Platforms for biomarker analysis using high-throughput approaches in genomics, transcriptomics, proteomics, metabolomics, and bioinformatics

B. Alex Merrick, Robert E. London, Pierre R. Bushel, Sherry F. Grissom, and Richard S. Paules

Summary

Global biological responses that reflect disease or exposure biology are kinetic and highly dynamic phenomena. While high-throughput DNA sequencing continues to drive genomics, the possibility of more broadly measuring changes in gene expression has been a recent development manifested by a diversity of technical platforms. Such technologies measure transcripts, proteins and small biological molecules, or metabolites, and respectively define the fields of transcriptomics, proteomics and metabolomics that can be performed at a cell-, tissue-, or organism-wide basis. Bioinformatics is the discipline that derives knowledge from the large quantity and diversity of biological, genetic, genomic and gene expression data by integrating computer science, mathematics, statistics and graphic arts. Gene, protein and metabolite expression profiles can be thought of as snapshots of the current, poorly-mapped molecular landscape. The ultimate aim of genomic platforms is to fully map this landscape to more completely describe all of the biological interactions within a living system, during disease and toxicity, and define the behaviour and relationships of all the components of a biological system. The development of databases and knowledge bases will support the integration of data from multiple domains, as well as computational modelling. This chapter will describe the technical platform...
methods involving DNA sequencing, mass spectrometry, nuclear magnetic resonance combined with separation systems, and bioinformatics to derive genomic and gene expression data and include the relevant bioinformatic tools for analysis. These genomic, or omics platforms should have wide application to epidemiological studies.

Introduction

The sequencing of the human genome stands as one of the major scientific achievements of the twentieth century. It embodies a defining moment in modern biology by which most high-throughput technologies are compared for size scope, and complexity. Beginning in 1990, it took roughly a decade for the first draft of the human genome to be completed. By 2003, about 99% of all gene-containing regions were described, numbering about 20,500 genes (1), although some regions of the genome, such as centromeres, telomeres and gene deserts, continue to undergo characterization and study. Data from the human genome project has provided a generalized human map of the three billion nucleotides comprising the DNA of a few human subjects. However, studies on the variations (polymorphisms) in human DNA sequences are currently underway; samples from 270 individuals of multiethnic backgrounds are being used in a consortium called the International HapMap Project for haplotype mapping (http://www.hapmap.org) (2). The goal is to identify the patterns of single nucleotide polymorphism (SNP) groups, called haplotypes or haps, among individual human beings. In addition, interpretation of the human genome has been greatly enhanced by the DNA sequencing of many other genomes that allow comparison of genetic organization, evolution and function. Nearly 300 genomes have been completely sequenced and range from unicellular organisms, like E. coli and S. cervisiae, to model invertebrate organisms, such as Drosophila melanogaster and C. elegans, to several mammalian species for which the completed and ongoing genome projects are all available online (3).

Although the conception of the idea for sequencing the human genome is relatively recent, the project could not have occurred without the preceding decades of biological and technological developments. Particularly noteworthy of these contributions are early cytogenetics and chromosomal studies at the beginning of the twentieth century by Morgan and colleagues, the discovery of DNA structure by Watson and Crick in 1953, DNA cloning in 1973 by Berg and Cohen, the DNA sequencing reaction in 1975 by Sanger, reverse transcriptase in 1970 and restriction endonucleases in 1971, and the polymerase chain reaction (PCR) by Mullis in 1983 (4). A new century now begins with an era of “omics,” those fields describing a multitude of genomic functions aimed at further deciphering the biological meaning of sequences in the human genome. The purpose of this chapter is to describe the technical platforms in genomics, transcriptomics, proteomics, metabolomics and bioinformatics that could be useful in epidemiologic studies. These analytical platforms favor high sample throughput and generation of large data sets.

Omes and omics

Gene expression constantly changes during health, adaptation, toxicity, disease and aging. While the genetic blueprint of an individual is relatively static, the various levels of gene expression to form and operate a complex organism are dynamically regulated, structurally complex and spatially determined. At any point in time, only a portion of a genome is expressed in specific cells and tissues. At the mRNA level of gene expression, the transcriptome represents all genes transcribed at any one moment, and the proteome is the complement of proteins making up cells and tissues. Small molecules and metabolites comprise the metabolome. The global study of each gene expression level is suffixed with “omics,” such as transcriptomics, proteomics and metabolomics. Figure 7.1 suggests a sequence of gene expression based on genomic DNA sequences that are dynamically reflected in changes of transcripts, proteins and metabolites. Each level of gene expression (represented by upward, curved, dotted lines) has the opportunity to feed back and influence other levels reflective of highly integrated, multicellular processes in cells, tissues and organisms. Studies of each gene expression area utilize very different technical platforms to maximize large scale coverage of the transcriptome, proteome and metabolome. Technical platforms may involve mass parallel analysis using robotics, miniaturization, automation and computer processing. The integration of the many levels of gene expression is often referred to as systems biology by bioinformatics or computational biology. Bioinformatics represents an applied field of mathematics to biochemistry and molecular biology using statistics, computer science and artificial intelligence to design algorithms to derive biological meaning from gene expression data.
Genomics

Chromosomal abnormalities are responsible for many developmental defects and malignancies, and include rearrangements in genomic DNA or changes in copy number, such as deletions, duplications and amplifications. Identification of genomic changes and mutations that underlie disease rely on comparisons of DNA sequences between affected and unaffected individuals. Finding disease-causing chromosomal abnormalities by genomic analysis is confounded by the fact that many sequence polymorphisms are functionally irrelevant and produce no observable biological consequence. Detection of many disease-causing mutations, such as those in p53 in the Li-Fraumeni syndrome (5) and ATM in ataxia telangiectasia (6), have been found by sectional resequencing of genomic DNA or PCR-amplified DNA or RNA. However, de novo sequencing of individuals using automated Sanger-based capillary electrophoresis systems has so far been practical for only small regions of the human genome-containing candidate genes.

Recent advances in nucleic acid sequencing technologies using massive parallel sequencing, called next-generation sequencing, now allow sequencing of much larger genomic intervals (7). Sequencing of entire genomes can take place within a matter of several weeks, in a comprehensive search for chromosomal aberrations and mutations that affect phenotype. DNA sequencing does have inherent advantages in achieving single-base resolution and importantly for de novo analysis of samples without the prior knowledge of existing DNA sequence required for fabricated sequence platforms (8). New sequencing technologies that are high-throughput and low-cost while maintaining high accuracy and completeness are in continued development (9). New platforms often integrate real-time (RT) PCR and may incorporate microelectrophoresis, sequencing by hybridization, mass spectrometry, high-density oligonucleotide arrays, or incorporation of nanopore technology to bring within sight the goal of routine human genome sequencing (10) for personalized medicine (11).

There are several alternatives to whole-genome assessment of chromosomal abnormalities that do not involve traditional DNA sequencing. Alternative platforms involve selective genotyping of very focused genomic loci (short sections of DNA) that are potentially related to disease susceptibility. Haplotype mapping data have been extremely useful in providing candidate genomic regions with high polymorphic variation. Hundreds of thousands of loci can be very rapidly genotyped using BeadArray platforms, a technology based upon direct hybridization of whole-genome-amplified (WGA) genomic DNA to BeadArrays of locus-specific, 50mer oligonucleotide sequences (12). As such, genome-wide association studies (GWAS) comprise an important evolving field in genetic epidemiology in which more than 450 GWAS have been published, and the associations of greater than 2000 single nucleotide polymorphisms (SNPs) or genetic
loci have been reported so far (13). Equally as important are high-density oligonucleotide microarrays for SNP detection in linkage analysis of susceptibility genes often used in cancer studies (14) or pharmacogenomics (15). In addition, fluorescence in situ hybridization (FISH) provides a visual map to examine all the chromosomes of a patient for abnormalities using fluorescent probes for specific genes or whole chromosome probes (16). Although high-throughput sequencing methodologies have been developed to accommodate the demand for sequence output, they consume large amounts of a valuable and potentially limiting genomic DNA. WGA can potentially remove DNA as a limiting factor for genomic analyses (17) that include multiple displacement amplification (MDA), primer extension preamplification (PEP), and degenerate oligonucleotide primed PCR (DOP) (18). However, genomic amplification technologies generate a certain level of replication error in sequence, which should be considered during verification studies. In summary, bead or chip DNA arrays or FISH platforms exemplify whole-genome technologies for high-throughput, compared to DNA sequencing for detection of genetic variation that can be linked to disease-based foci, protein, biomarkers and pharmacogenomic responses.

Transcriptomics

Transcriptomics studies the full, global complement of mRNA molecules expressed in cells and tissues. Some 20 500 genes are present within the human genome, of which about 10–15 000 are expressed at any one time in any particular tissue (19). Many expressed genes are necessary to perform basic functions of the cell, regardless of cell type or tissue, but a proportion of the expressed genes contribute to a cell’s unique phenotype and specialized functions. Beginning around 1989, DNA microarrays, consisting of thousands of high-density cDNAs or oligonucleotides on support surfaces (called chips), were introduced and have evolved into powerful and versatile platforms for transcriptomic analysis (20–23). Spotted microarrays, either cDNAs or oligonucleotides, have been used extensively since the late 1990s, particularly by academia-based research scientists (see Figure 7.2). Commercial oligonucleotide arrays provide highly reproducible platforms representing the entire genome. Oligonucleotides from 25–70 bp in length are arrayed by either spotting pre-synthesized oligonucleotides directly onto glass, chemically synthesizing directly onto glass substrates (e.g. Agilent Technologies, Inc.),

Figure 7.2. Platforms for gene expression in transcriptomics, proteomics and metabolomics. Transcriptomic platforms are cDNA or oligonucleotides bound to glass slides or microbeads for analysis of mRNA. Metabolic platforms are nuclear magnetic resonance (NMR) or mass spectrometry (MS) instruments for small biologic molecules or metabolites. Proteomic platforms can be gel-based or liquid chromatography-based (e.g. linear column gradients or multidimensional chromatography (MuDPIT)) for separation of proteins before identification (ID) by mass spectrometry. Use of stable isotopes greatly facilitates protein quantitation (ICAT, isotope coded affinity tags; iTRAQ, isobaric tags for relative and absolute quantitation; SILAC, stable isotope labelling by amino acids in culture), while non-isotopic or label-free methods can also be used that include spectral counting and ion precursor signal measurement. Retentate chromatography mass spectrometry (i.e. surface enhanced laser detection ionization (SELDI)) has been used for rapid profiling of biofluid samples using chemically reactive surfaces for separation and MALDI for generating protein mass spectra. Alternatives for MS-based proteomics involve affinity arrays, such as antibody microarrays or fluorescently tagged antibody bound bead suspensions (i.e. Luminex technology).
or by synthesizing directly onto quartz wafers by photolithographic technology (e.g. Affymetrix, Inc.) (24). In addition, oligonucleotides can be covalently linked with microbeads that can then be used in a 96-well microtiter dish format or on a glass substrate (e.g. Illumina, Inc.). With ever increasing technological advances, microarrays have progressed from chips with only several hundred probes to modern DNA chips reflecting expression of thousands, to even millions, of features per array.

The strength of any gene expression analysis, and the ability to determine expression profiles as potential biomarkers of exposure or effect, is dependent on proper experimental design and careful execution to minimize sources of variance and error and maximize useful biological information. Thus it is critical, in any microarray experiment, to design proper controls to include with samples of biological interest. Proper sampling and integrity of the RNA obtained from those samples are vital in determining the success of the analysis. Once RNA is isolated, proper labelling, hybridization, washing and scanning can dramatically influence the integrity of the resulting data. Since transcript expression is generally expressed in relative terms of a fold change of a particular gene expressed in one sample relative to a value in a control or normal sample, it is critical that the investigator structure the experiment in such a way as to minimize variations other than the one variable to be tested. Fortunately, commercial microarray providers have added increasingly more stringent quality control measurements in the production facilities. This has resulted in high reproducibility and low variation in microarrays coming from a manufacturer. Investigators have been allowed to shift resources away from multiple analyses of single samples to focusing on expanding the numbers of experimental samples. In turn, this has resulted in significant improvements in the confidence of results from microarray experiments. In fact, several large consortium efforts have demonstrated that comparable biological affects could be revealed in carefully controlled experiments in which multiple commercial microarray platforms, as well as rigorously quality-controlled spotted cDNA arrays, were used to analyse the same biological material (25–27).

Additional technological improvements have allowed for the reduction in the starting amounts of mRNA required in the labelling processes for the commercial platforms. Most labelling protocols used a single, PCR-based linear amplification of sample mRNA, which is used to incorporate a nucleotide conjugated with a fluorescent dye, biotin, or some other chemical modification. This amplification step has reduced the starting material for a sample to be analysed to only a few micrograms of total mRNA or less. Furthermore, protocols have been developed for additional rounds of PCR-based amplification of starting mRNA samples that make it possible to analyse very small quantities in the range of nanograms, and even picograms, of mRNA. These developments have facilitated gene expression profiling of samples derived from laser capture microdissection (LCM), for example, as well as biopsy samples, and other clinically derived samples that are limited in quantity. In addition, recent technological developments, particularly using bead-based microarrays (e.g. Illumina BeadChip), have opened up the possibility of using formalin-fixed, paraffin-embedded material for gene expression analysis.

An accessible, biological material fluid of principal interest to several clinical research scientists is blood. Many researchers are interested in testing the utility of gene expression profiling of peripheral blood leukocytes to generate biomarkers as surrogates for other tissues or organs affected in disease or injury processes (28). The utility of this approach has been demonstrated in studies of inflammatory responses and diseases in both animal models and humans, of neurological disorders, of angiomylipoma (AML) and renal cell cancers, and of cardiac injury (29–36). Recent studies have used gene expression analysis of blood samples to generate molecular profiles as biomarkers of exposure and exposure-induced injury to arsenic, benzene, tobacco smoke and hepatotoxic levels of acetaminophen (37–41).

A common application in transcriptomics, useful in epidemiology studies, is to compare transcript outputs between normal and diseased tissues in what has been termed transcript profiling or expression profiling. Transcript expression studies can query all known or predicted genes in an organism, providing an abundance of information that represents a snapshot of the expression status of a tissue at any given time. One can gain considerable insight into molecular mechanisms from properly structured microarray experiments, both on the level of individual genes and on the level of biological pathways and processes. The potential for mRNA degradation makes expression profiling most applicable to freshly isolated tissues, cultured cells or flash-frozen tissue sections, but not paraffin-embedded tissue. While microarray approaches
can be used to interrogate the entire genome on a single microarray chip, focused arrays representing distinct gene subsets have been used to focus upon changes in specific pathways or processes. These include both glass slide-based microarrays (e.g. the National Institute of Environmental Health Sciences’ Human ToxChip (42)), and PCR-based gene expression analyses (e.g. SuperArrays).

DNA arrays can also reflect epigenetic effects upon gene expression. Epigenetics is defined as heritable changes in gene expression that are not due to DNA sequence alterations. Methylation is the most common epigenetic change and is detected by bisulfite conversion, methylation-sensitive restriction enzymes, methyl-binding proteins, and anti-methylcytosine antibodies. Combining these techniques with DNA microarrays and high-throughput sequencing has made the mapping of DNA methylation feasible on a genome-wide scale. Genomic DNA methylation occurs particularly at cytosines in clusters of cytosine-guanine dinucleotides, or CpG islands (p is the phosphodiester bond between C and G bases). Methylation of CpG islands in promoter regions frequently results in gene silencing, which normally occurs during development (43), but is often observed as an early alteration in some cancers by causing inactivation of tumour suppressors genes, such as von Hippel-Lindau disease (VHL), inhibitor of cyclin-dependent kinase 4a (p16INK4a), and breast cancer gene 1 (BRCA1) (44).

DNA microarrays have also been developed for expression beyond profiling. In addition to SNP and comparative genomic hybridization (CGH) applications mentioned in the previous section, genome-wide localization of transcription factor binding sites can be accomplished by chromatin immunoprecipitation (ChIP) analysed on a microarray chip that forms the so-called ChIP-on-chip technique (45). The method can be innovatively combined with different types of DNA arrays, such as SNP chips, to form “ChIP-on-SNP” (46). The future for array technologies will also bring about a revolution in clinical DNA diagnostics (47), develop pharmaceuticals in pharmacogenomics (48), and personalized medicine (49).

**Proteomics**

The field for describing protein expression on a global scale is proteomics, which aims to detail the structure and functions of all proteins in an organism over time. The wide application of proteomics has generated great interest in many established disciplines of exposure biology and medicine, including the field of epidemiology (50,51). Chemical or toxicant exposure can bind to or modify proteins, produce changes in protein expression, and dysregulate critical biological pathways and processes that lead to toxicity and disease, which in theory should be detectable by proteomic analysis. Primary aims in proteomic analysis are the discovery of key modified proteins, the determination of affected pathways, and the development of biomarkers for association with and eventual prediction of disease.

The complexity of a proteome, represented by the total protein expression of a specific cell, organ, tissue or biofluid, presents numerous challenges for comprehensive analysis. Proteins are more complex than nucleic acids, and therefore proteomic analysis involves measurement of just some of the many attributes of proteins during any single expression analysis (52). Proteins exhibit many attributes of interest to biomarker development in epidemiology studies, including determination of protein sequence identity, quantity, post-translational modifications (PTM), protein–protein interactions, structure and function. Some of the challenges in proteomic analysis include: defining the identities and quantities of an entire proteome in a particular spatial location, such as serum or subcellular structures like mitochondria; the existence of multiple protein forms and complexes; the evolving structural and functional annotations of the human and rodent proteomes; and integration of proteomics data with transcriptomics or other expression data. Primary aims of proteomic analysis are to achieve maximal proteome identification, quantitative high-throughput protein measurement, timely analysis, and discovery-oriented platforms. Proteomic platforms represent combinations of technologies that describe protein attributes by the separation, quantitation and identification of all proteins in a biological sample. Proteomic analysis includes four broad categories of proteomic platforms: mass spectrometry has played a central role in proteomic platform development in large part because of its sensitive and versatile ability to identify proteins; the ability to separate proteins greatly determines the designation of platform type by gel-based separation or liquid chromatophic separation linked to mass spectrometry (53); solid phase adsorption, based on partitioning of peptides and proteins due to specific chemical properties, has been exploited in rententate chromatography combined with mass spectrometry; and finally, affinity chromatography, which
sorts and identifies proteins in one reaction, is exemplified by use of antibodies in various formats (54). The following proteomic platforms represent some of the primary technologies being used for separating, identifying, and quantifying proteins during toxicoproteomic studies (Cf. Figure 7.2).

2D PAGE and DIGE

Two-dimensional polyacrylamide gel electrophoresis (2D PAGE) systems have been combined with mass spectrometry in an established and adaptable platform. Since 1975, 2D PAGE has been the most commonly used proteomic platform to separate and comparatively quantitate protein samples (55). Current state-of-the-art 2D gels use immobilized pH gradient (IPG) gels to separate proteins by charge. They are then resolved by mass spectrometry using sodium dodecyl sulfate (SDS) gel electrophoresis for effective separation of complex protein samples in μg to mg quantities. Either visible stains, such as Coomassie Blue or silver, or fluorescent staining are used for sensitive protein detection. After electronic alignment (registration) of stained proteins in 2D gels by image analysis software, intensities of identical protein spots are compared among treatment groups and a ratio (fold change) is calculated for each protein. A relatively new variation of the 2D PAGE technique, difference gel electrophoresis (DIGE), allows an investigator to measure three samples per gel that have been labelled with Cy2, Cy3 and Cy5 fluorescent dyes, which reduces some of the error associated with electronic registration during multiple gel alignment. This strategy allows for direct comparison of samples on one gel for better reproducibility and quantitation than conventional image analysis for comparison of multiple 2D gels. Thus, separation of proteins by 2D gels, using single stains or multiple fluor (i.e. DIGE), can be combined with mass spectrometry for ready protein identification to form a versatile and discovery-oriented platform for use in proteomic studies (56). In addition, some protein samples are sufficiently limited in their protein content that a simple size separation (one dimension) by SDS–PAGE can be used to identify protein bands of interest by mass spectrometry in 1D-Gel-MS.

Multidimensional LC-MS/MS

Proteomic platforms incorporating liquid chromatography (LC) as the primary means of separation (versus gel-based separations) have become the preferred means of analysis. There are many different types of LC separations often termed “multidimensional.” In this proteomic platform, LC is used to separate protein digests by exploiting different biophysical properties of proteins before identification by tandem mass spectrometry (MS/MS). One of the most notable multidimensional LC-MS/MS platforms is Multidimensional Protein Identification Technology (MuDPIT). MuDPIT attempts to identify all proteins in a sample by two-dimensional separation of protein digests by charge (strong anion exchange matrix) and hydrophobicity (C18 column) with online LC immediately before entry into a tandem mass spectrometer (MS/MS) for protein identification (57). The platform has also been called shotgun proteomics, as entire protein lysates are trypsin digested into thousands of peptide fragments without any prior fractionation before separation and identification. Advantages of this newer platform are the potential for detection and identification of low abundance proteins that may not be observed in gel staining-based methods. One drawback is that LC-MS/MS platforms, like MuDPIT, are only semiquantitative and somewhat low-throughput in capacity.

The issue of protein quantitation in proteomics is an important one, since changes in protein expression may be a matter of altering existing gene and protein expression rather than turning them on (induction) or off (repression), which makes quantitation of proteins crucial in normal and diseased or control and experimental states. The use of stable isotopes, as detailed in the section below, takes advantage of the high resolution power of mass spectrometers to discriminate protein samples stable-isotopically labelled for quantitative comparison. However, rapid developments are being made using “label-free” approaches to quantitation if sample mass spectral data are sufficiently detailed (58). The two main approaches used are ion precursor signal intensities and spectral counting. Though they deliver relative sample quantitation, versus absolute protein measurement, they are simpler and less costly than stable isotopic methods, but not as precise. Label-free protein quantitation methods should be considered at the outset of designing LC-MS/MS proteomic studies and weighed against the considerable advantages and choices of stable isotopic approaches.

Stable isotope LC-MS/MS platforms: ICAT, iTRAQ and SILAC

A primary goal of proteomics is to comprehensively analyse all proteins in a sample, or as many
as possible. However, the prospect of quantifying protein levels for comparison among protein samples has been a difficult aspect of proteomic analysis. Protein quantitation can be considered either in relative terms as a proportion of treatment (test) samples compared to control samples, or in absolute terms as the number of molecules (moles) or concentration (molarity). Internal standards are useful, but not realistic, for complex protein samples of unknown composition in most proteomic studies. Since many proteomic platforms are based in mass spectrometry, comparison of intensity signals seems the most direct means for comparative measurements; however, intensities are subject to many interfering factors. Quantitation by mass spectrometry has generally been regarded as semiquantitative under the best of circumstances.

The use of stable isotopes for tagging proteins has made great strides in proteomics for determining the relative amounts of proteins among samples (59). Stable isotopes of an element differ in mass due to the number of neutrons, but have the same elemental and chemical characteristics as the element. Stable isotopes are not radioactive. Common stable elements and their stable isotopes are 1H and 2H; 12C and 13C; 14N and 15N; 16O and 18O; 32S and 34S. A unique feature of high-resolution mass spectrometers is the ability to finely distinguish between small differences in mass, even to the point of resolving the relative abundance of stable isotopes in otherwise identical samples. Several proteomic platforms for protein quantitation and identification have been built around the use of isotopic tagging of proteins (isotope coded affinity tagging (ICAT)), peptides (isotope tag for relative and absolute quantitation (iTRAQ), or metabolic incorporation of isotopically tagged amino acids in cell culture (stable isotope labelling with amino acids in cell culture (SILAC)). The applications of stable isotopes in proteomics have been recently reviewed for their sensitive detection of proteins in a quantitative and comparative fashion (60). As mentioned above, continuing improvements in spectral counting for use in LC-MS/MS platforms should have wide utility as a versatile, isotope-free method of protein quantitation when stable isotope use is not feasible (61).

**SELDI-TOF mass spectrometry**

Retentate chromatography-mass spectrometry (RC-MS) is a high-throughput proteomic platform that creates a laser-based mass spectrum (based on matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry) from a chemically-absorptive surface. The principle of this approach is the adsorptive retention of a subset of sample proteins on a thin, chromatographic support (i.e. hydrophobic, normal phase, weak cation exchange, strong anion exchange or immobilized metal affinity supports). The absorptive surfaces are placed on thin metal chips which can be inserted into a MALDI-type mass spectrometer. The laser rapidly desorbs proteins from each sample on a metal chip to create a mass spectrum profile. RC-MS can be performed upon any protein sample, but thus far this platform has found greatest utility in the analysis of serum and plasma for disease biomarker discovery (62). The lead commercial platform of RC-MS proteomic platforms is the surface-enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI-TOF-MS) instrument (63). Analysis of samples is relatively rapid (100/day), and only a few μl of sample is necessary. A downside is that protein identification of peaks is not readily accomplished without additional conventional separation and analysis.

**Antibody arrays**

Protein microarrays represent a promising new proteomic tool that closely emulates the design for parallel analysis of DNA microarray technology (64). Protein microarray formats can be divided into two types of multianalyte sensing formats: forward phase arrays and reverse phase arrays. In the forward phase array format, the analyte is captured from the solution phase by different capture molecules, such as an antibody immobilized on a substratum (i.e. glass slides). In contrast, the reverse array format immobilizes the individual test samples in each array spot so that, for example, hundreds of different patient blood or tissue samples are arrayed and probed with one detection protein, such as an antibody and a single analyte endpoint are measured for comparison across multiple samples. Such microarray studies have been carried out for metastatic ovarian cancer (65).

Many different types of capture molecules can be arrayed, including peptides (i.e. peptide substrates for kinases on phosphorylation arrays), proteins (protein–protein interaction arrays) and oligonucleotides (i.e. transcription factor binding arrays to oligonucleotides), but the most prevalent are antibody arrays in the forward phase format. Antibody arrays can directly separate proteins from complex biological fluids like plasma, serum or cell lysates by affinity binding to specific antigenic sites on target proteins.
Generally, current commercial antibody array platforms fall into three classes based on the targeted proteins: cytokine/chemokine arrays, cellular function protein arrays and cell signalling arrays. Although not all proteins for any given cell type or biofluid (i.e. blood, serum, plasma, urine, cerebral spinal fluid) are currently represented on antibody arrays, they do provide a rapid screen for protein alterations that may be relevant to tissue injury or disease (54). Antibodies can be placed in ordered array on glass slides or on a fluorescent microbead format (i.e. Luminex technology) for multiplexed separation, identification and quantitation (66).

For example, in a study investigating the chemotherapeutic and radiotherapy of patients with rectal cancer, 40 tumour samples were analysed by DNA microarray and plasma samples were analysed by antibody (Luminex) bead microarray platforms. Using a kernel-based method with Least Squares Support Vector Machines to predict rectal cancer regression grade, investigators found that combining and integrating of microarray and proteomics data improved predictive power leading to the best model based on five genes and 10 proteins with an accuracy of 91.7%, sensitivity of 96.2% and specificity of 80% (67). In a different approach, a molecular epidemiological study used SELDI-TOF MS for in vivo studies of humans exposed to benzene. By using two sets of 10 exposed and 10 unexposed subjects, researchers identified with chemically-reactive surfaces and validated with antibody-coated chips three differentially expressed proteins in the serum of benzene-exposed individuals, two of which were identified as PF4 and CTAP-III, both members of the CXC-chemokine family (68). The same can be done with peptides (instead of antibodies) as the affinity ligand. This method was applied in the development of two diagnostic antibodies against avian influenza detection for epidemiologic studies, in which the epitopes of two monoclonal antibodies (mAbs) against avian influenza nucleoprotein (NP) were found using truncated NP recombinant proteins and peptide array techniques (69).

Future developments in proteomics will see incorporation of more sophisticated methods of quantitation in proteomic analysis (70), combining higher data density LC-MS/MS platforms with stable isotope labelled peptides, spectral counting, and parallel use of complementary proteomic platforms, such as tissue arrays (71). Study designs that remove abundant proteins from biofluids, enrich subcellular structures, and include cell-specific isolation from heterogeneous tissues will greatly increase differential expression capabilities. Advancement in mechanistic insights and biomarker development using proteomics will be furthered by completely defining plasma (serum) proteome and circulating microparticles in humans and rodent species as accessible biofluids (72). Some of the representative biomarkers and patterns of protein and message expression are shown in Table 7.1. Reviews on using proteomics to develop biomarkers and further mechanistic insights have been published (73–75).

### Metabolomics

Analogous to the genomic characterization of cellular DNA and the proteomic characterization of the set of proteins expressed at a given time, cells also can be

<table>
<thead>
<tr>
<th>Biomarker/Signature</th>
<th>Identification</th>
<th>Condition</th>
<th>Reference No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>KIM-1</td>
<td>DNA microarray</td>
<td>Renal toxicity</td>
<td>(145)</td>
</tr>
<tr>
<td>Adipsin</td>
<td>DNA microarray</td>
<td>GI toxicity, functional gamma secretase inhibitors (FGSIs)</td>
<td>(146)</td>
</tr>
<tr>
<td>CXC-chemokines</td>
<td>SELDI</td>
<td>Benzene exposure</td>
<td>(68)</td>
</tr>
<tr>
<td>Troponin I,T</td>
<td>2D gel-MS</td>
<td>Myocardial ischemia, infarction</td>
<td>(147)</td>
</tr>
<tr>
<td>Aminopeptidase-P</td>
<td>2D gel-MS</td>
<td>Radioimmunotherapy</td>
<td>(148)</td>
</tr>
<tr>
<td>Annexin A1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lipid-gene signature</td>
<td>DNA microarray</td>
<td>Drug-induced phospholipidosis</td>
<td>(149)</td>
</tr>
<tr>
<td>Multigene blood signature</td>
<td>DNA microarray</td>
<td>Systemic sepsis</td>
<td>(32)</td>
</tr>
</tbody>
</table>
characterized in terms of the set of low-molecular-weight metabolites (typically < 1500 D) that comprise the cellular "metabolome." The cellular metabolome provides a functional readout of the cellular proteome (Cf. Figure 7.1). Although the analysis of homogeneous cell populations in culture, receiving identical nutrients and oxygenation, and exposed to the same levels of excreted waste products, represents the most ideal system for metabolomic characterization, the approach has been extended to the analysis of extracts and fluids derived from higher organisms. Urinary and blood metabolites have been among the most frequent targets for metabolomic characterization, but analyses of other fluids, such as cerebrospinal fluid (CSF), bronchoalveolar lavage fluid (76) and saliva (77), and of cellular extracts also have been performed. Typical $^1$H NMR spectra illustrating the different metabolite composition of urine, blood, and saliva are shown in Figure 7.3. Currently, metabolomic characterization is being used for a wide range of objectives in human nutrition and toxicology, and for the development of pharmaceuticals and agricultural products. The underlying objectives of these studies include: discovery of metabolite signatures as prognostic indicators, diagnostics or biomarkers of disease states; establishing toxicological markers for drug development and environmental toxicology; understanding mechanisms of metabolic diseases; and correlation of metabolite phenotypes (metabotype) with genotype and environmental input (e.g. nutrition).

The screening of neonates for genetic disorders in intermediary metabolism is an application that predates the more recent interest in metabolomic characterization. The recent reviews of analytical approaches for clinical diagnosis of metabolic disorders (78,79) summarize methods for metabolite analysis and provide good examples of the application of MS to metabolomic analysis (Cf. Figure 7.2). Mass spectrometric analysis typically requires preparation of the metabolic components using either gas chromatography (GC) after chemical derivatization, or LC, with the newer method of ultra-performance liquid chromatography (UPLC) increasingly used. The use of capillary electrophoresis (CE) coupled to MS also has shown some promise. It was reported that a combination of approaches for metabolite extraction produced over 10 000 unique metabolite features, indicating both the complexity of the human metabolome and the potential of metabolomics in biomarker discovery (80).

Other more specialized techniques, such as Fourier transform infrared (FTIR) spectroscopy and arrayed electrochemical detection, have been used in some cases (81,82). The main limitation of FTIR is the low level of detailed molecular identification that can be achieved. In one study, MS was also employed for metabolite identification (82).

The extensive application of nuclear magnetic resonance (NMR) for metabolic characterization of urine and other body fluids has been developed primarily by Nicholson and coworkers (83–85). The primary advantage of the NMR approach (Cf. Figure 7.2) is the lack of required preparatory separations, which in turn leads to an unbiased and potentially quantitative measure of the constituents of the sample. Alternatively, while in principle it

Figure 7.3. Typical $^1$H NMR spectra illustrating the metabolite composition of urine (A), saliva (B) and plasma (C). Used with permission from (77).

TMAO, trimethylamine oxide; DMA, dimethylamine.
is possible to obtain a molecular mass corresponding to each unfragmented metabolite observed in a mass spectrum, in an NMR spectrum, molecular information is distributed among the resonances of a compound, and the position of these resonances can depend critically on pH, salt concentration, divalent ions and other physical parameters. Additional caveats discussed in various reviews include loss of more volatile metabolites, metabolite contributions derived from intestinal bacteria (86), potential bacterial growth in stored fluids, and other factors (84,87). Some have proposed using dimethylamine and its nitroso metabolite as biomarkers for small bowel bacterial overgrowth (86), while others have suggested monitoring 4-hydroxyphenylacetate as a potential screening method for small bowel disease and bacterial overgrowth syndromes (88). Significant levels of ethanol in plasma or urinary samples generally indicate either bacterial contamination of the sample or small bowel bacterial overgrowth (87). Such conditions typically accompany renal failure or other serious illnesses.

In studies of chemical toxins or pharmaceutics, the xenobiotic and its metabolites and conjugates typically constitute an important source of variation of the metabolome. While of interest from a metabolic perspective, these compounds are not directly indicative of organ toxicity or therapeutic response, so that in general these metabolites are not relevant to the study. One approach to dealing with this issue involves stable isotope labelling of the compounds under study, so that the compound and metabolites derived from it will exhibit characteristic features in the NMR or mass spectrum. Alternatively, if the test compound and all of its metabolites and conjugates can be identified, these can simply be ignored or eliminated from the analysis. One general source of toxicity resulting from the administration of high levels of test compounds is the depletion of sulfur-containing amino acids that results from the excretion of glutathione and cysteine conjugates. Thus, it was reported that rats receiving high levels of acetaminophen excreted significant amounts of pyroglutamic acid. This effect of excess acetaminophen was prevented/reversed by supplementation with methionine (89). The glutathione analogue ophthalmic acid recently has been found to accumulate after high dosage with acetaminophen, and may also function as a biomarker for oxidative stress and glutathione depletion (90).

Various multivariate analyses of metabolite composition have been applied to detect differences among subject groups, such as those receiving different treatments or different chemical exposures. This type of analysis, termed “metabonomics” by the Nicholson group, most frequently utilizes principal component analysis (PCA) of the spectral data to reveal clustering behaviour that differentiates treated from control subjects. The clusters of data points in PC plots reveal the uniformity of the control and treated groups, as well as the extent to which the treated group yields a distinct metabolic phenotype. Since the axes of the PC plot are dependent on the data set and do not correspond to independent variables, the ability of different laboratories to utilize the published information is limited. Interestingly, a recent study that evaluated statistical methodology for the analysis of gene expression data found that the use of PCA to reveal clustering behaviour generally degrades cluster quality, and concluded that PCA was only useful in special cases (91). Hence, there is a critical need for the identification of metabolic biomarkers that provide universally quantifiable indications of organ function and toxicity.

Another critical issue related to the identification of biomarkers is the need to separate acute and chronic effects of illnesses or toxins. It is not unusual for a particular metabolite to become elevated during the acute phase of a toxic response, but to become depressed as the chronic effects become significant. Alternatively, in the absence of chronic effects, the metabolite level may return to pretreatment values. For this reason, plots of data obtained at different times after dosage, or trajectory plots for individual subjects, can provide critical information on the time-dependent response. Several of the issues discussed above are illustrated in studies identifying the association of the oxidative stress biomarker 8-oxoguanosine with Parkinson's disease (92). In this study, elevation of the 8-oxoguanosine level was observed in cerebrospinal fluid (CSF), but not in the serum of patients with Parkinson. Further, there was a significant negative correlation between the level of the biomarker and the duration of the disease. Finally, as indicated in Table 7.2, 8-oxoguanosine is also elevated in other conditions, e.g. amyotrophic lateral sclerosis (93). Another important limitation on the identification of some biomarkers relates to the chemical reactivity, which can deplete the free metabolite pool and lead to heterogeneous adduct formation and difficulties of detection. Homocysteine, which has long been linked to cardiovascular disease, provides one example (94).
Much of the early NMR work evaluating the effects of various toxins noted changes in the levels of tricarboxylic acid cycle (TCA) metabolites and other abundant molecules that may be present in the nutrient source and that are not organ-specific (85). More recently, several more specific metabolomic biomarkers have been correlated with various diseases or treatments (Table 7.2). Notably, most of the biomarkers that have been related to metastatic growth are proteins, but new metabolomic markers continue to be developed. The metabolite 12(S)-hydroxyeicosatetraenoic acid (12(S)-HETE) has been demonstrated to play a pivotal role in experimental melanoma invasion and metastasis, suggesting that 12-lipoxygenase expression may be important in early human melanoma carcinogenesis (95–97). Changes in phosphate-containing metabolites and in phospholipid composition have been correlated with tumour stage and metastatic spread. In studies of extracts from tumour cell lines, elevations in phosphorylcholine or other membrane-related phosphomonoesters have frequently been observed, and suggested to be correlated with metastatic potential (98,99). Increases have been observed in aspartyl-4-phosphate.

Table 7.2. Biomarkers recently identified using metabolomics

<table>
<thead>
<tr>
<th>Biomarker</th>
<th>Sample Analysis</th>
<th>Condition</th>
<th>Reference No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAN; 2-PY</td>
<td>NMR – human and rodent urine</td>
<td>Type 2 diabetes mellitus</td>
<td>(110)</td>
</tr>
<tr>
<td>NAN; 2-PY</td>
<td>NMR – rat urine</td>
<td>peroxosome proliferation</td>
<td>(111,112)</td>
</tr>
<tr>
<td>ADMA</td>
<td>MS – human blood plasma</td>
<td>Renal failure; atherosclerosis</td>
<td>(108,109)</td>
</tr>
<tr>
<td>Ophthalmic acid</td>
<td>MS – mouse serum, liver extract</td>
<td>Acetaminophen-induced hepatotoxicity</td>
<td>(90)</td>
</tr>
<tr>
<td>Pyroglutamic acid</td>
<td>NMR – rat urine</td>
<td>APAP-induced deficiency of sulfur- amino acids</td>
<td>(89)</td>
</tr>
<tr>
<td>3-nitrotyrosine</td>
<td>HPLC - human CSF</td>
<td>Amyotrophic lateral sclerosis</td>
<td>(150)</td>
</tr>
<tr>
<td>8-oxoguanosine</td>
<td>HPLC – human CSF</td>
<td>Alzheimer’s disease</td>
<td>(93)</td>
</tr>
<tr>
<td>8-oxoguanosine</td>
<td>HPLC – human CSF</td>
<td>Parkinson’s disease</td>
<td>(92)</td>
</tr>
<tr>
<td>Modified nucleosides</td>
<td>LC-IT-MS of human urine</td>
<td>Breast cancer</td>
<td>(102)</td>
</tr>
<tr>
<td>12(S)-HETE*</td>
<td>HPLC - tumor cell extracts</td>
<td>Human melanoma</td>
<td>(95–97)</td>
</tr>
<tr>
<td>Aspartyl-4-phosphate</td>
<td>DESI-MS/NMR – murine urine</td>
<td>Lung cancer/ tumour growth</td>
<td>(100)</td>
</tr>
<tr>
<td>Phosphorylcholine</td>
<td>31P NMR – cell extracts</td>
<td>Breast cancer cell extracts</td>
<td>(98)</td>
</tr>
<tr>
<td>Depressed lysophosphatidyl choline levels</td>
<td>31P NMR – blood plasma</td>
<td>Renal cell carcinoma</td>
<td>(101)</td>
</tr>
<tr>
<td>Elevated xanthine, hypoxanthine, urate</td>
<td>GC-MS – human urine</td>
<td>Lesch-Nyhan syndrome</td>
<td>(103)</td>
</tr>
<tr>
<td>Glc-Gal-pyridinoline</td>
<td>HPLC - human urine</td>
<td>Synovial degradation – RA</td>
<td>(104-106)</td>
</tr>
<tr>
<td>4-hydroxyphenyl acetate</td>
<td>GC-LC – human urine</td>
<td>SBBO</td>
<td>(88)</td>
</tr>
<tr>
<td>Dimethylamine, nitrosodimethylamine</td>
<td>GC – human serum; GC – whole blood</td>
<td>SBBO</td>
<td>(86)</td>
</tr>
</tbody>
</table>

* In most reported studies, concentration or enzymatic activity of 12-lipoxygenase, rather than 12(S)-HETE, has been determined.

2-PY, N-methyl-2-pyridone-5-carboxamide; ADMA, NG,NG-dimethylarginine; APAP, Acetaminophen; DESI, Desorption electrospray ionization; GC-LC, Gas chromatography-liquid chromatography; GC-MS, Gas chromatography-mass spectrometry; HPLC, High performance liquid chromatography; LC-IT-MS, Liquid chromatography-ion trap-mass spectrometry; NAN, N-methylnicotinamide; NMR, Nuclear magnetic resonance; RA, rheumatoid arthritis; SBBO, small bowel bacterial overgrowth.
that may correlate with tumour growth (100). Phosphorus-31 NMR studies of blood plasma derived from patients with advanced renal cell carcinoma have been found to exhibit depressed levels of lysophosphatidylcholine (101). A significantly improved discrimination of breast cancer patients based on metabolomic analysis of modified nucleosides present in urine was recently reported (102).

A metabolomic approach using a combination of gas chromatography and MS has identified elevations in the levels of the metabolites xanthine, hypoxanthine, urate and guanine in patients with Lesch-Nyhan syndrome (103). The urinary cross link product, Glc-Gal-pyridinoline (104), has been identified as a biomarker for synovial degradation observed in osteoarthritis (105); treatment with ibuprofen lowers the level of this excreted metabolite (106). Endogenously formed N⁰,N⁰-dimethylarginine, also referred to as asymmetric dimethylarginine (ADMA), is a potent inhibitor of nitric oxide synthase (107). Plasma levels increase as a consequence of renal failure (108), and ADMA has been identified as a biomarker for atherosclerosis (109). Metabolomic analyses have identified several pyridine derivatives in urine from diabetic rats (110), and the same derivatives have shown up as biomarkers of peroxisome degradation (Table 7.2) (111,112). The presence of these compounds indicates a perturbation of the tryptophan-nicotinamide adenine dinucleotide (NAD) pathway.

As is typical for new technologies, some reports of putative biomarkers have proven controversial. Early identification of the partly-characterized metabolites CFSUM1 and CFSUM2, associated with chronic fatigue syndrome, were subsequently demonstrated to arise from incompletely derivatized pyroglutamic acid and serine (113,114), and the quantitative abnormalities of these metabolites in urine from patients with chronic fatigue syndrome/myalgic encephalomyelitis has been reported to be artefactual. Early analyses supporting the use of ¹H NMR of blood sera to diagnose coronary artery disease (115) have subsequently been found to be more equivocal than originally suggested and to compare unfavourably with angiography-based diagnosis (116).

As more specific biomarkers are identified, the power of this approach will continue to evolve, providing useful diagnostic information for pathological, environmental, toxicological, pharmaceutical and nutritional research, as well as enhancing the value of metabolomic analysis for basic research into mechanisms of toxicity. Future developments in mass spectrometry platforms in metabolomics will increase the detectable coverage of the metabolome in clinical specimens and experimental species, and permit better identification of metabolites in the process of converting raw data to biological knowledge (117).

Bioinformatics

The wealth of data generated through high-throughput omics approaches has become increasingly complex and too vast for conventional biomarker analysis strategies. Bioinformatics has played a crucial role in biomarker discovery and validation (Figure 7.4). It is a multidisciplinary field involving biology, computer science, mathematics and statistics to derive knowledge from biological, genomics and genetic data (118,119). Database systems, computational algorithms, statistical models, data mining methods and other analytical tools are typically employed in a bioinformatics framework to effectively manage, analyse and summarize the plethora of data. For example, proteomics approaches, such as SELDI-TOF and mass spectrometry in conjunction with bioinformatics tools, have greatly facilitated the discovery of new and better serum biomarkers to detect cancer (120).

The bioinformatics processes to translate omics data into clinically useful biomarkers can comprise a myriad of steps, beginning with initial analysis of the data to validation of the biomarkers (121). This multistep process typically involves discovery, data integration, predictive modeling, and delivery of the biomarkers to the clinic in a format to facilitate implementation (122). Bioinformatic analyses may take different approaches with several checkpoints along the way. A flowchart is described for the application of bioinformatics strategies to improve the identification of candidate biomarkers from cancer genome-wide expression analyses (123). The process proceeds with acquisition of gene expression data from cancer tissues, followed by the identification of candidate genes as biomarkers. The next step entails meta-mining public cancer data sets of the same type of pathophysiology to reduce the biomarker false-positive rate. The last step before use of the biomarkers in clinical trials involves validation of the candidates by RT–PCR, ELISA, tissue arrays, immunohistochemistry, and other types of bioassays.

What are these bioinformatics tools and processes and how are they used to aid in the discovery and validation of disease biomarkers? It is helpful to appreciate that a useful biomarker must be objective, highly accurate and very reliable in determining disease states and
assessing risk. In other words, they must generalize well (i.e. extend) to broad cases, exhibit significance in their reporting, and be precise in their utility. Unfortunately, omics data possesses systematic variation due to the experimental error in the data acquisition process (124). There are technical limitations in the sensitivity of detecting biomarkers that are lowly expressed or non-abundant; however, biomarkers do not need to be highly expressed or in large abundance. Thus, the challenge in biomarker identification is successfully mining omics data with inherent error to find the features that reliably, accurately and objectively relate to the pathophysiology of a disease (125).

Data normalization and dimension reduction (data condensing) techniques have been widely used to preprocess omics data before biomarker discovery. Standard ways of dealing with data normalization have been adopted for omics data. Robust Multiarray Average (RMA), loess and quantile normalization methods have seemed to pass the test of time (126–128). A systematic variation normalization (SVN) approach was developed specifically to remove systematic error from microarray gene expression data (129). Baseline subtraction, signal smoothing and normalization methodologies were employed in the preprocessing of mass spectrometry proteomics data to reduce the noise and to make the analysis of spectral data comparable (130). In general, once the omics data is made unbiased and adjusted to make fair comparisons across samples, all the data can be used to mine for biomarkers or be filtered first to remove uninformative or redundant information. Some believe that the inclusion of uninformative features in the biomarker selection process will severely degrade the performance of the predictor model (131). Thus, it has been suggested to remove variables that do not contribute to a biological response of interest before the selection of biomarkers. Filtering of the data can be based on a signal or relative level, fold change, a confidence level, standard deviation from the mean of the distribution, P-values, mutual information, full or partial correlations, or more elaborate methods. The underlying omics data should be rich enough to be narrowed down to a core set of features that best represent the biology of the system for biomarker selection. Caution must be taken with respect to the preprocessing of omics data for biomarker identification, as the use of a particular combination of methods will surely add variability to the set of indicators selected. In other words, the preprocessing and filtering steps are sources of variability in and of themselves. The US Food and Drug Administration-led MicroArray Quality Control Phase II (MAQC-II) consortium set out to address this by trying to understand the limitations of various bioinformatics data analysis methods in developing and validating microarray-based predictive models, and determining if best practices for development and validation of predictive models based on microarray gene expression and genotyping data can be derived for biomarker discovery and personalized medicine (132).

The development and utilization of classification and prediction methods for analysis of omics data have transcended the process
for identifying biomarkers from molecular signatures. One of the most widely used methods for classification is clustering. The process works by using a dissimilarity measure for the feature profiles in the omics data to iteratively form groups of samples that are tightly clustered. Features that cluster well together and can distinguish between groups of samples that differ in pathophysiology are considered to be potential biomarkers. Clustering and an F-test-like score based on within- and between-sample gene expression variance measures were used effectively to identify an intrinsic gene subset (i.e. a molecular portrait) that has a high predictive score for human breast tumours (133).

More sophisticated bioinformatics methods have been developed to identify potential biomarkers. A hybrid approach was developed based on the genetic algorithm (GA) and k-nearest neighbours (KNN) classifier that is capable of identifying gene and protein molecular signatures of diseases based on microarray and proteomics data, respectively (134,135). The GA serves as a search tool to choose small subsets of predictors, whereas the KNN functions as a non-parametric (no distribution model assumed) pattern recognition method to evaluate the discriminative ability of the subsets. More recently, a hybrid approach for biomarker discovery from microarray gene expression data was developed to distinguish between types of cancer (136). This approach is based on Fisher’s ratio (a measure for the linear discriminative power of variables) to select features “wrapped” with a classifier (hence, the procedure is called FR-Wrapper) to perform predictions. With these hybrid approaches, the two main objectives in biomarker discovery are met: 1) the identification of a small set of relevant indicators with minimum redundancy, and 2) the validation of the predictors using a classifier and cross-validation strategy. To balance false-positives and false-negatives in the selection of biomarkers, a clever method was proposed to use common peaks in mass spectrometry data as the predictive indicators (137). The procedure applies AdaBoost (a form of ensemble classifier training) to perform the classification and to select the informative common peaks.

Bioinformatics approaches to discover biomarkers can take on more sophisticated implementations. For instance, a dependence (interaction) network modeling scheme was suggested for identifying biomarkers from groups of genes or proteins (138). Very clear differences were observed in the dependence networks for cancer and non-cancer samples. On the other hand, a gene selection algorithm was used based on Gaussian processes to discover consistent gene expression patterns associated with ordinal clinical phenotypes (139). The method was able to identify subsets of genes as potential biomarkers for colon and prostate cancers. The integration of time-course microarray gene expression data with cytotoxicity measurements, by way of a partial least squares objective criterion, has been shown to be useful for identifying biomarkers in primary rat hepatocytes exposed to cadmium (140). The approach demonstrates the value of integrating omics data with associated biological data to glean more information about the biomarker’s diagnostic utility. More recently, a bioinformatics approach was introduced that takes into account the inherent correlation of genes when using gene expression data for biomarker discovery (141).

Finally, a host of techniques and software to integrate omics data are summarized, to shed additional light on the complex molecular interactions that take place on a systems biology level (142). Repositories of omics data from various studies that can be queried may bring about improved means for detecting biomarkers of a clinical process or phenotype than one which is isolated or from a small group of data sets (143). This realization has motivated the generators of omics data to store them in repositories for meta-mining purposes. Figure 7.5 represents a brief list of some of the publicly accessible databases that store, distribute and permit querying of omics data (a more comprehensive list is presented in (142)). A plasma proteome database at the Institute of Bioinformatics that stores comparisons of human plasma protein concentration levels along with their isoforms in normal and disease states should be useful for discovery of novel biomarkers (144).

Another database, ONCOMINE, stores a collection of curated cancer gene expression profiles integrated with a therapeutic target database and biological resources, such as Gene Ontology, so that the data can be mined for putative biomarkers (123).

The use of bioinformatics tools has increased mechanistic understanding and development of biomarkers in the analysis of massive genomics, proteomics and metabolomics data (Cf. Figure 7.4). Bioinformatics techniques will continue to be useful in organizing and extracting candidate biomarkers for chemical exposures and disease for epidemiology, clinical and experimental studies. However, mere access to sophisticated bioinformatics tools will be insufficient to grapple with
An ongoing and vigorous debate has emerged over the use and reproducibility of bioinformatics approaches and omics data for biomarker discovery and clinical applications (145,146). Clearly there is a need for rigorous quality control in the field of omics data for biomarker discovery and clinical applications (43,44). A fundamental understanding of the inherent problems and issues with omics data, and knowing how, where and when to apply which type of bioinformatics approach, are essential to effectively translating omics biomarkers into clinically useful diagnostic tools and epidemiological markers.

**Figure 7.5.** Online bioinformatics resources for various omics fields, including genomics, transcriptomics, proteomics, and metabolomics

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Resource</th>
<th>Host</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomics</td>
<td>Genomes OnLine Database (GOLD)</td>
<td>Genomesonline.org</td>
<td>Completed and ongoing genome projects</td>
<td><a href="http://www.genomesonline.org/index">http://www.genomesonline.org/index</a></td>
</tr>
<tr>
<td></td>
<td>EMBL-Bank</td>
<td>EMBL-EBI</td>
<td>Nucleotide sequence database</td>
<td><a href="http://www.ebi.ac.uk/">http://www.ebi.ac.uk/</a></td>
</tr>
<tr>
<td></td>
<td>ArrayExpress</td>
<td>EMBL-EBI</td>
<td>Repository for MIAME-compliant microarray data</td>
<td><a href="http://www.ebi.ac.uk/arrayexpress">http://www.ebi.ac.uk/arrayexpress</a></td>
</tr>
<tr>
<td></td>
<td>ArrayTrack™</td>
<td>FDA-NCTR</td>
<td>An integrated solution for managing, analysing and interpreting MIAME-compliant microarray gene expression data from pharmaco- and toxicogenomics studies</td>
<td><a href="http://www.fda.gov/ScienceResearch/BioinformaticsTools/Arraytrack/default.htm">http://www.fda.gov/ScienceResearch/BioinformaticsTools/Arraytrack/default.htm</a></td>
</tr>
<tr>
<td></td>
<td>ONCOMINE</td>
<td>University of Michigan</td>
<td>Database for mining and examining gene expression in cancer</td>
<td><a href="http://www.oncomine.org">http://www.oncomine.org</a></td>
</tr>
<tr>
<td></td>
<td>Environment, Drugs and Gene Expression (EDGE)</td>
<td>University of Wisconsin-Madison</td>
<td>Toxicology-related gene expression data for mapping transcriptional changes from chemical exposure</td>
<td><a href="http://edge.oncology.wisc.edu/">http://edge.oncology.wisc.edu/</a></td>
</tr>
<tr>
<td>Proteomics</td>
<td>World-2DPAGE</td>
<td>Expert Protein Analysis System (ExPASy)</td>
<td>A dynamic portal to query simultaneously worldwide proteomics databases</td>
<td><a href="http://ca.expasy.org/world-2dpage">http://ca.expasy.org/world-2dpage</a></td>
</tr>
<tr>
<td></td>
<td>Plasma Protein Database</td>
<td>Pandey Lab and Institute of Bioinformatics</td>
<td>A comprehensive resource for all human plasma proteins along with their isoforms</td>
<td><a href="http://www.plasmaproteomedatabase.org">http://www.plasmaproteomedatabase.org</a></td>
</tr>
<tr>
<td>Data Type</td>
<td>Resource</td>
<td>Host</td>
<td>Description</td>
<td>URL</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------------------------------------------</td>
<td>-------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>Open Proteomics Database (OPD)</td>
<td>University of Texas at Austin</td>
<td>For storing and disseminating mass spectrometry-based proteomics data</td>
<td><a href="http://apropos.icmb.utexas.edu/OPD">http://apropos.icmb.utexas.edu/OPD</a></td>
<td></td>
</tr>
<tr>
<td>Metabolomics</td>
<td>METLIN Metabolite Database</td>
<td>The Scripps Research Institute</td>
<td>A repository for mass spectral metabolite data</td>
<td><a href="http://metlin.scripps.edu">http://metlin.scripps.edu</a></td>
</tr>
<tr>
<td>Metabolomics</td>
<td>Human Metabolome Database (HMDB)</td>
<td>Genome Alberta and Genome Canada</td>
<td>Database of small metabolites found in the human body containing or linking chemical, clinical and molecular biology/chemistry data</td>
<td><a href="http://www.hmdb.ca/">http://www.hmdb.ca/</a></td>
</tr>
<tr>
<td>Integrated</td>
<td>Chemical Effects in Biological Systems (CEBS)</td>
<td>NIH-NIEHS</td>
<td>Integrates study design, clinical pathology and histopathology, gene expression and proteomics data from all studies and enables discrimination of critical study factors</td>
<td><a href="http://cebs.niehs.nih.gov/">http://cebs.niehs.nih.gov/</a></td>
</tr>
<tr>
<td></td>
<td>PharmGKB</td>
<td>Stanford</td>
<td>Collects, encodes and disseminates knowledge about the impact of human genetic variations on drug responses. Provides curated primary genotype and phenotype data, annotated gene variants, and gene-drug-disease relationships via literature review, and summarizes important PGx genes and drug pathways</td>
<td><a href="http://www.pharmgkb.org/">http://www.pharmgkb.org/</a></td>
</tr>
<tr>
<td></td>
<td>Kyoto encyclopaedia of genes and genomics pathways (KEGG)</td>
<td>Kyoto University, Japan</td>
<td>KEGG is an integrated database that relates genes to metabolic pathways; outlines functional relationships among genes; disease-associated genes; pharmaceuticals-targeted genes</td>
<td><a href="http://www.genome.jp/kegg">http://www.genome.jp/kegg</a></td>
</tr>
</tbody>
</table>
References


