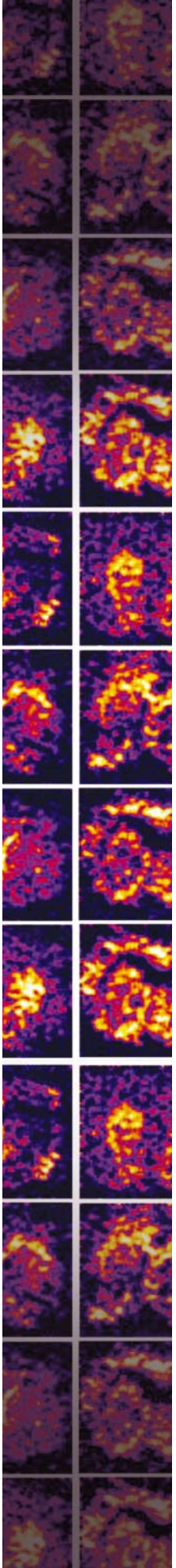

Biomedical informatics for proteomics

Mark S. Boguski* & Martin W. McIntosh†

**Human Biology Division, and †Public Health Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, PO Box 19024, Seattle, Washington 98109, USA (e-mail: mboguski@fhcrc.org; mmcintos@fhcrc.org)*

Success in proteomics depends upon careful study design and high-quality biological samples. Advanced information technologies, and also an ability to use existing knowledge to the full, will be crucial in making sense of the data. Despite its genome-scale potential, proteome analysis is at a much earlier stage of development than genomics and gene expression (microarray) studies. Fundamental issues involving biological variability, pre-analytic factors and analytical reproducibility remain to be resolved. Consequently, the analysis of proteomics data is currently informal and relies heavily on expert opinion. Databases and software tools developed for the analysis of molecular sequences and microarrays are helpful, but are limited owing to the unique attributes of proteomics data and differing research goals.

The subtitle of a recent conference on the Human Proteome Project asserted that "Genes Were Easy"[1]. Depending upon one's perspective, this statement might elicit feelings of hubris, envy or fear about the challenges and complexities of another ostensible paradigm shift in biomedical research. We have transitioned rapidly from the momentary comfort of a large, but finite and complete human genome to a seemingly infinite biological universe of post-transcriptional complexities[2–4].

Proteomics is often referred to as a 'post-genome' science, but its antecedents actually predate the Human Genome Project by two to three decades and developed along different intellectual lines[5,6]. Bioinformatics, although enjoying its ascendancy during the earlier days of genome sequencing[7,8], also traces its roots to a time long before the development of cloning and sequencing technologies, when protein primary structures were determined experimentally and not derived routinely and automatically from conceptual translations of coding DNA[9–11]. Although medical informatics[12] has until recently been largely detached from bioinformatics, the emergence of clinical genomics and proteomics increasingly requires the integrated analysis of genetic, cellular, molecular and clinical information and the expertise of pathologists, epidemiologists and biostatisticians.

Proteomics is the latest functional genomics[13] technology to capture our imagination and it is instructive to review some lessons learned during the earlier adoption of another functional genomics technology, namely gene expression analysis using microarrays and similar technologies[14,15]. Study design and sample quality, databases, data analysis and data standards are discussed with special emphasis on human plasma and serum proteomics (Box 1) because of the enormous potential of these studies to advance clinical diagnostics and therapeutic monitoring (see review in this issue by Hanash, page 226).

There are many implications of biomedical informatics for proteomics, including multiple platform technologies (for example, two-dimensional polyacrylamide gel electrophoresis, mass spectrometry, protein and antibody arrays), laboratory information-management systems, medical records systems, and documentation of clinical trial results for regulatory agencies. In the present work, we confine our discussions to mass spectrometry-based proteomics (and see accompanying review by Aebersold and Mann, page 198), and to study design and data resources, tools and analysis in a research setting.

Study design and sample quality

Potter[16] describes four study designs illustrating, for example, the difference between a clinical trial (experiment) and cross-sectional (observational) study (Fig. 1). There are critical differences between experimental biomedicine and epidemiological studies and in many microarray gene expression studies "the distinction between observational and experimental designs is not made"[16]. The same arguments hold true for proteomics studies. Indeed, most gene expression and proteomic analyses involving human specimens will, of necessity, be observational studies and this fact immediately raises the key issues of possible biases and confounding factors in the populations from which the samples are drawn. Plasma and serum proteomics, defined as the discovery and utilization of biomarkers in clinical blood specimens, provides an illustrative case in point.

Human plasma and serum proteomics may be particularly susceptible to observational biases because any confounding factor (such as smoking, diet or ascertainment bias) could conceivably cause a phenotypic response that might be confused with a specific characteristic of the disease process under study. Without careful sample ascertainment and/or the availability of detailed sample annotation, the conclusion of any such study can be misleading. For example, consider the task of acquiring serum samples in an attempt to identify, through proteomic analysis, diagnostic biomarkers that can differentiate cancer patients from healthy subjects. It is common and convenient to collect the disease specimens during surgery, whereas the control subjects do not typically donate their specimens in the operating room. Such a design results in complete confounding between specimen ascertainment and disease status and so it is impossible to determine whether any finding reflects a marker for disease or instead is a marker for pre-operative fasting, anaesthesia, psychological stress, or some other uncontrollable confounding phenomenon. Indeed, in quantitative terms, it is even known that systematic

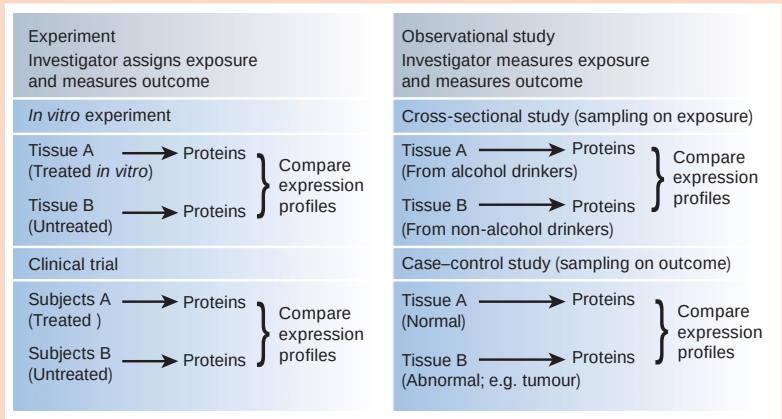

Figure 1 Experimental versus observational study. These study types differ by the manner in which an exposure, or treatment, is assigned to the study subjects. Assignment in experimental studies is under the control of the experimenter, whereas they have no control over treatment assignment in observational studies. Many clinical genomics studies are retrospective and observational, relying upon data from patient medical records to provide information on the relevant phenotype as well as pre-analytical variables and potential confounders. (Figure reproduced from ref. 16.)

differences in tourniquet application time, exercise, and whether the sample is obtained while the subject is sitting or recumbent can each individually induce a change in total protein concentrations by ± 10% (ref. 17).

It is at least as important for proteomics researchers to understand basic epidemiology as it is to understand complex analytic algorithms. Although pre-analytical errors can conceivably occur in samples of any biological nature or origin, here we have highlighted human serum proteome analysis because the nature of protein discovery work in sera means that confounding variables that affect any tissue or pathway may complicate study findings. Moreover, unlike traditional studies with low-dimensional measurements, none of the analytic methods we found allow any possibility to adjust for confounding even if confounders are annotated. Presently, careful design and specimen ascertainment may be the only way to have confidence in study findings with human subjects.

Two additional issues, associated with specimens of almost any kind, are sample quality and number. Quality involves both the preservation of molecular features (such as intact and representative messenger RNAs and proteins) and the assurance of both inter- and intra-sample homogeneity. For example, Huang *et al.*[18] have shown that the duration of ischaemia associated with surgical resection of tissues has significant effects on gene expression. Craven and Bank[19,20] describe some aspects of sample heterogeneity in proteomics, and methods (such as laser-capture microdissection) to address it.

In referring to functional genomics technologies and their relevance in clinical medicine, Margolin has admonished[21] that "Scientists...need to avoid the tendency, often driven by the high price of some of the newer techniques, of running under-controlled experiments or experiments with fewer repeated conditions than would have been accepted with standard techniques." The same caveat applies for proteomics research, but perhaps even more so because a framework to estimate efficient sample sizes has yet to be determined, and the nature of the technology creates substantial challenges to progress in achieving this goal. For example, with microarrays, because the number of interrogations is determined pre-experimentally by the number of genes or gene-specific probes on the array, the confidence in declaring one or a group of genes as differentially expressed can be quantified (using statistical *P* values) by reporting or controlling the rate of false identification[22]. Proteomics discovery with complex mixtures like sera has no such *a priori* enumeration of targets, and the discovery procedure is iterative and far more informal. The lack of a described procedural structure at this time makes it difficult to make any statement about the confidence of any finding.

Protein databases

Collections of protein sequences date back to the 1960s[23], preceding GenBank by nearly 20 years[24]. Since the early 1990s, important utilitarian goals of protein databases have included minimal redundancy, maximal annotation and integration with other databases[25]. These principles continue to be stressed today[26]. For both historical and practical reasons, current molecular sequence databases are designed to represent a comprehensive 'parts list' of an organism's genome, that is, the genes and all of the proteins they encode, and protein 'families' are usually classified according to their evolutionary history inferred from sequence homology. These databases are thus excellent tools for gene discovery, comparative genomics and molecular evolution, but there is much work to be done to even minimally serve the needs of proteomics and integrative biological science[27,28].

Today's principal protein databases emphasize molecular and cellular features and annotation and are not well suited to represent physiology. For example, there are approximately 500 known human serum proteins[17,29] with extensive information about normal and abnormal 'reference' values in health and disease[30]. But simple searches of the popular protein databases, SWISSPROT and LocusLink[26], using the terms 'human serum' and 'serum' yielded only 36–44 and 81–268 matches, respectively, and among the latter are many false positives. Similarly, 68–84 and 457–3,850 proteins are retrieved when 'human plasma' and 'plasma' are used to search SWISSPROT and LocusLink, respectively, again with many false positives.

This is just one example of the fact that there is no reliable or satisfying way to retrieve groups of proteins based upon well-known pathways or functional classifications (for example, coagulation, complement fixation or proteinase inhibitors). Furthermore, annotations about post-translational modifications are sparse and difficult to locate in any consistent way, although some progress is being made[31]. There is also the challenge of distinguishing annotations based upon modifications predicted from protein motifs compared with those based upon direct experimental evidence.

A more ideal database for plasma proteome studies would classify proteins from a functional, rather than an evolutionary, viewpoint (perhaps based upon an updated version of the Putnam classification, as discussed in ref. 32). Such a database would also annotate protein concentrations (and other practically measurable attributes) compared with normal ranges of values in reference samples. Attention to emerging data standards (Box 2) will also be important.

Protein identification by database searching

Until recently, the overarching purpose of database similarity searching was the sensitive detection of sequence homologues, regardless of the species or remoteness of the relationship, in order to infer similarity of function from similarity of sequence and/or to study the evolution of protein families or domains. The specific aims of most proteomics studies are different and therefore require different strategies and tools. For example, in the analysis of human serum,

Box 1
Glossary

Case-control and cohort study. These observational studies differ in the way study subjects are selected. Case-control studies select study subjects based on presence (cases) or absence (controls) of the phenotype (for example, disease) of interest. Cohort studies select participants based on the presence or absence of a risk factor of interest and subjects are followed over time for the development of an outcome of interest.

Confounder/Confounding. A confounder is a variable that distorts an apparent relationship between an exposure and a phenotype of interest. Confounding occurs when the relationship between an outcome (for example, disease) and an exposure of interest cannot be distinguished from other variables that also correlate with the outcome.

Plasma and serum. Plasma is the fluid, non-cellular portion of blood; serum is the protein solution remaining after blood or plasma has been allowed to coagulate. Serum thus lacks clotting-factor proteins. Blood samples are often treated with preservatives, anticoagulants and other additives prior to transport and storage or processing. These are some of the pre-analytical variables that may affect subsequent analyses.

Pre-analytical variables. These refer to those factors, both known and unknown, that may be present in a subject or may arise in any of the steps prior to a laboratory test and data analysis. Examples include genotype, physiological attributes such as age, gender, reproductive status, lifestyle effects (for example, diet or smoking), drugs and specimen collection, handling and processing protocols. Uncontrollable variables must be well understood in order to be able to separate their effects from the object or process under study. Most errors in clinical laboratory tests are known to occur in the pre-analytical phase[55,56].

Randomized clinical trial. An experimental study in which treatments are randomly assigned to subjects as a method to prevent treatment choice from being confounded.

one is interested in identifying proteins that are not normally present and/or variances in the concentrations of the normal constituents. The object of a database search in this case is to find an exact, or nearly exact, match between subsequences (peptide fragments) of serum components and those proteins encoded by the large (but finite) human genome. Weak similarities and interspecies matches are not pertinent, except in the case of 'foreign' proteins encoded by infectious organisms and parasites that may be released into the circulation. Statistical significance is important, but not in the sense of the probability that two sequences are related by chance. Rather, one is seeking an answer to the question of whether the presence or absence of a particular protein, at a particular concentration, deviates significantly from a normal range of values. If the condition is met, one is then interested in attempting to demonstrate a significant correlation between this protein and a risk factor or outcome of interest. One must bear these purposes in mind when attempting to use existing databases and search tools in a proteomics context.

There are several approaches that utilize mass spectrometry for protein and peptide analysis[33,34]. These include analytic peptide-mass fingerprinting, *de novo* sequence interpretation and comparative analysis of actual spectra with predicted spectra of peptide sequences from a protein database. The first and third of these methods use comparisons against a database and the reliability of any database search depends on the accuracy and resolution of data, quality of the sequence database, and that of the scoring algorithm used. The accuracy of the input data is affected by many factors that are unique to mass spectrometry compared with DNA sequencing and conceptual translations to protein[33,34]. For example, co- and post-translational modifications[2,35] of amino acid residues obviously affect the masses of real peptides and cannot be predicted consistently or reliably for virtual peptide sequences, although some search engines use error-tolerant heuristics in an attempt to take potential modifications into account[36,37]. Additionally, the effects of inaccuracy and discontinuity in both expressed sequence tag data[38] and genomic data[39], and thus in their encoded peptide sequences, have received some attention.

Selecting which of many candidate spectra is correct involves scoring the similarity of the observed and predicted spectra. Detailed consideration of specific scoring algorithms is beyond the scope of this review and more specific descriptions are found elsewhere[40–44]. In general, each scoring algorithm designates a quantity related to the probability that the candidate peptide could have produced the observed spectrum by chance. When the number of peptides to identify is small it is feasible for a skilled operator to evaluate all high-scoring candidate peptides manually and make assignments using their expert opinion. For each possible peptide spectrum this score is commonly used to rank the candidate peptides. But manual scoring for complex mixtures is not feasible. Instead, it is common to use the scoring algorithm to rank the candidates and assign only the highest scoring of all. This of course makes automated proteomics highly dependent on the quality of the scoring algorithm used.

Moreover, automated identification based on ranking peptides by their scores is not directly analogous to the well established procedure of ranking expressed genes on microarrays based on their *P* values[45,46], because peptide scores are not true *P* values even though they may fall between 0 and 1. Consider a simple example of three candidate peptides labelled PA, PB and PC, which together produce any of the three possible spectra S1, S2 and S3. We consider PB intermediate to PA and PC in that it may produce any of the three spectra (each with a probability of 1/3), but PA may produce either S1 or S2 (each with a probability of 1/2) but not S3 (probability of 0). Likewise, PC may produce either S2 or S3 (each with a probability of 1/2) but not S1 (probability of 0). Now with each possible experimental observation — S1, S2 or S3 — consider which peptide will achieve the highest-ranking score. The score for PB can never achieve a value higher than 1/3 because each of the three spectra are equally likely and the values of PA and PB will score either 0 or 1/2. Indeed, because all observations of at least one PA or PC will score 1/2, peptide PB will never achieve the highest rank (Table 1). Thus, even when considering a mixture rich in peptide PB, the latter will never achieve the highest rank but will instead be misidentified as PA or PC.

Automated peptide identification can be improved if the number of peptide choices (that is, the complexity of the mixture) is reduced. For instance, in the example above, peptide PB could possibly achieve the highest rank if either PA or PC is eliminated as a possible alternative. In simple mixtures, human operators can reduce the complexity by auditing the highest-ranking peptides and using their informal expert opinion to eliminate some of the highest-ranking peptides. Because manual review such as this is not feasible with highly complex mixtures such as sera, some investigators have begun to develop methods to formalize expert opinion and use it in more complex scoring algorithms that can automatically eliminate, or reduce in rank, peptides that would otherwise achieve a high rank. For example, Bafna *et al.*[40] give an example of how experienced spectrometrists, recognizing in the spectrum so-called neutral losses of water or ammonia from side chains of amino acids, can distinguish among peptide candidates that possess similar high

Table 1 **Scoring and peptide ranking in a simple mixture of three peptides**

Peptide	Possible spectra	Peptide score (probability) for observed spectra		
		S1	S2	S3
PA	S1, S2	1/2	1/2	0
PB	S1, S2, S3	1/3	1/3	1/3
PC	S2, S3	0	1/2	1/2
Peptide assignment		PA	PA or PC	PC

Box 2
Data standards

There are numerous examples in information management and processing where the existence of multiple and/or specialized file formats has hindered accessibility, information exchange and integration. The functional genomics (microarray) field provides a pertinent model for the development of standards that greatly enhance the opportunities for data access and exchange, data integration and meta-analysis[57,58]. Adherence to the MIAME standard (for 'minimum information about a microarray experiment') for microarray data is now required for manuscript submission to all *Nature* journals[59] and *Science* also supports this "evolving standardization"[60]. The Human Proteome Organisation is currently engaged in a proteomics standards initiative[61] to develop formats for mass spectrometry and protein–protein interaction data and annotation. These formats use eXtensible Markup Language (XML), which is an Internet standard for describing structured and semi-structured data. An earlier standard, ASN.1, has been used by the National Center for Biotechnology Information for years to transfer and integrate structured data and has more recently been utilized by data resources such as BIND[62]. Most of the main database providers now make their data (for example, sequences, structures, gene expression profiles and PubMed records) available in XML. Nearly all software vendors are implementing a standard suite of extensions based on XML and web services that make it easy to publish and exchange XML data. This common software base will revolutionize the way data is accessed and used online by liberating data from the software applications that created it[63,64].

scores. Bafna *et al.* go on to describe an approach to formalize this and other expert opinions and include them in a complex scoring algorithm. The information required to implement the algorithm is substantial, such as the need to specify the probability of peptide-fragmentation patterns (which are instrument dependent), but such information may be essential to achieve the goal of better operator-independent peptide identifications in complex mixtures.

Another challenge to automating proteomics with complex mixtures is to decide when even the best match of a scoring algorithm is simply not good enough. Better yet would be an approach to state the certainty that an identified match is correct, because not all assignments between peptides and a candidate will be correct. Indeed, with complex mixtures the peptides without correct assignment will likely greatly outnumber those with correct assignments[47]. Establishing a criteria for acceptance overall therefore becomes the main focus of automated proteomics. What criteria should be used to decide whether to accept or reject the assignment deemed 'most-likely' by the scoring algorithm? It is generally assumed that higher-scoring assignments are more likely to be correct than lower-scoring assignments, and so it is common to designate a single score threshold above which all assignments will be accepted. But unlike true *P* values, the score value conveys no information about the actual quality of the match and so it is not possible to directly ascertain the performance characteristics of any specific choice of threshold. For example, the dogma of accepting hypotheses based on *P* values less than 0.05 means, by definition, 5% of all false tests will be misidentified as true. Without such an interpretation of a scoring algorithm, the quality of a match based on automatic scoring cannot be assessed and errors cannot be controlled.

It is essential that some agreed-upon criteria be developed for reporting the quality of any peptide assignment. Any specific threshold could be characterized by its sensitivity (the rate of accepting accurate peptides assignments) and its specificity (the rate of rejecting inaccurate peptide assignment). Of course the specificity and sensitivity of any threshold will depend on the mixture and the sequence database, because these will affect the distribution of scores among true and false matches. One proposal to control the performance of automated matching has been given by Keller *et al.*[47] who make the observation that it may be possible to determine the sensitivity and specificity of any assignment using unsupervised, model-based clustering techniques. Keller *et al.* estimate the reference distributions of the correct and incorrect assignments within any experiment. Importantly, their proposal identifies thresholds in an experiment- and database-dependent manner so that a series of experiments can use comparable criteria. In essence, Keller *et al.* describe an approach that may allow a scoring algorithm to be converted into *P* value-like quantities that can then be used to control error rates.

Pattern matching without protein identification

Recently, substantial attention has been given to using chromatography-based proteomics to measure the concentration of low molecular weight peptides in complex mixtures, such as plasma or sera. These technologies commonly use time-of-flight (TOF) spectroscopy with matrix-assisted or surface-enhanced laser desorption/ionization, to produce a spectrum of mass-to-charge (m/z) ratios that can be analysed in order to identify unique signatures from its chromatography pattern. Each m/z value of the spectrum reflects the abundance of possibly many peptides having a similar mass. Thus, with complex mixtures, these TOF methods are not able to identify individual peptides.

When used with complex mixtures, analysis methods are intended to identify peaks, or features, of the spectrum that can segregate identifiable groups; in this way they are similar to unsupervised learning approaches commonly used when evaluating expression arrays[48–51]. However, because of experimental variation of those spectra, expression array clustering methods are appropriate only if alignment and peak identification and selection algorithms are first used. Adam *et al.*[48] take this approach to ensure that the features they identify as important are actual peaks. Another approach, used for example by Petricoin *et al.*[49,50], is to avoid peak identification all together and accommodate experimental variation in the clustering algorithm. By ignoring peak identification, the resulting classification may produce an algorithm more suitable for prediction, but the features identified may not correspond to actual peaks at all and so this approach may be less useful if eventual peptide identification is the goal.

Even though the TOF algorithms have not yet led to peptide identification, this factor does not greatly limit their utility for identifying newer and far more accurate approaches for medical diagnostics, because diagnosing disease is a problem of prediction rather than of aetiology. Algorithms that have potential clinical relevance have already been identified by Petricoin *et al.*[50] and Adam *et al.*[48] for diagnosing ovarian and prostate cancer, respectively. The excitement surrounding these TOF technologies is also due in part to their requiring only very small volumes of specimen (typically less than 50 μl) to generate their spectra. This is especially true for studies that rely on limited, and therefore precious, supplies of archival specimens[52]. The efficiency of the TOF approaches, and their demonstrated ability to generate highly accurate diagnostic tests in case-control studies, may provide considerable advantages for this technology compared with others for the development of medical diagnostics.

Conclusions and future challenges

Proteomics is a powerful, post-genome paradigm that seeks to describe and explain what Erwin Chargaff called the "immensely diversified phenomenology" of cells and organisms[53]. Beyond the enumerations and characterizations of different proteomes lies the elucidation of macromolecular interactions, complexes and networks. Informatics will play a crucial role in working towards these goals. Should we be optimistic? To paraphrase

N. K. Hayles, "... annotations, insofar as they represent informational patterns abstracted from their instantiation in a biological substrate, can never fully capture the embodied actuality, unless they are as prolix and noisy as the body itself"[54]. Well, we shall do the best we can. □

doi:10.1038/nature01515

1. Cambridge Healthtech Institute Conference on Human Proteome Project, 2–4 April 2001, McLean, Virgina <http://www.healthtech.com/2001/hpr/index.htm> (2001).
2. Krishna, R. G. & Wold, F. Post-translational modification of proteins. *Adv. Enzymol. Relat. Areas Mol. Biol.* **67,** 265–298 (1993).
3. Keegan, L. P., Gallo, A. & O'Connell, M. A. The many roles of an RNA editor. *Nature Rev. Genet.* **2,** 869–878 (2001).
4. Maniatis, T. & Tasic, B. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418,** 236–243 (2002).
5. Dayhoff, M. O. & Eck, R. V. MASSPEC: a computer program for complete sequence analysis of large proteins from mass spectrometry data of a single sample. *Comput. Biol. Med.* **1,** 5–28 (1970).
6. Anderson, N. G., Matheson, A. & Anderson, N. L. Back to the future: the human protein index (HPI) and the agenda for post-proteomic biology. *Proteomics* **1,** 3–12 (2001).
7. Boguski, M. S. Bioinformatics. *Curr. Opin. Genet. Dev.* **4,** 383–388 (1994).
8. Boguski, M. S. The turning point in genome research. *Trends Biochem. Sci.* **20,** 295–296 (1995).
9. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8,** 357–366 (1965).
10. Dayhoff, M. O. Computer aids to protein sequence determination. *J. Theor. Biol.* **8,** 97–112 (1965).
11. Doolittle, R. F. Some reflections on the early days of sequence searching. *J. Mol. Med.* **75,** 239–241 (1997).
12. Shortliffe, E. *et al.* (eds) *Medical Informatics: Computer Applications in Health Care and Biomedicine* (Springer, New York, 2000).
13. Hieter, P. & Boguski, M. Functional genomics: it's all how you read it. *Science* **278,** 601–602 (1997).
14. Duyk, G. M. Sharper tools and simpler methods. *Nature Genet.* **32**(Chipping Forecast II Suppl.), 465–468 (2002).
15. Kohane, I. S., Kho, A. T. & Butte, A. J. *Microarrays For an Integrative Genomics* (Massachusetts Institute of Technology Press, Cambridge, MA, 2003).
16. Potter, J. D. At the interfaces of epidemiology, genetics and genomics. *Nature Rev. Genet.* **2,** 142–147 (2001).
17. McClatchey, K. D. (ed.) *Clinical Laboratory Medicine* (Lippincott, Philadelphia, 2002).
18. Huang, J. *et al.* Effects of ischemia on gene expression. *J. Surg. Res.* **99,** 222–227 (2001).
19. Craven, R. A. & Banks, R. E. Laser capture microdissection and proteomics: possibilities and limitation. *Proteomics* **1,** 1200–1204 (2001).
20. Craven, R. A. & Banks, R. E. Use of laser capture microdissection to selectively obtain distinct populations of cells for proteomic analysis. *Methods Enzymol.* **356,** 33–49 (2002).
21. Margolin, J. From comparative and functional genomics to practical decisions in the clinic: a view from the trenches. *Genome Res.* **11,** 923–925 (2001).
22. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57,** 289–300 (1995).
23. Dayhoff, M. O. & Eck, R. V. *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Silver Spring, MD, 1966).
24. Smith, T. F. The history of the genetic sequence databases. *Genomics* **6,** 701–707 (1990).
25. Bairoch, A. & Boeckmann, B. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* **19**(Suppl.), 2247–2249 (1991).
26. Maglott, D. R. *et al.* NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* **28,** 126–128 (2000).
27. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25,** 25–29 (2000).
28. Bader, G. D. *et al.* BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Res.* **29,** 242–245 (2001).
29. Adkins, J. N. *et al.* Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. *Mol. Cell. Proteomics* **1,** 947–955 (2002).
30. Kratz, A. & Lewandrowski, K. B. Case records of the Massachusetts General Hospital. Weekly clinicopathological exercises. Normal reference laboratory values. *N. Engl. J. Med.* **339,** 1063–1072 (1998).
31. Jung, E. *et al.* Annotation of glycoproteins in the SWISS-PROT database. *Proteomics* **1,** 262–268 (2001).
32. Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1,** 845–867 (2002).
33. Chakravarti, D. N., Chakravarti, B. & Moutsatsos, I. Informatic tools for proteome profiling. *Biotechniques* **32**(Comput. Proteomics Suppl.), S4–S15 (2002).
34. Liebler, D. C. *Introduction to Proteomics* (Humana, Totowa, NJ, 2002).
35. The Association of Biomolecular Resource Facilities. Delta Mass: A Database of Protein Post Translational Modifications <http://www.abrf.org/index.cfm/dm.home> (2002).
36. Wilkins, M. R. *et al.* High-throughput mass spectrometric discovery of protein post-translational modifications. *J. Mol. Biol.* **289,** 645–657 (1999).
37. Creasy, D. M. & Cottrell, J. S. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* **2,** 1426–1434 (2002).
38. Choudhary, J. S. *et al.* Matching peptide mass spectra to EST and genomic DNA databases. *Trends Biotechnol.* **19**(Suppl.), S17–S22 (2001).
39. Choudhary, J. S. *et al.* Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* **1,** 651–667 (2001).
40. Bafna, V. & Edwards, N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* **17**(Suppl.) S13–S21 (2001).
41. Eng, J., McCormack, A. & Yates, J. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989 (1994).
42. Fenyo, D. Identifying the proteome: software tools. *Curr. Opin. Biotechnol.* **11,** 391–395 (2000).
43. Field, H. I., Fenyo, D. & Beavis, R. C. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* **2,** 36–47 (2002).
44. Perkins, D. N. *et al.* Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567 (1999).
45. Efron, B. & Tibshirani, R. Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **23,** 70–86 (2002).
46. Pepe, M. S. *et al.* Selecting differentially expressed genes from microarray experiments. *Biometrics* (in the press).
47. Keller, A. *et al.* Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392 (2002).
48. Adam, B. L. *et al.* Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* **62,** 3609–3614 (2002).
49. Petricoin, E. F. III *et al.* Serum proteomic patterns for detection of prostate cancer. *J. Natl Cancer Inst.* **94,** 1576–1578 (2002).
50. Petricoin, E. F. *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359,** 572–577 (2002).
51. Qu, Y. *et al.* Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin. Chem.* **48,** 1835–1843 (2002).
52. Pepe, M. S. *et al.* Phases of biomarker development for early detection of cancer. *J. Natl Cancer Inst.* **93,** 1054–1061 (2001).
53. Judson, H. *The Eighth Day of Creation: Makers of the Revolution in Biology* expand. edn (Cold Spring Harbor Laboratory Press, New York, 1996)
54. Hayles, N. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics* (University of Chicago Press, Chicago, 1999).
55. Bonini, P. *et al.* Errors in laboratory medicine. *Clin. Chem.* **48,** 691–698 (2002).
56. Narayanan, S. The preanalytic phase. An important component of laboratory medicine. *Am. J. Clin. Pathol.* **113,** 429–452 (2000).
57. Spellman, P. T. *et al.* Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3,** 46 (2002).
58. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genet.* **29,** 365–371 (2001).
59. Editorial. Coming to terms with microarrays. *Nature Genet.* **32,** 333–334 (2002).
60. Ball, C. *et al.* Standards for Microarray Data. *Science* **298,** 539 (2002).
61. Orchard, S., Kersey, P., Hermjakob, H. & Apweiler, R. The HUPO proteomics standards initiative meeting: towards common standards for exchanging proteomics data. *Comp. Funct. Genom.* **4,** 16–19 (2003).
62. Bader, G. D. & Hogue, C. W. BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **16,** 465–477 (2000).
63. Abiteboul, S., Buneman, P. & Suciu, D. *Data on the Web: From Relations to Semistructured Data and XML* (Morgan Kaufmann, San Francisco, 2000).
64. Coyle, F. *XML, Web Services, and the Data Revolution* (Addison-Wesley, Boston, 2002).

Acknowledgements We thank L. Hartwell, J. Potter and G. Omenn for stimulating discussions and J. Gray, J. Pounds and L. Geer for valuable suggestions and critical readings of the manuscript.