a major element of data integration, as they ensure interoperability of databases and allow simultaneous querying of multiple genomes. In contrast to free text annotation, structured terminologies are computationally tractable, making them a standard tool for comparing genomes, analyzing genetic networks, and linking annotation to literature. The general principles of biological ontologies have been recently reviewed by Bard and Rhee.[58]

Wide use of ontologies in biology was provoked by the advent of the genomic era. Tellingly, the first functional role catalog of proteins was developed as part of the *Escherichia coli* genome annotation effort long before the genome was completely sequenced. As all subsequent ontologies adopted in molecular bioinformatics, this catalog was hierarchically organized according to the complex hierarchical nature of biological knowledge. This first version of the catalog, reflecting the annotation status as of 1992,[59] covered 1717 *E. coli* gene products, contained only six most general categories (Intermediary metabolism, Biosynthesis of small molecules, Macromolecule metabolism, Cell structure, Cellular processes, Other functions), and had up to three levels of hierarchy. For example, nine gene products known at that time were classified as belonging to the top category "Biosynthesis of small molecules", with their second and third classification levels being defined as "Amino acids" and "Histidine", respectively. The most up-to-date version of the Riley catalog, MultiFun,[60] contains 10 major categories and up to five hierarchical levels.

The Riley approach was later further developed and substantially extended in the MIPS Functional Catalog (FunCat) to accommodate the much more complicated biology of *S. cerevisiae*.[61] It was later adapted and used to annotate plant and animal genomes as well as a number of prokaryotic organisms. The modern version of FunCat[62] has a total of 1307 individual categories. The 27 top level categories serve as the origin to a hierarchical tree-like

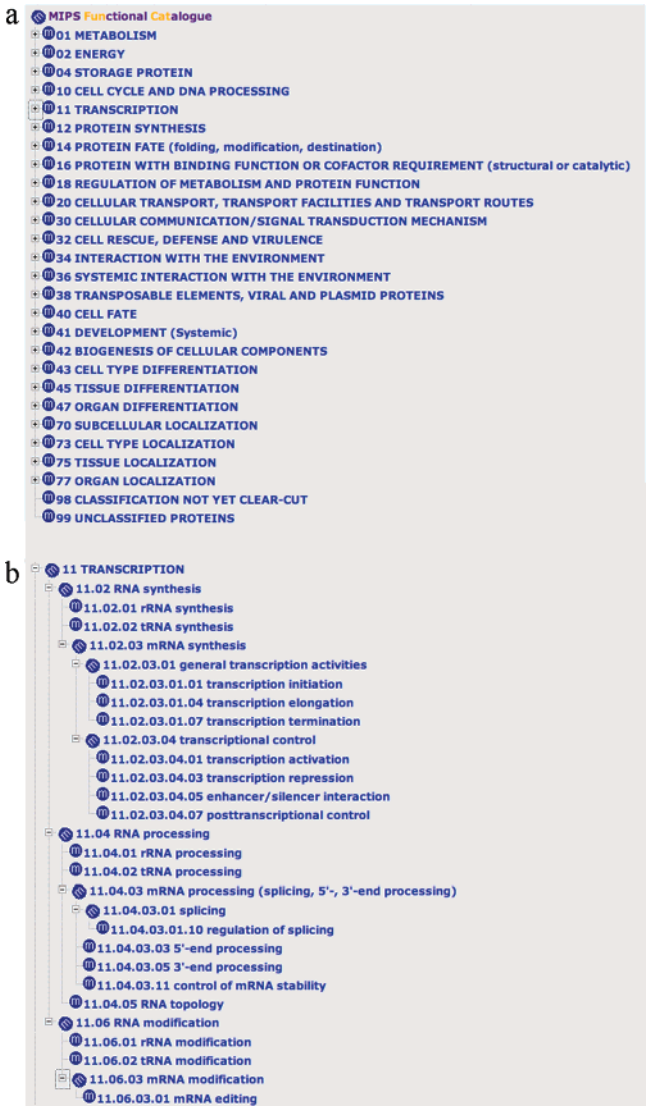

Figure 4. Structure of the MIPS Functional Catalog: (a) all top-level categories; (b) the category "11. Transcription" fully expanded.

structure which may contain up to six subcategories (Figure 4). An essential novel feature of FunCat is its multidimensionality, meaning that any protein can be ascribed to multiple categories. The most valuable aspect of the FunCat project is the ongoing manual assignment of functional categories to gene products in the expanding set of genomes. Hand-curated FunCat annotaton is currently available for four eukaryotic organisms (*S. cerevisiae*, *N. crassa*, *A. thaliana*, and *H. sapiens*), four eubacterial organisms (*Helicobacter pylori*, *L. monocytogenes*, *L. innocua*, and *B. subtilis*), and one archaeal organism (*T. acidophilum*). These genomes collectively code for 78 883 proteins, out of which FunCat assignments are available for 48 638 gene products, or more than 60%.

The Gene Ontology (GO) has become a community standard for annotating genomes of multicellular organisms. GO describes biological roles of genes and gene groups in terms of attributes defined by three major branches: molecular function, biological process, and cellular component.[63] In contrast to FunCat, which is organized as a simple tree, GO is built on the Directed Acyclic Graph (DAG) architecture, in which a child node may have multiple parents. An even more important difference is that GO is a much more detailed classification than FunCat: it currently contains 9805

Table 2. On-line Genome Databases

genome database	no. of genomes[a] eucaryotic	eubacterial	archaeal	URL	ref
Ensembl	33			www.ensembl.org	209
UCSC Genome Browser	30			genome.ucsc.edu	208
PEDANT	63	409	31	pedant.gsf.de	238
Comprehensive Microbial Resource		329	26	cmr.tigr.org	210
IMG	13	356	28	img.jgi.doe.gov	217
MaGe		27		www.genoscope.cns.fr/agc/mage	214
manndb		82		manndb.llnl.gov	213
SEED	562	643	37	theseed.uchicago.edu	90
MAGPIE	3	116	14	magpie.ucalgary.ca	202
MicrobesOnline		304	26	www.microbesonline.org	212
PUMA2	1047	611	46	compbio.mcs.anl.gov/puma2	207
KEGG	85	388	29	www.genome.jp/kegg	72

[a] These numbers are not comparable with each other, as they refer both to completely sequenced and incomplete genomes, which, in some cases, may include only a few genes.

biological process terms, 7076 molecular function terms, and 1574 cellular component terms. Manual GO assignments are available for 158 107 gene products from 30 genomes (The Gene Ontology Project, 2006), including the most important model organisms, such as fly and mouse. Furthermore, the UniProt database has adopted the GO nomenclature and is providing GO assignments for a quickly growing set of proteomes, including human.[64] However, most of these assignments are made electronically, and only 1% of database entries have been associated with GO terms manually.

High-quality functional categorization of gene products is arguably the most important aspect of genome annotation for interpreting results of high-throughput experiments in a biological context. For example, functional commonalities between coexpressed genes identified by microarray experiments are typically established by delineating significantly enriched or depleted ontology terms in respective gene lists.[65] Functional categories are also a useful level of abstraction for analyzing protein interaction data.[54]

3.4. Cellular Localization

A comprehensive source of information on protein subcellular location is the UniProt database.[66] The UniProt annotation has been used to train influential algorithms for predicting cellular localization from a sequence, such as TargetP[67] and PSORT.[68] However, many UniProt assignments are marked as "potential", "probable", or "by similarity", indicating they are not explicitly corroborated by experimental data. PSORTdb[69] uses the UniProt annotation as a starting point to create a manually curated high-quality dataset called ePSORTdb, which currently describes localization of over 2000 bacterial proteins.

Many genome databases also contain manually annotated localization assignments. One special aspect of both FunCat and GO systems is that they consider protein localization in the cell as a functional category. For example, the FunCat top-level category "70. Subcellular localization" distinguishes 25 different types of possible localizations, some of which are prokaryote-specific (e.g., 70.37 prokaryotic nucleoid), some of which are universal (e.g., 70.32 flagellum), some of which are only relevant for fungi (e.g. 70.29 bud/growth tip), while others apply to different eukaryotic organisms, including plants. This means that for all organisms with manual FunCat assignments, careful annotation of protein localization is also available. OrganelleDB[70] consolidates available localization data for 25 000 eukaryoic proteins derived from original model genome databases, such as those listed in Table 1, as well as from the Gene Ontology resource.

3.5. EC Numbers and Metabolic Pathways

The hierarchical classification of enzymatic reactions known as the EC (Enzyme Commission) system is one of the most widely used bio-ontologies established long before the genome sequencing era. Each EC number is a unique code which describes enzyme activity at four progressively finer levels of detail. Enzyme classification with a plethora of associated information is available via the BRENDA resource,[71] where it is collected based on a literature survey. EC numbers represent an integral part of functional annotation in main-stream databases such as Uniprot.

At the whole genome level, systematic assignment of EC numbers creates the basis for organizing individual reactions into metabolic pathways. The KEGG pathway database[72] maintains a growing set of manually drawn metabolic maps (currently 306) which can be considered as general, organism-independent enzyme networks. Over 40.000 organism-specific pathways have been computationally generated by mapping EC numbers assigned to gene products from a specific genome to the generic maps. A somewhat different concept underlies the MetaCyc database,[73] where manually curated objects are representative of experimentally determined metabolic pathways rather than generic multispecies maps as in KEGG. The latest version of MetaCyc contains 700 pathways from over 600 species. Finally, the Reactome knowledge base[74] is primarily devoted to human metabolic and regulatory pathways. It pursues an interesting community-based annotation model. A panel of editors first determines priority areas of annotation and then invites established bench scientists to annotate a specific information module. Completeness and consistency of data is enforced by specially designed software. Annotation submitted by researchers is further refined by Reactome staff members.

3.6. Databases of Orthologs

The NCBI's COG database[75] can be viewed as a large collection of phylogenetic profiles. It is one of the most widely used genomic datasets because of its enormous usefulness for comparative genomics and protein function prediction. Orthologs are initially inferred automatically by all-against-all similarity comparison of gene products and subsequent identification of mutually consistent best bi-directional similarity hits. The resulting ortholog groups are

being manually curated to exclude possible erroneous assignments. In particular, multidomain proteins are manually split into individual domains. The prokaryotic and eukaryotic versions each contain nearly 5000 COGS, with the former covering 75% of gene products in completely sequenced genomes of microorganisms, and the latter covering 54% of gene products in seven model eukaryotic organisms. Some major genome resources have developed their own orthology databases. For example, KEGG introduced the KO system, which combines orthologous relationships with pathway information.[72]

4. From Gene-By-Gene Annotation to Hierarchical Modular Representation of Proteomes

Proteins in the cell typically perform their function by engaging either in direct physical interactions or in functional interactions by getting involved in the same cellular process with other proteins. The complex structure of functional relationships can be mathematically described by networks where proteins or genes are nodes and connections between the nodes reflect different kinds of associations. Investigating the topology of protein interaction, metabolic, signaling, and transcriptional networks allows researchers to reveal the fundamental principles of molecular organization of the cell and to interpret genome data in the context of large-scale experiments. Such analyses have become an integral part of the genome annotation process: annotating genomes today increasingly means annotating networks.

Modularity has emerged as one of the major organizational principles of cellular processes. Functional modules are defined as molecular ensembles with an autonomous function.[76,77] Proteins or genes can be partitioned into modules based on shared patterns of regulation or expression, involvement in a common metabolic or regulatory pathway, or membership in the same protein complex or subcellular structure. For example, within protein complexes, a certain part of their constituent components form large cores characteristic for a given complex while other, typically smaller, groups of proteins systematically occur together in different complexes, forming stable functional modules which can be flexibly used in the cell in a variety of functional contexts.[78] Modular representation and analysis of cellular processes allows interpretation of genome data beyond single gene behavior. In particular, modules represent a convenient framework for studying the evolution of living systems.[79−82]

Novel manual annotation strategies provide for joint curation of gene groups belonging to the same functional module, often across many genomes, in order to improve the quality of assignments and, importantly, to increase the speed of annotation work. For example, the MIPS annotation group is moving from annotating individual protein interactions to experimentally derived multiprotein complexes. The MIPS Mammalian Protein Complex Database currently contains literature-curated data on 122 complexes involving 643 gene products.[83]

Algorithmically, modular architectures can be defined as densely interconnected groups of nodes on biological networks[84] (Figures 5 and 6). Automated tools to delineate statistically significant subnetworks based on a wide spectrum of protein and gene attributes allow researchers to attach functional roles to previously uncharacterized genes involved in known modules as well as to discover previously unknown modules.[85−87] Alternatively, one can define functional modules as groups of genes assigned to the same MIPS FunCat category.[88] Dependent on the desired degree of detail, this can be done at different levels of functional hierarchy, but

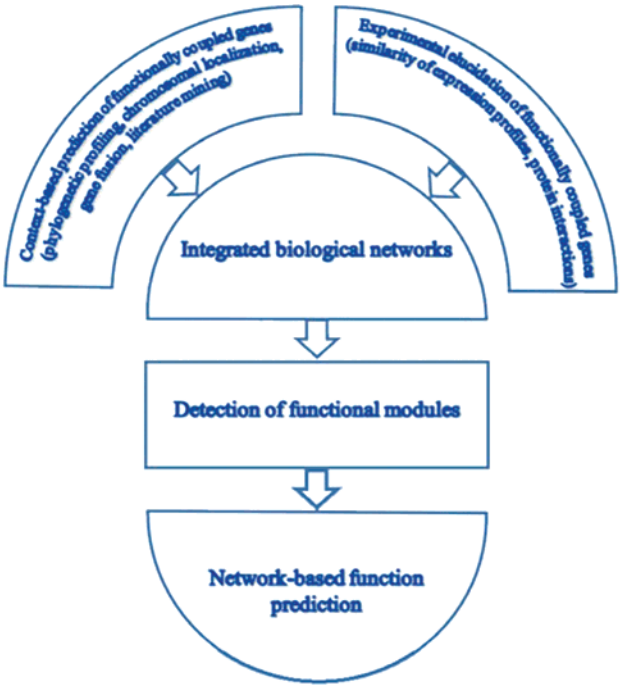

Figure 5. General overview of the function prediction process. Computational predictions and experimental data are used to build a network of functional associations between proteins. Densely interconnected groups of proteins on biological networks often correspond to functional modules.

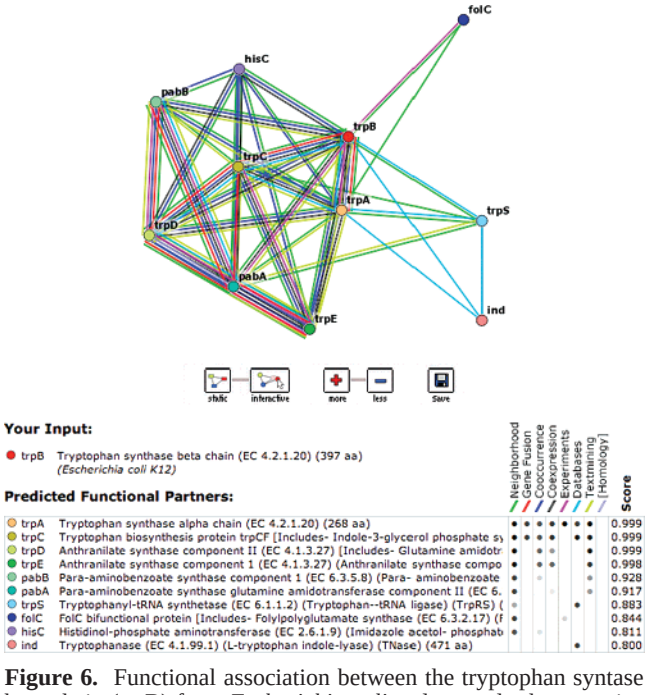

Figure 6. Functional association between the tryptophan syntase beta chain (trpB) from *Escherichia coli* and several other proteins predicted by different context-based methods by the STRING[154] server. Nodes on the graph represent proteins functionally coupled with trbB while edges of different color represent different types of functional association: chromosomal neighborhood (green edges), gene fusion (red edges), phylogenetic profiles (blue edges), conserved coexpression of genes (black edges), experimental data (e.g., protein−protein interactions), database annotation (e.g., KEGG), and literature mining.

generally the most natural choices are categories at the second-highest level which reflect specific cellular processes such as "ribosome biogenesis" or "oxidation of fatty acids". Second-level FunCat modules were used by Petti and Church[89] in their study of coregulated functional module pairs within the *S.cerevisiae* transcriptional network. In fact, systematic assignment of FunCat categories to all gene products in a given genome essentially means defining a hierarchy of functional modules.

A conceptually similar approach has been adopted in the SEED annotation environment.[90] SEED operates with subsystems which are defined as a set of related functional roles and are thus a more general concept than functional modules. Subsystems can correspond to metabolic pathways, structural complexes, or any other cellular processes and components. In contrast to traditional gene-by-gene annotation of one single genome, SEED enables human experts to curate specific subsystems in multiple genomes simultaneously, achieving much higher annotation quality.

5. Experimental Annotation of Genomes

It is obvious that any detailed biochemical experiment that sheds light on the function of a yet uncharacterized genomic ORF helps to illuminate the inner workings of the cell and ultimately contributes to genome annotation. Most such studies are problem-oriented, hypothesis-driven investigations motivated by the particular interests of individual research groups. On the other hand, whole-genome experiments published in the scientific literature are essentially discovery-oriented endeavors in which one high-throughput method, such as the two-hybrid system for detecting protein interactions, is applied to elucidate the global structure of biological networks. What is still largely missing, however, are experimental studies motivated by bioinformatics predictions. For a sizable fraction of genomic ORFs, up to 30−40% of all genes,[91] no experimental information is available, and no meaningful function prediction can be made based on homology information or context-based techniques. These ORFs are labeled "conserved hypothetical" if they have homologs in multiple genomes, or just "hypothetical" if they occur in only one genome (in which case they are also often called genomic ORFans). These latter alone probably constitute as much as 15% of microbial gene complements, although many of them may be artifacts of gene prediction and not be expressed.[92]

It has recently been argued that directed experimental investigation of hypothetical proteins would be extremely beneficial for the genomics community. Roberts[93] proposed to form a consortium of bioinformaticians with the goal to assemble a comprehensive list of hypothetical, conserved hypothetical, and putatively misannotated proteins. The list would be prioritized, in particular giving stronger preference to conserved hypothetical ORFs because of their presumed higher functional importance and also because they have higher chances to actually encode proteins. Possible criteria for setting priorities on functional targets are phyletic distribution (with preference given to particularly ubiquitous genes), essentiality, availability of three-dimensional structures and information on expression and binding, as well as practical considerations, such as the likeliness of a protein yield to solubilization, expression, and purification studies.[94] With regard to the latter task, reasonably accurate computational algorithms to predict the experimental behavior of proteins from their sequences are beginning to emerge.[95] This approach bears similarity to the target selection process generally adopted in structural genomics efforts,[96] where proteins for experimental structure determination are picked based on a broad spectrum of bioinformatics criteria, typically including the requirement for target proteins to have no related known structure. In the second step, experimental labs throughout the world would be invited to pick targets of interest from the list and conduct their thorough experimental characterization using dedicated funding. In a complementary approach, one would focus on so-called "orphan functions"—those functions that are biochemically characterized but for which no gene has yet been found.[97]

While experimental genome annotation as outlined above is still in its infancy, the first efforts in this direction are already underway. The *Shewanella* Federation has launched a project to characterize hypothetical proteins (which constitute 40% of genes in this organism) in a multistage program which involves expression studies on genes and proteins, generating tentative functional predictions, and then experimentally verifying them for high-priority targets. To aid the prioritization process, a novel classification scheme for functional assignments was developed[98] that takes into account both bioinformatics criteria (availability of homologs, known motifs) and available biochemical knowledge. Out of 538 hypothetical proteins, exact biochemical function could be ascribed by similarity searches to only 3% of gene products, and an additional 3% obtained well-defined functions with unknown specificity. Further, two categories encompass a total of 28% of genes for which only coarse function assignments could be made based on a variety of annotation attributes, ranging from conserved sequence motifs to protein interaction data, and for an additional 13%, some partial functional insights could be derived. Finally, conserved expressed proteins and *Shewanella* specific expressed proteins constitute 35% and 17% of the dataset, respectively. This functional breakdown of hypothetical proteins in a prokaryotic organism is extremely illuminating and will definitely serve as a model for other similar studies.

Directed experimentation is also a valid option for verifying gene predictions made by automatic means. The ENCODE (Encyclopedia Of DNA Elements) project[99] is aimed at identification and comprehensive experimental characterization of all genetic elements in the human genome. Other efforts are focusing on validating *ab initio* gene predictions by microarray and PCR experiments.[100]

6. Automatic Annotation of Genomes

6.1. Annotation Transfer by Homology

In spite of the growing sophistication of bioinformatics methods, an overwhelming majority of computer-based functional assignments for proteins continue to be made by the traditional approach, which involves imputing annotation from one or several previously annotated gene products to the query protein based on a sufficiently significant degree of similarity. In the UniProt database, more than half of the functional descriptions have the status "By similarity". All automated genome analysis systems use BLAST[101] or FASTA[102] searches as their primary functional prediction engine.

In general, annotation transfer by homology (ATH) can give spurious results either because the available similarity is not sufficient to warrant the transfer of information from the source to the target sequence or because the annotation

of the source sequence is already wrong. Provided that the source annotation is correct, ATH is a reliable option when it comes to annotating genomes which are only minor variations (e.g., closely related strains or nearly identical viruses) of already known genomes.[103] But what is the accuracy of ATH for the cases of lower sequence similarity levels which are typical in genome annotation practice?

Quantifying the correlation between protein sequence similarity and similarity of function is much more difficult than quantifying that between sequence similarity and structural relatedness. Resemblance of protein structural folds can be objectively measured, for example by calculating the root mean square deviation between optimally superimposed structures and then comparing it with sequence alignment scores.[104,105] Wilson et al.[106] made an early attempt to quantify the functional relatedness of individual structural domains based on a combination of GO assignments and enzyme classes. They defined four possible scenarios for any pair of domains: (i) general similarity, either both domains are enzymes or both are nonenzymes; (ii) same functional class, both domains share the same top-level enzyme class or GO category; (iii) same precise function, both domains share the exact same enzyme classes and GO categories; (iv) no functional similarity at all. On average, 20% sequence identity was found to be required for general functional similarity, 25% identity indicated a common functional class, and 40% identity typically guaranteed a precise functional match. Devos and Valencia[107] in their influential study arrived at similar conclusions by calibrating the reliability of ATH using Riley functional classes of *E.coli* proteins (see above), EC numbers, SwissProt keywords, and structurally defined active sites. Multidomain proteins were shown to be much less functionally conserved, thus requiring substantially higher sequence identity levels for reliable transfer of annotations.[108] Finally, Rost[109] and Tian and Skolnick[110] argued that enzyme classes are even less conserved than reported in the previous publications if no requirement of shared structural folds is imposed on sequences and database bias is removed by clustering sequences to correct for unequal representation of large families. According to Rost, even extremely low BLAST E-values (below 10^{-50}) may not be sufficient for confident transfer of complete four-digit EC numbers. Moreover, at sequence identities below the 50% level, it is not even always possible to distinguish enzymes from their nonenzymatic homologs.[111]

The core methodological issue in ATH is the possibly accurate identification of true orthologs which, apart from some special cases such as nonorthologous gene displacement, perform equivalent, albeit not necessarily completely identical, functions in different organisms.[112] While, for many, especially prokaryotic, protein families and sequenced genomes, ortholog databases such as COG readily provide precomputed and verified information on orthologous relationships, accurate *de novo* identification of orthologs poses a significant algorithmic challenge. In complex eukaryotic genomes, finding an ortholog is complicated by one-to-many or even many-to-many orthologous relationships.[113] The operational definition of orthologs frequently used in whole-genome comparison involves, at the very minimum, the requirement of a highly significant global alignment over the major part of the protein length and, furthermore, the presence of a best bidirectional or even triangular (involving three genomes) similarity hit, although much more sophisticated methods properly handling paralogous relationships exist (e.g., ref 114; see also ref 115 for comparison and discussion).

Over the past decade, phylogenomic approaches have been introduced which infer functions of proteins by considering their evolutionary relationships with other genes.[116,117] Phylogenomic function inference for a given query protein involves multiply aligning this protein with its homologs found by a database search, calculating a phylogenetic tree, and then tracing the evolutionary history of the query protein, in particular paying attention to potential duplication events and taking into account apparent paralogs. If properly done, phylogenomics inference of function is very accurate and avoids many typical errors intrinsic to straightforward ATH.[118] Nevertheless, an overwhelming majority of functional predictions are still being done by the highest BLAST match approach. Brown and Sjölander[119] attribute this state of affairs to general unawareness about the actual error rate of ATH, understandable inertia of genomic software designers in embracing novel techniques, and a much higher level of sophistication and computational complexity of phylogenomics analyses, in particular a strong dependence on the quality of multiple sequence alignments and tree-building procedures. For large, divergent, and complex eukaryotic families, complete automation of phylogenomic inference may not be easily possible and some degree of manual curation of the alignments and trees may be required. Recently, more automated and statistically rigorous methods have been reported,[120] and attempts to model human logic while performing phylogenomics analyses using expert knowledge have been undertaken.[121]

6.2. Automatic Functional Class Definitions

One special and particularly important flavor of automatic annotation involves attachment of GO or FunCat labels and EC numbers to genomic proteins, which is currently seen as one of the core elements of genome annotation and is heavily relied upon in virtually all modern genome analysis systems. While functional roles (as well as the first three digits in EC numbers) are group characteristics and thus provide much coarser functional assignments than specific SwissProt-like description lines of individual proteins, they have the great advantage of belonging to a generally accepted controlled vocabulary and thus making predictions made by different researchers for different organisms, experiments, and conditions easily understandable and directly comparable.

Naïve GO term extraction from the best BLAST-matched sequences[122,123] is an efficient high-throughput approach, but it suffers from all typical pitfalls of transitive sequence annotation, as discussed above. More advanced methods attempt to alleviate the problem by combining homology-based inference with orthogonal sources of information, including text mining and predictions of protein cellular localization,[124] and by applying machine learning techniques (specifically, Support Vector Machines) to the annotation attributes associated with top-scoring BLAST hits.[125] The UniProt database relies on existing mappings of GO terms to other extensively annotated protein attributes, particularly InterPro domains (Interpro2go) and keywords (spkw2go), which are available from the Gene Ontology consortium and are being frequently updated.[64] For example, InterPro domains are reliably annotated with GO information, and HMM-based InterPro searches are very sensitive and specific. Thus, transferring GO information from strong InterPro hits,

where available, is a much more reliable option than using the best BLAST hit.

For enzyme classes, beyond copying EC numbers from the best database search hit, advanced consensus methods explore conserved residues of the active site determining enzyme function. Tian et al.[126] identify residue positions determining the functional specificity of enzymes by comparing sequence profiles of homofunctional families annotated in SwissProt with the same EC number and broader supersets of these families which also include additional representatives found by database searching and not necessarily sharing the same EC number. Such sets of functionally determining residues provide very high discrimination power for four-digit EC numbers and were reported to outperform KEGG assignments when applied to the *E. coli* genome.

An alternative approach consists in predicting functional classes from sequences without any reliance on homology information. Natural language processing methods that extract GO terms from millions of PubMed abstracts show great promise in this respect.[127,128] The ProtFun method[129,130] uses neural networks trained on a large variety of protein attributes directly predictable from sequences (cellular localization, post-translational modifications, secondary structure, and others) to associate Riley functional roles, GO terms, and EC numbers with gene products.

The main difficulty in meaningful assignment of functional classes lies in the hierarchical organization of various gene ontologies. Most of the computational methods developed so far are limited to predicting assignments at one selected depth in the hierarchy. As noted by Barutcuoglu et al.,[131] constructing individual classifiers for each functional class (e.g., GO category) violates the fundamental requirement that the genes belonging to this class must also belong to the parent class on the hierarchy. In what appears to be the first application of multilabel prediction techniques to gene function analysis, the authors directly incorporate the constraints of the hierarchical classifications by combining multiple single-class classifiers in a Baysian framework. Applied to GO annotation, this method yielded substantially improved prediction accuracy, especially for deeper (more specific) nodes of the GO graph. Several other modern approaches exploit child−parent relationships between GO terms and the general topology of the GO graph.[132−134] The GOtcha algorithm[133] proceeds by first making BLAST-based assignments of individual leaf nodes of the GO hierarchy to the query sequence. These assignments are then propagated to parent nodes, appropriately weighted, all the way to the root node of a given GO ontology (molecular function, biological process, and cellular component), such that if a specific parent node has several child nodes picked by BLAST, the GO term corresponding to this node will be assigned with much higher confidence than each of the child nodes, as the parent node is corroborated by several independent similarity hits. In particular, if several BLAST hits point to nodes on one of the three GO ontologies, then the top-level root node of that ontology will be assigned to the query sequence with the highest confidence, as it has the highest support. GOtcha was shown to provide much higher specificity and selectivity compared to the best BLAST hit method.

A new class of recently developed methods leverages biological networks of different types to propagate knowledge from confidently annotated sequences to uncharacterized proteins. In principle, this approach is logically connected to deriving network modules (see above) because ascribing hypothetical genes to a well-defined functional module is a powerful method of function prediction. On the most basic level, a certain node of the network can be ascribed to one or more functional roles that most frequently occur in the annotation of other nodes to which it is immediately connected[135] or of those nodes that are separated from the target node by a certain predefined number of edges.[136] More sophisticated algorithms introduce weighting of edges according to the reliability of the underlying experimental data they are derived from and explicitly take into account the topological properties of the network.[137,138] Another important principle guiding propagation of knowledge is that interconnected nodes that often correspond to functionally coupled proteins are expected to have at least partial commonality in their functional annotation.[139] Massjouni et al.[140] described the VIRtual Gene Ontology that relies on a functional linkage network constructed from protein interaction and gene expression data. For each node on the network corresponding to a hypothetical protein, functions of the neighboring nodes are mapped using the GAIN (Gene Annotation using Integrated Networks) algorithm,[141] which takes into account the weights assigned to connecting edges and the topological properties of the network neighborhood and operates under the constraint of maximal consistency of the assignments made for connected nodes.

6.3. Guilt by Association: Context-Based Function Prediction

The availability of complete genome sequences sparked a principally new group of computational approaches to gene function prediction. For many genes that cannot be characterized by homology searches, useful functional hints can be delineated from their genomic context (Figures 5 and 6). Several algorithms of this type were pioneered nearly simultaneously as soon as the number of finished genomes became sufficiently high (over 10) to allow for statistically significant inference. The conserved genomic neighborhood method[142,143] deduces functional coupling based on short-range colinearity between genes in different prokaryotic genomes. It exploits nonrandom proximity of genes involved in operons, which represents a specific complementary information signal not recognizable by sequence comparison. Phylogenetic profiling[144] relies on the correlation of protein occurrence across a set of genomes to predict functional associations. Similarity of evolutionary patterns shared by two proteins may indicate that they interact with each other directly or share a common functional role. The underlying idea is that many pathways or complexes require all their members to be present in order to fulfill their functions. The gene fusion approach[145] detects those pairs of proteins that are encoded by two different amino acid chains in one genome while constituting a single multidomain molecule in another genome. Further possibilities to identify genes whose functions may be related include linguistic methods that exploit co-occurrence of gene names in literature abstracts as well as identification of conserved coexpression.[146] Because genomic contexts are more conserved between closely related species than between distant ones, more recent implementations of context based methods incorporate phylogenetic analyses in order to take into account evolutionary distances between the genomes compared[147,148] and thus achieve an improved signal-to-noise ratio.

A new algorithmic twist involves the application of context based methods to individual protein domains rather than full-length protein chains. A large variety of widely spread interaction domains that mediate molecular interactions are frequently combined in proteins in a complicated mosaic fashion[149] and often represent major functional entities in cellular interaction networks. Clustering protein domains with similar phylogenetic profiles allows researchers to build domain interaction networks which provide clues for describing molecular complexes.[56] Similarly, the Domain Teams method[150] considers chromosomal neighborhoods at the level of conserved domain groups.

From today's point of view, employing context based methods on a large scale for genome annotation is essentially equivalent to joining protein nodes on a global network of functional associations by edges either derived by the different computational techniques described above or supported by experimental data. Each edge gets ascribed a numerical score reflecting the confidence of the underlying computational or experimental evidence. This general approach has been implemented in several highly valuable resources (e.g., Predictome,[151] Prolinks,[152] Phydbac,[153] STRING[154]) that systematically maintain collections of functional links between gene products for a large number of genomes and provide software tools to navigate and analyze gene association networks. The most comprehensive system, STRING, currently allows exploring various types of genomic context for 800.000 genes from 200 organisms. Integrated context analysis systems represent efficient large-scale multigenome annotation tools. They have been used to provide functional assignments for hypothetical genes,[155,156] to identify missing genes in known metabolic pathways,[157,158] to reconstruct novel, experimentally uncharacterized metabolic pathways,[159] and to study gene−phenotype relationships.[160]

7. Assessing and Improving the Quality of Automatic Genome Annotation

7.1. Errors, Errors Everywhere

The core problem in automating genome analysis is without a doubt the notoriously high level of errors made by unsupervised algorithms. It is extremely difficult to reproduce computationally the complex decision process of a human curator, who, while making a decision on a particular assignment, will survey literature, carefully analyze available alignments, and heavily rely on his specific experience and intuition. Typical sources of annotation errors have been reviewed before.[161,162] In addition to fundamental difficulties in annotation transfer by homology, as discussed above, dubious assignments may be caused by spurious similarity hits stemming from compositionally biased protein sequences and failure to take into consideration multidomain organization of proteins. Further complications include wrong gene models and unrecognized pseudogenes. Annotation errors systematically pollute sequence databases, leading to the gradual deterioration of the total corpus of available information and undermining further analysis efforts.[163]

7.2. Annotation Benchmarks

Ideally one would wish to be able to evaluate the specificity and sensitivity of automatic functional assignments pretty much the same way it is done in other areas of bioinformatics, such as protein secondary structure prediction, where generated predictions are rigorously compared with experimentally determined structures. However, objective assessment of the quality of functional annotation is a much more difficult task due to the scarcity of trusted data that could be used as a standard of truth. Even for the best experimentally characterized genomes, such as *E. coli* and *S. cerevisiae*, comparison of predicted and "known" functional information expressed in natural language is far from trivial, as the notion of protein function is elusive, is semantically ambiguous, and may depend on a particular cellular context. For most of the ORFs in newly sequenced genomes, and even for many important model organisms, no experimental verification of predictions can be made. Brown and Sjölander[119] estimated that only 3% of nontrivial UniProt annotations have experimental confirmation.

At least three different approaches to assess the quality of automatic function assignments, apart from direct experimental verification, are conceivable. The early estimates of the error rate of gene annotation were made by comparing assignments made manually by different groups and/or by automatic systems, most notably GeneQuiz.[164] While no statement can be made in the majority of cases where the predictions agree (they may be all wrong or all correct), those cases where they disagree point to potential errors. Just recently, the same approach has been applied by Mi et al.[165] to compare functional annotation of the *D. melanogaster* genome made by the GO consortium and by the Celera Genomics team, which used its own in-house Panther ontology.[166] An additional benefit of such comparisons is in assessing the overlap between the different annotation schemes and ontologies and identifying the knowledge gaps that need to be filled.

The second approach relies on estimating the consistency of annotation for sequence-similar gene products across several different genomes. According to Devos and Valencia,[107] the chance of error varies greatly dependent on the particular type of the annotation attribute. In three prokaryotic genomes—*M. genitalium*, *H. influenzae*, and *M. jannaschii*—the percentage of erroneous predictions made by the GeneQuiz system[167] was estimated to be 2% for the first EC digit, around 20% for SwissProt keywords, over 30% for the last EC digit, and even higher for substrate-binding sites.

Finally, annotation results can be benchmarked against an accepted high-quality standard of truth which is believed to be largely "correct". To assist such comparisons, Iliopoulos et al.[168] classified errors arising in the process of transitive (similarity based) annotation into seven classes, scored by their severity in descending order: false positive (7), overprediction (6), domain error (5), false negative (4), underprediction (3), undefined source (2), and typographical error (1). One important feature of this scale is that it assigns a significantly higher weight to over- rather than underpredictions. The rationale for this choice is that overpredictions have a much higher potential to pollute sequence databases with wrong information. The authors reported comparable general quality levels of the originally published annotation of the *Chlamydia trachomatis* genome and automatic assignments produced by the GeneQuiz system,[167] as judged by careful manual inspection. At the same time, the overlap between these two annotation efforts in terms of completely correct assignments was only 51%, highlighting different biases in human and machine-generated judgment.

For four bacterial genomes, a specially designed functional annotation benchmark set[169] has been made available which includes hand-curated MIPS FunCat assignments as well as a number of precalculated protein attributes (with their associated scores), such as BLAST similarity hits, InterPro domain assignments,[170] similarity-derived fold assignments, and so on. This dataset can be used for assessing the performance of machine learning techniques for predicting protein function from sequences.

7.3. Annatomics: Data Mining in Genome Annotation

Automatic generation of genome annotation is in some sense comparable to high-throughput experimental techniques, such as genome sequencing or two-hybrid essays. Similar to the familiar notions of genome, proteome, transcriptome, interactome, and a dozen or more other -omes, the author would like to coin the term *annatome* to describe the entire body of annotation data accumulated in today's genomic databases. Just as any other omics results, *annatomics* data may be imprecise, inconsistent, and wrong. A difficult and timely challenge faced by bioinformatics is to design intelligent systems aimed at improving the overall quality of machine-generated annotation.

Is it possible to reduce errors in genome annotation? The total annatome can be considered to be a collection of records, one per each of the 6 million genes known today, containing a varying number of attributes, ranging from just a few minimal descriptors (length, pI) for hypothetical proteins to dozens of annotation items (motifs, EC numbers, localization, structural folds, etc) for better characterized proteins. Analyzing the current annatome data by data mining techniques may help researchers find interesting statistical trends in this large collection of records and may point to potential spurious assignments.

7.4. Automated Correction of Annotation Errors

One way to deal with the problem of error correction is to detect inconsistencies in the annotation of related proteins forming a sequence cluster. This approach is conceptually similar to knowledge propagation in biological networks, as described above. Based on the observation that more than 95% of proteins have more that two annotation attributes, with 10 being the average number, Kaplan et al.[171] implemented a system that represents protein−keyword relationships in biological databases in the form of a hierarchical graph, each node of which symbolizes proteins sharing unique combinations of keywords. While analyzing protein sets attributed to the same functional category by automated annotation methods, observing certain proteins occupying areas on the graph that are distinct from the main bulk of the collection clearly points to potential false annotations. More generally, one can define a score which indicates how similar are the sets of annotations for any given pair of proteins.[172] Functionally related proteins are naturally expected to have more similar annotation than unrelated ones. Based on the defined similarity measure, proteins are clustered into groups with homogeneous annotation, the so-called property clusters. This method can be used to detect false positive annotation by any given automatic method aimed, for example, at detection of conserved sequence motifs. The idea is to find those proteins that share the same annotation, e.g., a sequence motif, from the test method and at the same time form disjoined subsets as a result of clustering in the space of other annotated features. Alternatively, annotation errors can be identified by comparing protein groupings obtained by sequence and annotation clustering.[173] Again, the underlying assumption is that the more sequence-similar proteins are, the higher chance they have to share functional annotation.

Another promising tactic in intelligent filtering and improvement of biological annotation is through knowledge discovery techniques aimed at detecting common patterns, rules, or anomalies. In addition to being widely used for mining biological literature[174,175] and experimental data,[176] rule-based techniques have been applied to predict protein annotation features from a set of other annotation features.[177] The RuleMiner algorithm[178] extracts characteristic annotation features associated with protein function groups defined by sequence similarity, shared conserved motifs, and a common taxonomic distribution. Rules learned from the annotation of such groups may be applied to classifying yet uncharacterized instances, and they can be combined with the results of similarity-based annotation transfer using knowledge-based voting procedures.[179] Major protein annotation efforts routinely use rule-based procedures for checking the integrity of information, finding minor errors, and automating trivial annotation procedures which do not require human intervention.[180−182] Uninformative pieces of information (e.g., description lines containing only words such as "hypothetical", "putative", and "unknown" transferred from the best similarity hit) can be filtered out using simple lexical analyses based on specially prepared vocabularies.[167,183]

A more sophisticated approach to this problem involves automatic learning of rules from a highly curated and reliable database, such as Swiss-Prot, and then using these rules to further improve annotation either in the same database or in another automatically generated database, such as TrEMBL. Kretschmann et al.[184] applied the C4.5 data mining algorithm to derive decision trees representing the knowledge on Swiss-Prot keywords. Rules obtained in this fashion combined with information on sequence groups gleaned by sequence analysis can be applied both for consistency checks within Swiss-Prot and for generating keywords for new TrEMBL entries with high accuracy. Conversely, exclusion rules for a specific protein group (e.g., sharing the same sequence motif) can be generated by the C4.5 algorithm to detect contradicting annotation items, as implemented in the Xanthippe postprocessing system.[185]

Rules can be extracted more efficiently from a very large database using the formalism of association rule mining and the well established *Apriori* algorithm.[186] Association rules are simple implications that can be formulated in the form $(A_1 \& ... \& A_n) \Rightarrow Z$, where $A_1 ... A_n$ (the left-hand side of the rule) and Z (right-hand side) are different features, and the rule means "database entries that possess all features $A_1 ... A_n$ are likely to possess feature Z". Each rule is characterized by its coverage, the number of entries in the database that possess all features $A_1 ... A_n$, its support, the number of entries satisfying both the left and the right sides of the rule simultaneously, and its strength, which is essentially the probability that a given database entry will satisfy the right side of the rule given that it satisfies the left side of the rule. Association rules have found their application in bioinformatics for identifying pairs of related GO terms,[187] interpreting gene expression data,[188] and investigating relationships between different types of genomics data.[189]

Artamonova et al.[190] applied association rule mining for identifying errors in protein annotation data. The strategy is to find rules with the strength very close, but not equal, to 1.0, which means that such rules have a minor number of exceptions, and then to identify all proteins that constitute exceptions from strong rules. Applied to the SWISS-PROT database, this approach yielded 7396, 4956, and 4046 rules with strength greater than 0.95 and coverage over 50 which were not fulfilled exactly once, twice, or three times; these rules typically infer keywords and Intepro domains from mixed left-hand side annotation items. In order to test whether exceptions from strong rules actually correspond to annotation errors, subsequent releases of the SWISS-PROT database were compared and additional manual verification was conducted. It was indeed found that exceptions from strong rules get corrected substantially more often than the rest of the annotation. For unsupervised annotation automatically generated by the PEDANT genome analysis system, the total fraction of exceptions from strong rules classified by manual analysis as errors was as high as 68%. It was also found that most of the errors in the SWISS-PROT database are underpredictions (i.e., absence of features that would be expected based on association rules), consistent with the prudent manual annotation process adopted by SWISS-PROT, while in PEDANT errors are typically caused by overpredictions.

8. Computational Infrastructure for Genome Annotation

8.1. Tools To Support Distributed Genome Annotation

Specialized software tools to support decentralized annotation efforts have been made available (see ref 191 for a detailed comparison of several such systems). A growing number of annotation consortia are joining the GMOD collaborative effort,[192] which develops standardized reusable software components (browsers, query engines, and visualization and editing tools) and ontologies for creating and maintaining model organism databases. The Distributed Annotation System (DAS)[193] is a simple client-server architecture that allows different groups to provide their annotation tracks to a central server based on the DAS XML specification. Each annotation attribute has a description line associated with it and is characterized by its coordinates relative to the reference sequence (e.g., chromosome) being visualized. DAS is the central mechanism for displaying annotation tracks within the Ensemble genome analysis system (see ref 194 for an overview of the Ensemble system). It is also being used by the European Virtual Institute for Genome Annotation to integrate annotations produced by the members of the BioSapiens Network.[195] More advanced annotation systems, such as Otter,[196] ASAP,[197] Manatee (manatee-.sourceforge.net), and PeerGAD,[191] support decentralized manual annotation and data exchange between multiple users and sites and have built-in versioning and history mechanisms.

8.2. Local Manual Annotation Tools and Viewers

Stand-alone software packages for genome display, manipulation, and editing are also available. Artemis[198] is a popular Java tool particularly suitable for visualizing DNA-level features (GC content, codon usage, etc.) of compact genomes; it can also be used remotely over the Internet when run as a Java applet. To assist the gene editing process, an excellent sequence annotation editor—Apollo—has been developed.[199]

8.3. Genome Annotation Pipelines and On-line Resources

Arguably, the first ever genome analysis system was the computer program used to automate the annotation of the yeast chromosome fragment coding for just 182 protein products[200] and later developed into the full-blown software package GeneQuiz.[201] Several other software suites were developed in the middle of the 1990s (MAGPIE,[202] PEDANT,[203] Genotator,[204] AceDB),[205] and the first Web-based genome databases were made available. Pioneering work on metabolic reconstruction from genome data resulted in integrated environments such as WIT/PUMA,[206,207] KEGG,[72] and MetaCyc.[73] At the next stage of genome sequencing, large organizations such as EBI, the Sanger Center, and the University of California created comprehensive Web portals to disseminate data produced by the human genome project. Both the UCSC Genome Browser[208] and Ensemble[209] systematically integrate all available information for multiple complex eukaryotic genomes, including alternative gene models, pseudogenes, isoforms, repeats, transcripts, genetic markers, SNPs, as well as functional annotation. Additional evidence tracks provided by external groups can also be accommodated.

Finally, as a reaction to the virtual explosion of the number of genomes available, there has recently been a proliferation of new annotation systems, each with design specifics reflecting the purposes and scientific interests pursued by the authors as well as their background. The genome analysis systems available today differ with respect to the particular software technology they are using (novel data standards, workflows), the means of deployment and data delivery (stand-alone software packages, Web-based resources, distributed annotation systems), the type of data (ESTs, genomic sequences), organism type (microbes, mammalian genomes), the scientific question they are designed to answer (e.g., pathway reconstruction), or the user community they target (small research groups, large institutions).

Many of these new resources focus on the annotation of microbial genomes and offer bacterial-specific annotation features, such as operon prediction;[210−217] the latter system has recently been extended to handle metagenomic data.[218] A useful feature comparison of many current systems can be found in ref 215.

The SwissProt team is using its own annotation pipeline HAMAP to produce annotation of entire microbial proteomes.[181] Capitalizing on quality rather than coverage, HAMAP produces completely automatic annotation only to those gene products that can be reliably attributed to a well-characterized protein family based on high sequence similarity, appropriate sequence length, and characteristic features; otherwise, proteins are subjected to further manual annotation.

Annotation systems for EST and cDNA sequences[183,219−221] start processing by grouping sequences into clusters and stress DNA-level similarity searches against major nucleotide-sequence databanks, EST collections, and other gene indices in order to derive possibly informative gene names, application of specialized gene finding methods, as well as frameshift detection, repeat and vector masking, and se-

quence quality analysis and visualization. Most of these developments are directly motivated by specific large-scale annotation projects.

Some systems cater to the needs of small research groups that are not able to invest significant resources in maintaining a bioinformatics infrastructure. As a complement to large multipurpose resources, the easily configurable GANESH system was developed by Huntley et al.[222] to support detailed annotation of selected genomic regions. Alternatively, as a result of the significantly higher computing power available today, it has become possible to create annotation Web servers that accept entire bacterial chromosomes from external users. In order to use the Annotation Service Engine of TIGR (http://www.tigr.org/edutraining/training/annotation_engine.shtml), initial contact by e-mail needs first to be established, while the BASys server runs in a fully automatic fashion and typically returns results within 24 h.[223]

In recent years, the idea of protocol-based genome data processing has been popularized, which draws parallels between the organization of routine bioinformatics analyses and experimental lab work. Just as wet experiments are carefully planned and then executed following a defined sequence of steps, tools such as BioPipe[224] and APAT[225] allow for the creation of customized workflows from standard modules, which typically include XML parsers for a variety of input and output formats, wrappers for running external applications, interfaces to SQL databases and batch processing systems, and facilities for transporting the results to the end user via standard exchange protocols, such as Web services. In contrast to conventional integrated genome analysis systems, protocol-based analysis pipelines have the advantage of being highly configurable and flexible, but their users are required to have a good understanding of software technologies as well as substantial system administration and bioinformatics skills. Recent releases of major genome analysis systems, such as PEDANT, have also been equipped with workflow-based process management (Frishman et al., in preparation).

9. Perspective: Genome Annotation for Systems Biology

The goal of systems biology research is to understand and exploit the relationships between individual molecular components of the cell and the overall behavior of biological systems. The principal approach to this problem involves possibly precise reconstruction of various kinds of biological networks in which genes are involved in order to understand their basic organizational principles and to learn how to build mathematical models capable of predicting how networks react to perturbation. Genome annotation represents the crucial first step in this process, and the quality of annotation data is a determining factor for the success of any subsequent model building and simulation effort. Wrong gene models, incorrectly predicted functional specificities of proteins, missing enzymes in metabolic pathways, corrupted biological networks due to erroneously derived functional associations between genes, as well as limited accuracy and resolution of experimental high-throughput techniques affect the quality of genome-scale metabolic reconstruction.

Beyond the mere quality problems, canonical genome annotation as we know it today does not represent an adequate knowledge basis for developing methods and tools capable of predicting the outcomes of genome-scale experimentation and of guiding systemic interrogation. In the recent publications by leading systems biology groups, a need for introducing additional dimensions to genome annotation has been emphasized. In addition to developing a possibly complete molecular "parts lists" for each organism and the corresponding "wiring diagrams", information about the chemical states of individual components as well as chemical transformations between them needs to be incorporated in order to obtain stoichiometric matrices suitable for building metabolic flux models, monitoring functional states of biological networks, and linking them to phenotypic events.[226] Further dimensions are the spatial distribution of molecular components in the cell and changes in their structure and behavior over time.[227,228]

The ultimate test for the utility of genome annotation in the years to come will be the predictive power of models built upon it. In fact, curation and testing of models should be performed synergistically with curation of genomes as part of an iterative refinement process.[229] While general-purpose genome annotation systems capable of meeting this major challenge currently do not exist, the first pioneering studies demonstrating the feasibility of this approach have been conducted. Model-based improvement of genome annotation involves detecting discrepancies between model predictions and actual phenotype data, identifying molecular components and reactions that would reconcile predictions with observed cellular behavior, using a wide spectrum of context-based bioinformatics methods and literature mining to identify candidate ORFs for missing roles, and then experimentally testing new hypotheses.[228] Sequence-based bioinformatics analyses combined with flux simulations have helped to shed light on the function of currently uncharacterized genes.[230]

10. Conclusions and Outlook

The biological community is attempting to cope with the flood of genome data by using a two-tier strategy. On the one hand, careful manual annotation of selected model organisms is going on, capitalizing on community action as opposed to disparate isolated efforts. Highly curated genomic datasets, such as GO ontology or pathway databases, represent the essential backbone of today's genome informatics. At the same time, the main bulk of other genomic sequences are processed automatically by software tools of ever growing complexity and sophistication that vitally depend on manually curated information. As the speed of sequencing grows and bioinformatics methods mature, we envision closer integration of automatic software tools with sequencing machines, whereby the raw FASTA files they produce now will be replaced by structured biological knowledge. In a quest for personalized approaches to treat diseases, such close integration of sequence information and intelligent analytical tools will allow researchers to speed up the cycle of biological discovery from raw experimental data and its interpretation back to the lab bench. We may live to see the time when the gene function prediction will become a routine readout displayed on a lab monitor just as pH measurements are today.

11. Acknowledgments

The author is indebted to Kaj Albermann, Mark Borodovsky, Ulrich Gueldner, Jean Hani, Thomas Rattei, and Louise Riley for a critical reading of the manuscript and many useful comments. This work was partially supported by the Biosapiens Network of Excellence.

12. References

(1) George, D. G.; Barker, W. C.; Hunt, L. T. *Nucleic Acids Res.* **1986**, *14*, 11.
(2) Bairoch, A.; Boeckmann, B. *Nucleic Acids Res.* **1991**, *19 Suppl*, 2247.
(3) Tateno, Y.; Miyazaki, S.; Ota, M.; Sugawara, H.; Gojobori, T. *Nucleic Acids Res.* **2000**, *28*, 24.
(4) Choi, I. G.; Kim, S. H. *Proc. Natl. Acad. Sci. U.S.A* **2006**, *103*, 14056.
(5) Metzker, M. L. *Genome Res.* **2005**, *15*, 1767.
(6) Service, R. F. *Science* **2006**, *311*, 1544.
(7) Shendure, J.; Mitra, R. D.; Varma, C.; Church, G. M. *Nat. Rev. Genet.* **2004**, *5*, 335.
(8) Riley, M.; Abe, T.; Arnaud, M. B.; Berlyn, M. K.; Blattner, F. R.; Chaudhuri, R. R.; Glasner, J. D.; Horiuchi, T.; Keseler, I. M.; Kosuge, T.; Mori, H.; Perna, N. T.; Plunkett, G., III; Rudd, K. E.; Serres, M. H.; Thomas, G. H.; Thomson, N. R.; Wishart, D.; Wanner, B. L. *Nucleic Acids Res.* **2006**, *34*, 1.
(9) Collins, J. E.; Goward, M. E.; Cole, C. G.; Smink, L. J.; Huckle, E. J.; Knowles, S.; Bye, J. M.; Beare, D. M.; Dunham, I. *Genome Res.* **2003**, *13*, 27.
(10) Jensen, L. J.; Saric, J.; Bork, P. *Nat. Rev. Genet.* **2006**, *7*, 119.
(11) Zhang, Z.; Gerstein, M. *Curr. Opin. Genet. Dev.* **2004**, *14*, 328.
(12) Brown, J. R.; Sanseau, P. *Drug Discovery Today* **2005**, *10*, 595.
(13) Rajewsky, N. *Nat. Genet.* **2006**, *38 Suppl*, S8.
(14) Werner, T. *Brief. Bioinform.* **2003**, *4*, 22.
(15) Gelfand, M. S. *Curr. Opin. Struct. Biol.* **2006**, *16*, 420.
(16) Saghatelian, A.; Cravatt, B. F. *Nat. Chem. Biol.* **2005**, *1*, 130.
(17) Ureta-Vidal, A.; Ettwiller, L.; Birney, E. *Nat. Rev. Genet.* **2003**, *4*, 251.
(18) Stein, L. D. *Nat. Rev. Genet.* **2003**, *4*, 337.
(19) Guigo, R. *Comput. Chem.* **1997**, *21*, 215.
(20) Guigo, R.; Reese, M. G. *Nat. Methods* **2005**, *2*, 575.
(21) Korf, I.; Flicek, P.; Duan, D.; Brent, M. R. *Bioinformatics* **2001**, *17 Suppl 1*, 140.
(22) Nielsen, P.; Krogh, A. *Bioinformatics* **2005**, *21*, 4322.
(23) Brent, M. R. *Genome Res.* **2005**, *15*, 1777.
(24) Besemer, J.; Borodovsky, M. Gene finding. In *Systems Biology, Volume 1, Genomics*; Rigoutsos, I., Stephanopoulos, G., Eds.; Oxford University Press: Oxford, U.K., 2006; pp 118.
(25) Brent, M. R.; Guigo, R. *Curr. Opin. Struct. Biol.* **2004**, *14*, 264.
(26) Ashurst, J. L.; Chen, C. K.; Gilbert, J. G.; Jekosch, K.; Keenan, S.; Meidl, P.; Searle, S. M.; Stalker, J.; Storey, R.; Trevanion, S.; Wilming, L.; Hubbard, T. *Nucleic Acids Res.* **2005**, *33*, 459.
(27) Chen, N.; Lawson, D.; Bradnam, K.; Harris, T. W.; Stein, L. D. *Genome Res.* **2004**, *14*, 2155.
(28) Schoof, H.; Ernst, R.; Nazarov, V.; Pfeifer, L.; Mewes, H. W.; Mayer, K. F. *Nucleic Acids Res.* **2004**, *32*, 373.
(29) Misra, S.; Crosby, M. A.; Mungall, C. J.; Matthews, B. B.; Campbell, K. S.; Hradecky, P.; Huang, Y.; Kaminker, J. S.; Millburn, G. H.; Prochnik, S. E.; Smith, C. D.; Tupy, J. L.; Whitfied, E. J.; Bayraktaroglu, L.; Berman, B. P.; Bettencourt, B. R.; Celniker, S. E.; de Grey, A. D.; Drysdale, R. A.; Harris, N. L.; Richter, J.; Russo, S.; Schroeder, A. J.; Shu, S. Q.; Stapleton, M.; Yamada, C.; Ashburner, M.; Gelbart, W. M.; Rubin, G. M.; Lewis, S. E. *Genome Biol.* **2002**, *3*, RESEARCH0083.
(30) Curwen, V.; Eyras, E.; Andrews, T. D.; Clarke, L.; Mongin, E.; Searle, S. M.; Clamp, M. *Genome Res.* **2004**, *14*, 942.
(31) Shah, S. P.; McVicker, G. P.; Mackworth, A. K.; Rogic, S.; Ouellette, B. F. *Bioinformatics* **2003**, *19*, 1296.
(32) Burge, C.; Karlin, S. *J. Mol. Biol.* **1997**, *268*, 78.
(33) Krogh, A. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1997**, *5*, 179.
(34) Allen, J. E.; Salzberg, S. L. *Bioinformatics* **2005**, *21*, 3596.
(35) Hubbard, T.; Birney, E. *Nature* **2000**, *403*, 825.
(36) Wilkerson, M. D.; Schlueter, S. D.; Brendel, V. *Genome Biol.* **2006**, *7*, R58.
(37) Frishman, D.; Mironov, A.; Mewes, H. W.; Gelfand, M. *Nucleic Acids Res.* **1998**, *26*, 2941.
(38) Delcher, A. L.; Harmon, D.; Kasif, S.; White, O.; Salzberg, S. L. *Nucleic Acids Res.* **1999**, *27*, 4636.
(39) Besemer, J.; Lomsadze, A.; Borodovsky, M. *Nucleic Acids Res.* **2001**, *29*, 2607.
(40) Lomsadze, A.; Ter-Hovhannisyan, V.; Chernoff, Y. O.; Borodovsky, M. *Nucleic Acids Res.* **2005**, *33*, 6494.
(41) Birney, E.; Clamp, M.; Durbin, R. *Genome Res.* **2004**, *14*, 988.
(42) Mironov, A. A.; Novichkov, P. S.; Gelfand, M. S. *Bioinformatics* **2001**, *17*, 13.
(43) Schiex, T.; Gouzy, J.; Moisan, A.; de, O. Y. *Nucleic Acids Res.* **2003**, *31*, 3738.
(44) Lottaz, C.; Iseli, C.; Jongeneel, C. V.; Bucher, P. *Bioinformatics* **2003**, *19 Suppl 2*, 103.
(45) Dwight, S. S.; Balakrishnan, R.; Christie, K. R.; Costanzo, M. C.; Dolinski, K.; Engel, S. R.; Feierbach, B.; Fisk, D. G.; Hirschman, J.; Hong, E. L.; Issel-Tarver, L.; Nash, R. S.; Sethuraman, A.; Starr, B.; Theesfeld, C. L.; Andrada, R.; Binkley, G.; Dong, Q.; Lane, C.; Schroeder, M.; Weng, S.; Botstein, D.; Cherry, J. M. *Brief. Bioinform.* **2004**, *5*, 9.
(46) Salimi, N.; Vita, R. *PLoS Comput. Biol.* **2006**, *2*, e125.
(47) Ashburner, M.; Bergman, C. M. *Genome Res.* **2005**, *15*, 1661.
(48) Winsor, G. L.; Lo, R.; Sui, S. J.; Ung, K. S.; Huang, S.; Cheng, D.; Ching, W. K.; Hancock, R. E.; Brinkman, F. S. *Nucleic Acids Res.* **2005**, *33*, 338.
(49) Keseler, I. M.; Collado-Vides, J.; Gama-Castro, S.; Ingraham, J.; Paley, S.; Paulsen, I. T.; Peralta-Gil, M.; Karp, P. D. *Nucleic Acids Res.* **2005**, *33*, 334.
(50) Guldener, U.; Munsterkotter, M.; Kastenmuller, G.; Strack, N.; van Helden, J.; Lemer, C.; Richelles, J.; Wodak, S. J.; Garcia-Martinez, J.; Perez-Ortin, J. E.; Michael, H.; Kaps, A.; Talla, E.; Dujon, B.; Andre, B.; Souciet, J. L.; De, M. J.; Bon, E.; Gaillardin, C.; Mewes, H. W. *Nucleic Acids Res.* **2005**, *33*, D364.
(51) Liolios, K.; Tavernarakis, N.; Hugenholtz, P.; Kyrpides, N. C. *Nucleic Acids Res.* **2006**, *34*, D332.
(52) Salwinski, L.; Miller, C. S.; Smith, A. J.; Pettit, F. K.; Bowie, J. U.; Eisenberg, D. *Nucleic Acids Res.* **2004**, *32*, D449.
(53) Guldener, U.; Munsterkotter, M.; Oesterheld, M.; Pagel, P.; Ruepp, A.; Mewes, H. W.; Stumpflen, V. *Nucleic Acids Res.* **2006**, *34*, D436.
(54) von Mering, C.; Krause, R.; Snel, B.; Cornell, M.; Oliver, S. G.; Fields, S.; Bork, P. *Nature* **2002**, *417*, 399.
(55) Jansen, R.; Yu, H.; Greenbaum, D.; Kluger, Y.; Krogan, N. J.; Chung, S.; Emili, A.; Snyder, M.; Greenblatt, J. F.; Gerstein, M. *Science* **2003**, *302*, 449.
(56) Pagel, P.; Wong, P.; Frishman, D. *J. Mol. Biol.* **2004**, *344*, 1331.
(57) Mishra, G. R.; Suresh, M.; Kumaran, K.; Kannabiran, N.; Suresh, S.; Bala, P.; Shivakumar, K.; Anuradha, N.; Reddy, R.; Raghavan, T. M.; Menon, S.; Hanumanthu, G.; Gupta, M.; Upendran, S.; Gupta, S.; Mahesh, M.; Jacob, B.; Mathew, P.; Chatterjee, P.; Arun, K. S.; Sharma, S.; Chandrika, K. N.; Deshpande, N.; Palvankar, K.; Raghavnath, R.; Krishnakanth, R.; Karathia, H.; Rekha, B.; Nayak, R.; Vishnupriya, G.; Kumar, H. G.; Nagini, M.; Kumar, G. S.; Jose, R.; Deepthi, P.; Mohan, S. S.; Gandhi, T. K.; Harsha, H. C.; Deshpande, K. S.; Sarker, M.; Prasad, T. S.; Pandey, A. *Nucleic Acids Res.* **2006**, *34*, D411.
(58) Bard, J. B.; Rhee, S. Y. *Nat. Rev. Genet.* **2004**, *5*, 213.
(59) Riley, M. *Microbiol. Rev.* **1993**, *57*, 862.
(60) Serres, M. H.; Riley, M. *Microb. Comp. Genomics* **2000**, *5*, 205.
(61) Mewes, H. W.; Albermann, K.; Bahr, M.; Frishman, D.; Gleissner, A.; Hani, J.; Heumann, K.; Kleine, K.; Maierl, A.; Oliver, S. G.; Pfeiffer, F.; Zollner, A. *Nature* **1997**, *387*, 7.
(62) Ruepp, A.; Zollner, A.; Maier, D.; Albermann, K.; Hani, J.; Mokrejs, M.; Tetko, I.; Guldener, U.; Mannhaupt, G.; Munsterkotter, M.; Mewes, H. W. *Nucleic Acids Res.* **2004**, *32*, 5539.
(63) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. *Nat. Genet.* **2000**, *25*, 25.
(64) Camon, E.; Magrane, M.; Barrell, D.; Lee, V.; Dimmer, E.; Maslen, J.; Binns, D.; Harte, N.; Lopez, R.; Apweiler, R. *Nucleic Acids Res.* **2004**, *32*, D262.
(65) Beissbarth, T. *Methods Enzymol.* **2006**, *411*, 340.
(66) Wu, C. H.; Apweiler, R.; Bairoch, A.; Natale, D. A.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Mazumder, R.; O'Donovan, C.; Redaschi, N.; Suzek, B. *Nucleic Acids Res.* **2006**, *34*, D187.
(67) Emanuelsson, O.; Nielsen, H.; Brunak, S.; von Heijne, G. *J. Mol. Biol.* **2000**, *300*, 1005.
(68) Nakai, K.; Horton, P. *Trends Biochem. Sci.* **1999**, *24*, 34.
(69) Rey, S.; Acab, M.; Gardy, J. L.; Laird, M. R.; deFays, K.; Lambert, C.; Brinkman, F. S. *Nucleic Acids Res.* **2005**, *33*, D164.
(70) Wiwatwattana, N.; Kumar, A. *Nucleic Acids Res.* **2005**, *33*, D598.
(71) Schomburg, I.; Chang, A.; Ebeling, C.; Gremse, M.; Heldt, C.; Huhn, G.; Schomburg, D. *Nucleic Acids Res.* **2004**, *32*, D431.
(72) Kanehisa, M.; Goto, S.; Hattori, M.; oki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. *Nucleic Acids Res.* **2006**, *34*, D354.
(73) Caspi, R.; Foerster, H.; Fulcher, C. A.; Hopkinson, R.; Ingraham, J.; Kaipa, P.; Krummenacker, M.; Paley, S.; Pick, J.; Rhee, S. Y.; Tissier, C.; Zhang, P.; Karp, P. D. *Nucleic Acids Res.* **2006**, *34*, D511.
(74) Joshi-Tope, G.; Gillespie, M.; Vastrik, I.; D'Eustachio, P.; Schmidt, E.; de Bono, B.; Jassal, B.; Gopinath, G. R.; Wu, G. R.; Matthews, L.; Lewis, S.; Birney, E.; Stein, L. *Nucleic Acids Res.* **2005**, *33*, D428.
(75) Tatusov, R. L.; Fedorova, N. D.; Jackson, J. D.; Jacobs, A. R.; Kiryutin, B.; Koonin, E. V.; Krylov, D. M.; Mazumder, R.; Mekhedov, S. L.; Nikolskaya, A. N.; Rao, B. S.; Smirnov, S.; Sverdlov, A. V.; Vasudevan, S.; Wolf, Y. I.; Yin, J. J.; Natale, D. A. *BMC Bioinf.* **2003**, *4*, 41.

(76) Hartwell, L. H.; Hopfield, J. J.; Leibler, S.; Murray, A. W. *Nature* **1999**, *402*, C47−C52.
(77) Hofmann, K. P.; Spahn, C. M.; Heinrich, R.; Heinemann, U. *Trends Biochem. Sci.* **2006**, *31*, 497.
(78) Gavin, A. C.; Aloy, P.; Grandi, P.; Krause, R.; Boesche, M.; Marzioch, M.; Rau, C.; Jensen, L. J.; Bastuck, S.; Dumpelfeld, B.; Edelmann, A.; Heurtier, M. A.; Hoffman, V.; Hoefert, C.; Klein, K.; Hudak, M.; Michon, A. M.; Schelder, M.; Schirle, M.; Remor, M.; Rudi, T.; Hooper, S.; Bauer, A.; Bouwmeester, T.; Casari, G.; Drewes, G.; Neubauer, G.; Rick, J. M.; Kuster, B.; Bork, P.; Russell, R. B.; Superti-Furga, G. *Nature* **2006**, *440*, 631.
(79) Snel, B.; Huynen, M. A. *Genome Res.* **2004**, *14*, 391.
(80) Pereira-Leal, J. B.; Teichmann, S. A. *Genome Res.* **2005**, *15*, 552.
(81) Spirin, V.; Gelfand, M. S.; Mironov, A. A.; Mirny, L. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8774.
(82) Chen, Y.; Dokholyan, N. V. *Trends Genet.* **2006**, *22*, 416.
(83) Mewes, H. W.; Frishman, D.; Mayer, K. F.; Munsterkotter, M.; Noubibou, O.; Pagel, P.; Rattei, T.; Oesterheld, M.; Ruepp, A.; Stumpflen, V. *Nucleic Acids Res.* **2006**, *34*, D169.
(84) Snel, B.; Bork, P.; Huynen, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 5890.
(85) Spirin, V.; Mirny, L. A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12123.
(86) Yu, H.; Zhu, X.; Greenbaum, D.; Karro, J.; Gerstein, M. *Nucleic Acids Res.* **2004**, *32*, 328.
(87) Tanay, A.; Sharan, R.; Kupiec, M.; Shamir, R. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2981.
(88) Antonov, A. V.; Tetko, I. V.; Mewes, H. W. *Nucleic Acids Res.* **2006**, *34*, e6.
(89) Petti, A. A.; Church, G. M. *Genome Res.* **2005**, *15*, 1298.
(90) Overbeek, R.; Begley, T.; Butler, R. M.; Choudhuri, J. V.; Chuang, H. Y.; Cohoon, M.; de Crecy-Lagard, V.; Diaz, N.; Disz, T.; Edwards, R.; Fonstein, M.; Frank, E. D.; Gerdes, S.; Glass, E. M.; Goesmann, A.; Hanson, A.; Iwata-Reuyl, D.; Jensen, R.; Jamshidi, N.; Krause, L.; Kubal, M.; Larsen, N.; Linke, B.; McHardy, A. C.; Meyer, F.; Neuweger, H.; Olsen, G.; Olson, R.; Osterman, A.; Portnoy, V.; Pusch, G. D.; Rodionov, D. A.; Ruckert, C.; Steiner, J.; Stevens, R.; Thiele, I.; Vassieva, O.; Ye, Y.; Zagnitko, O.; Vonstein, V. *Nucleic Acids Res.* **2005**, *33*, 5691.
(91) Bork, P. *Genome Res.* **2000**, *10*, 398.
(92) Siew, N.; Fischer, D. *Structure* **2003**, *11*, 7.
(93) Roberts, R. J. *PLoS Biol.* **2004**, *2*, E42.
(94) Galperin, M. Y.; Koonin, E. V. *Nucleic Acids Res.* **2004**, *32*, 5452.
(95) Smialowski, P.; Martin-Galiano, A. J.; Cox, J.; Frishman, D. *Curr. Protein Pept. Sci.* **2007**, *8*, 121.
(96) Brenner, S. E. *Nat. Struct. Biol.* **2000**, *7 Suppl*, 967.
(97) Karp, P. D. *Genome Biol.* **2004**, *5*, 401.
(98) Kolker, E.; Picone, A. F.; Galperin, M. Y.; Romine, M. F.; Higdon, R.; Makarova, K. S.; Kolker, N.; Anderson, G. A.; Qiu, X.; Auberry, K. J.; Babnigg, G.; Beliaev, A. S.; Edlefsen, P.; Elias, D. A.; Gorby, Y. A.; Holzman, T.; Klappenbach, J. A.; Konstantinidis, K. T.; Land, M. L.; Lipton, M. S.; McCue, L. A.; Monroe, M.; Pasa-Tolic, L.; Pinchuk, G.; Purvine, S.; Serres, M. H.; Tsapin, S.; Zakrajsek, B. A.; Zhu, W.; Zhou, J.; Larimer, F. W.; Lawrence, C. E.; Riley, M.; Collart, F. R.; Yates, J. R., III; Smith, R. D.; Giometti, C. S.; Nealson, K. H.; Fredrickson, J. K.; Tiedje, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2099.
(99) Encode project consortium. *Science* **2004**, *306*, 636.
(100) Brzoska, P. M.; Brown, C.; Cassel, M.; Ceccardi, T.; Di, F. V.; Dubman, A.; Evans, J.; Fang, R.; Harris, M.; Hoover, J.; Hu, F.; Larry, C.; Li, P.; Malicdem, M.; Maltchenko, S.; Shannon, M.; Perkins, S.; Poulter, K.; Webster-Laig, M.; Xiao, C.; Young, S.; Spier, G.; Guegler, K.; Gilbert, D.; Samaha, R. R. *Genomics* **2006**, *87*, 437.
(101) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389.
(102) Pearson, W. R. *Methods Enzymol.* **1990**, *183*, 63.
(103) Tcherepanov, V.; Ehlers, A.; Upton, C. *BMC Genomics* **2006**, *7*, 150.
(104) Sander, C.; Schneider, R. *Proteins* **1991**, *9*, 56.
(105) Abagyan, R. A.; Batalov, S. *J. Mol. Biol.* **1997**, *273*, 355.
(106) Wilson, C. A.; Kreychman, J.; Gerstein, M. *J. Mol. Biol.* **2000**, *297*, 233.
(107) Devos, D.; Valencia, A. *Proteins* **2000**, *41*, 98.
(108) Hegyi, H.; Gerstein, M. *Genome Res.* **2001**, *11*, 1632.
(109) Rost, B. *J. Mol. Biol.* **2002**, *318*, 595.
(110) Tian, W.; Skolnick, J. *J. Mol. Biol.* **2003**, *333*, 863.
(111) Todd, A. E.; Orengo, C. A.; Thornton, J. M. *Structure* **2002**, *10*, 1435.
(112) Koonin, E. V. *Annu. Rev. Genet.* **2005**, *39*, 309.
(113) Storm, C. E.; Sonnhammer, E. L. *Bioinformatics* **2002**, *18*, 92.
(114) O'Brien, K. P.; Remm, M.; Sonnhammer, E. L. *Nucleic Acids Res.* **2005**, *33*, D476.
(115) Hulsen, T.; Huynen, M. A.; de Vlieg, J.; Groenen, P. M. *Genome Biol.* **2006**, *7*, R31.
(116) Yuan, Y. P.; Eulenstein, O.; Vingron, M.; Bork, P. *Bioinformatics* **1998**, *14*, 285.
(117) Eisen, J. A. *Genome Res.* **1998**, *8*, 163.
(118) Sjölander, K. *Bioinformatics* **2004**, *20*, 170.
(119) Brown, D.; Sjölander, K. *PLoS Comput. Biol.* **2006**, *2*, e77.
(120) Engelhardt, B. E.; Jordan, M. I.; Muratore, K. E.; Brenner, S. E. *PLoS Comput. Biol.* **2005**, *1*, e45.
(121) Gouret, P.; Vitiello, V.; Balandraud, N.; Gilles, A.; Pontarotti, P.; Danchin, E. G. *BMC Bioinf.* **2005**, *6*, 198.
(122) Zehetner, G. *Nucleic Acids Res.* **2003**, *31*, 3799.
(123) Hennig, S.; Groth, D.; Lehrach, H. *Nucleic Acids Res.* **2003**, *31*, 3712.
(124) Xie, H.; Wasserman, A.; Levine, Z.; Novik, A.; Grebinskiy, V.; Shoshan, A.; Mintz, L. *Genome Res.* **2002**, *12*, 785.
(125) Vinayagam, A.; del Val, C.; Schubert, F.; Eils, R.; Glatting, K. H.; Suhai, S.; Konig, R. *BMC Bioinf.* **2006**, *7*, 161.
(126) Tian, W.; Arakaki, A. K.; Skolnick, J. *Nucleic Acids Res.* **2004**, *32*, 6226.
(127) Raychaudhuri, S.; Chang, J. T.; Sutphin, P. D.; Altman, R. B. *Genome Res.* **2002**, *12*, 203.
(128) Camon, E. B.; Barrell, D. G.; Dimmer, E. C.; Lee, V.; Magrane, M.; Maslen, J.; Binns, D.; Apweiler, R. *BMC Bioinf.* **2005**, *6 Suppl 1*, S17.
(129) Jensen, L. J.; Gupta, R.; Blom, N.; Devos, D.; Tamames, J.; Kesmir, C.; Nielsen, H.; Staerfeldt, H. H.; Rapacki, K.; Workman, C.; Andersen, C. A.; Knudsen, S.; Krogh, A.; Valencia, A.; Brunak, S. *J. Mol. Biol.* **2002**, *319*, 1257.
(130) Jensen, L. J.; Gupta, R.; Staerfeldt, H. H.; Brunak, S. *Bioinformatics* **2003**, *19*, 635.
(131) Barutcuoglu, Z.; Schapire, R. E.; Troyanskaya, O. G. *Bioinformatics* **2006**, *22*, 830.
(132) Khan, S.; Situ, G.; Decker, K.; Schmidt, C. J. *Bioinformatics* **2003**, *19*, 2484.
(133) Martin, D. M.; Berriman, M.; Barton, G. J. *BMC Bioinf.* **2004**, *5*, 178.
(134) Verspoor, K.; Cohn, J.; Mniszewski, S.; Joslyn, C. *Protein Sci.* **2006**, *15*, 1544.
(135) Schwikowski, B.; Uetz, P.; Fields, S. *Nat. Biotechnol.* **2000**, *18*, 1257.
(136) Hishigaki, H.; Nakai, K.; Ono, T.; Tanigami, A.; Takagi, T. *Yeast* **2001**, *18*, 523.
(137) Nabieva, E.; Jim, K.; Agarwal, A.; Chazelle, B.; Singh, M. *Bioinformatics* **2005**, *21 Suppl 1*, i302.
(138) McDermott, J.; Bumgarner, R.; Samudrala, R. *Bioinformatics* **2005**, *21*, 3217.
(139) Vazquez, A.; Flammini, A.; Maritan, A.; Vespignani, A. *Nat. Biotechnol.* **2003**, *21*, 697.
(140) Massjouni, N.; Rivera, C. G.; Murali, T. M. *Nucleic Acids Res.* **2006**, *34*, W340.
(141) Karaoz, U.; Murali, T. M.; Letovsky, S.; Zheng, Y.; Ding, C.; Cantor, C. R.; Kasif, S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2888.
(142) Dandekar, T.; Snel, B.; Huynen, M.; Bork, P. *Trends Biochem. Sci.* **1998**, *23*, 324.
(143) Overbeek, R.; Fonstein, M.; D'Souza, M.; Pusch, G. D.; Maltsev, N. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2896.
(144) Pellegrini, M.; Marcotte, E. M.; Thompson, M. J.; Eisenberg, D.; Yeates, T. O. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 4285.
(145) Enright, A. J.; Iliopoulos, I.; Kyrpides, N. C.; Ouzounis, C. A. *Nature* **1999**, *402*, 86.
(146) van Noort, V.; Snel, B.; Huynen, M. A. *Trends Genet.* **2003**, *19*, 238.
(147) Zheng, Y.; Anton, B. P.; Roberts, R. J.; Kasif, S. *BMC Bioinf.* **2005**, *6*, 243.
(148) Barker, D.; Pagel, M. *PLoS Comput. Biol.* **2005**, *1*, e3.
(149) Pawson, T.; Nash, P. *Science* **2003**, *300*, 445.
(150) Pasek, S.; Bergeron, A.; Risler, J. L.; Louis, A.; Ollivier, E.; Raffinot, M. *Genome Res.* **2005**, *15*, 867.
(151) Mellor, J. C.; Yanai, I.; Clodfelter, K. H.; Mintseris, J.; DeLisi, C. *Nucleic Acids Res.* **2002**, *30*, 306.
(152) Bowers, P. M.; Pellegrini, M.; Thompson, M. J.; Fierro, J.; Yeates, T. O.; Eisenberg, D. *Genome Biol.* **2004**, *5*, R35.
(153) Enault, F.; Suhre, K.; Claverie, J. M. *BMC Bioinf.* **2005**, *6*, 247.
(154) von Mering, C.; Jensen, L. J.; Snel, B.; Hooper, S. D.; Krupp, M.; Foglierini, M.; Jouffre, N.; Huynen, M. A.; Bork, P. *Nucleic Acids Res.* **2005**, *33*, D433.
(155) Wolf, Y. I.; Rogozin, I. B.; Kondrashov, A. S.; Koonin, E. V. *Genome Res.* **2001**, *11*, 356.
(156) Doerks, T.; von Mering, C.; Bork, P. *Nucleic Acids Res.* **2004**, *32*, 6321.
(157) Osterman, A.; Overbeek, R. *Curr. Opin. Chem. Biol.* **2003**, *7*, 238.
(158) Chen, L.; Vitkup, D. *Genome Biol.* **2006**, *7*, R17.
(159) Date, S. V.; Marcotte, E. M. *Nat. Biotechnol.* **2003**, *21*, 1055.
(160) Korbel, J. O.; Doerks, T.; Jensen, L. J.; Perez-Iratxeta, C.; Kaczanowski, S.; Hooper, S. D.; Andrade, M. A.; Bork, P. *PLoS Biol.* **2005**, *3*, e134.

(161) Bork, P.; Bairoch, A. *Trends Genet.* **1996**, *12*, 425.
(162) Galperin, M. Y.; Koonin, E. V. *In Silico Biol.* **1998**, *1*, 55.
(163) Gilks, W. R.; Audit, B.; de Angelis, D.; Tsoka, S.; Ouzounis, C. A. *Bioinformatics* **2002**, *18*, 1641.
(164) Brenner, S. E. *Trends Genet.* **1999**, *15*, 132.
(165) Mi, H.; Vandergriff, J.; Campbell, M.; Narechania, A.; Majoros, W.; Lewis, S.; Thomas, P. D.; Ashburner, M. *Genome Res.* **2003**, *13*, 2118.
(166) Mi, H.; Lazareva-Ulitsky, B.; Loo, R.; Kejariwal, A.; Vandergriff, J.; Rabkin, S.; Guo, N.; Muruganujan, A.; Doremieux, O.; Campbell, M. J.; Kitano, H.; Thomas, P. D. *Nucleic Acids Res.* **2005**, *33*, D284.
(167) Andrade, M. A.; Brown, N. P.; Leroy, C.; Hoersch, S.; de Daruvar, A.; Reich, C.; Franchini, A.; Tamames, J.; Valencia, A.; Ouzounis, C.; Sander, C. *Bioinformatics* **1999**, *15*, 391.
(168) Iliopoulos, I.; Tsoka, S.; Andrade, M. A.; Enright, A. J.; Carroll, M.; Poullet, P.; Promponas, V.; Liakopoulos, T.; Palaios, G.; Pasquier, C.; Hamodrakas, S.; Tamames, J.; Yagnik, A. T.; Tramontano, A.; Devos, D.; Blaschke, C.; Valencia, A.; Brett, D.; Martin, D.; Leroy, C.; Rigoutsos, I.; Sander, C.; Ouzounis, C. A. *Bioinformatics* **2003**, *19*, 717.
(169) Tetko, I. V.; Brauner, B.; Dunger-Kaltenbach, I.; Frishman, G.; Montrone, C.; Fobo, G.; Ruepp, A.; Antonov, A. V.; Surmeli, D.; Mewes, H. W. *Bioinformatics* **2005**, *21*, 2520.
(170) Mulder, N. J.; Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Binns, D.; Bradley, P.; Bork, P.; Bucher, P.; Cerutti, L.; Copley, R.; Courcelle, E.; Das, U.; Durbin, R.; Fleischmann, W.; Gough, J.; Haft, D.; Harte, N.; Hulo, N.; Kahn, D.; Kanapin, A.; Krestyaninova, M.; Lonsdale, D.; Lopez, R.; Letunic, I.; Madera, M.; Maslen, J.; McDowall, J.; Mitchell, A.; Nikolskaya, A. N.; Orchard, S.; Pagni, M.; Ponting, C. P.; Quevillon, E.; Selengut, J.; Sigrist, C. J.; Silventoinen, V.; Studholme, D. J.; Vaughan, R.; Wu, C. H. *Nucleic Acids Res.* **2005**, *33*, D201.
(171) Kaplan, N.; Vaaknin, A.; Linial, M. *Nucleic Acids Res.* **2003**, *31*, 5617.
(172) Kaplan, N.; Linial, M. *BMC Bioinf.* **2005**, *6*, 46.
(173) Kunin, V.; Ouzounis, C. A. *BMC Bioinf.* **2005**, *6*, 24.
(174) Hu, Z. Z.; Narayanaswamy, M.; Ravikumar, K. E.; Vijay-Shanker, K.; Wu, C. H. *Bioinformatics* **2005**, *21*, 2759.
(175) Tamames, J. *BMC Bioinf.* **2005**, *6 Suppl 1*, S10.
(176) Michailidis, G.; Shedden, K. *J. Comput. Biol.* **2003**, *10*, 689.
(177) Eisenhaber, F.; Bork, P. *Bioinformatics* **1999**, *15*, 528.
(178) Yu, G. X. *J. Bioinf. Comput. Biol.* **2004**, *2*, 615.
(179) Yu, G. X.; Glass, E. M.; Karonis, N. T.; Maltsev, N. *Proteins* **2005**, *61*, 907.
(180) Fleischmann, W.; Moller, S.; Gateau, A.; Apweiler, R. *Bioinformatics* **1999**, *15*, 228.
(181) Gattiker, A.; Michoud, K.; Rivoire, C.; Auchincloss, A. H.; Coudert, E.; Lima, T.; Kersey, P.; Pagni, M.; Sigrist, C. J.; Lachaize, C.; Veuthey, A. L.; Gasteiger, E.; Bairoch, A. *Comput. Biol. Chem.* **2003**, *27*, 49.
(182) Wu, C. H.; Huang, H.; Yeh, L. S.; Barker, W. C. *Comput. Biol. Chem.* **2003**, *27*, 37.
(183) Kasukawa, T.; Furuno, M.; Nikaido, I.; Bono, H.; Hume, D. A.; Bult, C.; Hill, D. P.; Baldarelli, R.; Gough, J.; Kanapin, A.; Matsuda, H.; Schriml, L. M.; Hayashizaki, Y.; Okazaki, Y.; Quackenbush, J. *Genome Res.* **2003**, *13*, 1542.
(184) Kretschmann, E.; Fleischmann, W.; Apweiler, R. *Bioinformatics* **2001**, *17*, 920.
(185) Wieser, D.; Kretschmann, E.; Apweiler, R. *Bioinformatics* **2004**, *20 Suppl 1*, I342.
(186) Agrawal, R.; Srikant, R. *Fast algorithms for mining association rules*; Proc. of the 20th International Conference on Very Large Databases, 1994, pp 487−499.
(187) Bodenreider, O.; Aubry, M.; Burgun, A. *Pac. Symp. Biocomput.* **2005**, 91.
(188) Creighton, C.; Hanash, S. *Bioinformatics* **2003**, *19*, 79.
(189) Satou, K.; Shibayama, G.; Ono, T.; Yamamura, Y.; Furuichi, E.; Kuhara, S.; Takagi, T. *Pac. Symp. Biocomput.* **1997**, 397.
(190) Artamonova, I. I.; Frishman, G.; Gelfand, M. S.; Frishman, D. *Bioinformatics* **2005**, *21*, iii49.
(191) D'Ascenzo, M. D.; Collmer, A.; Martin, G. B. *Nucleic Acids Res.* **2004**, *32*, 3124.
(192) Stein, L. D.; Mungall, C.; Shu, S.; Caudy, M.; Mangone, M.; Day, A.; Nickerson, E.; Stajich, J. E.; Harris, T. W.; Arva, A.; Lewis, S. *Genome Res.* **2002**, *12*, 1599.
(193) Dowell, R. D.; Jokerst, R. M.; Day, A.; Eddy, S. R.; Stein, L. *BMC Bioinf.* **2001**, *2*, 7.
(194) Hammond, M. P.; Birney, E. *Trends Genet.* **2004**, *20*, 268.
(195) Reeves, G. A.; Thornton, J. M. *Hum. Mol. Genet.* **2006**, *15 Spec No 1*, R81.
(196) Searle, S. M.; Gilbert, J.; Iyer, V.; Clamp, M. *Genome Res.* **2004**, *14*, 963.
(197) Glasner, J. D.; Liss, P.; Plunkett, G., III; Darling, A.; Prasad, T.; Rusch, M.; Byrnes, A.; Gilson, M.; Biehl, B.; Blattner, F. R.; Perna, N. T. *Nucleic Acids Res.* **2003**, *31*, 147.
(198) Rutherford, K.; Parkhill, J.; Crook, J.; Horsnell, T.; Rice, P.; Rajandream, M. A.; Barrell, B. *Bioinformatics* **2000**, *16*, 944.
(199) Lewis, S. E.; Searle, S. M.; Harris, N.; Gibson, M.; Lyer, V.; Richter, J.; Wiel, C.; Bayraktaroglir, L.; Birney, E.; Crosby, M. A.; Kaminker, J. S.; Matthews, B. B.; Prochnik, S. E.; Smithy, C. D.; Tupy, J. L.; Rubin, G. M.; Misra, S.; Mungall, C. J.; Clamp, M. E. *Genome Biol.* **2002**, *3*, RESEARCH0082.
(200) Bork, P.; Ouzounis, C.; Sander, C.; Scharf, M.; Schneider, R.; Sonnhammer, E. *Protein Sci.* **1992**, *1*, 1677.
(201) Scharf, M.; Schneider, R.; Casari, G.; Bork, P.; Valencia, A.; Ouzounis, C.; Sander, C. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1994**, *2*, 348.
(202) Gaasterland, T.; Sensen, C. W. *Trends Genet.* **1996**, *12*, 76.
(203) Frishman, D.; Mewes, H. W. *Trends Genet.* **1997**, *13*, 415.
(204) Harris, N. L. *Genome Res.* **1997**, *7*, 754.
(205) Stein, L. D.; Thierry-Mieg, J. *Comput. Sci. Eng.* **1999**, *1*, 44.
(206) Overbeek, R.; Larsen, N.; Pusch, G. D.; D'Souza, M.; Selkov, E.; Kyrpides, N.; Fonstein, M.; Maltsev, N.; Selkov, E., Jr. *Nucleic Acids Res.* **2000**, *28*, 123.
(207) Maltsev, N.; Glass, E.; Sulakhe, D.; Rodriguez, A.; Syed, M. H.; Bompada, T.; Zhang, Y.; D'Souza, M. *Nucleic Acids Res.* **2006**, *34*, D369.
(208) Hinrichs, A. S.; Karolchik, D.; Baertsch, R.; Barber, G. P.; Bejerano, G.; Clawson, H.; Diekhans, M.; Furey, T. S.; Harte, R. A.; Hsu, F.; Hillman-Jackson, J.; Kuhn, R. M.; Pedersen, J. S.; Pohl, A.; Raney, B. J.; Rosenbloom, K. R.; Siepel, A.; Smith, K. E.; Sugnet, C. W.; Sultan-Qurraie, A.; Thomas, D. J.; Trumbower, H.; Weber, R. J.; Weirauch, M.; Zweig, A. S.; Haussler, D.; Kent, W. J. *Nucleic Acids Res.* **2006**, *34*, D590.
(209) Birney, E.; Andrews, D.; Caccamo, M.; Chen, Y.; Clarke, L.; Coates, G.; Cox, T.; Cunningham, F.; Curwen, V.; Cutts, T.; Down, T.; Durbin, R.; Fernandez-Suarez, X. M.; Flicek, P.; Graf, S.; Hammond, M.; Herrero, J.; Howe, K.; Iyer, V.; Jekosch, K.; Kahari, A.; Kasprzyk, A.; Keefe, D.; Kokocinski, F.; Kulesha, E.; London, D.; Longden, I.; Melsopp, C.; Meidl, P.; Overduin, B.; Parker, A.; Proctor, G.; Prlic, A.; Rae, M.; Rios, D.; Redmond, S.; Schuster, M.; Sealy, I.; Searle, S.; Severin, J.; Slater, G.; Smedley, D.; Smith, J.; Stabenau, A.; Stalker, J.; Trevanion, S.; Ureta-Vidal, A.; Vogel, J.; White, S.; Woodwark, C.; Hubbard, T. J. *Nucleic Acids Res.* **2006**, *34*, D556.
(210) Peterson, J. D.; Umayam, L. A.; Dickinson, T.; Hickey, E. K.; White, O. *Nucleic Acids Res.* **2001**, *29*, 123.
(211) Meyer, F.; Goesmann, A.; McHardy, A. C.; Bartels, D.; Bekel, T.; Clausen, J.; Kalinowski, J.; Linke, B.; Rupp, O.; Giegerich, R.; Puhler, A. *Nucleic Acids Res.* **2003**, *31*, 2187.
(212) Alm, E. J.; Huang, K. H.; Price, M. N.; Koche, R. P.; Keller, K.; Dubchak, I. L.; Arkin, A. P. *Genome Res.* **2005**, *15*, 1015.
(213) Zhou, C. L.; Lam, M. W.; Smith, J. R.; Zemla, A. T.; Dyer, M. D.; Kuczmarski, T. A.; Vitalis, E. A.; Slezak, T. R. *BMC Bioinf.* **2006**, *7*, 459.
(214) Vallenet, D.; Labarre, L.; Rouy, Z.; Barbe, V.; Bocs, S.; Cruveiller, S.; Lajus, A.; Pascal, G.; Scarpelli, C.; Medigue, C. *Nucleic Acids Res.* **2006**, *34*, 53.
(215) Bryson, K.; Loux, V.; Bossy, R.; Nicolas, P.; Chaillou, S.; van de Guchte, M.; Penaud, S.; Maguin, E.; Hoebeke, M.; Bessieres, P.; Gibrat, J. F. *Nucleic Acids Res.* **2006**, *34*, 3533.
(216) Tanaka, N.; Abe, T.; Miyazaki, S.; Sugawara, H. *BMC Bioinf.* **2006**, *7*, 368.
(217) Markowitz, V. M.; Korzeniewski, F.; Palaniappan, K.; Szeto, E.; Werner, G.; Padki, A.; Zhao, X.; Dubchak, I.; Hugenholtz, P.; Anderson, I.; Lykidis, A.; Mavromatis, K.; Ivanova, N.; Kyrpides, N. C. *Nucleic Acids Res.* **2006**, *34*, D344.
(218) Markowitz, V. M.; Ivanova, N.; Palaniappan, K.; Szeto, E.; Korzeniewski, F.; Lykidis, A.; Anderson, I.; Mavromatis, K.; Kunin, V.; Garcia, M. H.; Dubchak, I.; Hugenholtz, P.; Kyrpides, N. C. *Bioinformatics* **2006**, *22*, e359.
(219) Liang, F.; Holt, I.; Pertea, G.; Karamycheva, S.; Salzberg, S. L.; Quackenbush, J. *Nucleic Acids Res.* **2000**, *28*, 3657.
(220) Ayoubi, P.; Jin, X.; Leite, S.; Liu, X.; Martajaja, J.; Abduraham, A.; Wan, Q.; Yan, W.; Misawa, E.; Prade, R. A. *Nucleic Acids Res.* **2002**, *30*, 4761.
(221) Hotz-Wagenblatt, A.; Hankeln, T.; Ernst, P.; Glatting, K. H.; Schmidt, E. R.; Suhai, S. *Nucleic Acids Res.* **2003**, *31*, 3716.
(222) Huntley, D.; Hummerich, H.; Smedley, D.; Kittivoravitkul, S.; McCarthy, M.; Little, P.; Sergot, M. *Genome Res.* **2003**, *13*, 2195.
(223) Van Domselaar, G. H.; Stothard, P.; Shrivastava, S.; Cruz, J. A.; Guo, A.; Dong, X.; Lu, P.; Szafron, D.; Greiner, R.; Wishart, D. S. *Nucleic Acids Res.* **2005**, *33*, W455.

(224) Hoon, S.; Ratnapu, K. K.; Chia, J. M.; Kumarasamy, B.; Juguang, X.; Clamp, M.; Stabenau, A.; Potter, S.; Clarke, L.; Stupka, E. *Genome Res.* **2003**, *13*, 1904.

(225) Deevi, S. V.; Martin, A. C. *Bioinformatics* **2006**, *22*, 291.

(226) Palsson, B. *Nat. Biotechnol.* **2004**, *22*, 1218.

(227) Bork, P.; Serrano, L. *Cell* **2005**, *121*, 507.

(228) Reed, J. L.; Famili, I.; Thiele, I.; Palsson, B. O. *Nat. Rev. Genet.* **2006**, *7*, 130.

(229) Palsson, B. *Nat. Biotechnol.* **2000**, *18*, 1147.

(230) Feist, A. M.; Scholten, J. C.; Palsson, B. O.; Brockman, F. J.; Ideker, T. *Mol. Syst. Biol.* **2006**, *2*, 2006.

(231) Moszer, I.; Jones, L. M.; Moreira, S.; Fabry, C.; Danchin, A. *Nucleic Acids Res.* **2002**, *30*, 62.

(232) Chisholm, R. L.; Gaudet, P.; Just, E. M.; Pilcher, K. E.; Fey, P.; Merchant, S. N.; Kibbe, W. A. *Nucleic Acids Res.* **2006**, *34*, D423.

(233) Schwarz, E. M.; Antoshechkin, I.; Bastiani, C.; Bieri, T.; Blasiar, D.; Canaran, P.; Chan, J.; Chen, N.; Chen, W. J.; Davis, P.; Fiedler, T. J.; Girard, L.; Harris, T. W.; Kenny, E. E.; Kishore, R.; Lawson, D.; Lee, R.; Muller, H. M.; Nakamura, C.; Ozersky, P.; Petcherski, A.; Rogers, A.; Spooner, W.; Tuli, M. A.; Van, A. K.; Wang, D.; Durbin, R.; Spieth, J.; Stein, L. D.; Sternberg, P. W. *Nucleic Acids Res.* **2006**, *34*, D475.

(234) Grumbling, G.; Strelets, V. *Nucleic Acids Res.* **2006**, *34*, D484.

(235) Blake, J. A.; Eppig, J. T.; Bult, C. J.; Kadin, J. A.; Richardson, J. E. *Nucleic Acids Res.* **2006**, *34*, D562.

(236) de la Cruz, N.; Bromberg, S.; Pasko, D.; Shimoyama, M.; Twigger, S.; Chen, J.; Chen, C. F.; Fan, C.; Foote, C.; Gopinath, G. R.; Harris, G.; Hughes, A.; Ji, Y.; Jin, W.; Li, D.; Mathis, J.; Nenasheva, N.; Nie, J.; Nigam, R.; Petri, V.; Reilly, D.; Wang, W.; Wu, W.; Zuniga-Meyer, A.; Zhao, L.; Kwitek, A.; Tonellato, P.; Jacob, H. *Nucleic Acids Res.* **2005**, *33*, D485.

(237) Rhee, S. Y.; Beavis, W.; Berardini, T. Z.; Chen, G.; Dixon, D.; Doyle, A.; Garcia-Hernandez, M.; Huala, E.; Lander, G.; Montoya, M.; Miller, N.; Mueller, L. A.; Mundodi, S.; Reiser, L.; Tacklind, J.; Weems, D. C.; Wu, Y.; Xu, I.; Yoo, D.; Yoon, J.; Zhang, P. *Nucleic Acids Res.* **2003**, *31*, 224.

(238) Riley, L.; Schmidt, T.; Wagner, C.; Volz, A.; Artamonova, I.; Heumann, K.; Mewes, H. W.; Frishman, D. PEDANT genome database: ten years online. *Nucleic Acids Res. 35* (Database issue), D354.

(239) Collins, F. S.; Morgan, M.; Patrinos, A. *Science* **2003**, *300*, 286.

CR068303K