EMBO
reports

feature

Human intuition in the quantitative age

The role of mathematics in biology is vital, but does it leave room for 'old-fashioned' observation and interpretation?

Philip Hunter

Biologists first dipped their toes into the waters of mathematics when Gregor Mendel developed his laws of trait inheritance in the mid-nineteenth century. Since then, statistics and computation have come to play an important part in almost all aspects of applied and fundamental research. The power of number crunching has even led to questions about the role of traditional observation or insight in experimentation, with a concomitant concern that over-reliance on mathematics might lead researchers to lazily follow preconceived ideas, or to ignore inconvenient data that actually indicate that a theory needs modification.

Lars Jensen, a group leader in Disease Systems Biology at the Novo Nordisk Foundation (NNF) Centre for Protein Research at the University of Copenhagen in Denmark, takes this position. "It is, in my opinion, a risk that has always been there and still is there in the mathematical era," he commented. "If researchers have a preconceived idea about how the results of an experiment should be, they may be tempted to classify observations as outliers if they do not fit the expectations."

> **"If researchers have a preconceived idea about how the results of an experiment should be, they may be tempted to classify observations as outliers if they do not fit the expectations"**

As such, the risk that statistics trump observation has to be considered carefully, but should not turn back the tide of computation and analysis in biology. Jeremy Nicholson, head of the Department of Surgery and Cancer in the Faculty of Medicine at Imperial College London, UK, argues that only the application of mathematics can show if the results of an experiment are true. "The only proof of biological activity is either in statistics, which of course goes back a long way, or geometry, as used in physical anthropology," he said.

> **…almost every biological state […] has an associated pattern of relative molecular concentrations […] these signatures can be detected against the background of normal cellular function**

The role of mathematics in biological analysis is expanding, particularly with the advent of the various 'omics' fields. Multivariate statistics, for example, allows the simultaneous analysis of variables—such as the expression levels of several genes—which makes it possible to draw simple inferences from complex data sets. This analysis can be performed not only on the expression of the genes themselves, but also downstream, on the behaviour of the gene products; reflected, for example, in the molecular composition of a blood or tissue sample. This, Nicholson argues, has led to progress in the emerging field of surgical metabonomics, which he defines as a systems approach to examining changes in hundreds or thousands of low-molecular-weight metabolites in an intact tissue or biofluid. "Our biggest recent advances in thinking are in surgical metabonomics and real-time profiling," he said, adding that these techniques will have a huge impact on diagnosis and surgical procedures.

The key point is that almost every biological state—be it a specific cancer or a metabolic condition, such as diabetes—has an associated pattern of relative molecular concentrations in cells and tissues. In principle, these signatures can be detected against the background of normal cellular function. The data usually come from nuclear magnetic resonance or mass spectrometry analyses, which yield spectral peaks and troughs relating to the identity and relative proportions of the molecular constituents of the sample. The immediate objective is not to identify individual molecules, but to analyse the overall pattern of the components. The components are usually moieties of larger molecules, such as hydroxyl or amino groups, which yield characteristic peaks. However, because different molecules have groups in common, it is not immediately possible to identify the exact contents of a sample.

> **…the mathematical tools underlying many […] methods are based on Bayes theorem [which] allows mathematicians to calculate the probability of a prior event on the basis of […] data that emerges afterwards**

"A typical example is where one is looking for biomarkers of a disease," explained Tim Ebbels, a senior lecturer in computational bioinformatics at Imperial College London. "You compare profiles from normal people against those with the disease and ask the question: which molecules change in concentration between the two groups?"

In the past, this analysis would have been done using a statistical technique such as a *t*-test, which compares just two variables at a time. The limitation is obvious: the test cannot detect small differences in concentration between many molecules. This is where modern, so-called ‘latent-variable’ techniques step in. “Not only do [latent-variable techniques] allow one to spot groups of metabolites changing together—as you might expect if they are involved in the same pathway, for instance—but they also provide simple and intuitive visualizations of the data,” Ebbels explained. “Visualization is a key part of the discovery process and would be hard to do without these kinds of tools. For instance, how do you visualize the levels of hundreds of metabolites changing over hundreds or thousands of individuals? You cannot just plot a scatter plot of the levels of metabolite one versus metabolite two. You need a tool that reduces the dimensionality of the data—this is what latent-variable methods do.”

Recent advances have helped to take analysis to the next level by identifying individual molecules associated with a particular disease state from their spectral data. Such insight used to involve time-consuming literature searches to identify candidates associated with a particular pattern and specific experiments to provide proof. Now, the emerging technique of statistical correlation spectroscopy identifies precise patterns that appear repeatedly across a set of samples, which enables software to identify the molecules that are responsible with increasing levels of certainty.

> **“…when you crunch large data sets, you accept that here and there you do make a mistake […] The idea is that, given the thousands of datapoints you can crunch, the few mistakes do not change the big picture”**

To a large extent, it is simply the availability of complex data sets spanning many variables that has driven progress in analytical biology, rather than advances in mathematics itself. Indeed, the mathematical tools underlying many of the recently developed methods are based on Bayes’ theorem, which was developed by the English Presbyterian minister Thomas Bayes and published after his death in 1763, by the Royal Society of London. The theorem allows mathematicians to calculate the probability of a previous event on the basis of evidence or data that emerges afterwards. In multivariate analyses, it identifies the event or condition most likely to be associated with a particular complex data set spanning many variables. In metabonomics it can identify the disease associated with a particular distribution of molecules with a high degree of accuracy. “Modern mathematical techniques in biology is a large subject, but you can think of the progress being based on Bayesian approaches,” noted Gael Yvert from the Laboratory of Molecular and Cellular Biology at the École Normale Supérieure, Lyon, France.

> **Ultimately, whatever tools are available and whatever technological advances are made, innovation and originality of the human spirit will still determine what makes science brilliant**

Yvert applies multivariate analysis to genetic mechanisms that underlie phenotypic differences between individuals within a species, focusing on the yeast *Saccharomyces cerevisiae*. Bayesian methods have had a profound effect on both the design and analysis of Yvert’s experiments. Essentially, Yvert has been able to strike a balance between statistical power and the cost of doing an experiment, which has helped him to minimize the number of microarrays needed to establish a link between a trait and a genetic background.

For example, if you want to identify the genes responsible for a particular trait by analysing expression levels across the genomes of two yeast strains under different experimental conditions, Bayesian techniques can provide the answer to a certain level of probability, which can then be increased by repeating the experiment. For a given budget, that means either reducing the number of conditions tested or the number of yeast strains that can be analysed. This is necessary, Yvert pointed out, because it is pointless to compromise on statistical power, especially if the conclusions are unexpected. “It is often better to have many replicates than to explore more conditions, because then robust inferences are obtained,” he explained.

There are other situations, though, in which modern analytical methods help scientists to understand biological mechanisms by revealing uncertainty. This is the case with interactions between proteins, which often depend on precise alignments of sites involving specific configurations of atoms. Until recently, the structure of such binding sites was obtained using X-ray or electron crystallography. These techniques have yielded increasingly accurate information about the position of atoms within a single protein or protein complex, but this is only a snapshot; it cannot reveal the extent of atomic flexibility within a complex molecule. Additionally, the crystal structures are based on the average positions of molecules, which might not be sufficient to predict the behaviour of the protein during interaction or binding. However, Russ Altman, chair of the Department of Bioengineering and director of the programme in Biomedical Informatics at Stanford University in the USA, explained that if you also know how much freedom each atom has to move within the structure, it becomes possible to predict how the molecule will interact during binding. Altman has applied computational modelling to determine the degree of uncertainty of the position of various atoms within a protein molecule.

“I think the ability to represent the uncertainty in the position of individual atoms is critical and still not fully appreciated,” Altman said. “Crystal structures are fabulous, but crystals provide an environment that may encourage unrealistically low atomic positional deviations. The field of protein disorder and its importance for structure and function has exploded, and our work was an early indicator and demonstration of the importance of thinking about this.”

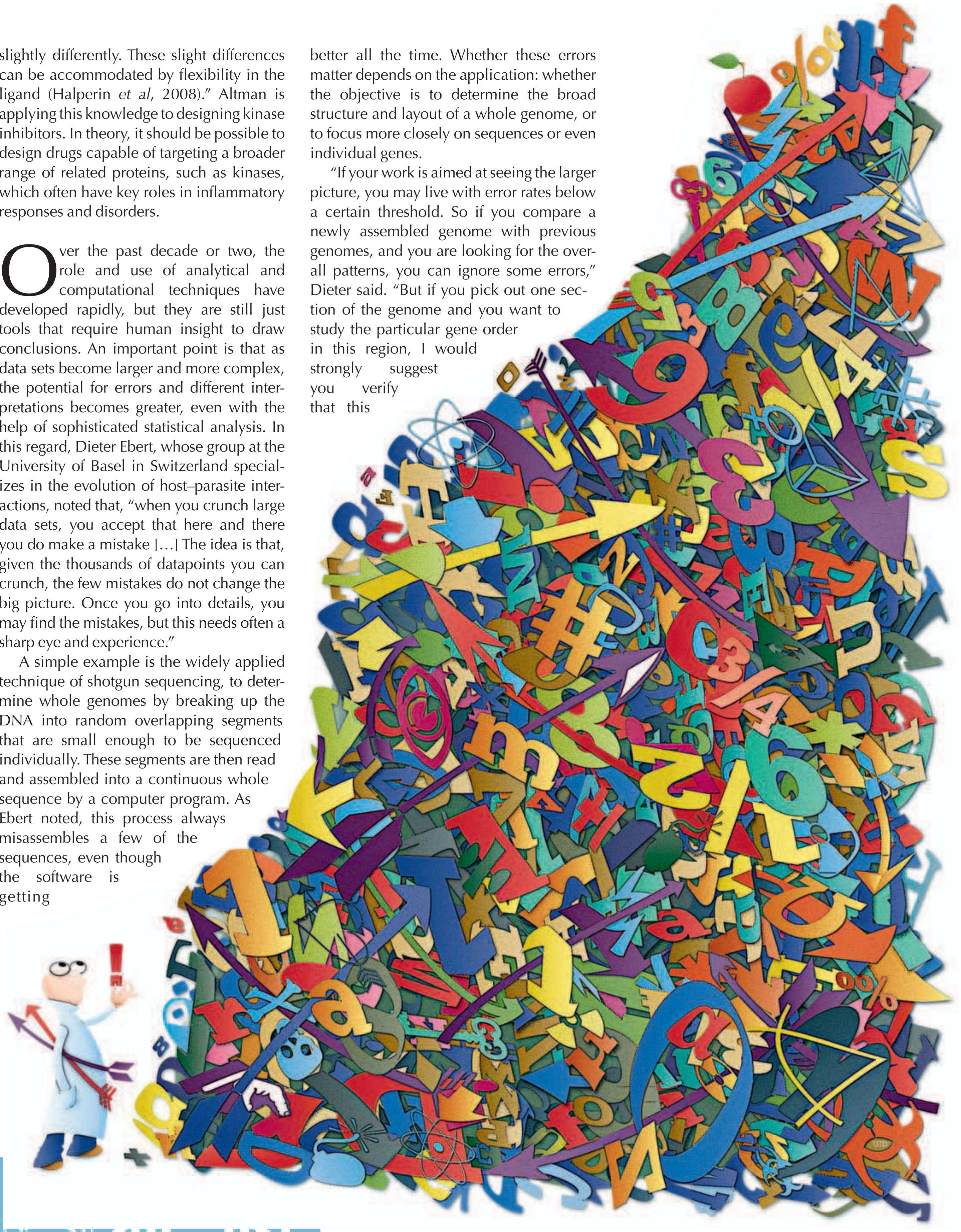
This has helped researchers to work out the detailed mechanisms of protein binding, which depend not only on the average positions of atoms, but also on their freedom of movement. “Our results suggest that certain atoms within proteins can be positioned with great certainty, while others have great uncertainty,” Altman commented.

This idea of representing the uncertainty of atoms has been scaled up to the study of binding between drugs and their targets, leading to the design of new therapeutic compounds that should be more effective. “We have recently shown that drug binding sites can be represented by a series of loosely interacting microenvironments,” Altman explained. “This representation allows us to recognize similar sites that might bind the same ligands, with similar microenvironments, but perhaps arranged slightly differently. These slight differences can be accommodated by flexibility in the ligand (Halperin *et al*, 2008).” Altman is applying this knowledge to designing kinase inhibitors. In theory, it should be possible to design drugs capable of targeting a broader range of related proteins, such as kinases, which often have key roles in inflammatory responses and disorders.

Over the past decade or two, the role and use of analytical and computational techniques have developed rapidly, but they are still just tools that require human insight to draw conclusions. An important point is that as data sets become larger and more complex, the potential for errors and different interpretations becomes greater, even with the help of sophisticated statistical analysis. In this regard, Dieter Ebert, whose group at the University of Basel in Switzerland specializes in the evolution of host–parasite interactions, noted that, “when you crunch large data sets, you accept that here and there you do make a mistake […] The idea is that, given the thousands of datapoints you can crunch, the few mistakes do not change the big picture. Once you go into details, you may find the mistakes, but this needs often a sharp eye and experience.”

A simple example is the widely applied technique of shotgun sequencing, to determine whole genomes by breaking up the DNA into random overlapping segments that are small enough to be sequenced individually. These segments are then read and assembled into a continuous whole sequence by a computer program. As Ebert noted, this process always misassembles a few of the sequences, even though the software is getting better all the time. Whether these errors matter depends on the application: whether the objective is to determine the broad structure and layout of a whole genome, or to focus more closely on sequences or even individual genes.

“If your work is aimed at seeing the larger picture, you may live with error rates below a certain threshold. So if you compare a newly assembled genome with previous genomes, and you are looking for the overall patterns, you can ignore some errors,” Dieter said. “But if you pick out one section of the genome and you want to study the particular gene order in this region, I would strongly suggest you verify that this



region was correctly assembled, even if the chance of misassembly is only 1 in 100."

The wider message is that as biology becomes more analytical and interdisciplinary, the skills required to design experiments and interpret results have inevitably changed. Nevertheless, one fundamental point remains: human skill and judgement are needed to determine whether a set of results confirms expectations, whether it indicates that further investigation is needed, or whether it requires revision of the existing orthodoxy. Ultimately, whatever tools are available and whatever technological advances are made, innovation and originality of the human spirit will still determine what makes science brilliant.

REFERENCE

Halperin I, Glazer DS, Wu S, Altman RB (2008) The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. *BMC Genomics* **16:** S2

Philip Hunter is a freelance journalist in London, UK.

EMBO *reports* (2011) **12,** 401–404.
doi:10.1038/embor.2011.57

Reading the tea leaves of Congress

With a new Republican majority in the House of Representatives, scientific research in the USA might face budget cuts

Melissa Suran

In 2009, Barack Obama became the 44th President of the USA, amid hopes that he would fix the problems created or left unresolved by his predecessor. However, despite his positive mantra, "Yes we can," the situation was going to get worse: the country was spiralling towards an economic recession, a collapsing residential real-estate market and the loss of millions of jobs. Now, the deficit lingers around US$14 trillion (US Department of the Treasury, 2011). In response to these hardships and the presence of a perceived 'socialist' president in office, a new political movement started brewing that would challenge both the Democrats and the Republicans—the two parties that have dominated US politics for generations. Known as the Tea Party, this movement has been gaining national momentum in its denouncement of the status quo of the government, especially in relation to federal spending, including the support of scientific research.

The name is a play on the Boston Tea Party, at which more than 100 American colonists dumped 45 tonnes of tea into Boston Harbour (Massachusetts, USA) in 1773 to protest against the British taxation of imported tea. Whereas the 18th century Boston Tea Party formed to protest against a specific tax, the Tea Party of the 21st century protests against taxes and 'big' government in general.

> **Tea Partiers tend to be especially critical when it comes to spending tax dollars on bank bailouts and health care, but anything goes when […]cutting […] public spending—even science**

Many view Tea Party followers as modern muckrakers, but supporters claim their movement is fundamentally about upholding the US Constitution. Tea Party Patriots, a non-partisan organization, considers itself to be the official home of the Tea Party movement. Fuelled by the values of fiscal responsibility, limited government and free markets, Tea Party Patriots believe, these three principles are granted by the Constitution, although not necessarily upheld by the administration.

"If you read the Constitution, the limits of government involvement in society [are] pretty well-defined and our government has gone farther and farther beyond the specific limits of the Constitution," said Mark Meckler, one of the co-founders of Tea Party Patriots. "Our Constitution is not designed as an empowering document, but as a limiting document… [and] was intended to be used as a weapon by the people against the government to keep it in the box." Tea Partiers tend to be especially critical when it comes to spending tax dollars on bank bailouts and health care, but anything goes when it comes to cutting back on public spending—even science. "We believe everything needs to be on the table since the government is virtually bankrupt," Meckler said. "We need to cut the waste, cut the abuse [and] get rid of the departments that shouldn't exist."

On 19 February 2011, the US House of Representatives, which is currently controlled by the Republicans, passed a federal-spending bill for the remainder of the 2011 fiscal year budget. Among other cuts, the bill called for billions of dollars to be slashed from the budgets of federal science agencies. If the bill is signed into law, the National Institutes of Health (NIH) will have $1.6 billion cut from its budget—a 5.2% decrease—and the Department of Energy (DOE) will experience an 18% cut in funding for its Office of Science. Other agencies targeted include the Environmental Protection Agency (EPA), the National Aeronautics and Space Administration (NASA), the National Institute of Standards and Technology (NIST) and the National Science Foundation (NSF; Anon, 2011; Cho, 2011). Although the US Senate, which has a Democratic majority, must consider the bill before any definite amendments to the budget are made, it is likely that there will be some cuts to science funding.

Although the House is in favour of science-related cuts, President Obama supports spending more on science education, basic research and clean-energy research. He has also proposed an 11.8% increase in the budget of the DOE, as well as an 8% increase in the NSF budget (Office of Management and Budget, 2011).

Joann Roskoski, acting assistant director of the Biology Directorate at the NSF, said her institute is strongly in favour of President Obama's budget proposal. "President Obama is a very strong supporter of fundamental research and STEM [science, technology, engineering and mathematics]