The Functional Analysis of Genomes

P.M. Selzer et al., *Applied Bioinformatics*, https://doi.org/10.1007/978-3-319-68301-0_6

6.1 The Identification of the Cellular Functions of Gene Products

The first human genome was published in 2001 by the Human Genome Project. At that time it was estimated that the number of human genes was in a range between 30,000 and 35,000. It is known today that the human genome is quite young from a phylogenetic point of view and shows an enormous difference between the number of genes and the size of the genome. It contains 19,000–20,000 genes (Ezkurdia et al. 2014), with an overall size of 3.3 Gigabases (see also ▶ Chaps. 4 and 7). Each human cell, except for sperm and eggs, has a complete set of genes. Obviously, however, a blood cell differs in its morphology and physiology from a liver cell. How, therefore, can these differences be explained if all cells have the same genetic material? The answer is simple. Not every gene is transcribed and expressed in every cell. It follows that only those proteins that are required are present in a cell at a given time during the cell's lifetime. The proteome of a cell or tissue is therefore dependent on the cell type and its current state.

In principle, the gene base order (the genotype) must therefore be altered via mutation for a modification of the gene expression and the resulting changes of the phenotype. But it has been shown in recent decades that environmental factors can influence the phenotype by alteration of the gene expression without adapting the nucleotide sequence of genes. This adaptation of gene expression is called epigenetic (Allis and Jenuwein 2016) and plays a fundamental role in the activation and inactivation of genes. The characteristic of this epigenetic modification is influenced by environmental factors like nutrition and stress. Nuclear DNA does not appear in free form but in the form of chromatin, which represents the basic building block of chromosomes. The basic repeat element of chromatin is the nucleosome, where DNA is wrapped around eight histone proteins. Based on the compact wrapping and the aggregation of individual nucleosomes, a gene can be active or inactive. In an active state it is called euchromatin and in an inactive state heterochromatin. This activation state can be influenced by the modification of single histone side chains. As an example, the acetylation of lysine side chains allows for interactions with bromodomain-containing proteins. The binding of these proteins increases the nucleosome accessibility and, therefore, the activity of transcription. In contrast, methylation leads to binding of chromodomain-containing proteins, which increases the nucleosome aggregation and decreases transcription (Allis and Jenuwein 2016). By now, a wide range of modifications and possible combinations are known, so that this is referred to as the histone code.

It also follows that knowledge of the genome and its genes is not sufficient to explain how a gene, a cell, or an organism works. To understand a complex biological system, one must study the regulation and expression of its genes, the function of expressed proteins, the quantitative occurrence of metabolites, and the effects of gene defects on an organism's phenotype. Besides the knowledge about genes, the function of the gene product must also be known. The study of this complexity is frequently termed systems biology, which tries to understand complete biological organisms and the dynamics behind a biological system as a whole. Its aim is to obtain an integrated picture of all regulatory processes at all levels, from the genome to the proteome and the metabolome, and from a single protein's behavior to the organelle and the biomechanics of the complete organism. Modern methods for the functional analysis of genomes (functional genomics) are called transcriptomics, proteomics, and metabolomics (◘ Fig. 6.1). These are usually high-throughput procedures that place heavy demands on data management and -analysis. These approaches are com-

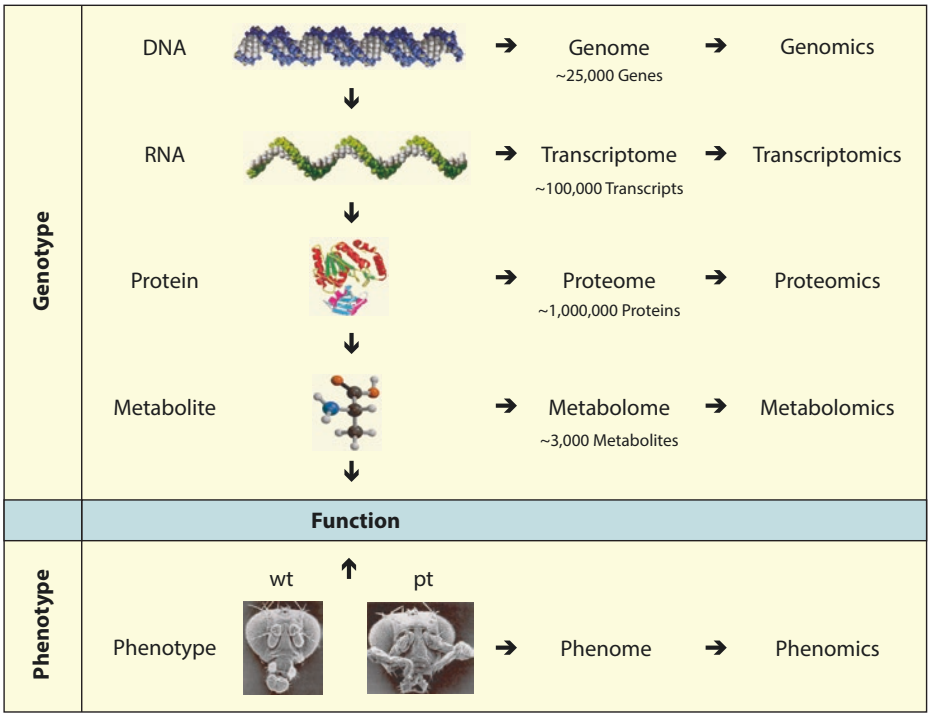

Fig. 6.1 Correlation between genotype and phenotype. From the genome via the transcriptome, proteome, and metabolome to the phenome. The example numbers in the genotype section are taken from *Homo sapiens*. The example in the phenotype section is taken from *Drosophila melanogaster* (Graphics of DNA, RNA, and metabolite from *Lehninger Biochemie*, 3rd Edition 2001, printed with permission from Springer-Verlag, Heidelberg, Germany. *D. melanogaster* microscopy images printed with permission from F. Rudolf Turner of Indiana University)

plemented by phenotypic analyses of model organisms and cells in vitro, also in a high-throughput format. Phenome describes all phenotypes and its analysis is called phenomics.

6.1.1 Transcriptomics

Unfortunately, the functions of many proteins based on nucleotide sequences alone are unknown. However, information regarding gene regulation and gene expression can offer insights into the functions of gene products in cells, tissues, and organisms. For example, because a gene is expressed exclusively in muscle cells, it can be inferred that the gene product is presumably important for the physiology of this cell type. Many techniques exist to analyze gene regulation and expression, e.g., the northern blot, which is a method for the detection of mRNA in agarose gels utilizing nucleic acid hybridization, or reverse transcriptase polymerase chain reaction (RT-PCR), a technique for the amplification of specific nucleotide sequences derived from mRNA. These methods, however, permit only the simultaneous analysis of just a few genes and are unsuitable for the efficient analysis of large amounts of data. Therefore, it became necessary to develop high-throughput procedures that permitted a more time-efficient analysis of whole transcriptomes.

6.1.1.1 DNA Microarrays

An example of a high-throughput method is the DNA microarray, which is well suited for the determination of cellular gene expression. Because one can create a profile of every cell based on the genes expressed, this method is also referred to as expression profiling. The solid support material of a DNA microarray can comprise a glass slide on which several thousand nucleic acid spots are placed next to one another (Fig. 6.2). Alternatively, other materials such as nylon membranes can be used as a support. Each DNA spot includes many copies of a unique single-stranded DNA, allowing for its unambiguous assignment to a specific gene (Holloway et al. 2002).

Many techniques are used for the production of DNA microarrays. In principle, one can distinguish between oligonucleotide arrays and cDNA arrays. For oligonucleotide arrays, short sequences of 20–50 nucleotides in length are synthesized directly on the support material (Fig. 6.2). The procedure involves photolithography, which was originally developed for semiconductor production and is still used in the computer industry. The glass slide is coated with linkers to allow for covalent bond formation with the nucleotides. The linkers are blocked with a photolabile protecting group to prevent the nucleotides from binding nonspecifically. By selectively applying a photo mask the photolabile protecting group is removed, thereby specifically activating selected array sectors. Then the surface of the array is incubated with a nucleotide solution that contains only one specific nucleotide, e.g., dATP. At the positions that had been activated by the photo mask, the nucleotide can now bind covalently to the linker of the support material. The nucleotides themselves are also blocked at the 5′ end with a photolabile protecting group, and these must be activated again before the following reaction can occur. Thus, by multiple repetitions and by the application of various masks, an oligonucleotide array of choice can be produced. This technique can produce densely packed microarrays with over 250,000 oligonucleotide spots per square centimeter. In 1994, Affymetrix became the first company to introduce a commercially available DNA chip [affymetrix].

In contrast, for cDNA arrays, considerably longer cDNA probes are placed on the array support (Fig. 6.2). First, the cDNAs are amplified to a length of several hundred nucleotides by means of PCR in the laboratory. These are then applied in tiny volumes as DNA spots onto the array support by means of a robot, after which they are immobilized, e.g., by ultraviolet light. A number of suppliers of spotting robots use slightly different procedures. One method is microspotting, whereby PCR products are applied with a capillary directly onto the array support. Another procedure is microspraying, whereby the cDNA solution is sprayed, much like from an ink-jet printer but without the nozzle ever touching the array support. A density of greater than 2500 DNA spots per square centimeter can be obtained with cDNA arrays.

The cDNA array technology is popular in many research laboratories because it is economical. Also, there is flexibility in the choice of the starting material (organism, tissues, or cells). Another microarray type is an oligonucleotide array, which is distinguished by an extreme density of high-quality spots. Because of the high density, for any given gene, several oligonucleotides can be placed on the array, permitting control of the results and increasing the precision of these arrays. The disadvantage of this technology is that the arrays are usually not produced in-house but must be purchased, often at considerable expense. Moreover, one is dependent on the arrays offered by the manufacturer which are not customizable.

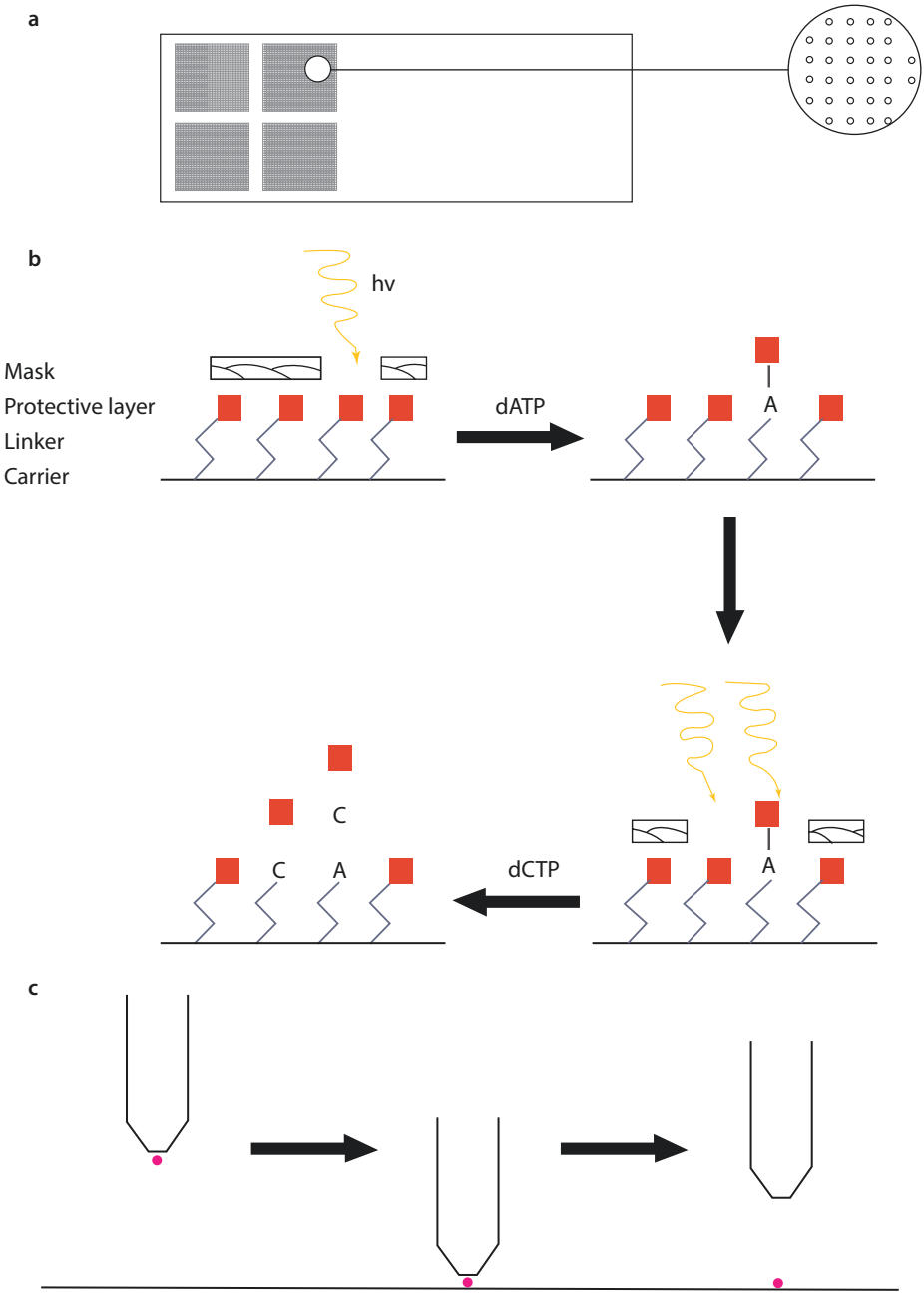

Fig. 6.2 DNA microarrays. **a** DNA microarray that consists of several thousand nucleic acid spots and arrayed in high density. **b** Schematic illustration of production of an oligonucleotide array using photolithography. **c** cDNA solutions are spotted during array production onto microarray plates by robots

The Performance of an Expression Profiling Experiment with cDNA

Many expression profiling studies compare the gene expression pattern of two different cell populations, e.g., that of healthy cells (cell type A) with tumor cells (cell type B) (Fig. 6.3). The first step is to isolate total RNA from both cell populations. The mRNA is transcribed into cDNA by the enzyme reverse transcriptase and simultaneously labeled by the incorporation of nucleotides that have been coupled to different fluorescent dyes. Usually, control cDNA (in this case from healthy cells) is labeled with Cy3 dye and sample cDNA (from the cancer cells) with Cy5 dye. Cy3 and Cy5 emit light in the green and red spectra, respectively. This method is referred to as direct labeling. In contrast, indirect labeling methods are used only if very small quantities of starting material are available. In this case, modified nucleotides are incorporated during cDNA synthesis that binds special dyes with a high affinity.

The labeled cDNA pools are mixed and denatured, and the single-stranded cDNAs are then incubated with the DNA microarray. DNA in the pools hybridizes to the complementary single-stranded DNA molecules making up the array. Laser activation of the microarray in the emission frequencies of the dyes followed by quantitative scanning of the emitted light measures the amount of bound cDNA. As a result, one gets two pictures, one in the green and one in the red wavelength range. If both are superimposed, a merged picture will result with colored spots (Fig. 6.3).

If genes are differentially expressed, i.e., in one cell population there are larger quantities of a specific mRNA, the spots will appear red or green. Spots will appear red if more Cy5-labeled cDNA is bound, i.e., an overexpression of those genes in the cancer cells compared to the controls. Conversely, spots will be fluorescent green if genes are less expressed in the cancer cells than in the controls. Spots appear yellow if red and green fluorescent cDNAs have hybridized to the spotted DNA in equal amounts. This means that the corresponding genes are expressed in the control and cancer cells at equal levels. Spots for which no complementary cDNAs are present in the pools appear black. Thus, it is obvious that the expression of a gene is a relative value between two samples; absolute quantities are not possible with cDNA arrays. This differs from oligonucleotide arrays that allow for an absolute quantification.

Interpretation of an Expression Profiling Experiment

Although the idea behind microarrays is simple, their use and the analysis of the results are more complex. This is due to the numerous sources of error, including statistical errors, based on stochastic fluctuations that cannot be influenced and systematic errors that lead to measurement deviations. Such systematic errors can be due to incorrect calibration of the instrument or changing environmental conditions (e.g., fluctuations in the temperature or atmospheric humidity) during its operation.

Errors can be minimized by proper experimental design. Statistical errors can be minimized by the repetition of experiments. Samples should be freshly prepared each time to ensure that each experiment is independent. Systematic errors can be minimized by a sophisticated experimental design and control experiments. One control experiment is dye swapping, in which cDNAs are labeled with a dye that is the opposite of that used in the original experiment (reciprocal labeling). Specifically, if in the original experiment the cDNA from the cancer and control cells was labeled with Cy5 and Cy3, respectively, then in the dye swapping control experiment, the cDNA of the cancer cells should be labeled with Cy3 and that of the control cells with Cy5. Because the same cDNA preparation is used for both the original and dye swapping control experiments

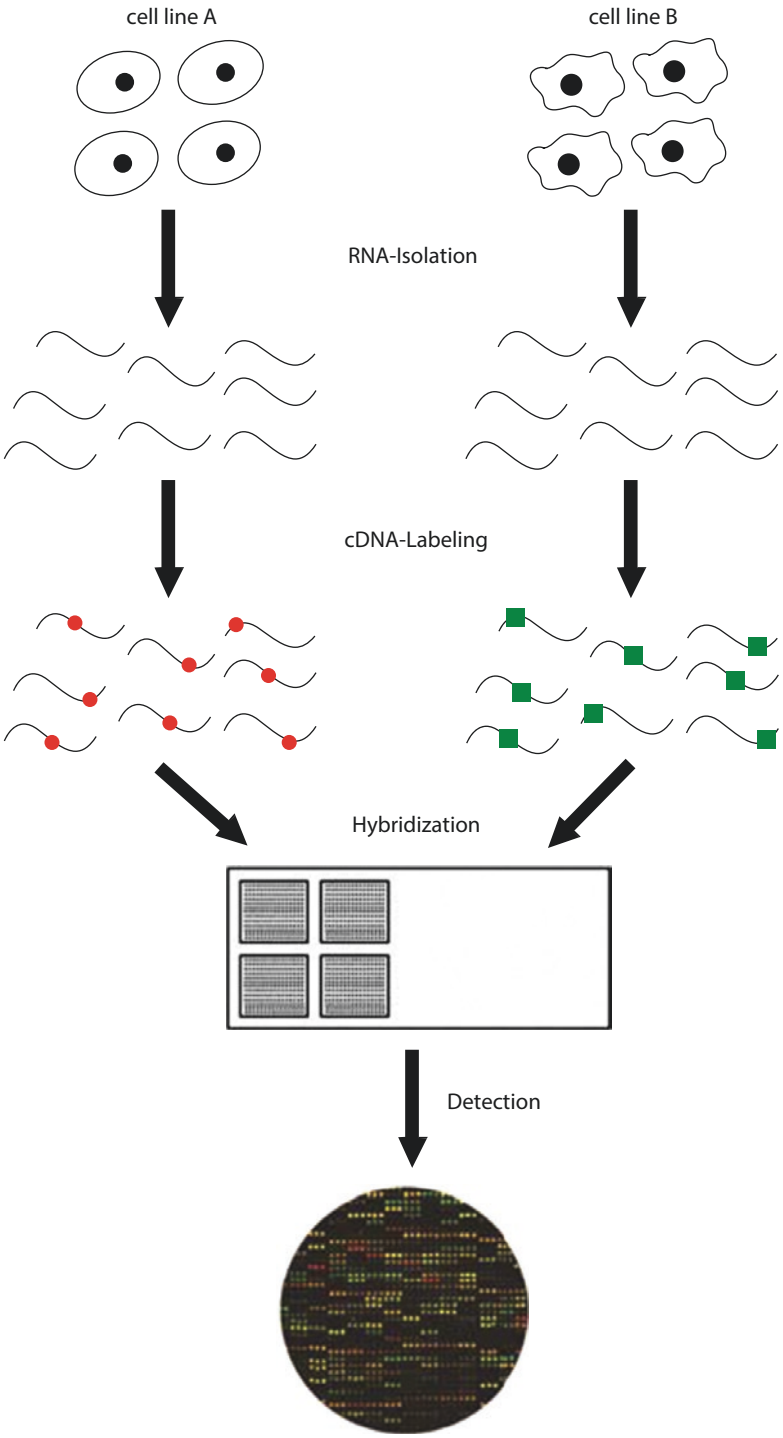

Fig. 6.3 Comparison of gene expression levels in two different cell lines as part of an expression profiling experiment using cDNA microarrays (Iron chip, printed with permission from M. Muckenthaler, EMBL Heidelberg, Germany)

and only the label differs, similar results should be obtained. Using the dye swapping control experiment, one can investigate whether an error occurred during labeling of the samples, and, if so, the extent of the error can be taken into account in the analysis of the results (Churchill 2002).

Interpretation of the data starts with an analysis of the figures made by the microarray scanner. The intensities of each spot must be measured to convert them into numeric values. This is a complex and comparatively difficult step. The many thousands of spots must be unambiguously identified. To do this, the peripheries of the spots and the fluorescence intensities in the two light channels must be measured and both then compared with the background. Atypical spots that have irregular shapes or contain clumps of red and green color can be marked and ignored for further analysis. All of these processes are usually carried out by the software of the microarray scanner.

Considering the huge number of providers of microarrays and supplies and the different protocols and complicated experimental setup (split into numerous individual steps), it is not surprising that microarray data contain systematic errors. Examples are the uneven distribution of the hybridization solution on an array, which leads to the nonhomogeneous staining of some areas of the array, or the different half-lives of the dyes, which can lead to inaccuracies when measuring spot intensities. To compensate for such systematic errors, the expression profiling values must be normalized. Normalization is based on the hypothesis that most genes are not differentially expressed in the samples. Normalization not only adjusts the results but also ensures the comparability of experiments carried out on different days or in different laboratories. There are numerous algorithms for normalization, and they all have advantages and disadvantages. The choice of algorithm depends on the experience and preference of the researcher (Quackenbush 2001).

It has long been a subject of discussion whether microarray platforms from different suppliers can be compared at all. In spite of these concerns, several researchers have shown that comparison is actually possible with an adequate experimental setup. However, standardized protocols and adequate controls are essential (Ji and Davis 2006). To oversee quality control, two consortia have been formed with members from academic research groups, the microarray industry, and US agencies. The MicroArray Quality Control Project [maqc] establishes standard controls that are aimed at facilitating the comparison of microarray experiments. The External RNA Controls Consortium (ERCC) [ercc] has similar aims. The ERCC develops external RNA controls that are added to the experimentally isolated RNA before cDNA synthesis. In this way, the extent to which the results of a microarray experiment agree with defined minimal criteria can be verified.

The next step in data analysis is the identification of genes for which expression is significantly different between the two samples. For simplicity in early microarrays, it was assumed that all those genes for which expression in the samples varied by at least twofold were differentially expressed. Today more complex statistical procedures are used to identify those genes with significant differences in expression levels. These methods have the advantage of identifying genes with low yet significant differences in expression levels. After these statistical analyses, a final number of genes that is differentially expressed is obtained. Importantly, the results should be validated by independent methods, such as northern blot analysis (Slonim 2002).

The determination of the differential expression of individual genes is not the only interesting aspect of microarrays, however; also of interest is the recognition of patterns in gene expression profiles. The idea is that genes that belong to a pathway or react in concert to a given environmental stimulus are coregulated and, therefore, display a similar expression profile. Using cluster analysis all genes with similar expression profiles can be combined into groups or clusters. ◘ Figure 6.4 shows such an analysis for 164 bacterial genes that are divisible into 13 clusters. Cluster analyses provide valuable insights into the function of proteins. If genes for whose products no function is currently known clustered with well-characterized genes, then coregulated expression could indicate a similar function or a common pathway to those unknown gene products. The unknown proteins could then be specifically examined for these properties.

Each expression profiling experiment generates an enormous amount of data. One experiment can include dozens of microarrays, which in turn consist of several thousand spots. Therefore, the resulting several hundred thousand or even millions of measurements must be managed and analyzed using special databases in which the data can be saved and retrieved at any time. Example databases are the Gene Expression Omnibus of the National Center for Biotechnology Information (NCBI) [geo] and the ArrayExpress of the European Bioinformatics Institute (EBI) [arrayexpress]. In addition to results, one can also find unprocessed raw data as well as the protocols and conditions under which the experiments were performed. These data should comply with the minimum information about a microarray experiment [miame] protocol in which the minimum requirements for an explicit interpretation and reliable reproduction of the microarray experiments are defined (Brazma et al. 2001).

In summary, performing microarray experiments, inclusive of the bioinformatic component, is complex and places high demands on the experimenter. Luckily, a variety of software solutions exist that simplify the analysis of the data. A known commercial program for the analysis of microarray data is the GeneSpring GX collection of Agilent Technologies [agilent]. Frequently used software packages that were developed in the academic environment are Bioconductor [bioconductor], the TM4 suite [tm4], and GenePattern [genepattern].

Besides expression profiling there are a variety of other applications for microarrays (Gershon 2005) that have gained increasing importance, for example in tumor medicine. The optimal treatment of a cancer patient is critically dependent on a diagnosis that is as precise as possible, which at present is based on a combination of clinical and histopathological data. In some cases, however, an exact diagnosis is difficult because tumors frequently have atypical properties. In such cases microarrays can help classify tumors according to their gene expression profiles. An example is acute leukemia. This cancer of leukocytes can be subdivided into acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) using clinical and morphological data for diagnostics. The distinction of these subtypes is essential because each is treated with different chemotherapeutics. An initial study (Golub et al. 1999) examined whether reliable results could be obtained by molecular diagnostics with the help of DNA microarrays compared to classical methods. The gene expression profiles from patients with a known diagnosis were analyzed and then compared with those from patients with an unknown diagnosis. The result demonstrated that the microarray diagnostic tool was reliable. In addition, a patient with a diagnosed atypical acute leukemia was also examined. Here

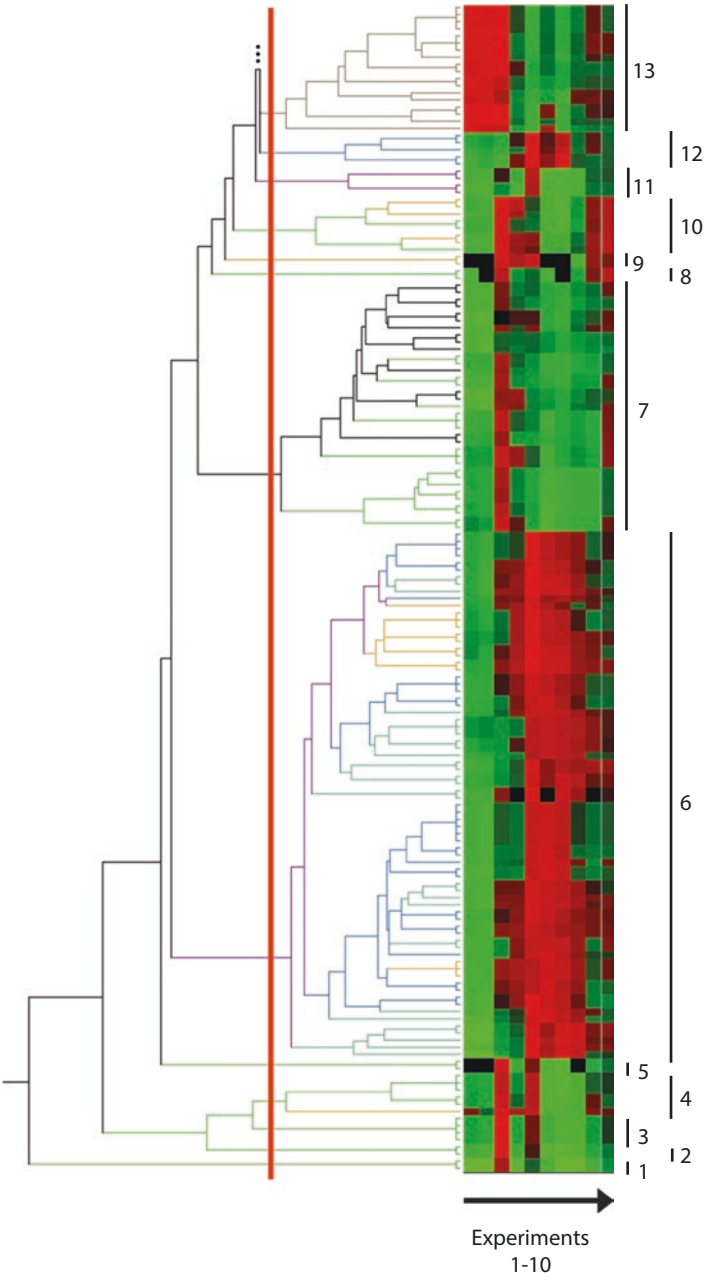

Fig. 6.4 Clustering of genes with similar expression profiles. Expression of 562 bacterial genes was measured in 10 different experiments. The expression profiles were then compared, and genes with similar expression patterns were grouped into clusters. In this figure, 13 clusters (black bars) with overall 164 genes are shown. For instance, cluster 13 contains 18 genes, which are highly expressed in the first 3 experiments (red), but then subsequently expression decreases (green). The red bar represents the threshold selected to define a cluster

the microarray diagnostic tool showed that this patient's gene expression profile was completely different from those of other patients. Its profile pointed more to a cancer of muscle tissue than to an acute leukemia. Because cytogenetic examinations also disagreed with an acute leukemia diagnosis and favored a muscle tumor, the final diagnosis and chemotherapy were changed accordingly. Thus, the classification of tumors based on DNA microarrays provides validated support to the standard diagnostic techniques (Golub et al. 1999).

Another important field for the application of microarray technology is toxicology. Toxicological analyses are designed to identify the damaging consequences of chemical substances on cells. For example, a potential new antibiotic might not only kill the infectious bacterium but also damage the cells or whole organs of the patient. Therefore, any new potential drug is studied for its toxicological properties by comparing them with existing toxins. These comparisons include gene expression profiling in DNA microarrays. If overlaps in the expression profiles between the known toxins and new compound occur, then the new substance will be classified as being potentially toxic. The analysis of toxicological characteristics using DNA microarrays is also known as toxicogenomics.

6.1.1.2 Serial Analysis of Gene Expression

Like DNA microarray technology, serial analysis of gene expression [sage] is a high-throughput technology for measuring gene expression. SAGE facilitates the comparison of gene expression in different cells or tissues and, therefore, the identification of differentially expressed genes. SAGE also requires the isolation of total RNA from cells or tissues and the conversion of mRNA into cDNA using the virally sourced enzyme reverse transcriptase. The cDNA is not cloned, however, but instead is treated with certain restriction enzymes that cut the DNA at specific sites. This results in the generation of short DNA fragments from each individual cDNA pool with a length between 10 and 11 nucleotides, a tag. Despite being so short, a tag is usually sufficient to unambiguously identify a specific mRNA. The tags are connected into long serial molecules and subsequently cloned into plasmids for sequencing. In a SAGE experiment, the frequency with which a tag appears in a sample is used as a measure of the magnitude of expression of the corresponding mRNA. For example, if the gene tag is found 5 times in a sample from healthy cells but 20 times in a sample from cancer cells, then one assumes that this gene is approximately fourfold overexpressed in the cancer cells. SAGE results can be saved in the Gene Expression Omnibus [geo] database at the NCBI. There, the information about each tag can be found, including its DNA sequence, frequency in tissues or cells, and the specific transcript from which the tag was derived [sage, sagemap].

The great advantage of SAGE over DNA microarrays is that all mRNA transcripts of a cell can be analyzed, including unknown transcripts (e.g., new splice variants). In the case of DNA microarrays, only mRNA transcripts are analyzed with existing cDNA spots on the microarray. Another advantage of SAGE is its steady reproducibility between experiments. One disadvantage of SAGE is the enormous amount of time needed to conduct high-throughput experiments. DNA microarrays, in contrast, show a high flexibility, and in the age of genome sequencing they allow for analyzing genes of the whole genome in a few experiments. SuperSAGE represents an improvement in

these methods that seems to compensate for the drawbacks by using different restriction enzymes that produce bigger tags (Matsumura et al. 2006). Millions of tags can now be analyzed in combination with next-generation sequencing.

6.1.2 Proteomics

The quantification of mRNA by DNA microarrays or SAGE provides important information about potential cellular functions of gene products. Measuring mRNA alone, however, is not sufficient to completely and precisely describe complex biological systems. Ultimately, cellular activities like metabolic processes are mediated by proteins of the proteome and not by genes of the genome or mRNA of the transcriptome. Analogous to the DNA microarray technology, therefore, high-throughput procedures have been developed for the parallel functional analysis of proteins, i.e., proteomics. Proteomics is classified into two categories: classical or quantitative proteomics and functional proteomics. Classical proteomics deals with the identification and quantification of proteins in cell lysates, whereas the aim of functional proteomics is the determination of protein function.

The Human Proteome Project [hpp] is an international consortium of several research groups that is comparable to the Human Genome Project. The aim is the systematic analysis and characterization of the human proteome for a better understanding of human biology on the cellular level. This should lead to improved medicinal applications (i.e., improved therapy and diagnosis of diseases). An important part of the project deals with a chromosome-based proteome analysis to analyze and understand the function of every single gene. A cooperation of different research groups in the fields of genomics, transcriptomics, proteomics, and metabolomics is needed to achieve this (◘ Fig. 6.5).

6.1.2.1 Classical Proteomics

Classical proteomics is similar to expression profiling, which is why it is also termed protein profiling. Both technologies permit the molecular fingerprinting of a cell based on the genes expressed at the mRNA or protein level. By comparing two or several such fingerprints, differentially expressed genes and proteins can be identified. Both technologies have advantages and disadvantages. Protein profiling detects the proteins that ultimately perform cellular functions. Also, the quantitative modifications in a protein's composition based on either a new synthesis or breakdown (protein turnover) can be measured. Other advantages of protein profiling are the ability to verify posttranslational modifications (e.g., phosphorylation and glycosylation) and to determine the protein composition of cellular compartments (e.g., of a mitochondrion or nucleus). One disadvantage, however, is that not all proteins are soluble, particularly transmembrane proteins, and therefore cannot be detected. A second limitation is the limit of detection such that weakly expressed proteins can be missed. In contrast, complete genomes can be analyzed in a few DNA microarray experiments, yet the assumption in expression profiling that the quantity of mRNA stochastically reflects that of the protein is often unwarranted. Moreover, the quantity of mRNA cannot provide information about protein turnover. Therefore, where possible, both expression and protein profiling should be performed as complementary techniques.

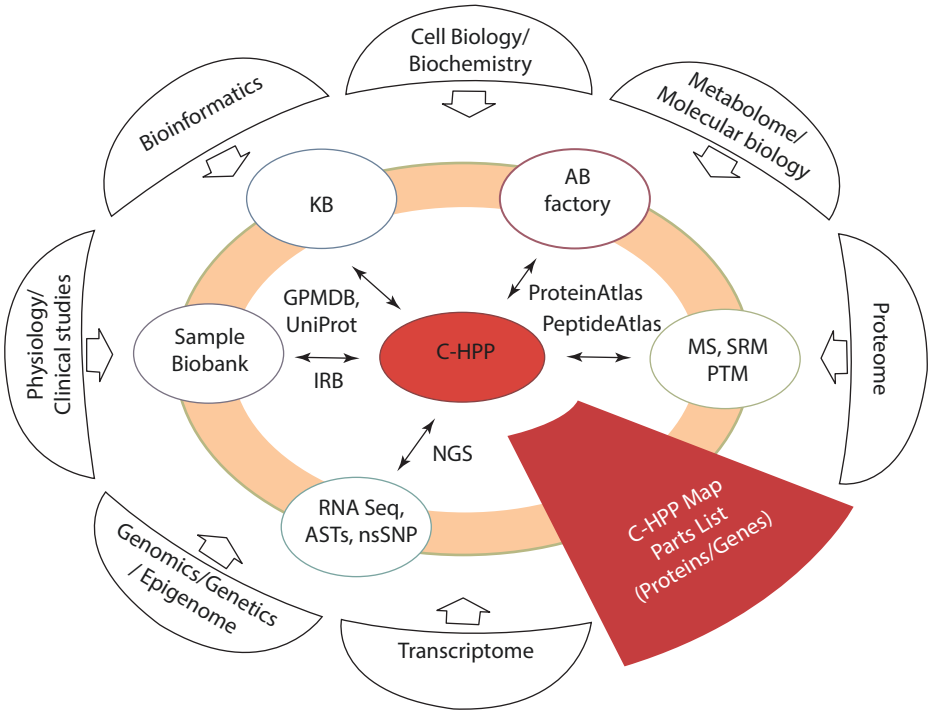

Fig. 6.5 Chromosome-based part of the Human Proteome Project (C-HPP). MS: mass spectroscopy, AB: antibody, KB: knowledge base (Taken from ► http://www.c-hpp.org/)

A common procedure for protein profiling combines two-dimensional gel electrophoresis (2D gel electrophoresis) with mass spectroscopy. In 2D gel electrophoresis, cell proteins are first separated through a separating matrix (e.g., a polyacrylamide gel) according to their individual charges generated by an electrical field. Separation is, therefore, possible owing to two inherent properties of proteins, charge and mass. The charge of a protein depends on its amino acid composition, e.g., cytochrome c contains many basic amino acids and is, therefore, positively charged at neutral pH. The net charge a protein carries depends on the pH of its surroundings, and the pH at which both the positive and negative charges of a protein are equal (i.e., a net charge of zero) is called the isoelectric point (pI). Accordingly, a protein will not migrate in the electrical field when its pI equals the pH of its surroundings. Because each protein has a characteristic pI value, one can separate a protein mixture in a pH gradient using an electrical field. This method, called isoelectric focusing, is used in 2D gel electrophoresis as the first dimension for the separation of proteins. In the second dimension, proteins are separated only according to their molecular mass. Peptides with a low molecular mass move faster than larger proteins through the pores of the polyacrylamide gel. In this way, up to 10,000 different proteins can be separated in high-resolution 2D gels. After separation, proteins are made visible using different staining procedures (e.g., silver staining or staining with fluorescent dyes) (Fig. 6.6). The gels are then digitized and evaluated with bioinformatic methods. Programs such as Melanie [melanie] at the

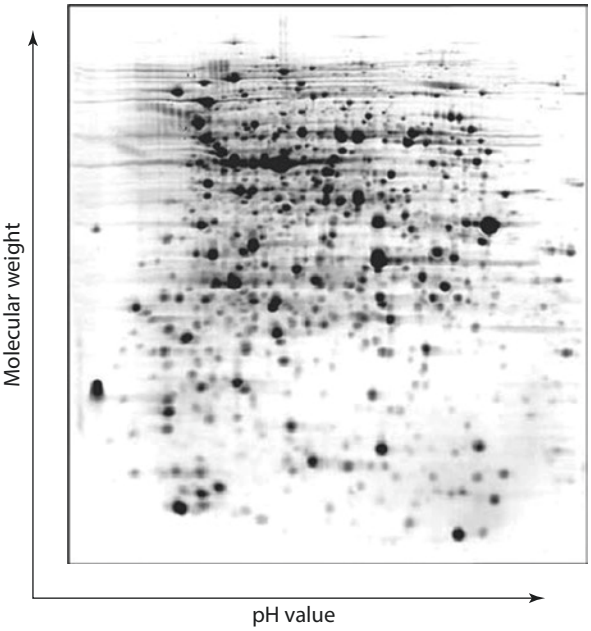

Fig. 6.6 Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE). A protein lysate of a bacterium was separated along a pH gradient (pH 3–10) in the first dimension and by molecular mass in the second dimension. The resolved proteins were then visualized by silver staining

Expasy proteomics server allow for the automatic detection and precise quantification of protein spots. Furthermore, Melanie allows the comparison of several 2D gels. Colocalized protein spots on different gels are identified and their quantitative differences measured based on spot intensity. Melanie also contains algorithms for normalization and statistical analyses with which the significance of the results can be judged to identify differentially expressed proteins.

The bioinformatic evaluation of 2D gels yields a list of expressed proteins for which only the isoelectric points and molecular masses are known. While the identity of some of these proteins can be determined using this information, for most proteins a partial determination of amino acid sequence is required. This sequence is compared with a protein database, and if the protein already exists, the identity can be confirmed.

Various techniques are used for the determination of amino acid sequence. A reliable method is sequencing by Edman degradation, for which, however, relatively large amounts of protein are required. An advance in protein analytics is mass spectroscopy–based analysis of peptides via matrix-assisted laser desorption/ionization–time of flight (MALDI–TOF). MALDI–TOF is sensitive enough to require only picomolar amounts of protein. Stained protein spots are excised from the polyacrylamide gel and incubated with a protease (e.g., trypsin), which hydrolyzes each protein into a specific peptide pattern. The peptides are extracted from the gel and analyzed after MALDI in a TOF spectrometer. For each peptide a specific peptide mass spectrum is generated (Fig. 6.7). At the same time, all proteins in a database are digested into peptides in silico based on the same cleavage specificity of trypsin, and the theoretical mass spectra of these fragments are calculated. The experimentally determined MALDI–TOF mass spectra are

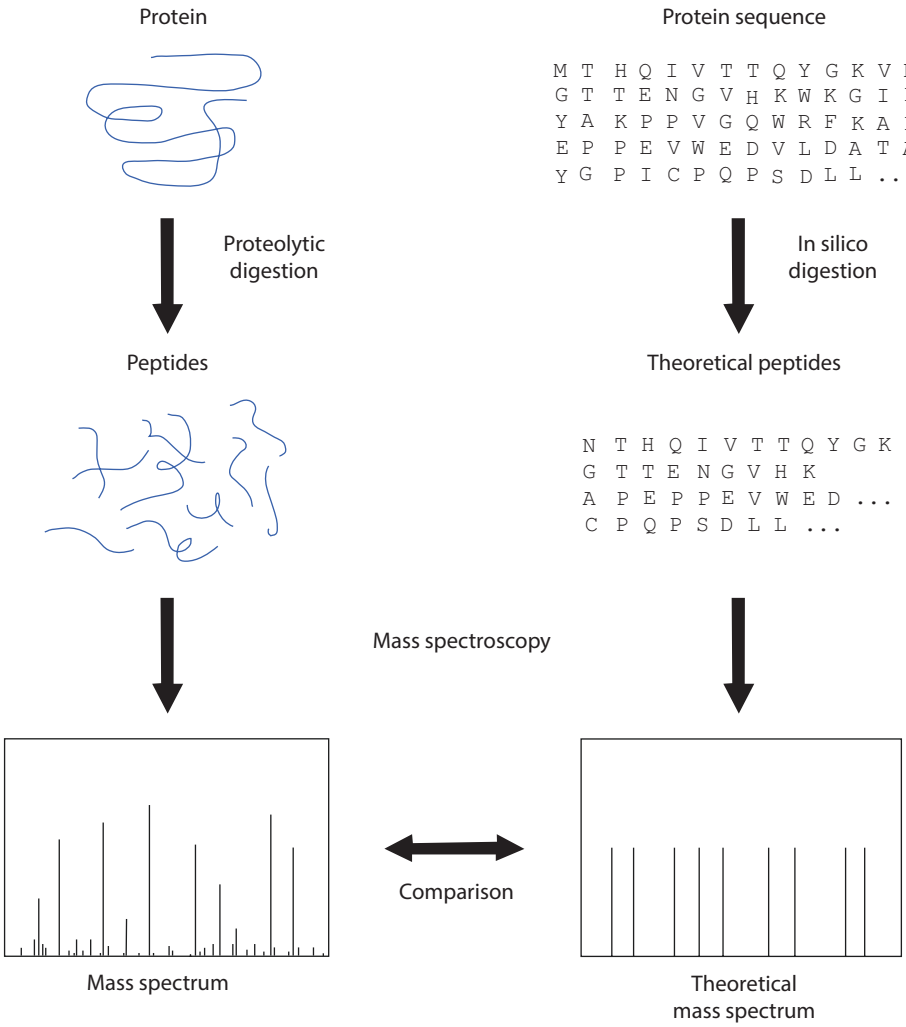

Fig. 6.7 Identification of proteins by cross-referencing data from mass spectroscopy experiments and mass spectra that were theoretically computed

then compared with the theoretical spectra, and those mass spectra that are identical are selected. Because a MALDI–TOF mass spectrum can result from more than one protein, definitive identification of the protein requires the spectra of several peptides. Thus, if several of the mass spectra determined by MALDI and those determined theoretically agree, then the experimentally analyzed protein is the same protein as that identified in the database.

An alternative protein ionization technique is electrospray ionization (ESI). ESI is sensitive and particularly suited to the analysis of high-molecular-mass compounds like proteins. The advantage of ESI over MALDI is that one can couple ESI to a liquid

chromatographic (LC) system. The latter can fractionate protein solutions and, at least with samples of moderate complexity (i.e., with a limited number of different proteins), replace the laborious 2D gel electrophoresis. By the direct coupling of an LC system to the mass spectrometer (LC/MS), protein identification is accelerated. The disadvantages of ESI are its strong sensitivity to alkali contamination and the somewhat more problematic assignment of distinct masses.

Other developments have also occurred in the field of mass spectroscopy (Griffin et al. 2001). In tandem mass spectroscopy (MS/MS), two mass spectrum analyzers are run consecutively, which greatly improves the sensitivity and selectivity of the system. For example, protein samples are ionized by ESI. Then, in the first spectrometer, ions of a given mass are selected and excited for further fragmentation, and detailed analysis is performed in the second spectrometer. Because of the combination of analyzers, therefore, an initial chromatographic separation may be unnecessary. In practice, however, these systems are frequently coupled as part of an LC-MS/MS system or even of a 2D LC-MS/MS system, which further increases sensitivity and selectivity.

6.1.2.2 Functional Proteomics

The aim of functional proteomics is to elucidate the function of proteins, e.g., identify protein–protein interactions. Many cellular processes are governed by such interactions, and their identification is an important topic for the understanding of protein function overall. Examples are the allosteric inhibition of enzymes, the regulation of signal transduction cascades by protein kinases, and the assembly of structural protein complexes to form the cytoskeleton. Numerous methods allow the analysis of such interactions, such as affinity chromatography and the yeast two-hybrid system. Their applications, however, are usually confined to studying the interactions of a limited number of proteins. In the meantime, these methods have advanced to the point where they can be used to dissect protein–protein interactions in complete proteomes (Fig. 6.8). In this context, the term interactome of an organism applies, and this type of research is also called interactomics.

To detect the interaction of two fusion proteins, the yeast two-hybrid system is commonly used (Fig. 6.9). Protein X, for which an interacting protein is sought, is coupled to the DNA-binding domain of a transcription factor. Protein X is then mixed with the expression products translated from a cDNA library (arbitrary protein Y) that have been fused to the cognate transcription factor's activating domain. Neither X nor Y alone is capable of forming a complete and functional transcription factor. Only when proteins X and Y interact are both domains brought together, and a functional transcription factor results that can activate the transcription of reporter genes. Their expression can be measured by activity tests and is thus indicative of an interaction between proteins X and Y. Using yeast two-hybrid, the whole proteome of baker's yeast (*Saccharomyces cerevisiae*) was analyzed for protein–protein interactions, leading to 4540 interactions of 3278 different proteins (Ito et al. 2001). An analysis of a large proportion of the human proteome was also tested for protein–protein interactions. Approximately 2800 protein–protein interactions for 1549 proteins were identified (Rual et al. 2005).

Tandem affinity purification (TAP) is another technology that is suitable for the analysis of multiprotein complexes. This technique is based on the combination of affinity chromatography and mass spectroscopy. The target gene is modified so that the gene product is labeled with a short peptide sequence or tag that facilitates the isolation

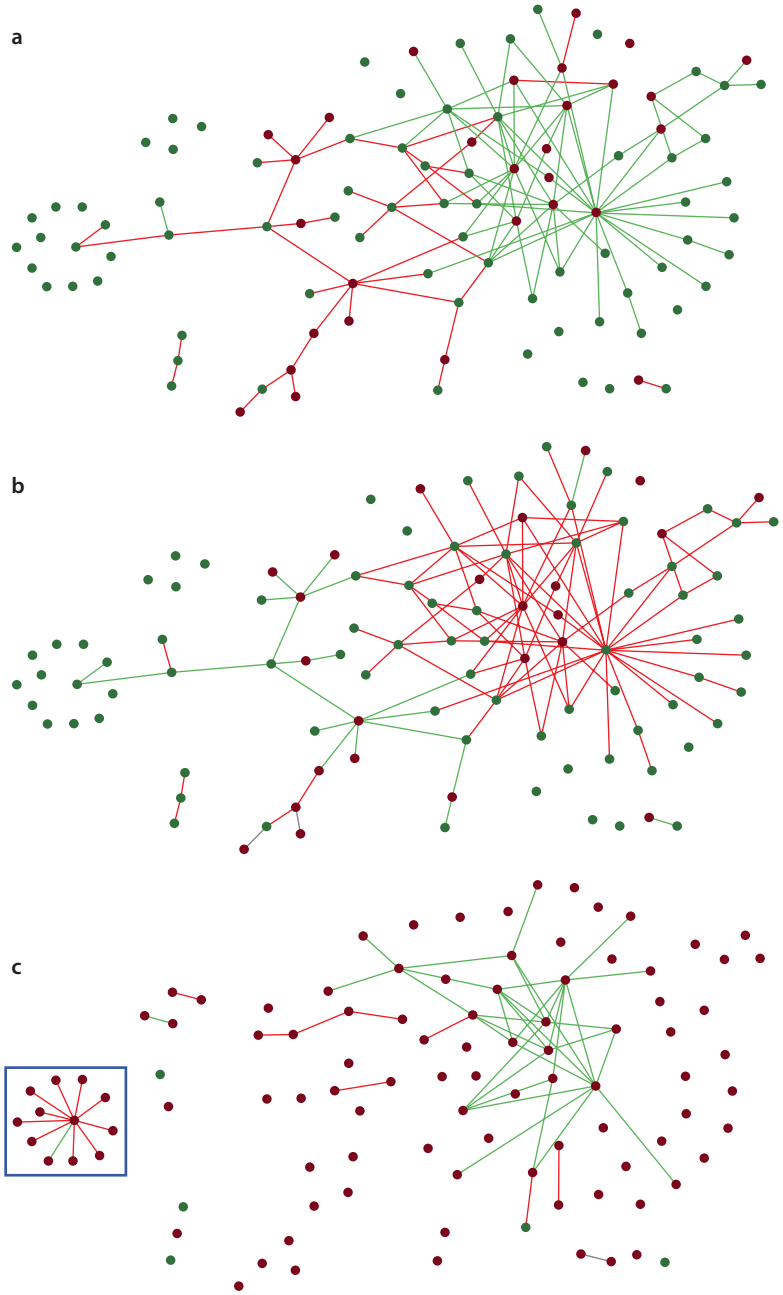

Fig. 6.8 Effects of pharmaceuticals on molecular networks. **a** Molecular network composed of proteins and lipids in healthy patients. Most connections are labeled in green, indicating a negative correlation between analytes. **b** Molecular network in diseased groups. The majority of correlations are labeled in red representing a change from a healthy to a diseased condition. **c** Molecular network in drug-treated patients. Many of the green links seen in healthy patients have been restored. However, in a second pathway, new network links appear (blue box). This is due to the off-target effects of the drug treatment (Printed with permission from BG Medicine Inc. USA)

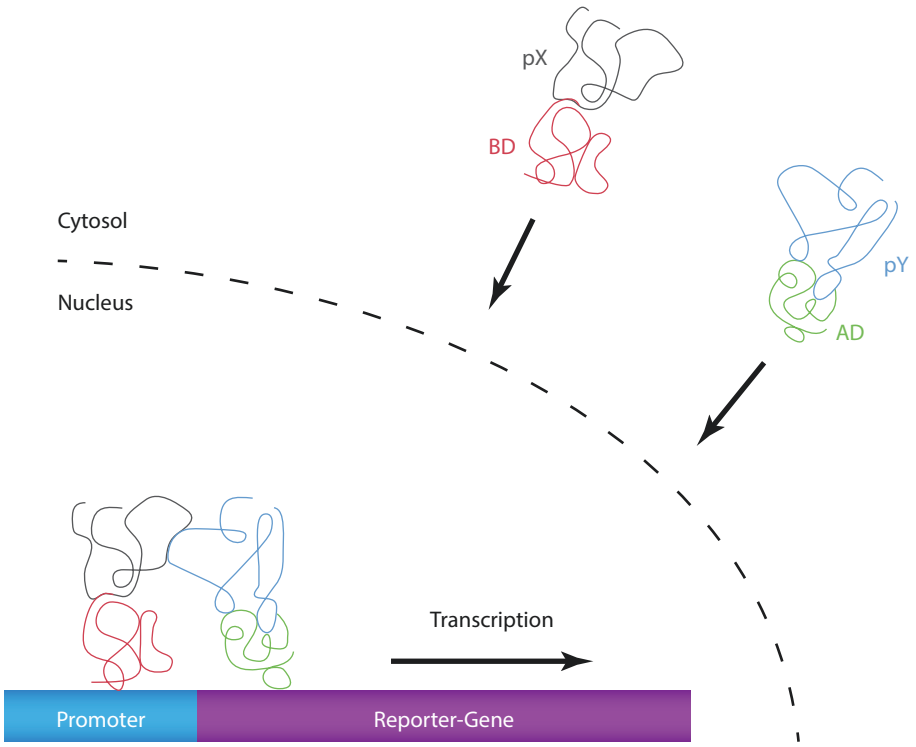

Fig. 6.9 Identification of protein–protein interactions using yeast two-hybrid system. Transcription of a reporter gene can only be activated when a fusion protein composed of the DNA binding domain of a transcription factor (BD) and a random protein X (pX) interact with a second fusion protein containing the cognate transcription factor's activating domain (AD) and a random protein Y (pY)

of the labeled protein from the protein lysate. The procedure is gentle and simultaneously copurifies those interacting cellular proteins that were bound to the labeled protein. The isolated multiprotein complex is then separated by gel electrophoresis, and the individual components are analyzed by mass spectroscopy. In this way, 232 different multiprotein complexes could be identified in the yeast *Saccharomyces cerevisiae*. Some of the multiprotein complexes consist of over 40 individual components. Furthermore, a potential function could be assigned to some unknown proteins based on their interaction with proteins with well-characterized known cellular functions (Gavin et al. 2002).

As is the case for every high-throughput experiment, the large quantity of data generated by interactomics requires the development of special databases. Example interactome databases are the IntAct Molecular Interaction Database [intact] and STRING [string]. To ensure that all relevant data of an experiment are included in the databases, the minimal information required for reporting a molecular interaction experiment (MIMIx) protocol regulates the minimal requirements for the storage of protein–protein interaction data (Orchard et al. 2007).

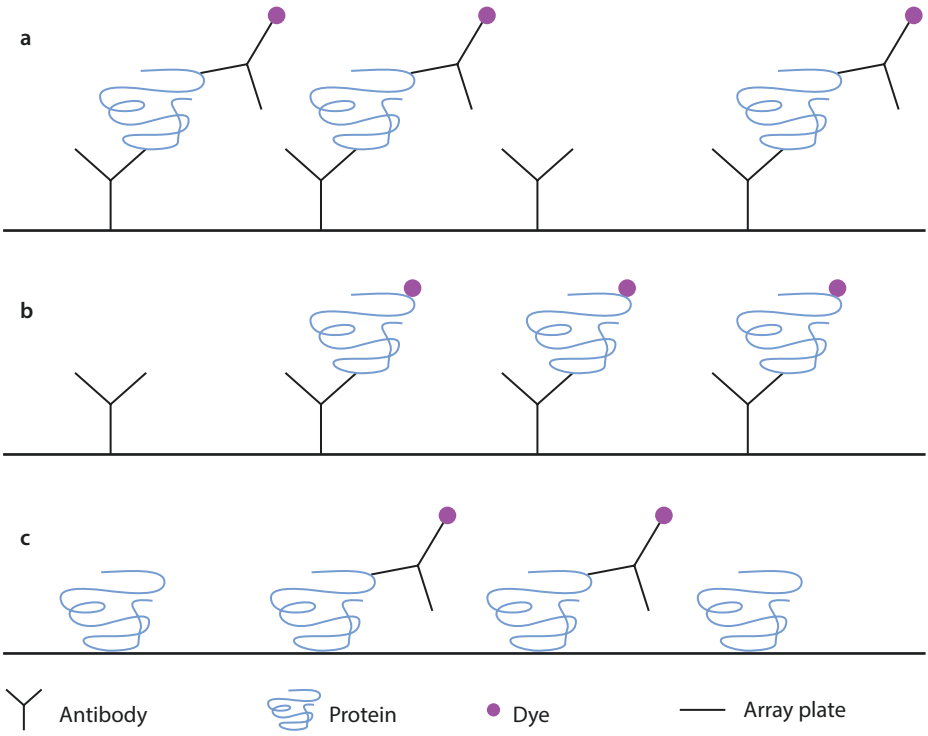

Fig. 6.10 Protein arrays. **a** In a sandwich assay, antibodies that are attached to array plates selectively bind to antigenic proteins after incubation with a protein lysate. Detection of the captured protein is performed with a secondary antibody that binds to a different antigenic site on the protein. **b** In an antigen capture assay, the antigenic proteins are labeled directly prior to incubation with the protein array, thereby dispensing with the need for a labeled secondary antibody for detection. **c** In direct or reverse-phase arrays, proteins are directly coupled to array plates and detected with labeled antibodies

6.1.2.3 Protein Arrays

An alternative method for the analysis of proteomes is based on protein array technology (Eisenstein 2006). Protein arrays are built similarly to DNA microarrays. Spots are applied at high density to a coated glass plate or membrane. These spots consist of reagents that have a high protein-binding affinity (e.g., antibodies). Protein arrays are also suitable for the generation of a protein profile, whereby three different variants of protein arrays are distinguished (MacBeath 2002):

- One variant is the sandwich assay (Fig. 6.10a), whereby antibodies are directly coupled to the protein arrays. The arrays are then incubated with a protein lysate. If a protein is present in the lysate for which an antibody has been spotted onto the array, then the protein will bind to that antibody. The detection of this binding is carried out with a secondary antibody that is directed against the same protein but to a different epitope than the primary antibody. The secondary antibody is labeled (e.g., with an enzyme that catalyzes a visually detectable reaction) to allow for the detection and quantitation of the binding.

- The second variant is the antigen capture assay (Fig. 6.10b). As above, the primary antibodies are directly bound to the matrix. It differs from the sandwich assay in that proteins in the lysate are already labeled (e.g., with fluorescent dyes). With this assay two cell lysates can be compared by labeling the proteins of the respective lysates with different dyes. Both lysates are mixed and incubated with the protein array. Depending on the amount of labeled bound protein, one or other lysate will contain more labeled protein. The basic concept of this procedure is analogous to that of an expression profiling experiment.
- In the third variant, the direct or reverse-phase assay, proteins and not antibodies are coupled to the protein arrays followed by the use of labeled antibodies. This way, proteins that interact with the antibodies are identified (Fig. 6.10c).

Protein arrays can also identify protein–protein interactions, as described in the previous section. Unlike the yeast two-hybrid system and TAP, however, it is an in vitro method. Protein interactions are analyzed outside the cell and under in vitro conditions – conditions that may lead to interactions that do not occur in vivo. On the other hand, protein arrays have the advantage that they can be produced in large quantities, allowing for multiple repetitions of experiments and modification of the conditions (pH, temperature, protein concentration, availability of ions, and cofactors). Moreover, with such arrays, thousands of proteins and even whole proteomes can be analyzed at the same time. For example, proteins from baker's yeast *Saccharomyces cerevisiae* were sought that could interact with the calcium-binding protein calmodulin. The array contained 5800 of the possible 6200 yeast proteins (Zhu et al. 2001). Thirty-nine proteins were identified as potential interacting partners, of which just six had already been described as calmodulin-binding proteins. The example highlights how protein arrays can define novel protein–protein interactions. Furthermore, protein arrays can also aid the detection of protein interactions with glycosides, lipids, nucleic acids, or other general ligands.

6.1.3 Metabolomics

Comparison of tumor cells with normal cells shows how striking it is that in the former, metabolic enzymes are frequently overexpressed. This should not be surprising, however, because cancer cells grow faster and thus have a greater need for metabolites. The notion is, therefore, that by quantifying cellular metabolites cells can be profiled in a similar way as with either microarray or proteomic techniques. The total metabolite pool of a cell is called the metabolome, and the research field dealing with metabolic profiling is termed metabolomics (Fig. 6.1). Metabolomics is a relatively new research area, although in 1970 Robinson and Pauling had already described experiments to identify and quantify the metabolites in human urine. The Human Metabolite Database [hmdb] contains all metabolites that can be found in the human body or at least should presumably occur. The latter is based on known metabolic pathways, but the final evidence is still pending. The database contains over 42,000 entries about metabolites that are linked with over 5600 protein sequences (Wishart et al. 2013). These entries comprise peptides, lipids, amino acids, nucleotides, carbohydrates, organic acids, vitamins, minerals, food additives, pharmaceutical agents, toxins, pollutants, and any other chemical substances with a molecular

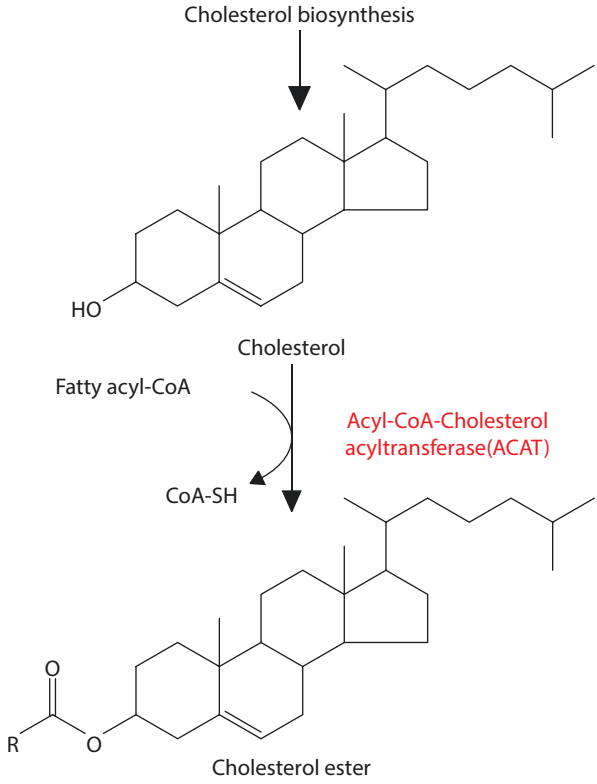

Fig. 6.11 Cholesterol ester biosynthesis catalyzed by Acyl-CoA-cholesterol acyltransferase

mass less than 2000 Dalton (Da). This list illustrates why the definition of the metabolome is so difficult compared with the genome, transcriptome, or proteome since depends not only on the genome but also the substance uptake from the environment (e.g., via food or pollution). Thus, the database does contains both endogen and exogen metabolites.

Despite the fact that the metabolome is rather small compared to the genome, transcriptome, or proteome, the technical demands required for metabolomics are particularly high. The reason for this lies in the extreme diversity of the various physical and physicochemical properties of the metabolites to be measured. Some metabolites are relatively small and hydrophilic (e.g., vitamin C), while others have a much higher mass and are nonpolar (e.g., cholesterol esters) (Fig. 6.11). At present, no single technology exists to identify and quantify all metabolites simultaneously. However, technological progress over the last few years has resulted in methods that can measure a small number of metabolites in parallel. Usually, the relative quantities of the metabolites in two different samples are compared to each other, similarly to the approach with DNA microarrays. In addition, more sensitive equipment and adequate standards also allow for the absolute quantification of metabolites in a single assay.

Two methods are primarily employed to measure metabolites: nuclear magnetic resonance (NMR) and mass spectroscopy. Sensitive NMR spectroscopy can generate physical, chemical, electronic, and, especially, structural data from molecules and metabolites. The method more frequently employed, however, is mass spectroscopy. Usually a chro-

matographic step (e.g., gas chromatography (GC)) to separate metabolites is performed first. Then, with the help of highly specialized equipment, more than 4000 raw data peaks can be measured, corresponding to approximately 1800 metabolite peaks (Kell 2006). Metabolomic experiments generate huge amounts of data, which must be analyzed and converted to biologically useful knowledge.

Many researchers have expressed the opinion that metabolomics describes the functions within a cell better than genomics, transcriptomics, or proteomics. They justify their opinion by the cell processes, pointing out that genes encode transcripts; transcripts in turn encode proteins, and these are eventually responsible for the production of metabolites. Therefore, metabolites are at the end of the information chain and, thus, closely connected to their function. A further argument is the amplification of information. It has been experimentally determined that even small changes in the concentration of a few enzymes can lead to significant changes in the concentration of many metabolites (Raamsdonk et al. 2001). The reasons for this are that the synthesis and turnover of metabolites in general are catalyzed by several enzymes, and one metabolite can be involved in many different reactions. In this connection one may also speak of metabolic networks (◘ Figs. 7.3 and 7.4).

The strength of metabolomics is that it confers the possibility to construct models of quantitative changes in the metabolome due to its networked structure. Indeed, many models have already been devised, particularly for well-studied organisms such as baker's yeast, *Saccharomyces cerevisiae*. For example, with the help of a metabolic model that represents 750 genes and 1149 reactions in baker's yeast, 4154 growth phenotypes were predicted. A comparison with experimental results showed that the model had, in fact, correctly predicted 83% of the phenotypes (Duarte et al. 2004). The generation of such metabolic models overlaps to some extent with another area, namely systems biology, which is described in more detail in ▶ Sect. 6.2.

Electronic noses, which are already available as portable devices, are another application of metabolite analysis (Koczulla et al. 2011). Nanocomposite sensors are built into electronic noses for the detection of small amounts of molecular gases, acids, bases, and many other molecules. Patterns for different compositions can be retrieved and analyzed by computational methods using combinations of different sensors. Cyranose 320, for example, is an electronic nose from Sensigent [sensigent] that can be used for the analysis of humans breathing air. In a study with 30 patients, it was possible to distinguish between the breathing air of patients with non-small-cell lung cancer, chronic obstructive pulmonary disease (COPD), and healthy patients (Dragonieri et al. 2009). In another study three different bacterial strains, including methicillin-resistant *Staphylococcus aureus* (MRSA) and methicillin-susceptible *S. aureus* (MSSA) strains, were detected and differentiated (Dutta and Dutta 2006).

6.1.4 Phenomics

The phenotype or physical appearance is the sum of all extrinsic visible features of an individual (◘ Fig. 6.12). It refers to both morphological and physiological properties. Consequently, the visible and measurable properties of an organism or cell that are based on interactions of the genotype with the environment constitute the phenotype

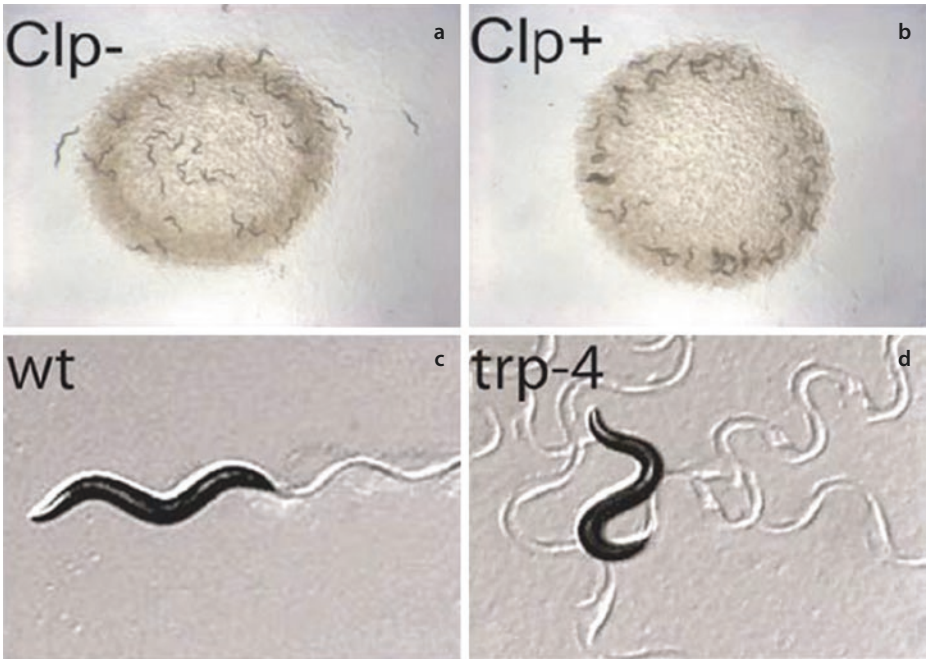

Fig. 6.12 Phenotypes in the roundworm *Caenorhabditis elegans*. **a** Most strains are solitary feeders and do not show a clumping phenotype (Clp−). **b** Some strains aggregate on the border, recognized as the clumping phenotype (Clp+). The phenotype is caused by a naturally occurring genetic polymorphism in a single gene. **c** Phenotype of a moving wild-type worm (wt). **d** Phenotype of a trp-4 knockout worm with an abnormal body posture. The ion channel mutants have a greater frequency of body movement with more pronounced flexing (**a**, **b** Printed with permission from Marie-Anne Félix, Institut Jacques Monod, France. **c**, **d** Printed with permission from X. Z. Shawn Xu, University of Michigan Medical School, USA)

(Sect. 6.1.2). By this definition, therefore, metabolomics, being measurable, are also a representation of a phenotype that is based on interactions of the genotype with the environment. Many methods exist in the context of functional genomics that define protein function based on phenotypes. This research area is also called phenomics if it is carried out in a high-throughput format.

Initially, forward genetic screens were used in which genomes were randomly mutated, the resulting phenotypes recorded, and the genes responsible for the modified phenotype identified. Using this approach several thousand genes were identified and characterized. The arrival of sequenced whole genomes offered alternative approaches to performing genetic screens for those genes without an ascribed function. The strategy that links a distinct gene with its function is called reverse genetics.

As subsequent analysis, knockout experiments are often carried out whereby genes are selectively mutated ("switched off") so that no functional protein is encoded. The consequence can be an altered phenotype whose properties can then be accurately documented. If a gene encodes an essential protein, the resulting phenotype may be lethal, i.e., the cell or organism dies. Such knockout experiments are usually performed mainly in cell lines or in

model organisms such as the fruit fly *Drosophila melanogaster* [genedisruptionproject]. The disadvantage of this method is the complicated and time-consuming experimental approaches required, reflected by the fact that complete and comprehensive genome-wide knockout data are available for only a few organisms (e.g., baker's yeast).

Analogous to knockouts are "knockin" experiments to elucidate the function of gene products. In this case, genes are transfected into cells or organisms and then observed to determine whether they cause phenotypic changes. The knockin strategy is frequently used as additional proof of a protein's function. If the phenotypic change of a prior knockout can be reversed by a knockin experiment, then there is little doubt as to the protein's function. For example, a bacterium in which a specific flagellar protein has been knocked out is rendered immotile. If the same bacterial clone has the gene restored in a knockin experiment and subsequently recovers motility, then this is solid evidence that the protein is essential for proper flagellum function.

Unfortunately, knockout and knockin strategies are laborious and not amenable to high throughput. The discovery and experimental application of RNA interference (RNAi) has resulted in a revolution for reverse genetic screening. RNAi is an evolutionarily conserved mechanism that involves the repression of gene expression by double-stranded RNA (dsRNA) (Vanhecke and Janitz 2005). After gaining access to the cytoplasm of the cell, dsRNA molecules are first cut into lengths of 21–25 nucleotides, termed small interfering RNAs (siRNAs), by the enzyme dicer (Fig. 6.13). The single-stranded siRNA is then loaded into the enzyme complex called the RNA-induced silencing complex (RISC). The activated enzyme complex, guided by the siRNA strand, binds specifically to the complementary mRNA, which is cut by the endonuclease activity of the RISC. In this way, the expression of the target gene is specifically blocked, preventing translation of the cognate protein. Because transcription blockade by RNAi may not always be complete, the term *gene knockdown* applies.

An advantage of RNAi technology is its efficiency. Experiments are fast, simple, cost-efficient, and, importantly, amenable to high-throughput formats. Numerous publications have analyzed complete genomes using RNAi. For example, 86% of all genes of the nematode *Caenorhabditis elegans* were examined by means of RNAi (Kamath et al. 2003). Approximately 10% of the targeted genes led to a change in phenotype, of which approximately one-third were already known. In another study, new modulators of p53 that causes cell cycle arrest in human cells were searched for by RNAi. Of the 8000 genes analyzed, five new modulators were discovered (Berns et al. 2004).

Unfortunately, not all RNAi results are absolutely reliable. For instance, it is known that the efficiency of RNAi is massively dependent on the incorporated nucleotide sequence. In some cases, the target mRNA is either only partly degraded or not at all, leading to a false negative result. The experimenter will see no change in the phenotype and infer that the gene product has no important function. Importantly, such data should be checked by an independent method, such as RT-PCR, to determine whether the target RNA has in fact been degraded. Conversely, RNAi can also generate false positive results. In this case, the siRNAs produced by the nuclease dicer can hybridize with more than one target mRNA, which in turn leads to the degradation of several mRNAs. Therefore, changes in phenotypes cannot be assigned unambiguously and, in the worst case, might lead to incorrect functional predictions of gene products.

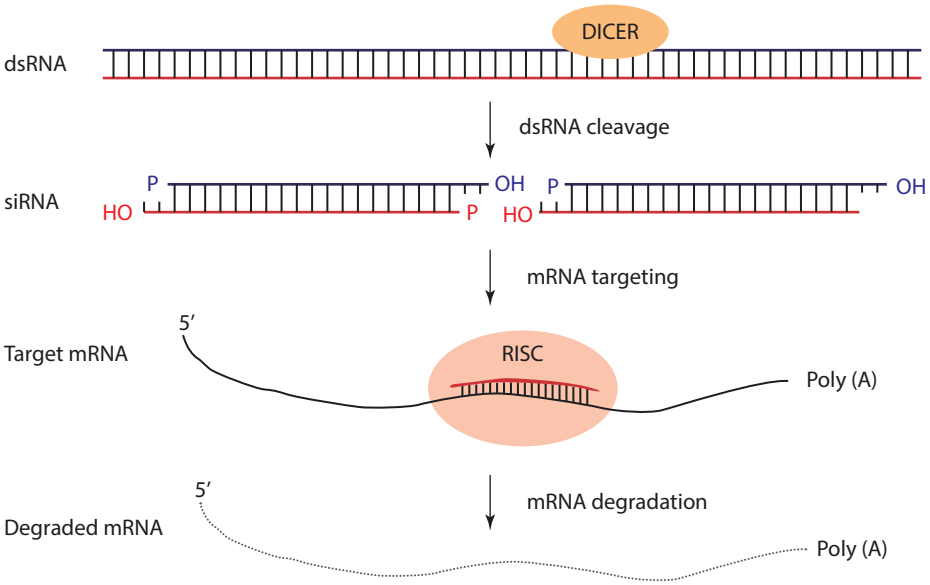

Fig. 6.13 Specific degradation of mRNA by RNA interference (RNAi). A type III ribonuclease (dicer) binds to and cleaves double-stranded RNA (dsRNA) into 21–25 base pair duplexes, termed small interfering RNA (siRNA). The siRNA is incorporated into the multiprotein complex called RNA-induced silencing complex (RISC), which also contains an RNAse. RISC unwinds the siRNA, releases the sense strand, and facilitates hybridization of the antisense strand of the siRNA to the complementary strand of the cognate messenger RNA (mRNA). The binding activates the nuclease activity in the RISC, leading to cleavage of the target mRNA. The damaged mRNA is then degraded significantly, reducing the expression of the target gene

The PhenomicDB is a very interesting and integrated database in which phenotypes from different organisms that were generated by various methods (e.g., knockout, knockin, knockdown) are integrated into one database and linked to genotypic data. Furthermore, databases like the Human Genome Variation Database [hgvdb] store human genotype–phenotype relations (Brookes and Robinson 2015).

6.2 Systems Biology

The foregoing discussion on high-throughput procedures has established genomics, transcriptomics, proteomics, metabolomics, and phenomics as important technologies that facilitate the functional determination of gene products. Like all high-throughput experiments, however, these approaches produce false negative and false positive results. False negative results can lead to information being missed, whereas false positive results might lead the experimenter in the wrong direction. Therefore, to identify valid results, an idea was put forth regarding the integration of all available data from the aforementioned technologies and analyze them together (Fig. 6.14). This integration of experimental data improves the reliability and the generation of more reliable hypotheses. The research field that focuses on the integration of various high-throughput data is known as systems biology because it analyzes entire biological systems. Systems biol-

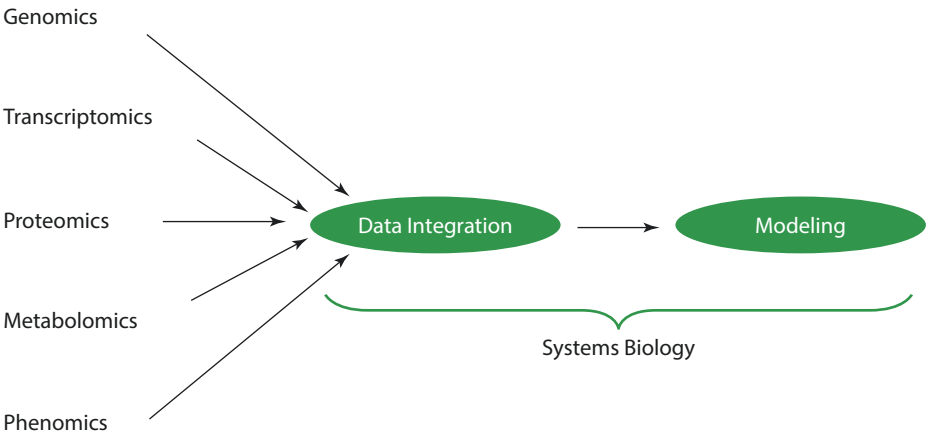

Fig. 6.14 Systems biology integrates data derived from different experimental technologies and generates computational models

ogy aims to produce as accurate a picture as possible of all the regulatory processes within a cell or organism by analyzing the interactions between the component parts of the biological system, e.g., metabolic pathways, organelles, cells, and tissues.

An example of a systems biology approach is the analysis of phagosomes, which are special organelles found in phagocytosing cells (e.g., macrophages). After phagocytosis, particles such as bacteria are transported into phagosomes, where they are destroyed. In a study by Stuart et al. (2007), the phagosome of a cell line derived from the fruit fly *Drosophila melanogaster* was analyzed. Proteins of the phagosome were identified by classical proteomics methods. Construction of a protein–protein interaction networks complemented the results, which were finally validated by RNAi experiments. With the help of this systems biology approach a detailed model of the phagosome was built and new regulatory proteins and pathways associated with phagocytosis identified.

However, systems biology frequently goes beyond the mere description and interpretation of experimental data. The ambitious aim is to develop computer models that simulate biological systems and predict consequences upon changing parameters (e.g., changing the concentration of a specific metabolite). One of the first mathematical models in biology was published in 1952 by Alan Hodgkin and Andrew Huxley, which explained the transmission of action potentials. Since then, the increasing availability of high-quality data (both quantitative and qualitative) and greater computer capacities have allowed for more realistic models to be developed. For example, a model has been generated to simulate glycolysis in baker's yeast, *Saccharomyces cerevisiae*. Compared with experimental data, most metabolite concentrations were correctly predicted within a maximal deviation of two (Teusink et al. 2000).

Even more demanding are computer models that perform complete cell simulations (Ishii et al. 2004). A well-known model is the E-cell system that developed a "virtual bacterium" consisting of 127 essential genes from the genome of *Mycoplasma genitalium* (Fig. 6.15). This bacterium has fewer than 500 genes and is therefore excellently suited

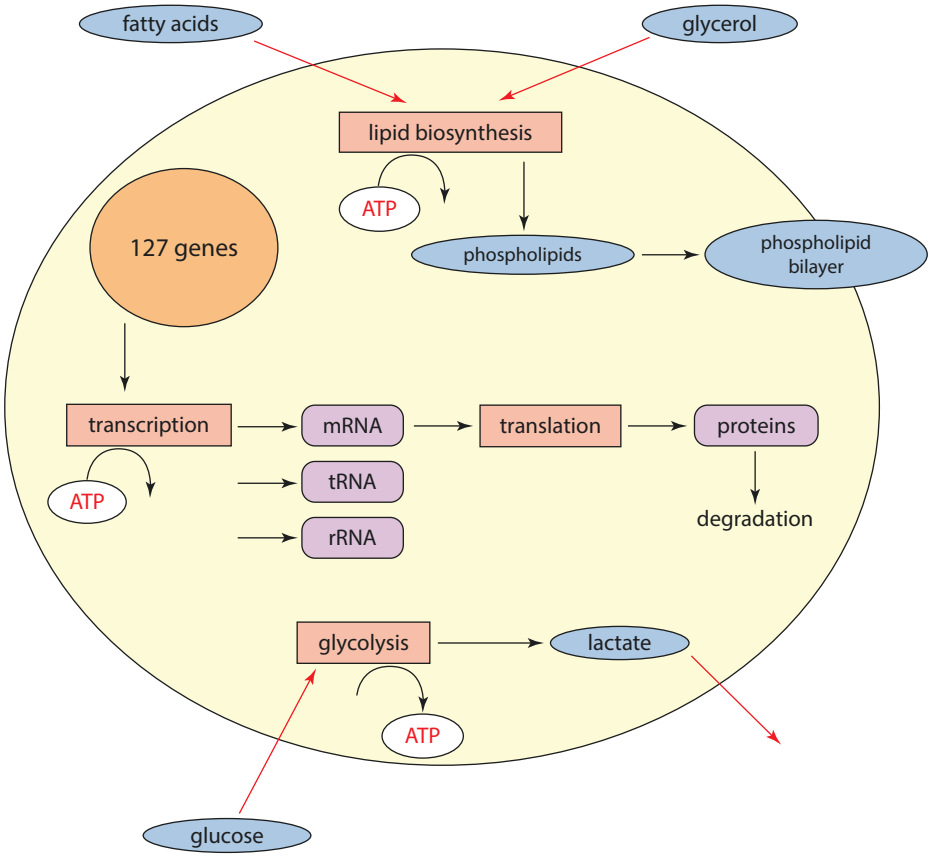

Fig. 6.15 Overview of metabolism in E-cell model. The model cell has pathways for glycolysis and phospholipid biosynthesis, transcription, and translation

for the construction of a cell model. With the model, the transport of extracellular glucose through the cell membrane was simulated, in addition to the metabolism of the sugar and the accompanying ATP production. The model also produced a surprise. When the concentration of extracellular glucose was set to zero, the model predicted a temporary increase in the intracellular ATP concentration before a final drop. This was contrary to the expectation that the ATP concentration would drop immediately upon depletion of the glucose. After much speculation, the conclusion was that the model's prediction was correct. During glycolysis, two molecules of ATP are generated from each molecule of glucose. Following more detailed examination, it became apparent that in the first part of glycolysis, two molecules of ATP are spent before the production of four ATP molecules in the second part of the reaction. At the moment when the glucose concentration is lowered to zero, the consumption of ATP molecules stops before the generation of new ATP molecules, which are then consumed. Thus, the model recognized the short temporal shift and correctly predicted the temporary increase in ATP concentration.

In 2012, a computational model of the human pathogen *Mycoplasma genitalium* was presented that simulates the whole cell including all molecular components and their interactions (Karr et al. 2012). The model aimed at describing a complete cell cycle of a single cell and predicting observable cellular behavior. It was based on a complete genome with 525 genes and a detailed analysis of over 900 data sources, including primary sources, books, and a database. Overall, data sets of 192 wild-type and 3011 knockout cells were calculated on a cluster with 128 nodes. The calculations were finally validated by experimental data that were not part of the model development. Deep insights into previously unseen and unobservable cell processes were made possible by the use of this model, like in vivo rates of protein–DNA interactions.

The emergence of systems biology has been accompanied by the development of a dedicated exchange format for the representation of biological models, the Systems Biology Markup Language (SBML). SBML is an XML-based computer-readable format in which biological networks are exactly described. The central idea of SBML was the creation of a standardized format that permits the simple exchange of data between many different software applications. Therefore, each calculated model can be tested in different software environments without additional effort. At the same time, specialized databases have been established in which computer models can be stored and are accessible to all interested scientists. One example database is the BioModels Database at EBI [biomodels].

6.3 Exercises

Exercise 6.1

In the GEO Datasets database, find the entry GDS1399. GDS1399 is a microarray experiment that examines the effect of distinct gene mutations on global gene expression in *Escherichia coli*. The DAM mutant lacks the enzyme DNA adenine methyltransferase (DAM). This enzyme transfers methyl groups to sites with a characteristic short sequence in the *E. coli* genome and thereby exerts a significant influence on the regulation of gene expression. An *E. coli* strain without any genetic alteration is referred to as the wild type.

1. How many replicates of the wild type and DAM mutants were used in the experiment?
2. Determine the number of genes whose expression in the DAM mutant is increased or decreased. Use the option *compare 2 sets of samples* and analyze a twofold or more expression compared to the wild-type strain.
3. For how many genes is the expression in the DAM mutant significantly different from that in the wild type? Use the *two-tailed t-test (A* vs. *B)* similar to ► Exercise 6.1-2 with a 0.050 level of significance.

Exercise 6.2

Visit the web site of the Princeton University MicroArray database (PUMAdb, ► https://puma.princeton.edu/index.shtml). This database stores primary data, normalized data, and pictures of microarray experiments. You will find under *Help* a

number of descriptions and tutorials. Especially the section about data normalization provides a good overview on necessary data analysis. A *World Session* must be activated before access to the public data will be granted. Using *Standard Search* to search for a publication of van Brummeln et al. (2009) based on the organism *Plasmodium falciparum*. Deal with the available data.

Exercise 6.3

Go to the web page of the BROAD Institute and test the software GenePattern (► http://software.broadinstitute.org/cancer/software/genepattern/). A tutorial is provided that should allow a first analysis together with a look at the results within 10 min. Test the program using these data sets.

Exercise 6.4

Go to the Expasy home page and look for the software Swiss2DPage [swiss2dpage]. Use the function *Search by description, ID or gene* to search for the entry HSP60 (Heat Shock Protein 60). Select *CH60-HUMAN*.

1. Open the 2D–PAGE of the entry "HEPG2_HUMAN." The spots corresponding to HSP60 are marked in red. How many spots are found for HSP60? How can it be explained that several spots exist for one protein?
2. Next, click on the picture of the 2D electrophoresis from liver (LIVER_HUMAN). How many spots correspond to HSP60 in this case? Why are fewer spots found now?
3. Look at the protein lists for HEPG2_HUMAN and HEPG2SP_HUMAN (secreted proteins). Use the search *protein list*. Can HSP60 be found on this gel? Give reasons for the result.
4. What methods were used to identify the proteins in both protein lists from task 3?
5. Search in the protein list of HEPG2_HUMAN from task c for the unknown protein that represents Spot 106. Click on its SWISS accession number P31929. Look for the section *cross-references* and follow the link *UniProtKB/Swiss-Prot*. What partial amino acid sequence of the protein was identified by microsequencing? Unfortunately, the entry was marked in UniProt as obsolete. You can find the sequence using the button *history*.
6. Follow the link *graphical interface* on the start page for an overview about all gels together with identified proteins. Select *2D–Page of nucleolar proteins from Human HeLa Cells*. Click on the highlighted spot with the lowest molecular weight and a pI value of approx. 5.7. Which protein is it? What is the molecular weight of the protein? Follow also the link at the entry *External Data extracted from UniProtKB/Swiss-Prot*. What synonyms are there for this protein?

Exercise 6.5

Go to the protein–protein interaction database STRING (► http://string-db.org/). Enter `Thioredoxin reductase` in the search field *Protein Name* and

`Mycobacterium tuberculosis` in the search field *Organism*. Select *TrxB2* in the results list, and click *continue*. The network of feasible TrxB2 interactions will be shown as a search result. You can analyze the possible interactions by selecting *TrxC* and then *re-center network on this node* under the option *action*. Which other proteins show a direct molecular interaction with the highest confidence? Deal with View Settings and Data Settings and analyze particular interactions in detail.

Exercise 6.6

Go to the home page of the program PeptideMass (► http://www.expasy.org/tools/peptide-mass.html). Run an in silico digest of the human protein kinase src (accession number P12931) with the enzyme trypsin. How many peptides with a mass of >1000 Da are generated by this digest? What is the peptide mass of the largest peptide?

Exercise 6.7

Go to the Human Metabolome Database (► http://www.hmdb.ca/). Find out, which food small molecule is responsible for the occurrence of 1-methylxanthine in the human body. To which Origin is 1-methylxanthine counted? Analyze the whole metabolic process 1-methylxanthine arises from. What molecule is the precursor?

References

Allis CD, Jenuwein T (2016) The molecular hallmarks of epigenetic control. Nat Rev Genet 17:487–500

Berns K, Hijmans EM, Mullenders J et al (2004) A large-scale RNAi screen in human cells identifies new components of the p53 pathway. Nature 428(6981):431–437

Brazma A, Hingamp P, Quackenbush J et al (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29(4):365–371

Brookes AJ, Robinson PN (2015) Human genotype-phenotype databases: aims, challenges and opportunities. Nat Rev Genet 16(12):702–715

Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. Nat Genet 32: 490–495

Duarte NC, Herrgard MJ, Palsson BO (2004) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Res 14(7):1298–1309

Dutta R, Dutta R (2006) Intelligent Bayes Classifier (IBC) for ENT infection classification in hospital environment. Biomed Eng Online 5:65

Dragonieri S, Annema JT, Schot R et al (2009) An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD. Lung Cancer 64(2):166–170

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

Eisenstein M (2006) Protein arrays: growing pains. Nature 444(7121):959–962

Gavin AC, Bosche M, Krause R et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868):141–147

Gershon D (2005) DNA microarrays: more than gene expression. Nature 437(7062):1195–1198

Golub TR, Slonim DK, Tamayo P et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537

Griffin TJ, Goodlett DR, Aebersold R (2001) Advances in proteome analysis by mass spectrometry. Current Opin in. Biotech 12:607–612

Holloway AJ, van Laar RK, Tothill RW, Bowtell DL (2002) Options available from start to finish-for obtaining data from DNA microarrays II. Nat Genet 32:481–489
Ishii N, Robert M, Nakayama Y et al (2004) Toward large-scale modeling of the microbial cell for computer simulation. J Biotechnol 113(1–3):281–294
Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast interactome. Proc Natl Acad Sci U S A 98:4569–4574
Ji H, Davis RW (2006) Data quality in genomics and microarrays. Nat Biotechnol 24(9):1112–1113
Kamath RS, Fraser AG, Dong Y et al (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421(6920):231–237
Karr JR, Sanghvi JC, Macklin DN, Gutschow MV et al (2012) A whole-cell computational model predicts phenotype from genotype. Cell 150(2):389–401
Kell DB (2006) Systems biology, metabolic modelling and metabolomics in drug discovery and development. Drug Discov Today 11(23–24):1085–1092
Koczulla AR, Hattesohl A, Biller H et al (2011) Smelling diseases? A short review on electronic noses. Pneumologie 65(7):401–405
Matsumura H, Bin Nasir KH, Yoshida K et al (2006) SuperSAGE array: the direct use of 26-base-pair transcript tags in oligonucleotide arrays. Nat Methods 3(6):469–474
MacBeath G (2002) Protein microarrays and proteomics. Nat Genet 32:526–532
Orchard S, Salwinski L, Kerrien S et al (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat Biotechnol 25(8):894–898
Raamsdonk LM, Teusink B, Broadhurst D et al (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nat Biotechnol 19(1):45–50
Rual JF, Venkatesan K, Hao T et al (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437(7062):1173–1178
Quackenbush J (2001) Computational analysis of microarray data. Nature Rev. Genetics 2:418–427
Slonim DK (2002) From patterns to pathways: gene expression data analysis comes of age. Nat Genet 32:502–508
Stuart LM, Boulais J, Charriere GM (2007) A systems biology analysis of the Drosophila phagosome. Nature 445(7123):95–101
Teusink B, Passarge J, Reijenga CA et al (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. Eur J Biochem 267(17):5313–5329
Vanhecke D, Janitz M (2005) Functional genomics using high-throughput RNA interference. Drug Discov Today 10(3):205–212
Wishart DS, Jewison T, Guo AC et al (2013) HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res 41(Database issue):D801–D807
Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A et al (2001) Global analysis of protein activities using proteome chips. Science 293:2101–2105

Further Reading

agilent. http://www.genomics.agilent.com
affymetrix. http://www.affymetrix.com/
arrayexpress. http://www.ebi.ac.uk/arrayexpress/index.html
bioconductor. https://www.bioconductor.org/
biomodels. https://www.ebi.ac.uk/biomodels-main/
ecell. http://www.e-cell.org
ercc. http://jimb.stanford.edu/ercc/
geo. https://www.ncbi.nlm.nih.gov/geo/
genedisruptionproject. http://www.fruitfly.org/p_disrupt/index.html
genepattern. http://software.broadinstitute.org/cancer/software/genepattern/
hgvdb. http://www.hgvd.genome.med.kyoto-u.ac.jp/
hmdb. http://www.hmdb.ca/
hpp. http://www.thehpp.org/
intact. http://www.ebi.ac.uk/intact/
maqc. http://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/default.htm

melanie. http://world-2dpage.expasy.org/melanie/
miame. http://fged.org/projects/miame/
sage. http://www.sagenet.org/
sagemap. https://www.ncbi.nlm.nih.gov/projects/SAGE/
sensigent. http://www.sensigent.com/products/cyranose.html
string. http://string-db.org/
swiss2dpage. http://world-2dpage.expasy.org/swiss-2dpage/
tm4. http://www.tm4.org/