Science, medicine, and the future

Bioinformatics

Ardeshir Bayat

Bioinformatics, a new interdisciplinary science, is essential to managing, understanding, and harnessing clinical benefit from new genetic data

Centre for Integrated Genomic Medical Research, University of Manchester, Manchester M13 9PT

Ardeshir Bayat
MRC fellow

Correspondence to:
ardeshir.bayat@man.ac.uk

BMJ 2002;324:1018–22

An unprecedented wealth of biological data has been generated by the human genome project and sequencing projects in other organisms. The huge demand for analysis and interpretation of these data is being managed by the evolving science of bioinformatics. Bioinformatics is defined as the application of tools of computation and analysis to the capture and interpretation of biological data. It is an interdisciplinary field, which harnesses computer science, mathematics, physics, and biology (fig 1). Bioinformatics is essential for management of data in modern biology and medicine. This paper describes the main tools of the bioinformatician and discusses how they are being used to interpret biological data and to further understanding of disease. The potential clinical applications of these data in drug discovery and development are also discussed.

Summary points

- Bioinformatics is the application of tools of computation and analysis to the capture and interpretation of biological data
- Bioinformatics is essential for management of data in modern biology and medicine
- The bioinformatics toolbox includes computer software programs such as BLAST and Ensembl, which depend on the availability of the internet
- Analysis of genome sequence data, particularly the analysis of the human genome project, is one of the main achievements of bioinformatics to date
- Prospects in the field of bioinformatics include its future contribution to functional understanding of the human genome, leading to enhanced discovery of drug targets and individualised therapy

Methods

This article is based on personal experience in bioinformatics and on selected articles in recent issues of *Nature Genetics*, *Nature Genetics Reviews*, *Nature Medicine*, and *Science*. Key terms including bioinformatics, comparative and functional genomics, proteomics, microarray, disease, and medicine were used to search for relevant articles in the peer reviewed scientific literature.

Bioinformatics and its impact on genomics

Last year it was announced that the entire human genome had been mapped as a result of the efforts of the worldwide human genome project and a private genomic company.[1,2] However, in recent years, the scientific world has witnessed the completion of whole genome sequences of many other organisms. The analysis of the emerging genomic sequence data and the human genome project is a landmark achievement for bioinformatics.

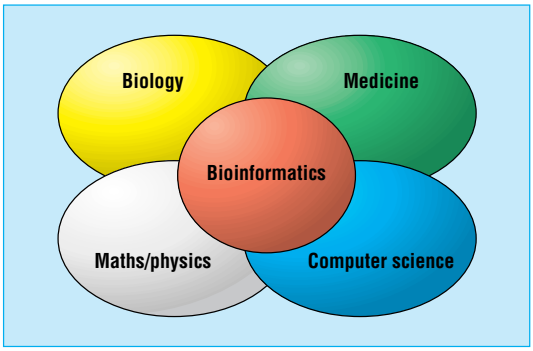

Fig 1 Interaction of disciplines that have contributed to the formation of bioinformatics

P+

An additional figure appears on bmj.com

A novel strategy for random sequencing of the whole genome (the so called "shot gun" technique) was used to sequence the genome of *Haemophilus influenzae* in 1995.[3] This was the very first complete genome of any free living organism to be sequenced. Other bacterial genomes, such as those of *Mycoplasma genitalium* and *Mycobacterium tuberculosis*, were sequenced soon after,[4,5] and the sequence of the plague bacterium *Yersinia pestis* was recently completed.[6] The sequence and annotation of the first eukaryotic genome, that of *Saccharomyces cerevisiae* (a yeast),[7] was followed by those of other eukaryotic species such as *Caenorhabtidis elegans* (a worm),[8] *Drosophila melanogaster* (fruit fly),[9] and *Arabdopsis thaliana* (mustard weed)[10] (see fig A on bmj.com). Sequencing of several other species, including

Useful bioinformatic websites (available freely on the internet)

- National Center for Biotechnology Information (www.ncbi.nlm.nih.gov)—maintains bioinformatic tools and databases
- National Center for Genome Resources (www.ncgr.org/)—links scientists to bioinformatics solutions by collaborations, data, and software development
- Genbank (www.ncbi.nlm.nih.gov/Genbank)—stores and archives DNA sequences from both large scale genome projects and individual laboratories
- Unigene (www.ncbi.nlm.nih.gov/UniGene)—gene sequence collection containing data on map location of genes in chromosomes
- European Bioinformatic Institute (www.ebi.ac.uk)—centre for research and services in bioinformatics; manages databases of biological data
- Ensembl (www.ensembl.org)—automatic annotation database on genomes
- BioInform (www.bioinform.com)—global bioinformatics news service
- SWISS-PROT (www.expasy.org/sprot/)—important protein database with sequence data from all organisms, which has a high level of annotation (includes function, structure, and variations) and is minimally redundant (few duplicate copies)
- International Society for Computational Biology (www.iscb.org/)—aims to advance scientific understanding of living systems through computation; has useful bioinformatic links

zebrafish, pufferfish, mouse, rat, and non-human primates, are either under way or nearing completion by both private and public sequencing initiatives.[11] The knowledge obtained from these sequence data will have considerable implications for our understanding of biology and medicine. As a result of comparative genomic and proteomic research, we will soon be able to not only locate each human gene but also fully understand its function.[12]

Bioinformatic tools

The main tools of a bioinformatician are computer software programs and the internet. A fundamental activity is sequence analysis of DNA and proteins using various programs and databases available on the world wide web. Anyone, from clinicians to molecular biologists, with access to the internet and relevant websites can now freely discover the composition of biological molecules such as nucleic acids and proteins by using basic bioinformatic tools. This does not imply that handling and analysis of raw genomic data can easily be carried out by all. Bioinformatics is an evolving discipline, and expert bioinformaticians now use complex software programs for retrieving, sorting out, analysing, predicting, and storing DNA and protein sequence data.

Large commercial enterprises such as pharmaceutical companies employ bioinformaticians to perform and maintain the large scale and complicated bioinformatic needs of these industries. With an ever-increasing need for constant input from bioinformatic experts, most biomedical laboratories may soon have their own in-house bioinformatician. The individual researcher, beyond a basic acquisition and analysis of simple data, would certainly need external bioinformatic advice for any complex analysis.

The growth of bioinformatics has been a global venture, creating computer networks that have allowed easy access to biological data and enabled the development of software programs for effortless analysis. Multiple international projects aimed at providing gene and protein databases are available freely to the whole scientific community via the internet.

Bioinformatic analysis

The escalating amount of data from the genome projects has necessitated computer databases that feature rapid assimilation, usable formats and algorithm software programs for efficient management of biological data.[13] Because of the diverse nature of emerging data, no single comprehensive database exists for accessing all this information. However, a growing number of databases that contain helpful information for clinicians and researchers are available. Information provided by most of these databases is free of charge to academics, although some sites require subscription and industrial users pay a licence fee for particular sites. Examples range from sites providing comprehensive descriptions of clinical disorders, listing disease susceptibility genetic mutations and polymorphisms, to those enabling a search for disease genes given a DNA sequence (box).

These databases include both "public" repositories of gene data as well as those developed by private companies. The easiest way to identify databases is by

Fig 2 Ensembl website: a genomic data search facility freely available on the internet. Ensembl is a joint project between the European Bioinformatic Institute and the Sanger Centre, which is capable of automatically tracking the sequenced pieces of the human genome and assembling and analysing them to identify genes and other features of interest to biomedical researchers

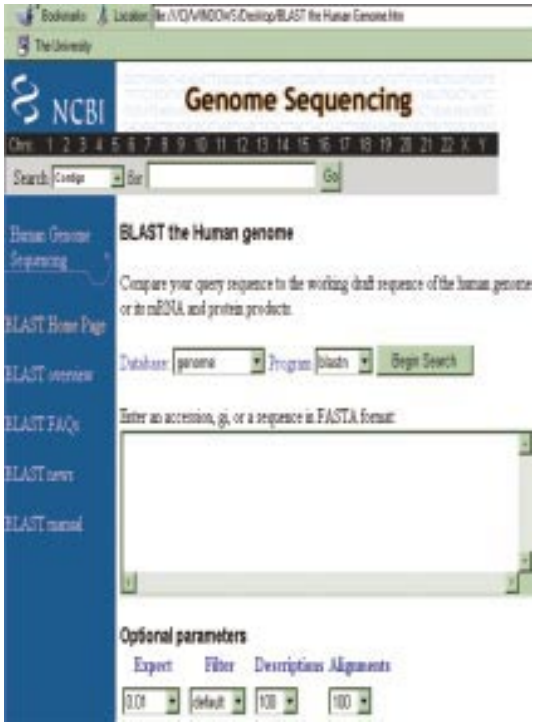

Fig 3 Web page illustrating freely available BLAST services run by the National Center for Biotechnology Information. BLAST (basic local alignment search tool) is a set of similarity search programs designed to explore all of the available DNA sequence databases

searching for bioinformatic tools and databases in any one of the commonly used search engines. Another way to identify bioinformatic sources is through database links and searchable indexes provided by one of the major public databases. For example, the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov) provides the Entrez browser, which is an integrated database retrieval system that allows integration of DNA and protein sequence databases. The European Bioinformatic Institute archives gene and protein data from genome studies of all organisms, whereas Ensembl produces and maintains automatic annotation on eukaryotic genomes (fig 2). The quality and reliability of databases vary; certainly some of the better known and more established ones, such as those above, are superior to others.

One of the simplest and better known search tools is called BLAST (basic local alignment search tool, at www.ncbi.nlm.nih.gov/BLAST/). This algorithm software is capable of searching databases for genes with similar nucleotide structure (fig 3) and allows comparison of an unknown DNA or amino acid sequence with hundreds or thousands of sequences from human or other organisms until a match is found. Databases of known sequences are thus used to identify similar sequences, which may be homologues of the query sequence. Homology implies that sequences may be related by divergence from a common ancestor or share common functional aspects. When a database is searched with a newly determined sequence (the query sequence), local alignment occurs between the query sequence and any similar sequence in the database. The result of the search is sorted in order of priority on the basis of maximum similarity. The sequence with the highest score in the database of known genes is the homologue. If homologues or related molecules exist for a query sequence, then a newly discovered protein may be modelled and the gene product may be predicted without the need for further laboratory experiments.

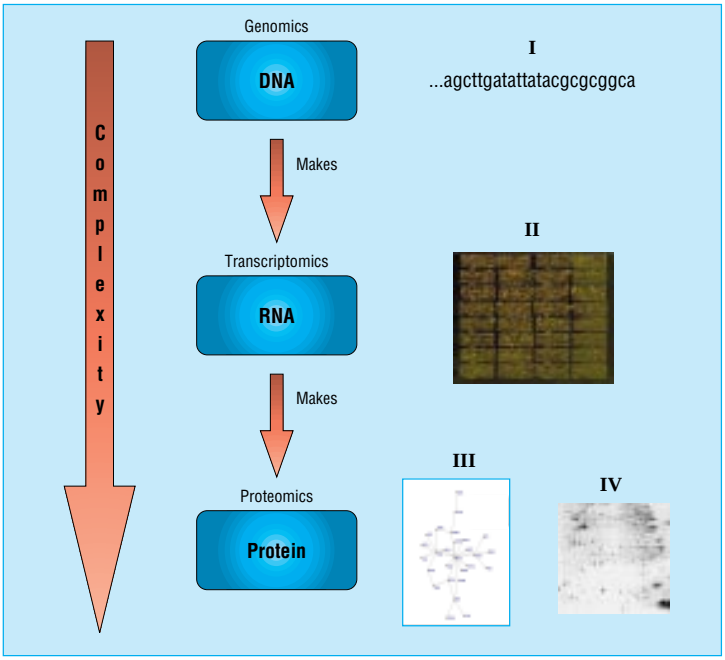

Fig 4 Schematic diagram representing complexity of genomic data processing. Analysis and interpretation of biological data considers information at every level from the genome (total genetic content) to the proteome (total protein content) and transcriptome (total messenger RNA content) of the cell. The images numbered I-IV to the right of the diagram represent relevant examples of DNA (image I is base pair nucleotides); RNA (image II is a microarray showing levels of gene expression); and protein (image III is a structure of a single protein; image IV is a two dimensional gel electrophoresis showing separation of all proteins of a cell—each spot corresponds to a different protein chain)

Functional genomics

Since the completion of the first draft of the human genome,[1 2] the emphasis has been changing from genes themselves to gene products. Functional genomics assigns functional relevance to genomic information. It is the study of genes, their resulting proteins, and the role played by the proteins.

Analysis and interpretation of biological data considers information not only at the level of the genome but at the level of the proteome and the transcriptome (fig 4). Proteomics is the analysis of the total amount of proteins (proteome) expressed by a cell, and transcriptomics refers to the analysis of the messenger RNA transcripts produced by a cell (transcriptome). DNA microarray technology determines the expression level of genes and includes genotyping and DNA sequencing. Gene expression arrays allow simultaneous analysis of the messenger RNA expression levels of thousands of genes in benign and malignant tumours, such as keloid and melanoma. Expression profiles classify tumours and provide potential therapeutic targets.[14]

Bioinformatic protein research draws on annotated protein and two dimensional electrophoresis databases. After separation, identification, and characterisation of a protein, the next challenge in bioinformatics is the prediction of its structure. Structural biologists also

use bioinformatics to handle the vast and complex data from *x* ray crystallography, nuclear magnetic resonance, and electron microscopy investigations to create three dimensional models of molecules.[15]

Other applications of bioinformatics

Apart from analysis of genome sequence data, bioinformatics is now being used for a vast array of other important tasks, including analysis of gene variation and expression, analysis and prediction of gene and protein structure and function, prediction and detection of gene regulation networks, simulation environments for whole cell modelling, complex modelling of gene regulatory dynamics and networks, and presentation and analysis of molecular pathways in order to understand gene-disease interactions.[16] Although on a smaller scale, simpler bioinformatic tasks valuable to the clinical researcher can vary from designing primers (short oligonucleotide sequences needed for DNA amplification in polymerase chain reaction experiments) to predicting the function of gene products.

Clinical application of bioinformatics

The clinical applications of bioinformatics can be viewed in the immediate, short, and long term. The human genome project plans to complete the human sequence by 2003, producing a database of all the variations in sequence that distinguish us all. The project could have considerable impact on people living in 2020—for example, a complete list of human gene products may provide new drugs and gene therapy for single gene diseases may become routine (www.ornl.gov/hgmis/medicine/tnty.html).

Basic bioinformatic tools are already accessed in certain clinical situations to aid in diagnosis and treatment plans. For example, PubMed (www.nlm.nih.gov) is accessed freely for biomedical journals cited in Medline, and OMIM (Online Mendelian Inheritance in Man at www3.ncbi.nlm.nih.gov/Omim/), a search tool for human genes and genetic disorders, is used by clinicians to obtain information on genetic disorders in the clinic or hospital setting. An example of the application of bioinformatics in new therapeutic advances is the development of novel designer targeted drugs such as imatinib mesylate (Gleevec), which interferes with the abnormal protein made in chronic myeloid leukaemia.[17] (Imatinib mesylate was synthesised at Novartis Pharmaceuticals by identifying a lead in a high throughput screen for tyrosine kinase inhibitors and optimising its activity for specific kinases.) The ability to identify and target specific genetic markers by using bioinformatic tools facilitated the discovery of this drug.

In the short term, as a result of the emerging bioinformatic analysis of the human genome project, more disease genes will be identified and new drug targets will be simultaneously discovered. Bioinformatics will serve to identify susceptibility genes and illuminate the pathogenic pathways involved in illness, and will therefore provide an opportunity for development of targeted therapy. Recently, potential targets in cancers were identified from gene expression profiles.[18]

Additional educational resources

Journals

- Specific bioinformatic journals exist (for example, www.bioinformatics.oupjournals.org), but papers from every area of science and medicine involving bioinformatic analysis are published in any biomedical journal. Examples include:
- The human genome (special issue). *Nature* 2001;409:813-933.
- The human genome (special issue). *Science* 2001;5507:1145-434.
- The human genome (special issue). *JAMA* 2001;286:2211-333.
- The human genome (special issue). *Scientific American* 2000;283:38-57.
- Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? *Method Inform Med* 2001;40:346-58.
- Online Lectures on Bioinformatics (www.lectures.molgen.mpg.de/)

Books

- Mount DW. *Bioinformatics: sequence and genome analysis*. Cold Spring Harbor Laboratory Press, 2001.
- Baxevanis AD, Ouellette BFF. *Bioinformatics: a practical guide to the analysis of genes and proteins*. 2nd ed. John Wiley and Sons, 2001.
- Lengauer T (Ed). *Bioinformatics*. Wiley-VCH Series, 2001. (Methods and principles in medicinal chemistry series.)
- Higgins D, Taylor W. *Bioinformatics*. Oxford University Press, 2000. (Practical approach series.)
- Baldi P, Brunak S. *Bioinformatics*. 2nd ed. MIT Press, 2001. (Adaptive computation and machine learning series.)

***BMJ* archive**

- Aitman TJ. DNA microarrays in medical practice. *BMJ* 2001;323:611-5.
- Mathew CG. Postgenomic technologies: hunting the genes for common disorders. *BMJ* 2001;322:1031-4.
- Stewart A, Haites N, Rose P. Online medical genetics resources: a UK perspective. *BMJ* 2001;322:1037-9.
- Savill J. Molecular genetic approaches to understanding disease. *BMJ* 1997;314:126-9.

In the longer term, integrative bioinformatic analysis of genomic, pathological, and clinical data in clinical trials will reveal potential adverse drug reactions in individuals by use of simple genetic tests. Ultimately, pharmacogenomics (using genetic information to individualise drug treatment) is likely to bring about a new age of personalised medicine; patients will carry gene cards with their own unique genetic profile for certain drugs aimed at individualised therapy and targeted medicine free from side effects.

Future directions

The practice of studying genetic disorders is changing from investigation of single genes in isolation to discovering cellular networks of genes, understanding their complex interactions, and identifying their role in disease.[19] As a result of this, a whole new age of individually tailored medicine will emerge. Bioinformatics will guide and help molecular biologists and clinical researchers to capitalise on the advantages brought by computational biology.[20] The clinical

research teams that will be most successful in the coming decades will be those that can switch effortlessly between the laboratory bench, clinical practice, and the use of these sophisticated computational tools.

I thank Tessa Richards, Dipak Roy, and Professor Bill Ollier for advice on the preparation of this manuscript and Andy Brass for providing me with some of the diagrams.

Funding: Medical Research Council.

Competing interests: None declared.

1 International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860-921.
2 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science* 2001;291:1304-51.
3 Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 1995;269:496-512.
4 Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al. The minimal gene complement of Mycoplasma genitalium. *Science* 1995;270:397-403.
5 Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* 1998;393:537-44.
6 Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, et al. Genome sequence of Yersinia pestis, the causative agent of plague. *Nature* 2001;413:523-27.
7 Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. *Science* 1996;274:546.
8 The C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* 1998;282:2012-8.
9 Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of Drosophila. *Science* 2000;287:2196-204.
10 Arabidopsis Genomics Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 2000;408:796-815.
11 Stein L. Genome annotation: from sequence to biology. *Nat Rev Genet* 2001;2:493-503.
12 Subramanian G, Adams MD, Venter JC, Broder S. Implications of the human genome for understanding human biology and medicine. *JAMA* 2001;286:2296-306.
13 Benton D. Bioinformatics—principles and potential of a new multidisciplinary tool. *Trends Biotech* 1996;14:261-312.
14 Maggio ET, Ramnarayan K. Recent developments in computational proteomics. *Trends Biotech* 2001;19:266-72.
15 Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, et al. Structural genomics: beyond the human genome project. *Nat Genet* 1999;23:151-7.
16 Tsoka S, Ouzounis CA. Recent developments and future directions in computational genomics. *FEBS Lett* 2000;480:42-8.
17 Druker BJ, Sawyers CL, Kantarjian H, Resta DJ, Reese SF, Ford JM, et al. Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. *N Engl J Med* 2001;344:1038-42.
18 Graeber TG, Eisenberg D. Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nat Genet* 2001;29:295-300.
19 Debouk C, Metcalf B. The impact of genomics on drug discovery. *Annu Rev Pharmacol Toxicol* 2000;40:193-208.
20 Butler D. Are you ready for the revolution? *Nature* 2001;409:758-60.

When I use a word
Meta-

Mr John Gleave, a neurosurgeon, has written to ask me the origin of the meta- in meta-analysis. The answer comes from Aristotle.

The Greek preposition μετά (meta) had several meanings, depending on whether it governed the accusative, genitive, or dative case. With the accusative it could mean coming into or among, in pursuit of, or coming after in place or time; with the genitive it could mean in the midst of, between, or in common with; and with the dative it could mean in the company of or over and above. It was also used as a prefix to express such notions as sharing, being in the midst of, succession, pursuit, reversal, and (most commonly) change. Examples of the last include metabolism, metamorphosis, and metaplasia.

In scientific English words its uses include "consequent upon" (as in the obsolete terms meta-arthritic, metapneumonic), "behind" or "beyond" in an anatomical sense (metabranchial, metacarpal, metaphysis), "coming later" (metaphase, which comes after prophase), or "changing" (metachromasia, a property of materials that stain a different colour from the stain used). In geology meta- is used to distinguish various types of metamorphic processes. And chemists use meta- to differentiate certain metameric chemical compounds (such as metacresol, paracresol, orthocresol).

And so to Aristotle. Some 250 years after his death, Aristotle's manuscripts came into the hands of Andronicus of Rhodes, who edited them. Andronicus called one set of papers *The Physics* (τὰ φυσικά), dealing as they did with natural science. Then he published a set of papers that he called *The Metaphysics* (τὰ μετὰ τὰ φυσικά), simply because it came after *The Physics*. However, because *The Metaphysics* dealt with what Aristotle called "primary philosophy," or ontology, metaphysics came to be misunderstood as "the science of that which transcends the physical."

As a result, the prefix meta- was then used to designate any higher science (actual or hypothetical) that dealt with more fundamental problems than the original science itself. This use first appeared in the early 17th century (John Donne, for example, writes about metatheology) but did not become really popular until the middle of the 19th century. Examples include metaethics (the study of the foundations of ethics, especially the nature of ethical statements) and metahistory (an inquiry into the principles that govern historical events).

Then, from about 1940, it became commonplace to prefix meta- to designate concern with basic principles. A metacriterion is a criterion that defines criteria. A metatheorem is a theorem about theorems. A metalanguage is a language that supplies terms for analysing a language; a metametalanguage does the same for a metalanguage. And Jean Tinguely described his machine-like sculptures as "metamechanical." (But a metaphysician is not a doctor's doctor.)

In these poststructuralist times we recognise many metaforms. *Mantissa*, a medical novel by John Fowles, is metafiction; Francois Truffaut's film *La Nuit Amercaine* is metacinema; several paintings by Magritte, notably *La Condition Humaine*, are meta-art; and John Cage's piano piece *4'33"* is metamusic.

So meta-analysis is an analysis of analyses, in which sets of previously published (or unpublished) data are themselves subjected as a whole to further analysis. In this statistical sense it was first used in the 1970s by GV Glass (*Educ Res* 1976;3(Nov):2). As he wrote, "The term is a bit grand, but it is precise and apt." Incidentally, meta-analysis should not be confused with metanalysis, which is the process whereby, for example, "a nadder" becomes "an adder" (see *BMJ* 1999;318:1758 and 2000;321:953).

I trust that this cures Mr Gleave's metagrobolism.

Jeff Aronson *clinical pharmacologist, Oxford*

We welcome articles up to 600 words on topics such as *A memorable patient, A paper that changed my practice, My most unfortunate mistake,* or any other piece conveying instruction, pathos, or humour. If possible the article should be supplied on a disk. Permission is needed from the patient or a relative if an identifiable patient is referred to.