On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning

Douglas B. Kell and Ross D. King

At present, the assignment of function to novel genes uncovered by the systematic genome-sequencing programmes is a problem. Many studies anticipate that this can be achieved by analysing patterns of gene expression via the transcriptome, proteome and metabolome. Thus, functional genomics is, in part, an exercise in pattern classification. Because many genes have known functional classes, the problem of predicting their functional class is a supervised learning problem. However, most pattern classification methods that have been applied to the problem have been unsupervised clustering methods. Consequently, the best classification tools have not always been used. Furthermore, the present functional classes are suboptimal and new unsupervised clustering methods are needed to improve them. Better-structured functional classes will facilitate the prediction of biochemically testable functions.

The following quotations provide a useful basis for discussion of this topic. According to Brent, 'Just as development of the telescope and microscope was followed by periods during which science was mostly done by observation rather than experiment, it is possible that the development of gene-expression monitoring and other functional genomic methods may presage a phase in which biology once again becomes more observational… Making new technology work may be easier than using it to discover truth'[1]. Everitt and Dunn provided further thought: 'It is often suggested that it is helpful to recognize that the analysis of data involves two separate stages. The first, particularly in new areas of research, involves data exploration, in an attempt to recognize any non-random pattern or data requiring explanation. At this stage, finding the question is often of more interest than seeking the subsequent answer, the aim of this part of the analysis being to generate possibly interesting hypotheses for later study… A confirmatory analysis becomes possible once a research worker has a well-defined hypothesis in mind'[2]. Furthermore, Everitt states that: 'Any classification is a division of objects into groups based on a set of rules – it is neither true nor false (unlike, for example, a theory) and should be judged largely on the usefulness of the results'[3].

What do we mean by a gene's functional class?

From the systematic genomic-sequencing programmes, recognition that the functions of up to half the genes uncovered are not known is perhaps the central feature of the post-genomic era[4,5]. Currently, the main approaches to the problem of assigning gene function seem to be based implicitly on the view that genes with similar function are likely to be similarly co-expressed ('guilt by association'[1,6]). Further, whole-genome approaches to the analysis of gene function at the level of the transcriptome[7–10], the proteome[11,12] and the metabolome[13] are now considered *de rigeur* for serious functional genomics. In particular, it is possible to knock out individual genes in appropriate model organisms systematically[14] and to compare the patterns of expression of all genes in strains wherein the functions of the gene knocked out (or overexpressed) are nominally known with those in which the gene knocked out (or overexpressed) is of unknown function. Thus, huge datasets are becoming available that we might seek to use as the 'input' to some kind of mathematical modelling or data analysis program, whose 'output' is an assignment to a certain class of gene function.

In view of the likely widespread adoption of these and related methods, we consider that the question of how best, in principle, to assign function to genes from such datasets might itself benefit from a wider consideration of the literature on the classification of objects using multivariate data. In particular, the ability to monitor all gene expression simultaneously using microarray technology makes it especially timely to revisit the question of what we actually mean by a gene's functional class, and our purpose here is to rehearse the relevant arguments.

Class assignment from multivariate data

The essential problem discussed here can be described with reference to Fig. 1, which illustrates a situation wherein a dataset is obtained from what might be a

D.B.Kell (dbk@aber.ac.uk) is at the Institute of Biological Sciences, University of Wales, Aberystwyth, UK SY23 3DD. R.D. King (rdk@aber.ac.uk) is at the Department of Computer Science, University of Wales, Aberystwyth, UK SY23 3DB.

 0167-7799/00/$ – see front matter PII: S0167-7799(99)01407-9

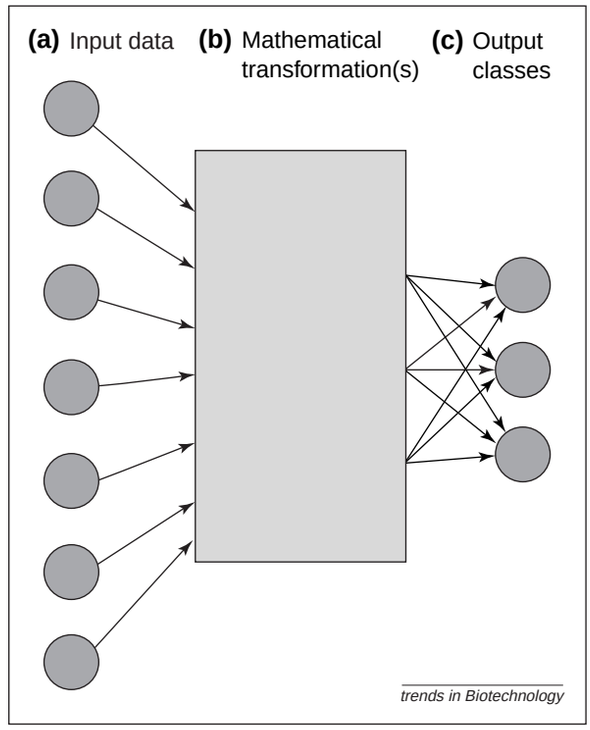

Figure 1

The class assignment problem. The problem is represented by **(a)** input data, whose choice one might hope to optimize; **(b)** mathematical transformations, which use the differential data in the input patterns to establish the correct functional class assignment(s) on the basis of those input patterns; and **(c)** output classes, into which one might wish to assign a sample whose measured properties are to be used as the inputs.

large number of samples, each of which has a large number of datapoints attached to it. These datapoints can be numerical values of gene or protein expression levels, metabolite concentrations or growth rate under certain conditions, binary valued properties (e.g. the ability or otherwise to grow on a particular substrate) or, for our general purposes, any phenotypic properties.

The outputs from this arrangement are several possible functional classes. Between the inputs and the outputs are a series of mathematical transformations that, if applied to the inputs, can be used to generate the outputs. The problem of functional genomics can then be reduced to: (1) choosing the optimal inputs; (2) finding the optimal output classes and their structure; and (3) thereby finding the correct mathematical relationships that use the differential data in the input patterns to establish the correct functional class assignment(s) on the basis of those input patterns. Although (1) and (3) are beginning to be developed, the question of how one might seek to optimize the output classes rationally has not yet been adequately considered.

Present functional classes

A typical functional classification of the type widely available on the Web is that for *Mycobacterium tuberculosis* (Box 1). Similar lists exist for *Escherichia coli*[15,16] and other model organisms (Table 1). Whatever the merits of any particular groupings chosen, we can state the following:

- organisms are sufficiently different that the class structure must be organism specific[17];
- there is clearly substantial arbitrariness in the existing class structures, which are largely based on our existing knowledge[18] of the relevant biochemistry;
- the class structure is based only on the known 'outputs', and is not derived from (let alone optimized against) the different patterns of the 'inputs' that might be observed experimentally;
- the implicit view from the structure and display of these databases is, in most cases, that there is an essentially (strictly) hierarchical relationship between classes; thus in Box 1, 'small-molecule metabolism' is divided into ten classes, each of which is then subdivided at least once, with no gene product appearing in more than one class;
- there are many more genes than we should like in the classes 'other', 'conserved hypotheticals' and 'unknown', and these classes are clearly likely to be highly heterogeneous (indeed, so heterogeneous as to be potentially valueless); and
- the semantics of what is meant by a 'function' need to be considered.

Although these problems are certainly (and explicitly) recognized by those who seek to categorize genes into functional classes, the main questions we ask are as follows.

- Are these the best groupings, how do we decide, and what does 'best' mean?
- What are the best methods for forming mathematical relationships between any input patterns we observe and any functional class(es) to which we may assign them?

Table 1. Examples of websites in which a listing of suggested functional classes has been given

Organism	Web address
Bacillus subtilis	http://bioweb.pasteur.fr/GenoList/SubtiList/help/classif-search.html
Caenorhabditis elegans	http://www.proteome.com/databases/WormPD/WormPDcategories/Functional_Categories.html
Escherichia coli	http://ecocyc.pangeasystems.com/ecocyc/ecocyc.html http://www.genome.ad.jp/dbget-bin/get_htext?E.coli.operon.kegg
Mycobacterium tuberculosis	http://www.sanger.ac.uk/Projects/M_tuberculosis/Gene_list/
Saccharomyces cerevisiae	http://www.mips.biochem.mpg.de/proj/yeast/catalogues/funcat/index.html http://www.proteome.com/databases/YPD/YPDcategories/Functional_Categories.html

Pattern classification methods

There are many areas of science in which pattern classification methods developed in statistics and artificial intelligence are important, and where the arrangement is exactly as shown in Fig. 1. The goal of pattern recognition is to classify objects of interest that possess particular attributes into several categories or classes. The functional genomics agenda is therefore to be seen, in part, as an exercise in pattern recognition.

Pattern classification methods can be grouped into two different categories: unsupervised and supervised learning methods[19–29]. If a set of multivariate observations is given with the aim of establishing the existence of classes in the input data, with no knowledge or care for an imposed class structure, we will be using clustering or unsupervised learning. Alternatively, if there is a defined class structure, the need is then to establish rules by which new objects are correctly classified into one or more of the existing classes. This supervised learning is often referred to as discrimination or multivariate calibration in the statistical literature, as the class structure is produced on the basis of known, correctly classified objects and their attendant properties.

By definition then, the problem of predicting gene functional class is (or may be cast as) a supervised learning problem because many genes have known functional classes. It is therefore surprising that most pattern classification methods that have been applied to the problem are unsupervised, for example, for transcriptome data[6,30–34] (see Ref. 35 for a related counterexample). This is unfortunate because these methods measure only what changes; however, for functional assignment we are interested not in what changes but in which changes matter.

One subclassification of statistical methods is given in Fig. 2. Other subclassifications discriminate between, for example, neural and statistical methods[25]; evolutionary computing techniques[36], various types of tree-based classifiers[37,38] and other machine-learning methods[29] are also important tools. Important early studies that assigned relationships between genes and phenotypes led to the one-gene–one-enzyme paradigm[39], which was perhaps implicitly translated more extensively than was appropriate into one-gene–one-phenotype, although increased interest in traits to which many genes contribute has spurred the analysis of quantitative trait loci[40]. However, where class membership is not strictly defined (i.e. where every object or gene could be a member of more than one class), our class structures must take this into account. Depending on the experimental setup, one approach might be to use the methods of 'fuzzy' computation[41–43]. Well-known examples of this type of behaviour include stimulons and regulons with overlapping specificities (class membership). The problem is particularly acute when the same gene has entirely different functions depending on the environmental conditions. Thus, glycerol kinase is catabolic when glycerol is the growth substrate, but is anabolic under most other conditions[15]. Optimal functional class structures are therefore not fixed, even for a particular organism, but depend on the environmental conditions.

Box 1. A typical set of gene functional classes, as taken from the *Mycobacterium tuberculosis* project (http://www.sanger.ac.uk/Projects/M_tuberculosis/Gene_list/). Classification is strictly hierarchical

Top level list

I. Small-molecule metabolism (1066)
- A. Degradation (163)
- B. Energy metabolism (292)
- C. Central intermediary metabolism (45)
- D. Amino acid biosynthesis (95)
- E. Polyamine synthesis (1)
- F. Purines, pyrimidines, nucleosides and nucleotides (60)
- G. Biosynthesis of cofactors, prosthetic groups and carriers (117)
- H. Lipid biosynthesis (65)
- I. Polyketide and non-ribosomal peptide synthesis (41)
- J. Broad regulatory functions (187)

II. Macromolecule metabolism (662)
- A. Synthesis and modification of macromolecules (215)
- B. Degradation of macromolecules (87)
- C. Cell envelope (360)

III. Cell processes (206)
- A. Transport/binding proteins (123)
- B. Chaperones/heat shock (16)
- C. Cell division (19)
- D. Protein and peptide secretion (14)
- E. Adaptations and atypical conditions (12)
- F. Detoxification (22)

IV. Other (469)
- A. Virulence (38)
- B. IS elements, repeated sequences and phage (135)
- C. PE and PPE families (167)
- D. Antibiotic production and resistance (14)
- E. Bacteriocin-like proteins (3)
- F. Cytochrome P450 enzymes (22)
- G. Coenzyme F420-dependent enzymes (3)
- H. Miscellaneous transferases (61)
- I. Miscellaneous phosphatases, lyases, and hydrolases (18)
- J. Cyclases (6)
- K. Chelatases (2)

V. Conserved hypotheticals (915)

VI. Unknowns (606)

The next level of subdivision for class I c

C. Central intermediary metabolism (45)
- General (13)
- Gluconeogenesis (2)
- Sugar nucleotides (14)
- Amino sugars (1)
- Sulphur metabolism (15)

The next level of subdivision for one of the subclasses of class c

Sugar nucleotides (14)
- Rv1512 *epiA* nucleotide sugar epimerase
- Rv3784 *epiB* probable UDP-galactose 4-epimerase
- Rv1511 *gmdA* GDP-mannose 4,6 dehydratase
- Rv0334 *rmlA* glucose-1-phosphate thymidyltransferase
- Rv3264c *rmlA2* glucose-1-phosphate thymidyltransferase
- Rv3464 *rmlB* dTDP-glucose 4,6-dehydratase
- Rv3634c *rmlB2* dTDP-glucose 4,6-dehydratase
- Rv3468c *rmlB3* dTDP-glucose 4,6-dehydratase
- Rv3465 *rmlC* dTDP-4-dehydrorhamnose 3,5-epimerase
- Rv3266c *rmlD* dTDP-4-dehydrorhamnose reductase
- Rv0322 *udgA* UDP-glucose dehydrogenase/GDP-mannose 6-dehydrogenase
- Rv3265c *wbbL* dTDP-rhamnosyl transferase
- Rv1525 *wbbl2* dTDP-rhamnosyl transferase
- Rv3400 – probable β-phosphoglucomutase

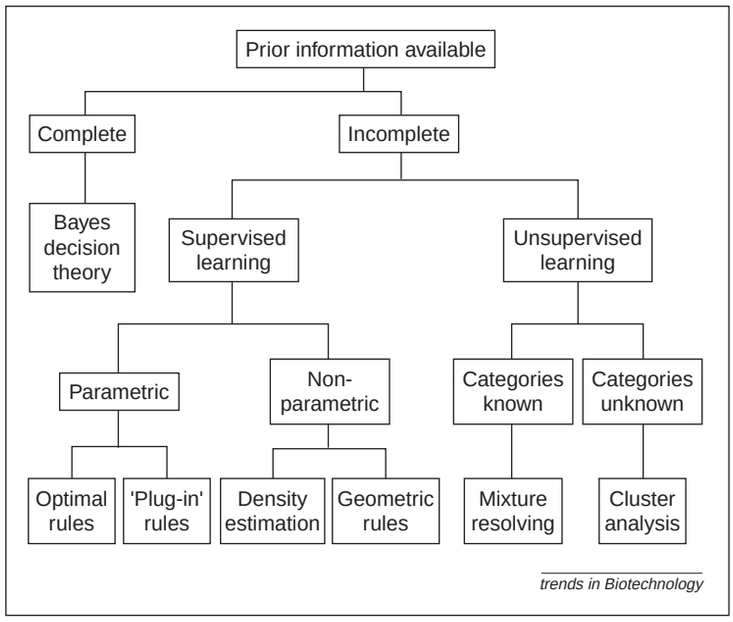

Figure 2

One categorization of methods that can be used for attacking problems in statistical pattern recognition[20].

What can (microbial) taxonomy tell us?

Taxonomy (particularly numerical taxonomy) is the field that is perhaps most relevant to a biologist concerned with the classification of objects on the basis of a series of properties. A particular point to be made about taxonomic methods, in general, is that they use experimentally observed data as the inputs and produce the (purportedly optimal) classes as the outputs. Numerical taxonomy is a form of unsupervised learning and uses classification programs to generate taxonomic classes. This contrasts with the current class structures available for functional assignments, which are generated by hand rather than by computer, and are not based *a priori* on whole-genome-derived phenotypic data or whole-genome-expression studies.

According to Gilmour, 'natural taxa are thought to be those that are most highly predictive overall – that is, not for a special purpose but in terms of several logically independent statements that can be made concerning its members'[44]. We agree with Everitt[3] that this should be the preferred aim; this is also the broad view of the algorithmic information theory[45,46], which is arguably the best theory of unsupervised learning.

Sokal[44] also states that 'cladists consider supraspecific taxa as real entities, not as classes, and therefore to them a natural taxon is a taxon that exists in nature independent of man's ability to perceive it', and that 'by contrast, phenetic numerical taxonomists consider supraspecific taxa as classes and are therefore not especially concerned with their reality'. Again this phenetic view seems most suitable for our purposes.

Although this is not the place to discuss in detail the axioms of numerical taxonomy[47,48], it is worth listing a few of those that we consider significant and potentially of general use for our present needs:

- a great many characters should be used – this provides robustness against misclassification;
- groups should be seen as polythetic, that is, the class membership should be based on the largest possible number of shared characters and not on the presence or absence of single characters (these two points are based on the statistical principle that all information relevant to a problem should be used in solving it; http://omega.albany.edu:8008/JaynesBook.html);
- equal weighting is to be given initially to all characters measured (the infamous statistical principle of indifference); and
- because we are not trying (and indeed are trying not) to conform to any predetermined class structure we must take an empirical, data-driven, operational approach[49].

Phylogenetic methods that are based on the analysis of macromolecular sequences[50,51] are bound up so intimately with the questions of evolution that they do not seem suitable for our purposes. Indeed, the biggest (and effectively insuperable) problem with hijacking classical taxonomic methods lies in their assumptions that a tree structure is appropriate, and that such trees must by and large be hierarchical, with each organism or output class appearing in only one place.

The structures needed for functional genomics are different because many genes are involved in many responses. To this end, it is worth mentioning that metabolic control analysis[52–57] tells us (*a priori*, and even under conditions in which the changes in enzyme activity are small) that: (1) at the metabolic level, all metabolites will change their concentrations in a fashion described by the concentration–control coefficients; and (2) these changes in concentration can be large. Similarly, experimental studies at the level of the proteome show that changes in the expression of single genes can affect the levels of potentially hundreds of proteins[58]; transcriptome studies also demonstrate (and in view of the above necessarily show) that ostensibly minor changes in physiological state are accompanied by changes in the levels of hundreds of transcripts[7]. For our purposes, better structures are directed acyclic graphs (DAGs).

Directed acyclic graphs and multiple classes

Existing functional class organizations assume a tree structure. In a tree, each 'child' node has a single connection to a 'parent' node (class) at a higher level of generality. However, this is unsuitable for functional genomics because single genes or groups of genes can have more than one function. The correct functional class organization should therefore be a DAG, which differs from a tree structure in that even though the levels of generality increase unidirectionally, there can be connections to more than one parent[59]. Thus, for a given set of environmental conditions, we can allow for the fact that a single parameter change leads, in the steady state, to multiple changes and therefore we might have an output class structure that reflects this. The conclusion of this analysis is that the unsupervised learning algorithms that we apply to functional genomics data should learn DAG classifications. However, to the best of our knowledge, no such unsupervised learning algorithms exist. Existing unsupervised learning algorithms can only learn either unstructured classes or tree unstructured classes. Although it might be possible to combine existing unsupervised methods to learn DAGs, new computer science and statistics seem to be needed.

The existence of more than one function for a gene also causes problems for supervised prediction methods because it is usually assumed by such methods that each example has only one correct class label, that is, you learn a (mathematical) function mapping from input attributes to a single class. Relatively little work has been done on multiclassification problems in statistics and computer science[36]. Gilbert *et al.* presented an example of a system producing multiple rules using metabolome data[60]. Ideally, the supervised learning program should exploit the whole classification tree when learning; currently this can only be done by inductive logic programming algorithms[61].

On model complexity

If we could know quantitatively all the interactions between the players in a cell, we might be able to provide a full mathematical model of them; this is usually considered to be the gold standard. Although we can expect significant progress resulting from the availability of large-scale phenotypic data, the problem with this type of approach (the life-sized kinetic model) is that it is enormously difficult to parametrize accurately[62,63], even though not all pathways are possible[64], despite simplifying assumptions being made[57], and especially given the experimental evidence that cell growth can be chaotic[65]. Explanations, like organisms[66], have an optimal size.

Problem areas

In a short overview, it is not possible to set out all the likely pitfalls, but it is worth mentioning two. According to Brenner[67], ~8% of assignments in fully sequenced microbial genomes might be incorrect. It is clear that supervised learning methods must be trained using correct output classes, therefore this is a significant problem and is known as class noise. Similarly, the statistical analysis of the reproducibility and the significance of all observed differences in gene-expression data is in its infancy[68], and is clearly an area where more-robust methods than those currently in use will be required[69]. That, however, is a problem common to almost any computational analysis using multivariate data[70].

What are our needs in the post-genomic era? Induction vs deduction

Scientific inference uses a combination of deductive and inductive reasoning. In deductive logic: given the truth of the axioms and observations, the answer must be true (axiom: all whales are blue; observation: Percy is a whale; deduction: therefore Percy is blue). By contrast, inductive logic seeks to generalize rules from examples[71,72] (observations: Percy is blue and a whale, George is blue and a whale, Anne is blue and a whale; induction: all whales are blue) and even if all the observations are true, inductive rules can be falsified (Moby Dick is a whale and is white). Both types of reasoning are needed, induction to form new hypotheses and deduction to test these hypotheses. Of course, in the deductive framework, confounding observations can topple cherished axioms too.

Metabolic control analysis and functional genomics share the same agenda in that they seek to relate the presence and activities of individual genes and gene products to higher level processes of cellular biochemistry and physiology[57]. However, they can be considered to differ in a philosophical sense because the former is essentially deductive in character (and as practised) whereas the latter is (of necessity) inductive, at least initially, because so many open reading frames are of unknown function.

Because of the flood of data expected in the post-genomic era, it will be necessary to transfer the burden of much of the deductive and inductive reasoning to computers. As discussed by Brent[1], advances in much of recent biology have been hypothesis driven, even though the flood of genomic data suggests that we can now expect many advances to be data driven. We need the power of automatic inductive reasoning to induce hypotheses ('rules') from data[72] regarding the function of unclassified genes. Then, after a tentative class assignment, we can return to the deductive mode ('if gene X is claimed to have function Y, then the best way to confirm this is using phenotypic tests Z'); this deductive step could also be automated.

Conclusions and recommendations

High-dimensional data from the transcriptome, the proteome and the metabolome in genetically marked strains provide important inputs to classification methods designed to predict the functional class(es) for the modified gene. However, most of these classifications are unsupervised and the problem of predicting functional class is best cast as a supervised pattern classification problem. This requires the pre-assignment of functional classes and known class members, but current lists of functional classes are not driven by data from whole-organism studies and are suboptimal for the purposes of functional genomics.

We consider that novel, unsupervised classification methods could improve the current lists of functional classes themselves, and that inductive methods of machine learning (based both on phenotypic and other data, including macromolecular sequences) provide the best initial approaches to assigning gene function.

Acknowledgments

We thank the UK BBSRC and EPSRC for financial support, and Steve Oliver and Roy Featherstone for useful and stimulating discussions.

References

1 Brent, R. (1999) Functional genomics: learning to think about gene expression data. *Curr. Biol.* 9, R338–R341
2 Everitt, B.S. and Dunn, G. (1991) *Applied Multivariate Data Analysis*, Edward Arnold
3 Everitt, B.S. (1993) *Cluster Analysis* (3rd edn), Edward Arnold
4 Hinton, J.C.D. (1997) The *Escherichia coli* genome sequence: the end of an era or the start of the FUN? *Mol. Microbiol.* 26, 417–422
5 Bork, P. *et al.* (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.* 283, 707–725
6 Chu, S. *et al.* (1998) The transcriptional program of sporulation in budding yeast. *Science* 282, 699–705
7 DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686
8 Brown, P.O. and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* 21, 33–37
9 Lemieux, B. *et al.* (1998) Overview of DNA chip technology. *Mol. Breed.* 4, 277–289
10 Schena, M. *et al.* (1998) Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* 16, 301–306
11 Wilkins, M.R. *et al.* (1997) *Proteome Research: New Frontiers in Functional Genomics,* Springer-Verlag

12 Blackstock, W.P. and Weir, M.P. (1999) Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* 17, 121–127

13 Oliver, S.G. *et al.* (1998) Systematic functional analysis of the yeast genome. *Trends Biotechnol.* 16, 373–378

14 Shoemaker, D.D. *et al.* (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet.* 14, 450–456

15 Riley, M. (1993) Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* 57, 862–952

16 Riley, M. and Labedan, B. (1996) *Escherichia coli* gene products: physiological functions and common ancestries. In *Escherichia coli and Salmonella: Cellular and Molecular Biology* (2nd edn) (Neidhardt, F. *et al.*, eds), pp. 2118–2202, American Society for Microbiology

17 Karp, P.D. *et al.* (1999) Integrated pathway-genome databases and their role in drug discovery. *Trends Biotechnol.* 17, 275–281

18 Michal, G. (1999) *Biochemical Pathways: an Atlas of Biochemistry and Molecular Biology*, Wiley

19 Duda, R.O. and Hart, P.E. (1973) *Pattern Classification and Scene Analysis*, Wiley

20 Jain, A.K. and Dubes, R.C. (1988) *Algorithms for Clustering Data*, Prentice Hall

21 Therrien, C.W. (1989) *Decision Estimation and Classification: an Introduction to Pattern Recognition and Related Topics*, Wiley

22 Rich, E. and Knight, K. (1991) *Artificial Intelligence*, McGraw-Hill

23 Weiss, S.H. and Kulikowski, C.A. (1991) *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems*, Morgan Kaufmann Publishers

24 Fukunaga, K. (1992) *Introduction to Statistical Pattern Recognition*, Academic Press

25 Michie, D. *et al.* (1994) Machine learning: neural and statistical classification. In *Ellis Horwood Series in Artificial Intelligence* (Campbell, J., ed.), Ellis Horwood

26 Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*, Clarendon Press

27 Livingstone, D. (1995) *Data Analysis for Chemists*, Oxford University Press

28 Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*, Cambridge University Press

29 Mitchell, T.M. (1997) *Machine Learning*, McGraw-Hill

30 Wen, X.L. *et al.* (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. U. S. A.* 95, 334–339

31 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868

32 Tamayo, P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2907–2912

33 Somogyi, R. (1999) Making sense of gene-expression data. *Pharmainformatics (a Trends Guide)* 17–24

34 Kao, C.M. (1999) Functional genomic technologies: creating new paradigms for fundamental and applied biology. *Biotechnol. Prog.* 15, 304–311

35 Golub, T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537

36 Bäck, T. *et al.* (1997) *Handbook of Evolutionary Computation*, IOP Publishing/Oxford University Press

37 Breiman, L. *et al.* (1984) *Classification and Regression Trees*, Wadsworth International

38 Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann

39 Beadle, G.W. and Tatum, E.L. (1941) Genetic control of biochemical reactions in *Neurospora*. *Proc. Natl. Acad. Sci. U. S. A.* 17, 499–506

40 Tanksley, S.D. (1993) Mapping polygenes. *Annu. Rev. Genet.* 27, 205–233

41 Bezdek, J.C. and Pal, S.K. (1992) *Fuzzy Models for Pattern Recognition: Methods that Search for Structures in Data*, IEEE Press

42 Li, H-X. and Yen, V.C. (1995) *Fuzzy Sets and Fuzzy Decision-making*, CRC Press

43 Ruspini, E.H. *et al.* (1998) *Handbook of Fuzzy Computation*, Institute of Physics Publishing

44 Sokal, R.R. (1985) *Computer-assisted Bacterial Systematics* (Goodfellow, M. *et al.*, eds), Society for General Microbiology, Academic Press

45 Wallace, C.S. and Boulton, D.M. (1968) An information measure for classification. *Comput. J.* 11, 185–195

46 Li, M. and Vitányi, P. (1997) *An Introduction to Kolmogorov Complexity and its Applications*, Springer-Verlag

47 Sokal, R.R. and Sneath, P.H.A. (1963) *Principles of Numerical Taxonomy*, Freeman

48 Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical Taxonomy*, Freeman

49 Kell, D.B. *et al.* (1998) Viability and activity in readily culturable bacteria: a review and discussion of the practical issues. *Antonie Van Leeuwenhoek Int. J. Gen. Mol. Microbiol.* 73, 169–187

50 Miyamoto, M.M. and Cracraft, J. (1991) *Phylogenetic Analysis of DNA Sequences*, Oxford University Press

51 Hillis, D.M. *et al.* (1996) *Molecular Systematics*, Sinauer Associates

52 Cornish-Bowden, A. and Cárdenas, M.L. (1990) *Control of Metabolic Processes*, Plenum Press

53 Ovádi, J. (1995) *Cell Architecture and Metabolic Channeling*, Springer-Verlag

54 Fell, D.A. (1996) *Understanding the Control of Metabolism*, Portland Press

55 Heinrich, R. and Schuster, S. (1996) *The Regulation of Cellular Systems*, Chapman & Hall

56 Teusink, B. *et al.* (1998) *Methods in Microbiology: Yeast Gene Analysis* (Tuite, M.F. and Brown, A.J.P., eds), Academic Press

57 Kell, D.B. and Mendes, P. (2000) Snapshots of systems: metabolic control analysis and biotechnology in the post-genomic era. In *Technological and Medical Implications of Metabolic Control Analysis* (Cornish-Bowden, A. and Cárdenas, M.L., eds), pp. 3–25, Kluwer Academic Publishers

58 Lee, K.H. *et al.* (1996) Deregulated expression of cloned transcription factor E2F-1 in Chinese hamster ovary cells shifts protein patterns and activates growth in protein-free medium. *Biotechnol. Bioeng.* 50, 273–279

59 Knuth, D.E. (1973) *The Art of Computer Programming: Fundamental Algorithms*, Addison-Wesley

60 Gilbert, R.J. *et al.* (1999) Genetic programming as an analytical tool for metabolome data, In *Late-breaking Papers of EuroGP-99* (Langdon, W.B. *et al.*, eds), pp. 23–33, Software Engineering

61 Muggleton, S.H. (1990) Inductive Logic Programming. *New Generat. Comput.* 8, 295–318

62 Mendes, P. and Kell, D.B. (1996) On the analysis of the inverse problem of metabolic pathways using artificial neural networks. *BioSystems* 38, 15–28

63 Mendes, P. and Kell, D.B. (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* 14, 869–883

64 Pfeiffer, T. *et al.* (1999) METATOOL: for studying metabolic networks. *Bioinformatics* 15, 251–257

65 Davey, H.M. *et al.* (1996) Oscillatory, stochastic and chaotic growth rate fluctuations in permittistatically-controlled yeast cultures. *BioSystems* 39, 43–61

66 Haldane, J.B.S. (1927) *Possible Worlds, and Other Essays*, Chatto & Windus

67 Brenner, S.E. (1999) Errors in genome annotation. *Trends Genet.* 15, 132–133

68 Hilsenbeck, S.G. *et al.* (1999) Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J. Natl. Cancer Inst.* 91, 453–459

69 Wittes, J. and Friedman, H.P. (1999) Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data. *J. Natl. Cancer Inst.* 91, 400–401

70 Kell, D.B. and Sonnleitner, B. (1995) GMP – good modelling practice: an essential component of good manufacturing practice. *Trends Biotechnol.* 13, 481–492

71 Oldroyd, D. (1986) *The Arch of Knowledge: an Introduction to the History of the Philosophy and Methodology of Science*, Methuen

72 Langley, P. *et al.* (1987) *Scientific Discovery: Computational Exploration of the Creative Processes*, MIT Press