Comparative Genome Analyses

P.M. Selzer et al., *Applied Bioinformatics*, https://doi.org/10.1007/978-3-319-68301-0_7

7.1 The Era of Genome Sequencing

The extraordinary achievements of genome-based biology in recent years can be explained for the most part by the technological progress in DNA sequencing as well as developments in hardware and software that have made it possible to store and annotate huge amounts of data. The total number of all freely accessible nucleotides in GenBank [genbank], the DNA sequence database at the NCBI, is 218 billion bases within 196 million DNA sequences (Release 215.00, August 2016). The number of all protein sequences in the world's largest nonredundant protein database UniprotKB [uniprotkb] at the EBI totals 65 million (September 2016).

The first completely sequenced genomes, from the microbial organisms *Haemophilus influenzae* (Fleischmann et al. 1995) and *Mycoplasma genitalium* (Fraser et al. 1995), were published in 1995. Today, 165,178 microbial genomes are being sequenced or are already sequenced (163,302 from bacteria and 1876 from archaebacteria) [gold] (August 2016). Among these are the complete genomes of both virulent and nonvirulent strains of the same bacterium, which facilitates the identification of virulence factors. It is assumed that within the next few years, all important pathogenic microorganisms of humans, animals, and plants will have been sequenced. This flood of data will lead to new possibilities in the production of antimicrobial agents, vaccines, and diagnostic tests, all of which should aid the ongoing fight against infectious diseases (Selzer et al. 2000).

Meanwhile, the complete genomes of 283 eukaryotic organisms are known. These include *Saccharomyces cerevisiae* (baker's yeast), *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (fruit fly), *Arabidopsis thaliana* (mouse-ear cress), *Takifugu rubripes* (tiger puffer), *Homo sapiens* (man), and *Mus musculus* (mouse). Furthermore, as of September 2016, 13,000 eukaryotic genome sequencing projects are under way. These data will eventually contribute to the decoding of the secrets of biology and thereby help combat serious diseases of humans and animals.

7.2 Drug Research on the Target Protein

Systematic research into active substances as novel drugs dates back to the second half of the nineteenth century. A prime example is acetylsalicylic acid, which was synthesized in 1897 by two chemists, Felix Hoffmann and Arthur Eichengrün of the company Bayer. It is now world famous under the trade name aspirin. It is still a disputed question as to which of the two chemists was the actual inventor of the synthesis of acetylsalicylic acid. Regardless, this substance has lost neither its economic nor scientific importance. Since then, the identification of active compounds, including those with bioactivity against infectious diseases, has been dominated by direct testing (screening) in biological systems, mostly laboratory animals. Many antibiotics in use today were discovered in the first half of the twentieth century. However, since around the 1960s the number of new drugs has steadily declined. There are a number of reasons for this, including the constant decline in the success rate of nontargeted screening, the increased costs for research and development, and higher required safety standards. Furthermore, in the area of infectious diseases, the situation has been worsened by the emergence and

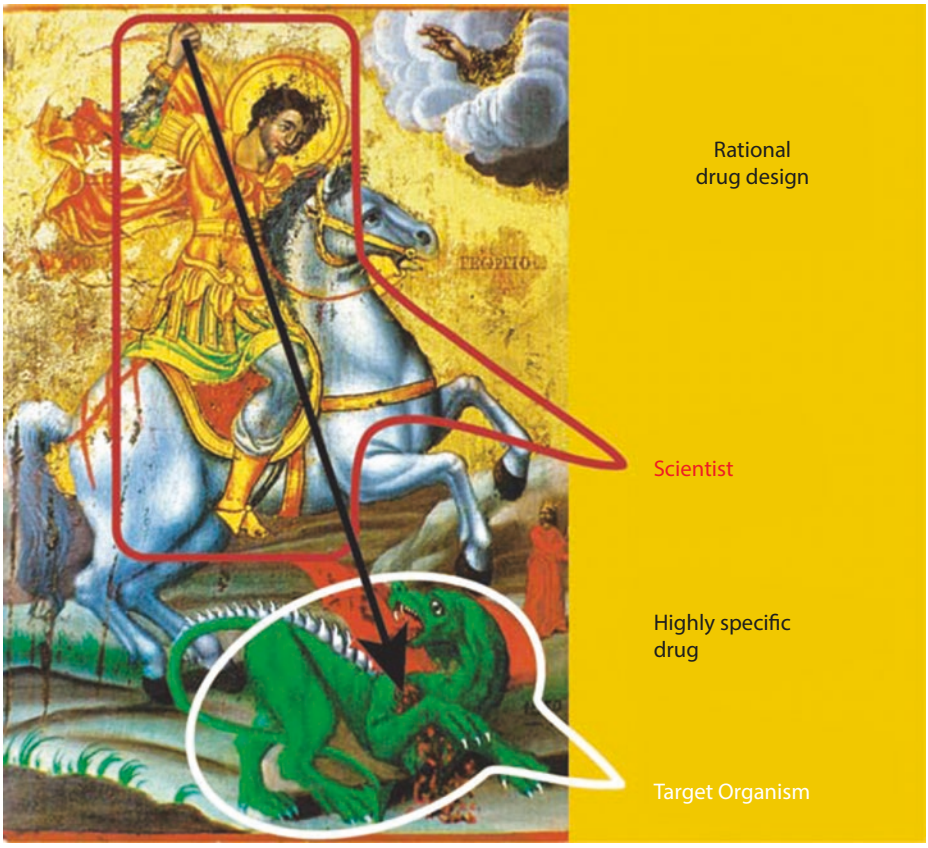


Fig. 7.1 The analogy of a Christian icon to a *target-based approach to drug development*. The icon shows Saint George as a dragonslayer. The dragon symbolizes the target organism that can only be killed by a precise blow to the heart (target protein). All other targets are irrelevant. Based on this realization, Saint George (the scientist) uses his horse (scientific tools) to guide his lance (a highly selective drug) to the target. The original icon is in Preveli Monastery, Crete (Greece)

increased spread of drug resistance. However, at about the same time, a new era of molecular research began in 1953 with the deciphering of the three-dimensional structure of the DNA double helix by James D. Watson and Francis H.C. Crick.

By sequencing whole genomes and the ensuing biological information, the approach to drug discovery has changed. Thus, in the target-based approach (Fig. 7.1), in which a target protein is used to search for new active compounds, the first step is to identify those proteins that are essential to the survival of the pathogenic organism. The second step is to find active chemical substances that influence the isolated target protein in the desired way. Only after such optimized chemical substances with the desired activity spectrum have been found using these in vitro methods will further testing be performed in a biological system (see also ▶ Chap. 5). For example, to develop a new antibiotic, an ideal prerequisite would be that the target protein is essential to the survival of the pathogenic bacteria under study and that the host

organism does not also possess the same or similar protein that may also be targeted, potentially resulting in toxicity. In this scenario, comparative whole genomic analysis would be well suited to identify pathogen-specific targets. Indeed, this approach was taken by Huynen et al. (1998) in their work on the genomes of three bacteria, *Escherichia coli*, *Haemophilus influenza*, and *Helicobacter pylori*. Orthologous proteins were identified either in all three or in two of the three organisms, in addition to species-specific proteins. For *H. pylori*, the major causative agent of gastric and duodenal ulcers, the authors predicted that 123 proteins were involved in interacting between the pathogen and host, i.e., the proteins represented potential targets for the development of an antibiotic. In pharmacological research, conserved targets usually lead to the development of broad-spectrum antibiotics, whereas with species-specific targets, narrow-spectrum antibiotics are generated.

Because of the increasing number of completely sequenced bacterial genomes, it is clearer which genes are generally conserved among bacteria and which are specific for certain bacterial species. However, it is not always easy to settle on the threshold of sequence similarity that blocks the pursuit of a target-based drug discovery approach due to potential toxicity arising from an unwanted interaction with the human protein counterpart. For example, bacterial dihydrofolate reductase has a sequence identity of 28% at the amino acid level to the corresponding human protein, yet the antibacterial drug, trimethoprim, is a very selective inhibitor of only the bacterial ortholog.

7.3 Comparative Genome Analyses Provide Information About the Biology of Organisms

Comparative genome analyses are frequently referred to as *comparative genomics*, whereby two or more genomes are compared to one another (Beckstette et al. 2004). The goal is to find similarities and differences between these genomes that yield information about the biology of the respective organisms. Another important aim of comparative genomics is the description of genome structure and the identification of coding and noncoding regions (Wei et al. 2002).

7.3.1 Genome Structure

Analysis of the structure of one or more genomes includes statistical measurements such as size and nucleotide composition, frequency of codon usage, and identification of conserved regions between two or more genomes. The percentage and frequency of guanine and cytosine (GC) content or adenine and thymidine (AT) content differ between groups of organisms and seem to have changed considerably in the course of evolution from microorganisms to multicellular organisms. Likewise, the codon usage for encoding identical amino acids is not the same in every organism (▶ Chaps. 1 and 3).

Many comparative studies of the genomes of humans and mice have shown that their organization, to a large extent, is similar. This indicates that, since the last com-

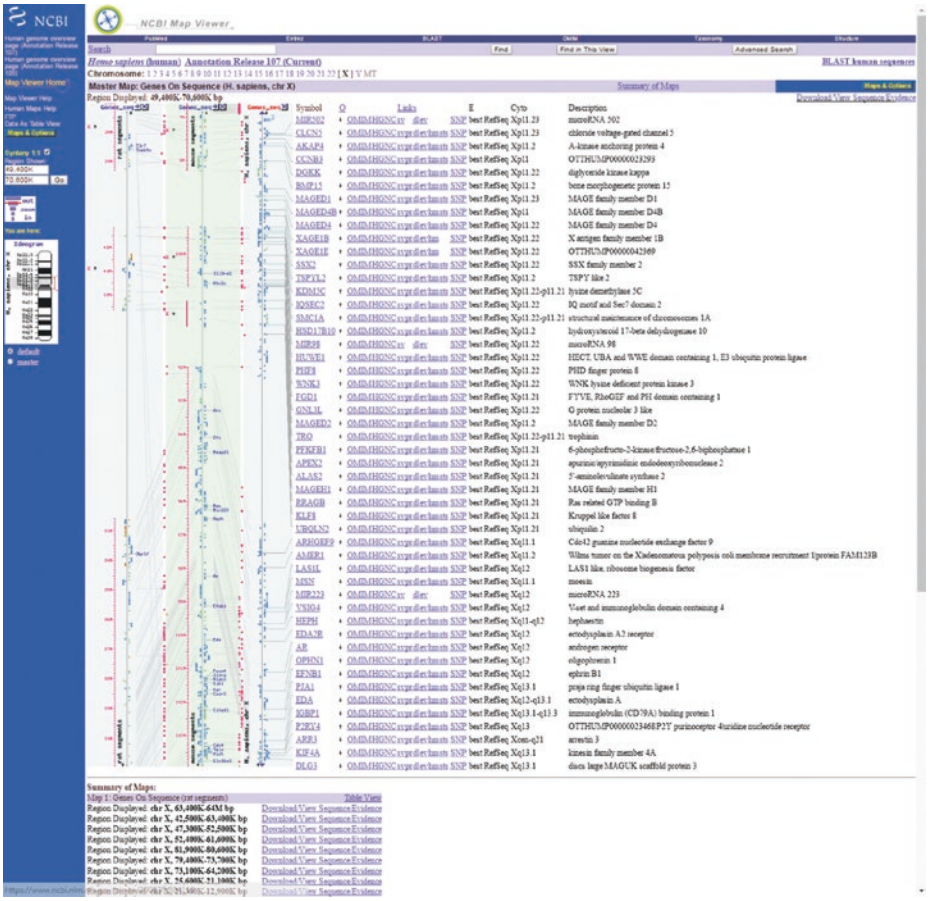


Fig. 7.2 Homology map of X chromosome of rat, mouse, and humans as taken from NCBI. Shown is a part of the detailed map for the X chromosome. The syntenic genes in this chromosome area are indicated by gray lines (Printed courtesy of the NCBI)

mon ancestor, the structural organization has been conserved. To describe such similarities between evolutionarily related chromosomal segments among species, various terms have been defined or broadened in their definition. If two or more genes lie on the same chromosome, then one speaks of *syntenic genes*, or *synteny*. This definition only applies, however, within a species. Between species, the definition was expanded such that when syntenic genes of orthologous proteins on a single chromosome are conserved across species, the term *conserved synteny* applies. The order of the genes on the chromosomes is not considered (Fig. 7.2). If, in addition, the order of the genes on the chromosomes is also conserved, then the regions are called *conserved segments* or *conserved linkages*.

With the growing number of completed eukaryotic genome sequences it has become apparent that conserved segments are present in all mammals. Although syntenic regions are observed between species such as humans and the puffer fish,

which separated approximately 450 million years ago, no larger blocks of conserved genome organization have been described thus far for such distantly related organisms (Frazer et al. 2003).

7.3.2 Coding Regions

The comparative analysis of coding regions between different genomes includes not just the identification of protein-encoding regions but also the direct comparison of the types and numbers of orthologous and paralogous proteins. The identification of genes in prokaryotes is comparatively simple because there are relatively few noncoding regions. Normally 85% of a bacterial genome encodes proteins or RNAs, with the smaller portion encoding regulatory units or noncoding regions. In contrast, the prediction of genes in eukaryotes is far more difficult because noncoding regions have increased in number over the course of evolution. Eukaryotic genomes possess a large number of intergenomic regions as well as a multitude of noncoding repeats. Furthermore, eukaryotic genes contain introns and exons, and different proteins frequently arise as a consequence of alternative splicing (► Chaps. 1 and 4). For instance, the genome of the prokaryote *Escherichia coli* has approximately 4300 genes at a genome size of 4600 kilobases (kb) with, on average, one gene for every kilobase in length. In contrast, the genome of the eukaryotic unicellular yeast *Saccharomyces cerevisiae* has approximately 6300 genes at a genome size of 12,000 kb, and the genome of the multicellular worm *Caenorhabditis elegans* contains approximately 19,000 genes at a genome size of 97,000 kb. Phylogenetically speaking, the human genome is very young and shows an enormous difference between the number of genes and its genome size: 19,000 to 20,000 genes at a total size of approx. 3.3 gigabases (Ezkurdia et al. 2014). There is no obvious connection between the size of the genome and the complexity of the organism, as demonstrated by the similar number of genes in the genome of *C. elegans* and humans. The relatively low number of protein-coding genes in the human genome can be understood considering posttranslational modifications like alternative splicing that allow for one gene coding for several proteins (► Chap. 4).

7.3.3 Noncoding Regions

The comparative analysis of noncoding regions, which in humans and other mammals can account for more than 97% of the genome, is still one of the greatest challenges of bioinformatics. Still, this area of genome analysis has received much attention in the last few years in the hope of identifying genomic regulatory units. For instance, it has already been shown bioinformatically that conserved noncoding regions have an accumulation of transcription factor binding sites. Furthermore, the probability of identifying such regulatory areas in noncoding regions increases when more than two genomes of closely related organisms are compared. It has already been shown that half of the noncoding regions identified in a comparison of the human and mouse genomes are also conserved in the genome of the dog.

7.4 Comparative Metabolic Analyses

For gene prediction, special emphasis has been placed on those genes that encode proteins involved in metabolism. Using gene prediction, it is possible to identify whether an organism possesses metabolic pathways, such as those in glycolysis or the citrate cycle, or whether alternative pathways are employed to generate energy. A comparison of two or more genomes at the level of their metabolic pathways can also be used to identify metabolic targets. This is particularly effective with prokaryotes because many genomes have already been sequenced. A number of software technologies are used to compare metabolomes: Encyclopedia of *Escherichia coli* Genes and Metabolism (EcoCyc) [ecocyc], Kyoto Encyclopedia of Genes and Genomes (KEGG) [kegg] (◘ Fig. 7.3), and the Reactome database [reactome] are among the best known.

The methods include manual and semiautomatic analyses. So far, however, there is no fully automatic analysis software that can calculate all the metabolic pathways. Furthermore, such databases are not always complete. Whereas initially the databases dealt mostly with metabolic pathways, over time regulatory mechanisms such as membrane transport, gene regulation, and signal transduction have also been incorporated (◘ Fig. 7.4).

In sequenced genomes, genes or proteins can be divided into orthologous groups. Accordingly, proteins that are either present or absent can be systematically identified and the resultant functional metabolic pathways constructed. If some required proteins are missing, either the corresponding metabolism is nonfunctional or other (including thus far unknown) proteins are involved. During the analysis of the genome of *Helicobacter pylori* it was noticed that neither glycolysis nor pentose phosphate metabolism was operational due to the absence of the requisite enzymes. Because both metabolic pathways generate protons and, therefore, lower pH, their operation would lead to an additional burden on an organism that already lives in the acidic environment of the stomach. In contrast, the genes coding for proteins that metabolize organic acids, such as those involved in anabolic gluconeogenesis, are present. Thus, *H. pylori*'s energy production seems to be fueled by amino acid degradation, and the substrates necessary are probably directly derived from the gastrointestinal tract.

To find specific metabolic pathways in KEGG, the genome must be compared with a reference genome. If the gene exists, it is highlighted in color. A sequence of colored rectangles therefore reflects the specific metabolic pathway in the studied organism (◘ Fig. 7.5). To be successful with this strategy, however, all alternatives must be known. It is often the case that a metabolic pathway does not show all the genes or proteins and is, therefore, considered incomplete. The reasons for this include that not all genes were predicted or some were predicted incorrectly or that current knowledge regarding the specific metabolic pathway is limited. It is also possible that one protein performs several functions and, thus, has a larger metabolic spectrum than originally suspected. Finally, alternative metabolic pathways that lead to the same biological result cannot be excluded.

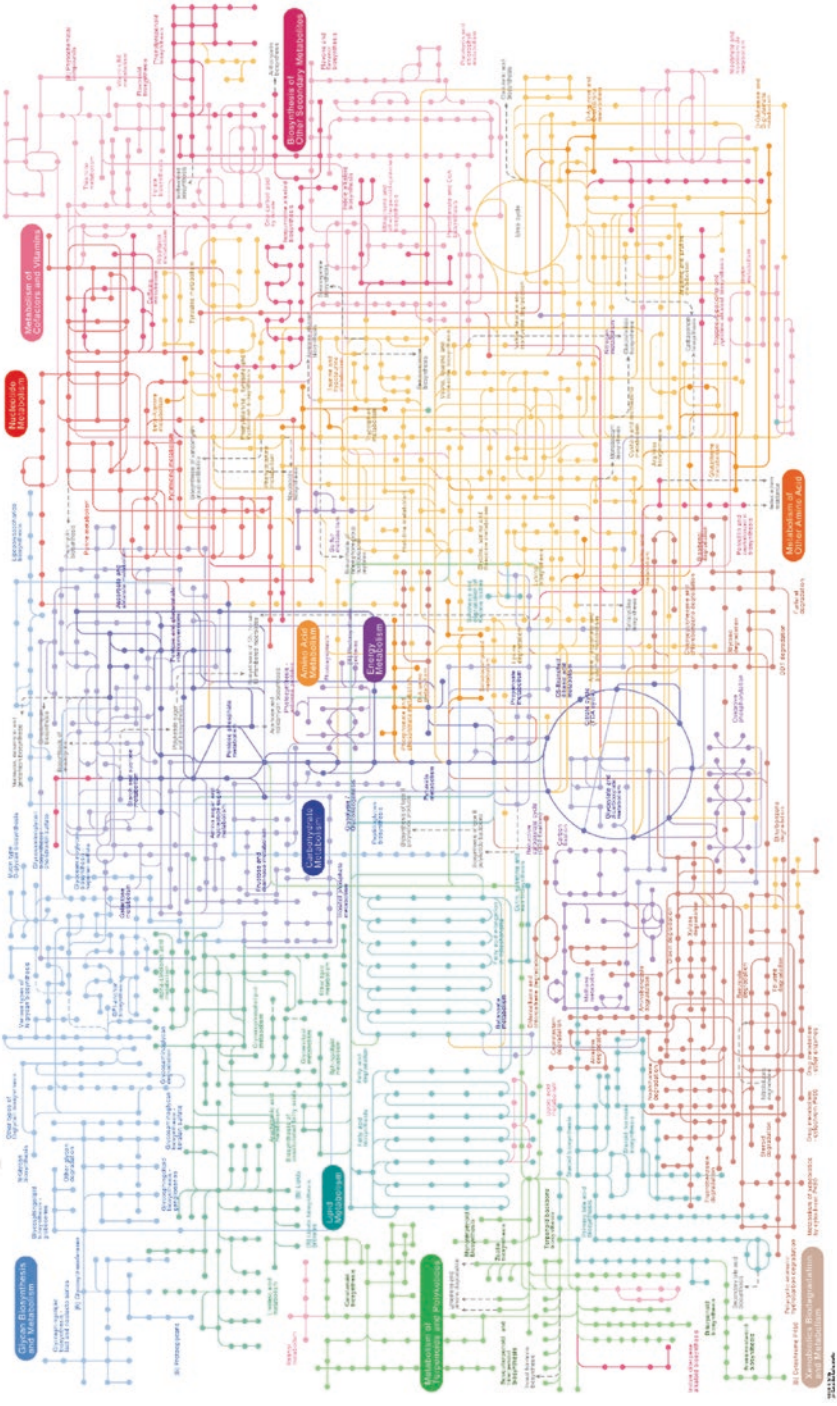


Fig. 7.3 Map of metabolic pathways in KEGG database (Printed courtesy of KEGG)

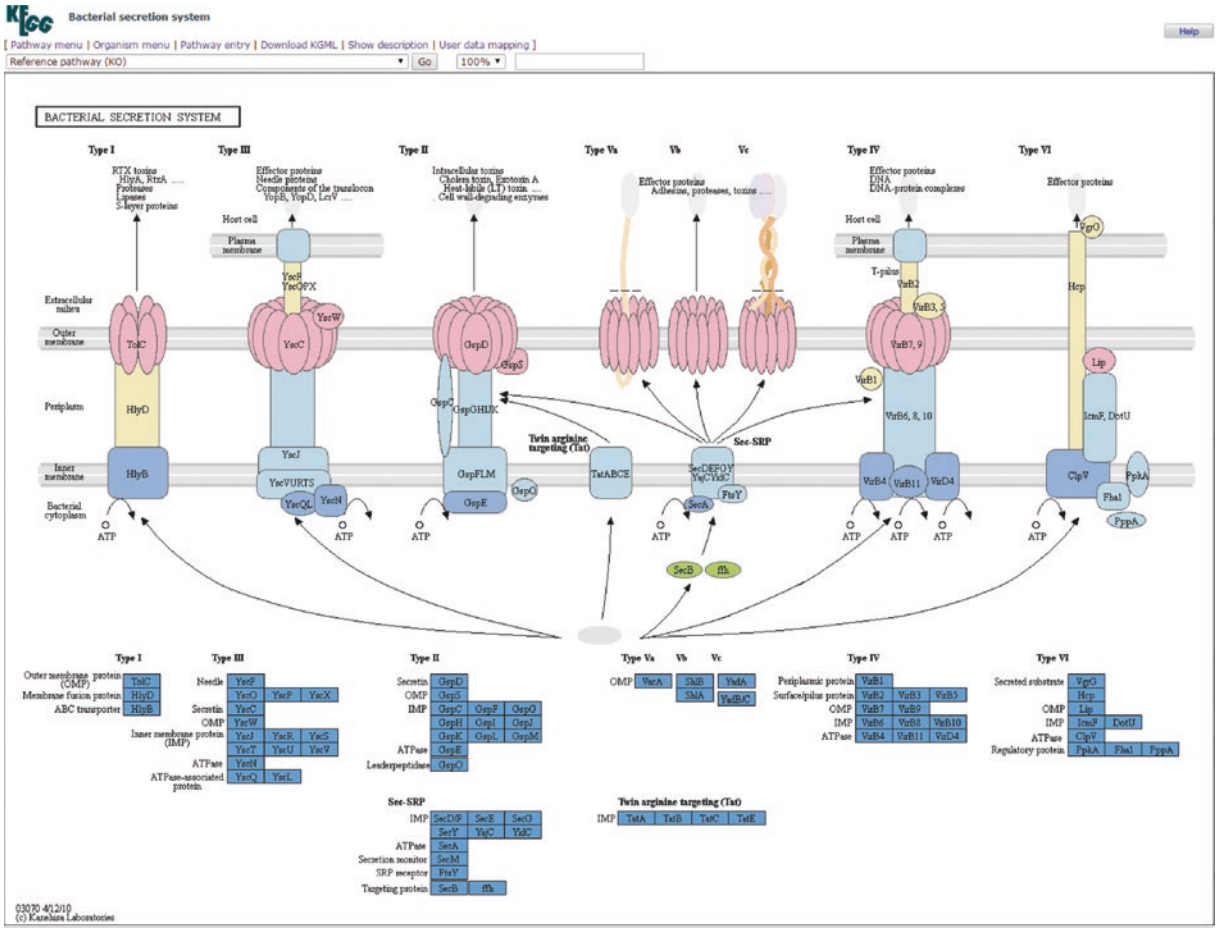
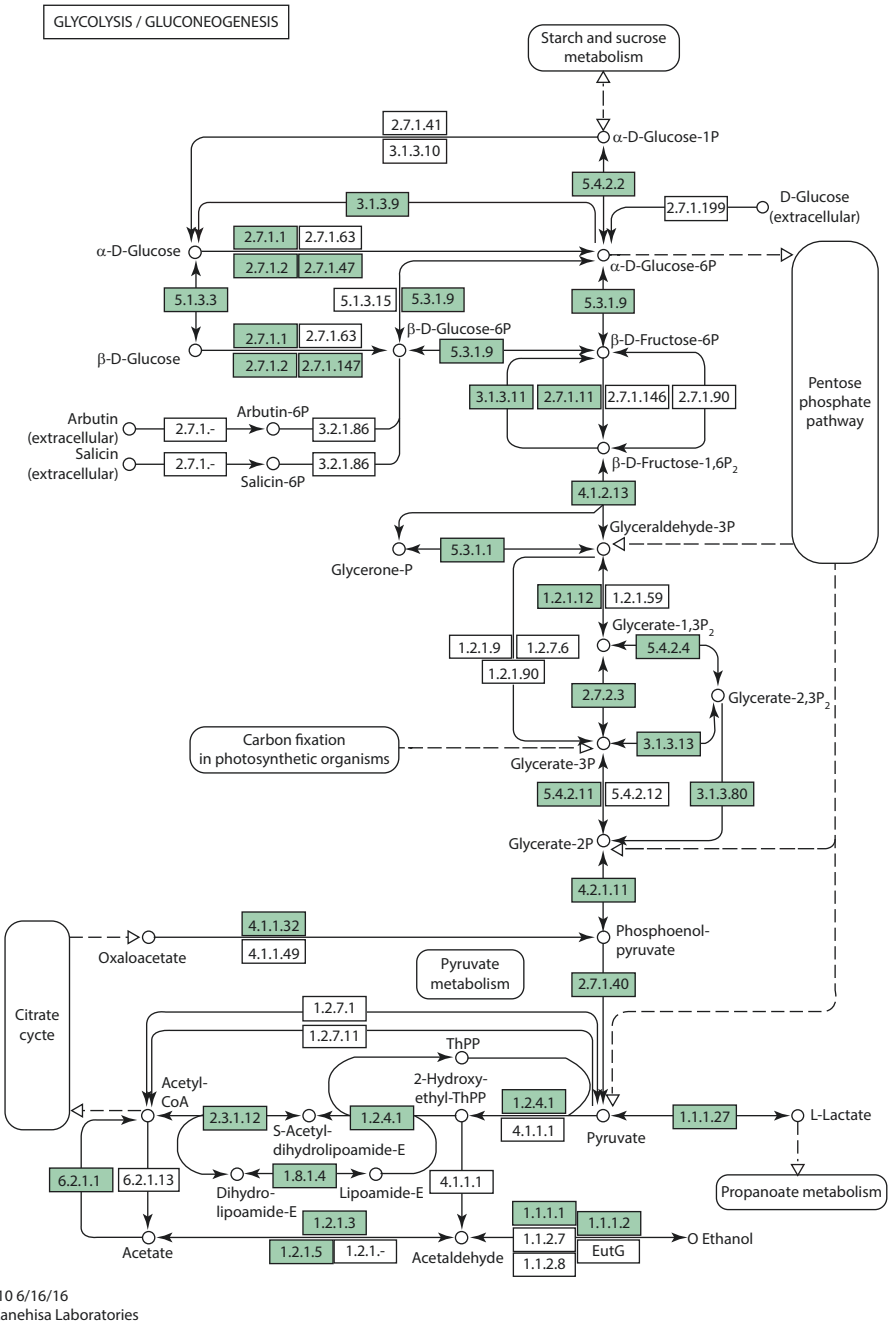


Fig. 7.4 Schematic representation of bacterial secretion pathways (Printed courtesy of KEGG)



00010 6/16/16
(c) Kanehisa Laboratories

Fig. 7.5 Metabolic map of glycolytic/gluconeogenic metabolism. The enzymes of this metabolic pathway known so far in humans are highlighted in color (Printed courtesy of KEGG)

7.4.1 Kyoto Encyclopedia of Genes and Genomes

KEGG is a product of the Japanese GenomeNet and is widely used for the analysis of metabolic pathways. Two of the three main databases, PATHWAY and LIGAND, deal with metabolic processes in cells and organisms. The third database, GENE, contains gene and protein information from sequencing projects and is comparable to other primary databases (Kanehisa et al. 2016). These databases are completed by BRITE, an ontology database for the description of biological relationships within pathways. Furthermore, KEGG offers information on experimental data from gene expression and yeast two-hybrid experiments (EXPRESSION). Another database, SSDB, contains information on groups of orthologous proteins.

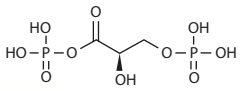
The most interesting databases are undoubtedly PATHWAY and LIGAND. PATHWAY contains graphical representations of metabolic pathways from a number of organisms, mostly prokaryotes, but also eukaryotes. The representations of the metabolic pathways are similar to those in the Biochemical Pathways Chart from Boehringer Mannheim [biochem-pathway]. The individual maps can be selected from a list or chart sorted according to the main metabolic pathways (Fig. 7.3). The known enzymes in reference pathways can be highlighted in color. This facilitates comparison of metabolic pathways between organisms. Figure 7.5 shows as an example of glycolysis/gluconeogenesis metabolism in humans. The enzymes drawn in green (small boxes) have already been described or are present in the human genome. The individual metabolic charts on the KEGG server are connected to the LIGAND database, a chemical database that contains the corresponding substances, enzymes, and reactions in the respective metabolic pathway. The small rectangular boxes with an enzyme number (enzyme classification, NC-IUBMB 1992) [enzyme] are for cross-referencing. The EC number consists of four blocks of numbers, each separated by a period. The first number describes one of the six functional groups (oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases), the two blocks following refer to further subclasses within the main class. The last block is a consecutive number of each of the enzymes in the particular subclass. Further cross-references are indicated by the circular symbols next to the substance names (e.g., β-D-glucose) as well as the rounded borders of other metabolic pathways. The latter do not lead to the LIGAND database, however, but to a detailed description of the respective metabolic paths. In the case of glycolysis/gluconeogenesis metabolism, for example, this leads to the citrate cycle or pentose phosphate metabolism.

By clicking on the circle at *Glycerate-1,3P2* a new window opens with an entry from LIGAND (Fig. 7.6). In addition to a unique substance number, the substance name and the empirical and constitutional formulas of the substance are given. What follows are cross-references to entries of reactions in which 1,3-bisphospho-D-glycerate is involved, to the metabolic pathways in which it operates, and to enzymes that are associated with the conversion of 1,3-bisphospho-D-glycerate. The CAS number in the field *DBLINKS* is a unique number given to every chemical substance by the Chemical Abstract Service [cas] upon first publication. Moreover, this field lists hyperlinks to other databases. The section *Structure* contains a graphical representation of the chemical structure and a number of buttons, which allow one to download the structure in various formats.

In addition to database queries via a graphical representation of metabolic pathways, LIGAND facilitates text searches for reactants or enzymes and searches for the substructures of more complex chemical structures.

KEGG COMPOUND: C00236

Help

Entry	C00236 Compound
Name	3-Phospho-D-glyceroyl phosphate; 1,3-Bisphospho-D-glycerate; (R)-2-Hydroxy-3-(phosphonooxy)-1-monoanhydride with phosphoric propanoic acid; D-Glycerate 1,3-diphosphate
Formula	C3H8O10P2
Exact mass	265.9593
Mol weight	266.0371
Structure	 C00236 Mol file KCF file DB search Jmol KegDraw
Reaction	R01061 R01063 R01512 R01515 R01517 R01660 R01662 R02188
Pathway	map00010 Glycolysis / Gluconeogenesis map00710 Carbon fixation in photosynthetic organisms map01060 Biosynthesis of plant secondary metabolites map01061 Biosynthesis of phenylpropanoids map01062 Biosynthesis of terpenoids and steroids map01063 Biosynthesis of alkaloids derived from shikimate pathway map01100 Metabolic pathways map01110 Biosynthesis of secondary metabolites map01120 Microbial metabolism in diverse environments map01130 Biosynthesis of antibiotics map01200 Carbon metabolism map01230 Biosynthesis of amino acids
Module	M00001 Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate M00002 Glycolysis, core module involving three-carbon compounds M00003 Gluconeogenesis, oxaloacetate => fructose-6P M00165 Reductive pentose phosphate cycle (Calvin cycle) M00166 Reductive pentose phosphate cycle, ribulose-5P => glyceraldehyde-3P M00308 Semi-phosphorylative Entner-Doudoroff pathway, gluconate => glycerate-3P M00552 D-galactonate degradation, De Ley-Doudoroff pathway, D-galactonate => glycerate-3P
Enzyme	1.2.1.12 1.2.1.13 1.2.1.59 2.7.1.106 2.7.2.3 2.7.2.10 2.7.4.17 3.6.1.7 5.4.2.4
Other DBs	CAS: 38168-82-0 PubChem: 3535 ChEBI: 16001 KNApSAcK: C00019552 PDB-CCD: X15[PDBj] 3DMET: B01197 NIKKAJI: J40.060B
LinkDB	All DBs
KCF data	Show

» Japanese version

DBGET integrated database retrieval system

Fig. 7.6 Database record in LIGAND database for β-D-glucose (Printed courtesy of KEGG)

7.5 Groups of Orthologous Proteins

Upon completion of a genome sequencing project, attention is turned to the analysis and classification of the predicted genes and the possible function of their gene products. The simplest approach is to compare unknown gene sequences with known genes and assign a function based on similarity. Some of the tools were already described earlier in this book. Because the comparison of whole genomes or proteomes with conventional methods is very laborious, however, commercial software packages have been developed that allow a comparison of large sequence data sets and the identification of common sequences, MUMmer, for instance (Delcher et al. 1999) [mummer].

In those cases where larger phylogenetic distances exist between organisms, direct sequence comparison is difficult owing to low sequence similarities. Another approach to phylogenetic classification of proteins, therefore, is by comparing orthologous and paralogous genes. Orthologous genes develop through the formation of species out of a common ancestor; paralogous genes develop through gene duplication. It is common sense that the function of orthologous genes is more conserved than that of paralogous genes because the evolutionary pressure is reduced on paralogous genes after gene duplication. This concept is called ortholog conjecture. Although this concept has been critically discussed recently (Studer and Robinson-Rechavi 2009; Nehrt et al. 2011), it is still considered valid and forms the backbone of most functional annotation methods (Huerta-Cepas et al. 2016). Therefore, the exact determination of orthology between proteins is of paramount importance. Unfortunately, the prediction of such relations is very difficult, analytically as well as informatically. The reasons for this are many and include nested duplications, genomic rearrangements, and horizontal gene transfers, which disguise the real relations.

Therefore, several complex systems for the classification of orthologous proteins have been developed. A well-known system was the COP database (Clusters of Orthologous Groups) at the NCBI [cog] (Wheeler et al. 2007). In addition to a text-based search, it was possible to compare sequences against the database and to predict the function of gene products. The quality was very high, owing to its manual curation. However, the database was a static system, meaning the number and kind of species used to build the clusters were predetermined and could not be influenced by the user. The COG database was discontinued in 2013.

A current database of orthologous protein groups is the eggNOG database [eggnog]. The database contains clusters of orthologous groups on several taxonomic levels together with functional annotations. In addition, the database entries are spiked with gene ontoloy (GO) entries, KEGG metabolic pathways, and information on SMART/Pfam domains. Currently the database contains 2031 eukaryotic and prokaryotic organisms (version v4.5, 2015). Moreover, 1655 prokaryotes have already been precompared to the database. For the cluster-building process, data from several primary databases are used, and – after a quality assurance step – all sequences are pairwise compared to each other using Smith–Waterman alignments. Interesting matches are stored and grouped into clusters with respect to taxonomic circumstances. The idea behind this last step is that the resolution of orthologous groups is critically dependent upon the taxonomic level under consideration. For instance, it can be reasonable to cluster a set of mammalian sequences together with some sequences of distantly related vertebrates. If we study the same set of mammalian sequences, however, on the taxonomic level of

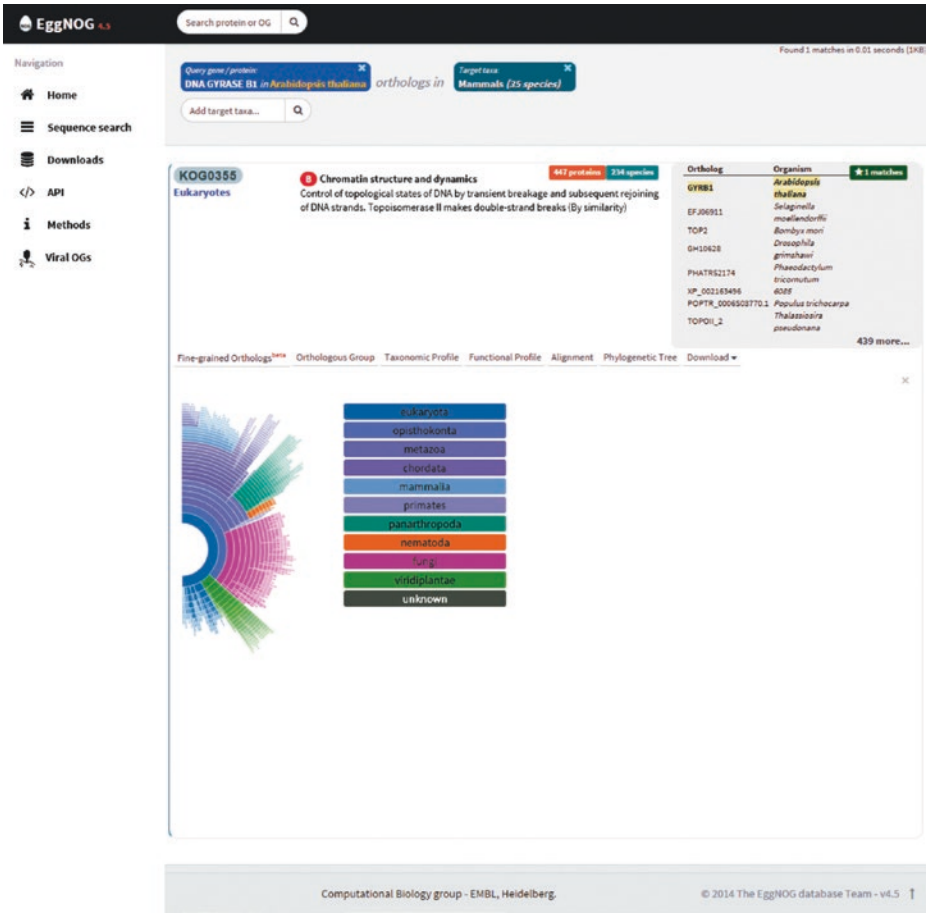


Fig. 7.7 Result of a database query of eggNOG database. The taxonomic profile is shown for the resulting orthologous group (Printed courtesy of EMBL)

primates, it can be reasonable to build two clusters. The eggNOG clustering is based upon the clustering used in the discontinued, manually curated COG database [cog], which lists the three kingdoms Eukaryotes, Bacteria, and Archaea: the COGs section contains all three kingdoms with a focus on Prokaryotes, KOGs contains the Eukaryotes, and arKOGs contains the Archaea. The clustering is thus done independently for each of the predefined taxonomic levels. Subsequently, inconsistencies that arise from incomplete proteomes or from assumptions of the heuristic algorithms are eliminated in an additional quality assurance step. In a last step, an automated, heuristic method is used to select the best matching annotation from different annotation databases. Such annotations are human readable, and it is thus very difficult to use them in a statistical analysis. Therefore, the COG database introduced a single letter classification, which is also used in the eggNOG database. Every orthologous group is assigned to a classification using a support vector machine.

The eggNOG database offers two query systems: a guided text search and a sequence query. For the guided text search, a search term – a protein or gene name – is entered. If

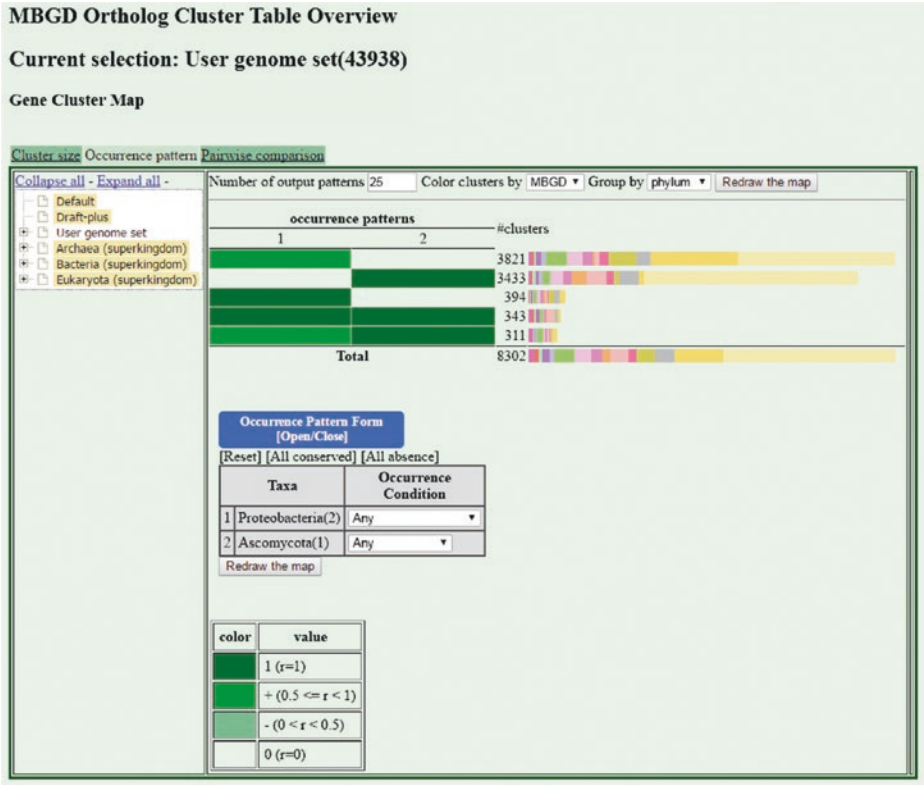


Fig. 7.8 Result of a *cluster analysis* of the MBGD [mbgd]. The organisms *E. coli* (Ecs), *H. pylori* (Hpj), and *S. cerevisiae* (Sce) were selected for calculating the underlying *cluster table* (Printed courtesy of MBGD)

there are entries for different organisms, the system prompts for the organism. Next, a list of target organisms can be selected. This can either be distinct organisms or all members of a clade. Based on the target organism list, eggNOG selects the taxonomic level. The results include hyperlinks to different visual representations, for example, there might be a display of a phylogenetic tree of the sequences (Fig. 7.7) that uses color markups for the source and target organisms, as well as for speciation and gene duplication. Moreover, it is possible to display a sequence alignment or a taxonomic or functional profile. The analysis comprises all members of the orthologous group; species that are not part of the query are hidden, however. In addition, it is possible to display a detailed view of the pairwise orthologs.

In the sequence query, it is not possible to select a target organism list. First one of the three possible kingdoms must be selected. The result list shows the resulting orthologous groups in all taxonomic levels. Each of the entries has the same visualization hyperlinks as described for the guided text search.

A similar system, the Microbial Genome Database (MBGD), facilitates the dynamic calculation of clusters according to user-defined parameters [mbgd] (Uchiyama et al. 2015) (Fig. 7.8). This approach takes into account that the classification of proteins into orthologous clusters can depend on the choice of organisms

and that a static set of clusters may unintentionally influence the results. The MBGD therefore provides a classification scheme rather than the static result of a classification. The cluster calculation depends on the user's parameter entries, either via orthology or homology criteria, and is based on precomputed similarity tables of all the proteins in the database. Besides text-based queries, MBGD offers a tool to evaluate and annotate one's own sequences.

7.6 Exercises

Exercise 7.1
How many genome sequencing projects are ongoing and how many have been completed?

Exercise 7.2
Go to the KEGG home page (► http://www.kegg.jp/) and display the metabolic map of glycolysis/gluconeogenesis metabolism.

Exercise 7.3
What enzymes catalyze the conversion of L-lactate to pyruvate? Does this conversion take place in humans? Does *Saccharomyces cerevisiae* make use of this metabolic step?

Exercise 7.4
How do the enzyme hyperlinks differ between the reference pathway and a species-specific map (e.g., *Homo sapiens*)?

Exercise 7.5
Display the chart of glycolysis/gluconeogenesis metabolism and compare the species-specific map of humans with that of *Helicobacter pylori 26,695*. What are the significant differences between the two maps? How can these differences be explained?

Exercise 7.6
Go to the NCBI BLAST home page and perform a BLAST search with the sequence Q9ZK41 against the Microbial Genome Database for the genomes of the following organisms or groups of organisms:
- *Staphylococcus aureus* RF122 (taxid: 273,036)
- *Streptococcus pneumoniae* D39 (taxid: 373,153)
- *Proteobacteria epsilon subdivision* (taxid: 29,547)
- How many reasonable hits are obtained for each organism?

Exercise 7.7
Go to the eggNOG database (► http://eggnogdb.embl.de/) and search for the orthologous group (OG) of the cyclind-dependent kinase CDK1 from *Homo sapiens*. Use the clade *Apicomplexa* for the target organisms. What taxonomic level is automatically chosen by eggNOG?

Exercise 7.8

Repeat ► Exercise 7.7 but select the *Marsupials* as the target organisms. What taxonomic level is chosen by eggNOG? How many sequences contain a PFAM kinase domain, and what is the frequency of occurrence for this domain? Are the respective marsupial sequences orthologs or paralogs? Are there paralogs within the selected taxonomic level?

Exercise 7.9

Go to the MBGD (► http://mbgd.genome.ad.jp/) and calculate a cluster table for the following organisms: *Staphylococcus aureus* RF122, *Escherichia coli* 536, and *Saccharomyces cerevisiae* S288C.

Exercise 7.10

From ► Exercise 7.9, how many clusters contain genes of all the selected organisms? Display these. To which functional category does the first cluster belong?

Exercise 7.11

Go back to the start page of the MBGD (► http://mbgd.genome.ad.jp/). In the overview of organisms, only those previously selected will be marked in red. Do a keyword search for the keyword *fructokinase*. How many entries are found?

References

Beckstette M, Mailänder JT, Marhöfer RJ, Sczyrba A, Ohlebusch E, Giegerich R, Selzer PM (2004) Genlight: interactive high-throughput sequence analysis an comparative genomics. J Integr Bioinform Yearbook 2004:79–94

Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL (1999) Alignment of whole genomes. Nucleic Acids Res 27:2369–2376

Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, Tress ML (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet 23:5866–5878

Fleischmann RD, Adams MD, White O et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fraser CM, Gocayne JD, White O et al (1995) The minimal gene complement *Mycoplasma genitalium*. Science 270:397–403

Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC (2003) Cross-species sequence comparisons: a review of methods and available resources. Genome Res 13:1–12

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H et al (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293

Huynen M, Dandekar T, Bork P (1998) Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. FEBS Lett 426:1–5

Kanehisa M, Yoto S, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–D462

NC-IUBMB (1992) Nomenclature Committee of the International Union of Biochemistry and molecular Biology, Enzyme Nomenclature 1992. Academic, Orlando

Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol 7:e1002073

Selzer PM, Brutsche S, Wiesner P, Schmid P, Müllner H (2000) Target-based drug discovery for the development of novel antiinfectives. Int J Med Microbiol 290:191–201

Studer RA, Robinson-Rechavi M (2009) How confident can we be that orthologs are similar, but paralogs differ? Trends Genet 25:210–216
Uchiyama I, Mihara M, Nishide H, Chiba H (2015) MBGD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. Nucleic Acids Res 43:D270–D276
Wei L, Liu Y, Dubchak I, Shon J, Park J (2002) Comparative genomics approaches to study organism similarities and differences. J Biomed Inform 35:142–150
Wheeler DL, Barrett T, Benson DA, Bryant SH et al (2007) Database resources of the National Center for Biotechnology. Nucleic Acids Res 35:D5–D12

Further Reading

biochem-pathway. http://web.expasy.org/pathways/
cas. http://www.cas.org/
cog. http://www.ncbi.nlm.nih.gov/COG/
ecocyc. http://ecocyc.org/
eggnog. http://eggnog.embl.de/
enzym. http://www.chem.qmw.ac.uk/iubmb/enzyme/
genbank. http://www.ncbi.nlm.nih.gov/Genbank/
gold. https://gold.jgi.doe.gov/
kegg. http://www.kegg.jp/
mbgd. http://mbgd.genome.ad.jp/
mummer. http://mummer.sourceforge.net/
reactome. http://www.reactome.org/
uniprotkb. http://www.uniprot.org/uniprot/