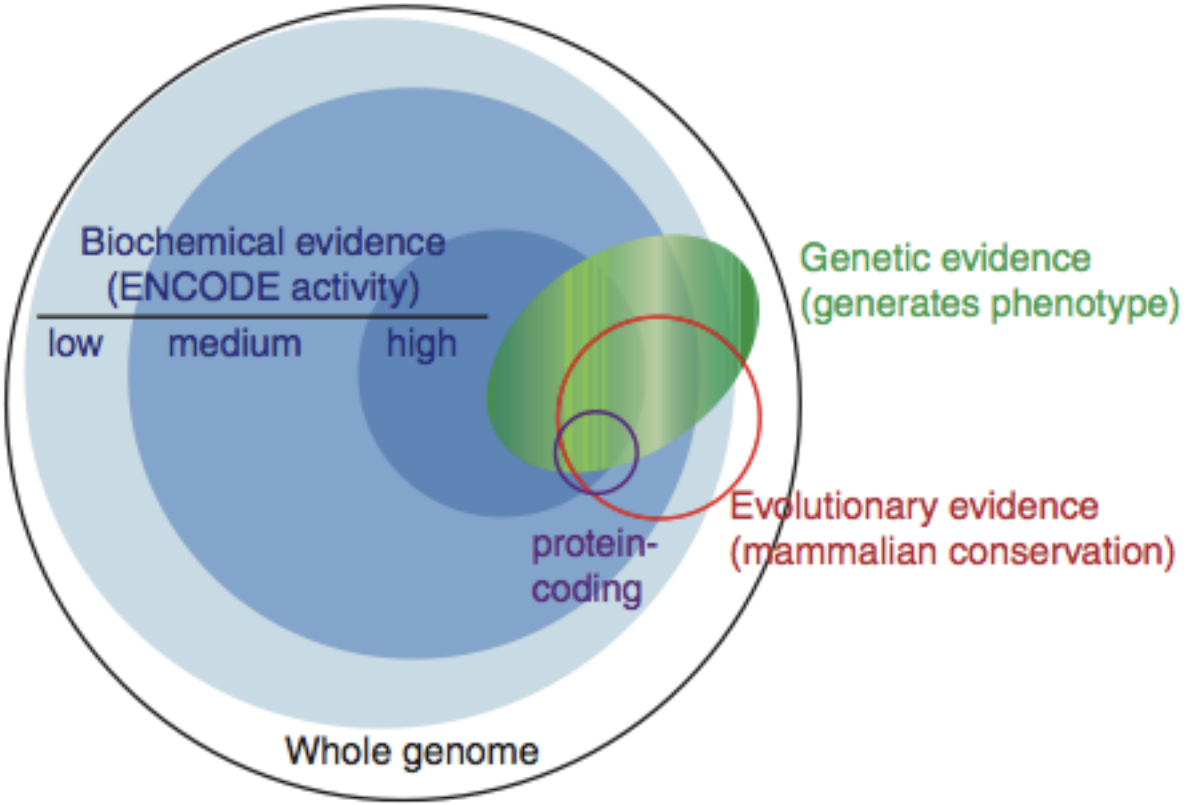
Distinguishing different approaches to function from definitions of function



Three circles corresponding to the magnitude of functional findings in ENCODE (see Chapter 8)

B&FG 3e
Fig. 14.17
Page 667

Functional genomics and DNA: integrating information

A goal of functional genomics is to provide integrated views of DNA, RNA, protein, and pathways. Many resources (such as those at Ensembl, EBI, and NCBI) offer this integrated view.

An example is the Frequency weighted links (FLink) tool at NCBI. Input a list of genes (or proteins or small molecules) and obtain a ranked list of biosystems.

Functional genomics and RNA

Surveys of RNA transcript levels across different regions (for multicellular organisms) and times of development provide fundamental information about an organism's program of gene expression.

As an example, the *Saccharomyces* Genome Database (SGD) offers many resources to describe gene expression in yeast. For each gene, an expression summary plots the log2 ratio of gene expression (*x* axis) versus the number of experiments (*y* axis; see **Fig. 14.3**, lower right). That plot is clickable, so experiments in which SEC1 RNA is dramatically up- or down-regulated can be quickly identified.

Outline : Functional genomics

Introduction
- Relation between genotype and phenotype

Eight model organisms
- *E. coli*; yeast; *Arabidopsis*; *C. elegans*; *Drosophila*; zebrafish; mouse; human

Functional genomics using reverse and forward genetics
- Reverse genetics: mouse knockouts; yeast; gene trapping; insertional mutatgenesis; gene silencing
- Forward genetics: chemical mutagenesis

Functional genomics and the central dogma
- Approaches to function; Functional genomics and DNA; …and RNA; …and protein

Proteomic approaches to functional genomics
- CASP; protein-protein interactions; protein networks

Perspective

Proteomic approaches to functional genomics

Basic features of proteins include their sequence, structure, homology relationships, post-translational modifications, localization, and function. In addition to the study of individual proteins, high throughput analyses of thousands of proteins are possible. We describe three approaches:

- identifying pairwise interactions between protein using the yeast two-hybrid system;
- identifying protein complexes involving two or more proteins using affinity chromatography with mass spectrometry; and
- analyzing protein pathways.

While protein studies have been studied in depth in a variety of model organisms, studies in *S. cerevisieae* are particularly advanced.

Proteomic approaches to functional genomics

We usually think of forward and reverse approaches in terms of genetics, but these terms can apply to proteomics.

Forward proteomics:

- Select experimental system (e.g. normal versus diseased tissue).
- Proteins are extracted and may be labeled with fluorescent dyes or other tags
- Proteins are separated and analyzed by techniques such as mass spectrometry.
- Spectra are analyzed and differentially regulated proteins are identified.
- These regulated proteins may reflect functional differences in the comparison of the original samples.

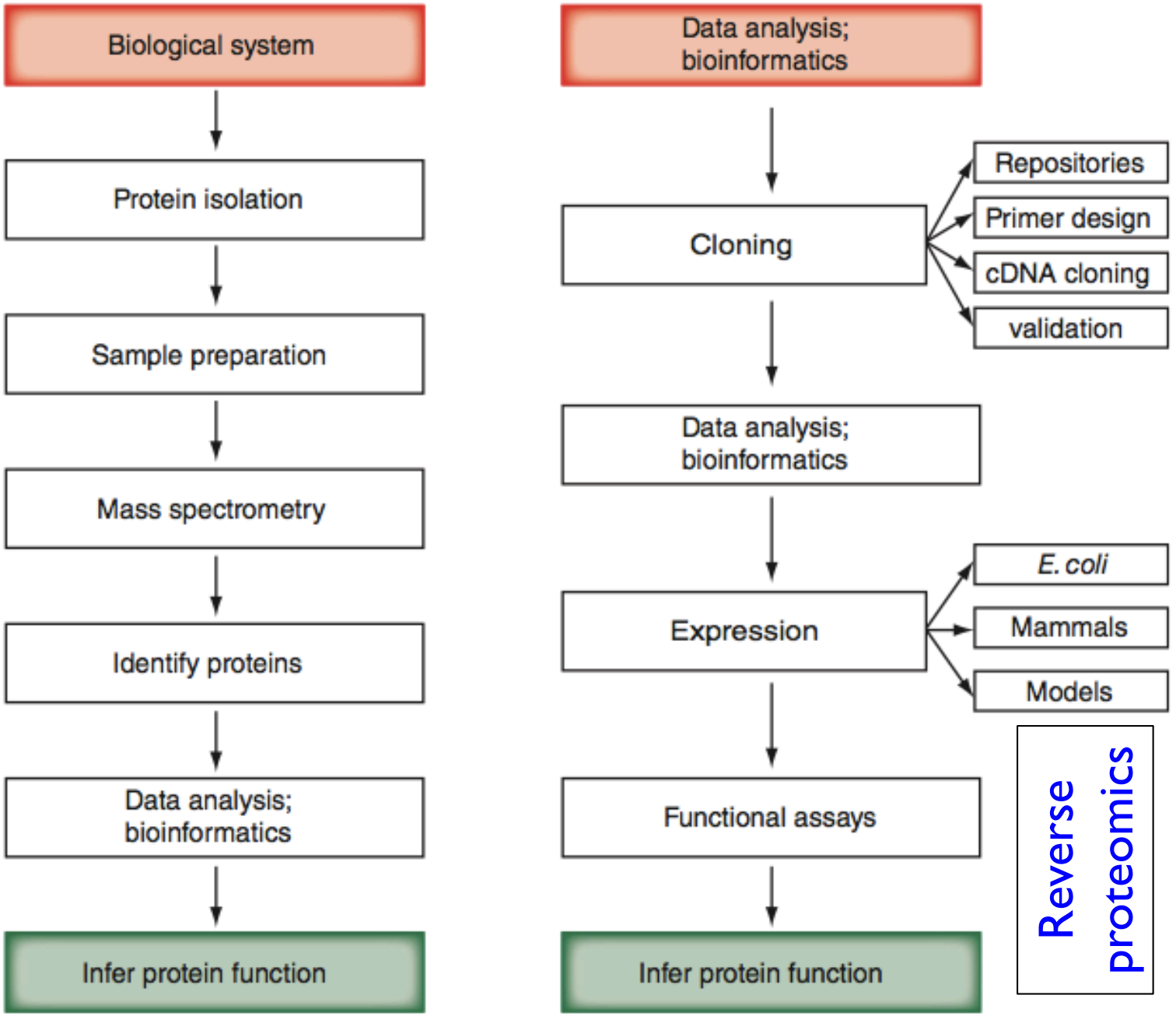
Proteomic approaches to functional genomics

We usually think of forward and reverse approaches in terms of genetics, but these terms can apply to proteomics.

Reverse proteomics:

- A genome sequence of interest is analyzed and genes, transcripts, and proteins are predicted.
- Complementary DNAs (cDNAs) are cloned based on information about open reading frames.
- cDNAs are validated by sequence analysis and expressed in systems such as *E. coli* (for the production of recombinant proteins), mammalian cells, or other model organism systems.
- Functional assays are performed; assays include the yeast two-hybrid system or other protein interaction assays.

Forward and reverse proteomics



B&FG 3e
Fig. 14.19
Page 671

Critical assessment of protein function annotation

The critical assessment of protein function annotation (CAFA) experiment is modeled on CASP. More than 48,000 protein sequences were released to 30 participating teams who predicted Gene Ontology (GO; Chapter 12) annotations.

The performances of various algorithms were assessed on a subset of 866 proteins for which "gold standard" GO annotation was available to the organizers, then used to assess the performance of prediction algorithms.

Critical assessment of protein function annotation

CAGI involved many challenges inherent in the nature of protein function:

- Protein function is defined at multiple levels, involving the role of a protein on its own and in pathways, cells, tissues, and organisms.
- Protein function is context dependent (e.g., many proteins change function in the presence of a signal such as calcium or a binding partner).
- Proteins are often multifunctional.
- Functional annotations are often incomplete and may be incorrect.
- Curation efforts map protein function to gene names, but multiple isoforms of a gene may have different functions.

Protein-protein interactions

Most proteins perform their functions in networks associated with other proteins and other biomolecules. As a basic approach to discerning protein function, pairwise interactions between proteins can be characterized.

Proteins often interact with partners with high affinity. (The two main parameters of any binding interaction are the affinity, measured by the dissociation constant K_D, and the maximal number of binding sites B_{max}.)

Protein-protein interactions

The interactions of two purified proteins can be measured with dozens of techniques such as the following:

- *Co-immunoprecipitation*: specific antibodies directed against a protein are used to precipitate the protein along with any associated binding partners.
- *Affinity chromatography*: a cDNA construct encodes a protein of interest in frame with glutathione *S*-transferase (GST) or some other tag. A resin to which glutathione is covalently attached is incubated with a GST fusion protein, and it binds to the resin along with any binding partners. Irrelevant proteins are eluted and then the specific binding complex is eluted and its protein content is identified.

Protein-protein interactions

- *Cross-linking with chemicals or ultraviolet radiation*: a protein is allowed to bind to its partners and then cross-linking is applied and the interactors are identified.
- *Surface plasmon resonance* (with the BIAcore technology of GE Healthcare): a protein is immobilized to a surface and kinetic binding properties of interacting proteins are measured.
- *Equilibrium dialysis and filter binding assays*, in which bound & free ligands are separated and quantitated.
- *Fluorescent resonance energy transfer* (FRET): two labeled proteins yield a characteristic change in resonance energy upon sharing a close physical interaction.

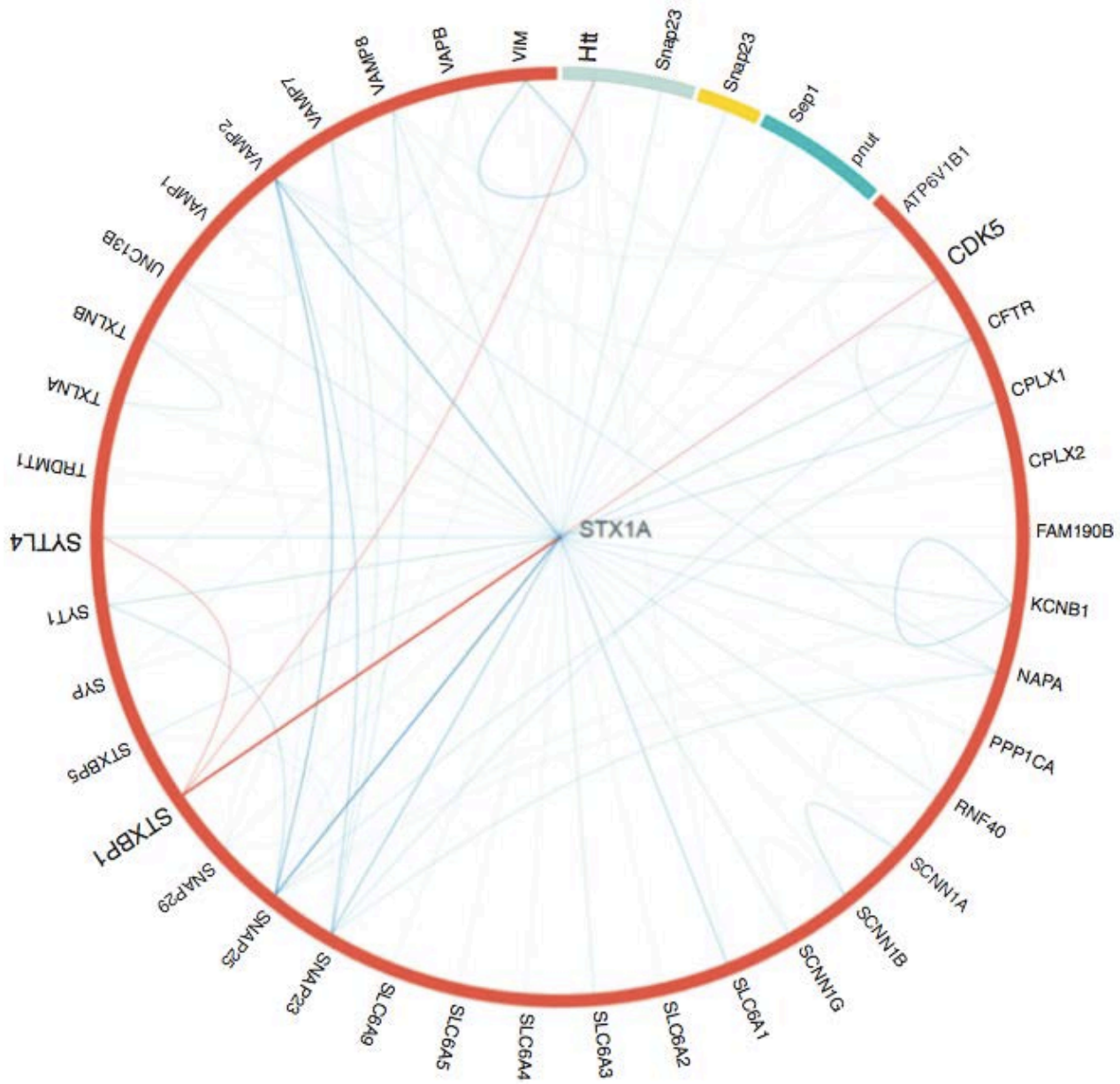
Protein–protein interaction databases

Database	Comment	URL
BioGrid	Repository for interaction datasets	http://www.thebiogrid.org/
Biomolecular Object Network Databank (BOND)	Requires log-in; formerly BIND	http://bond.unleashedinformatics.com/
Comprehensive Yeast Genome Database (CYGD)	From the Munich Information Center for Protein Sequences (MIPS)	http://mips.helmholtz-muenchen.de/genre/proj/yeast/
Database of Interacting Proteins (DIP)	From UCLA	http://dip.doe-mbi.ucla.edu/
Human Protein Reference Database (HPRD)	From Akhilesh Pandey's group at Johns Hopkins	http://www.hprd.org/
IntAct	At the European Bioinformatics Institute	http://www.ebi.ac.uk/intact/
Molecular Interactions (MINT) Database	Rome	http://mint.bio.uniroma2.it/mint/
PDZBase	Database of PDZ domains	http://abc.med.cornell.edu/pdzbase
Reactome	Curated resource of core human pathways and reactions	http://reactome.org/
Search Tool for the Retrieveal of Interacting Genes/Proteins (STRING)	Database of known and predicted protein–protein interactions	http://string.embl.de/

There are many prominent protein–protein interaction databases

B&FG 3e
Table 14.2
Page 677

Example of a protein-protein interaction database entry: BioGrid network map for syntaxin and its binding partners



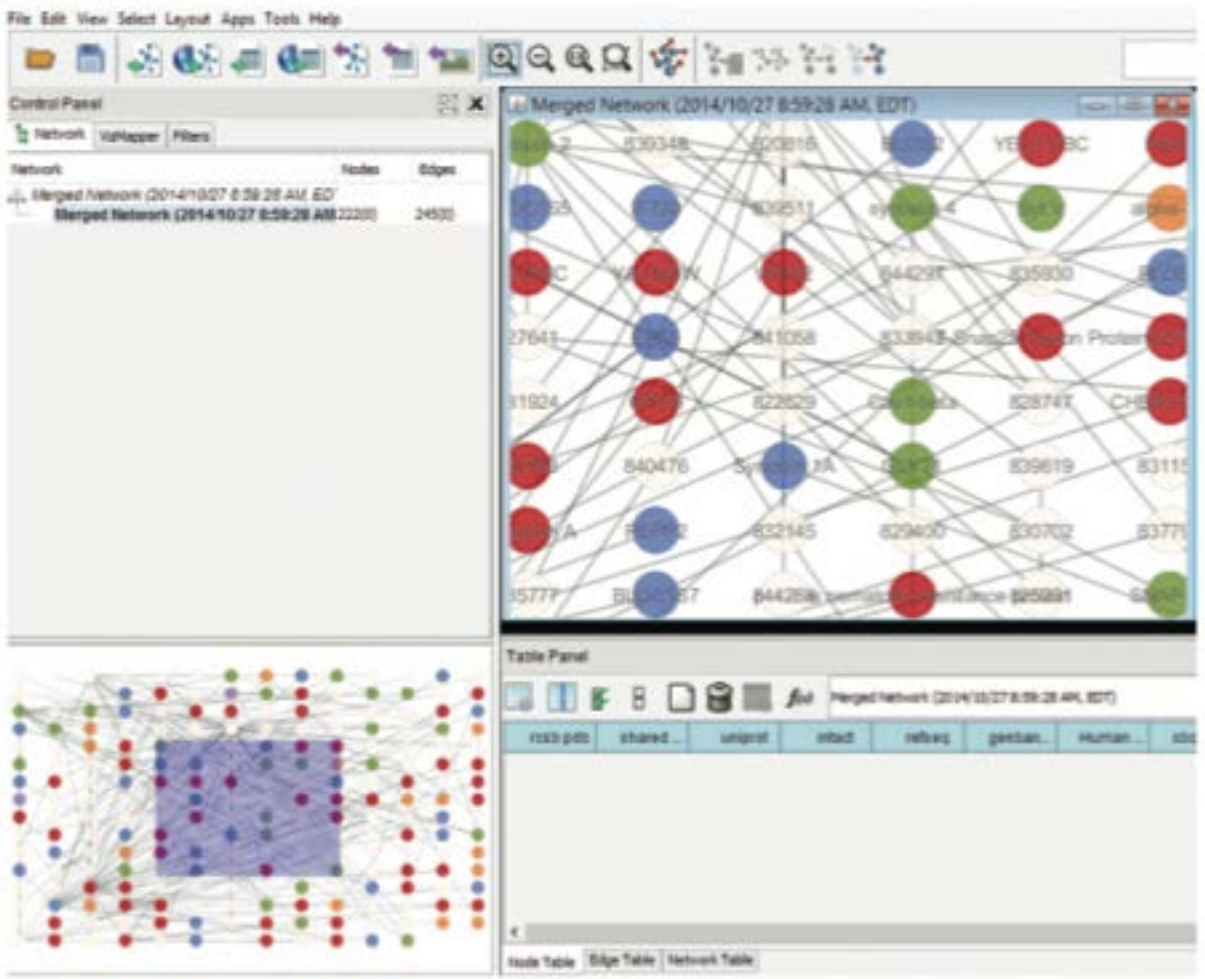
B&FG 3e
Fig. 14.20
Page 677

From pairwise interactions to protein networks

A typical mammalian genome has ~20,000 to 25,000 protein-coding genes, a subset of which (perhaps 10,000 to 15,000) are expressed in any given cell type. These proteins are localized to particular compartments (or are secreted) where many of them interact as part of their function.

Many databases show protein network data. We next show PSICQUIC and Cytoscape as examples.

Protein interaction networks



Zoom of Cytoscape diagram showing syntaxin binding partners

From pairwise interactions to protein networks

There are many issues regarding protein interaction networks.

- Assessment of accuracy. How likely is it that a false positive or false negative error has occurred? Benchmark ("gold standard") datasets are required that consist of trustworthy pathways.
- Choice of data. Many researchers integrate data from genomic sequences, expression of RNA transcripts, and protein measurements. But RNA and protein levels may be poorly correlated.
- Experimental organism. Function may be better conserved between paralogs than between orthologs!

From pairwise interactions to protein networks

There are many issues regarding protein interaction networks.

- Variation in Pathways. Some pathways (e.g. Krebs cycle) are characterized in great detail; many not. Some are transient, others stable.
- Categories of maps. Maps may be of metabolic pathways, physical and/or genetic interaction data, summaries of the scientific literature, or signalling pathways. Maps may be based on experimental data or inferred relationships.

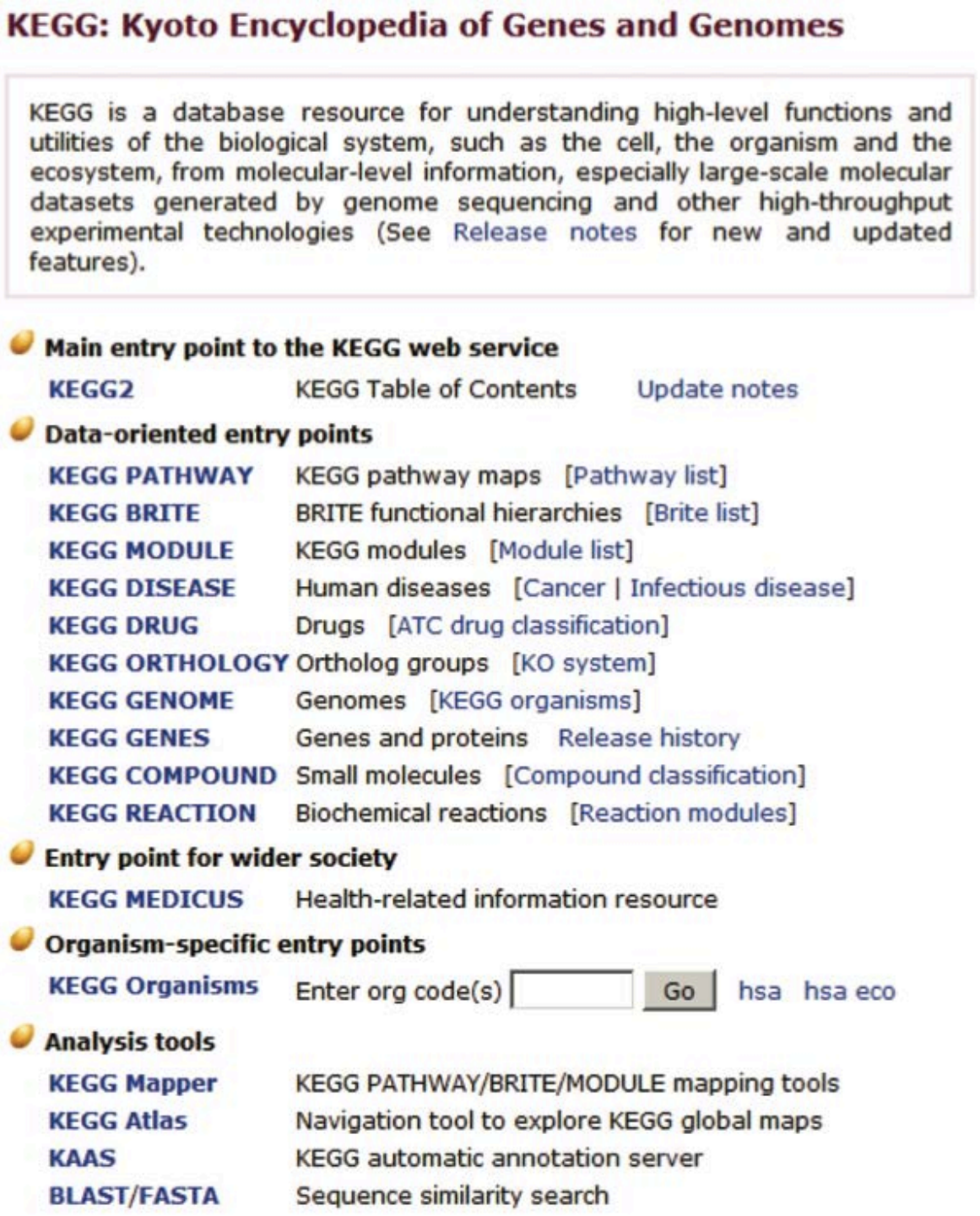
Pathways, networks, and integration: bioinformatics resources

There are many database resources.

- PathGuide lists >500 biological pathway resources.
- BioGRID database provides manual curation of ~32,000 publications describing physical and genetic interactions.
- MetaCyc is a database of metabolic pathways.
- Kyoto Encyclopedia of Genes and Genomes (KEGG) contains a detailed map of metabolism based on 120 metabolic pathways, with links to various organisms.
- KEGG pathways are a collection of manually drawn maps in six areas: metabolism; genetic information processing; environmental information processing; cellular pro- cesses; human diseases; and drug development.

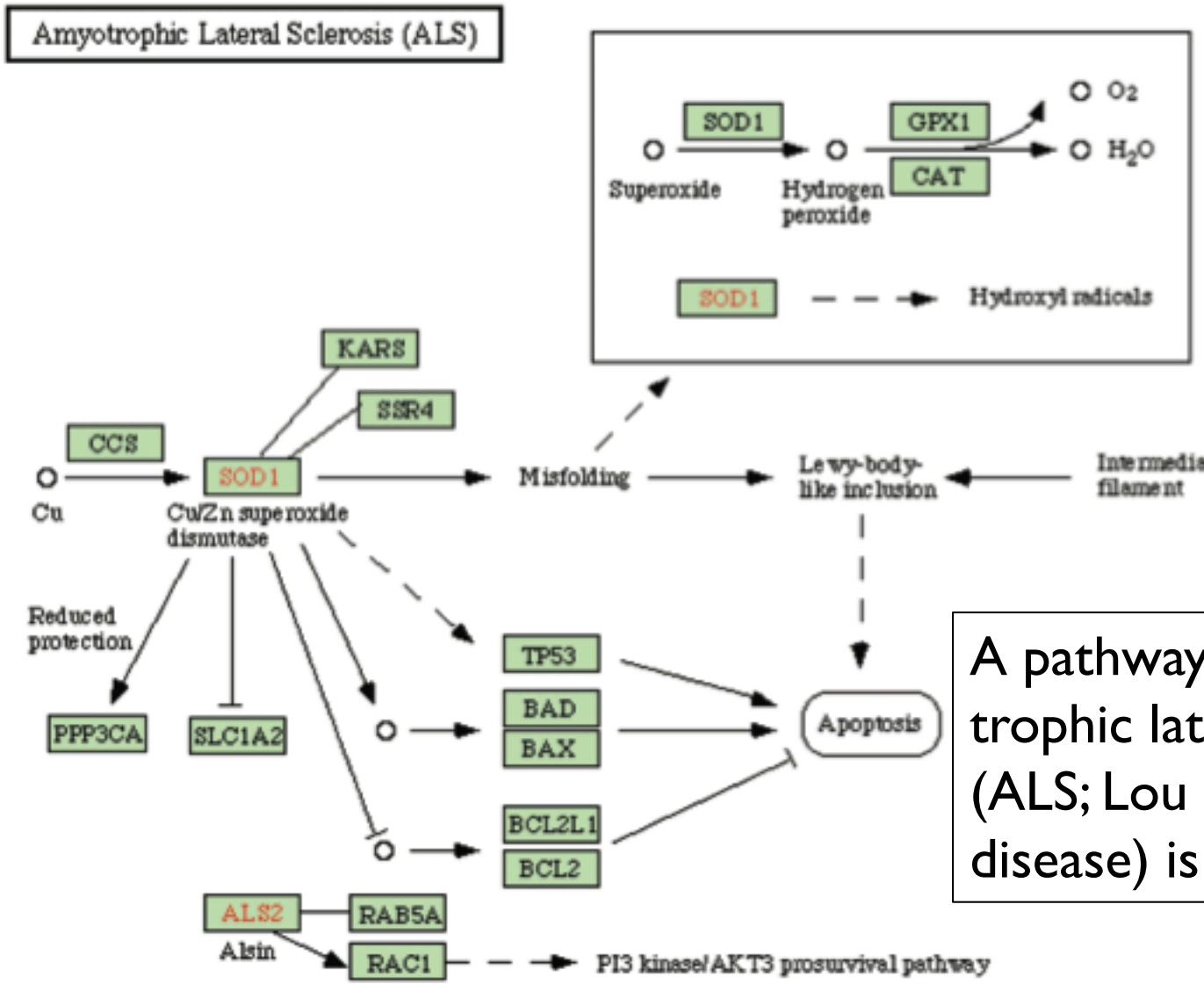
KEGG database

KEGG includes pathway maps, data for a broad range of organisms, and a variety of analysis tools.



B&FG 3e
Fig. 14.22
Page 683

KEGG database: disease pathways



A pathway for amyotrophic lateral sclerosis (ALS; Lou Gehrig's disease) is shown.

Fig. 14.24
Page 685

Perspective

The field of functional genomics is broad, and can be considered using many different categories.

- What type of organism do we wish to study? We highlighted eight model organisms, although many other models are commonly used.
- What type of questions do we want to address: natural variation or experimental manipulations used to elucidate gene function?
- What type of experimental approach do we wish to apply (e.g., forward versus reverse genetics)?
- What type of molecules do we wish to study (i.e., from genomic DNA to RNA to protein or metabolites)?
- What types of biological questions are we trying to address?

Perspective

- We are beginning to confront a problem that is perhaps even harder than identifying genes: identifying their function. Function has many definitions, as discussed in Chapters 8 and12 and in more detail here.