Curr. Issues Mol. Biol. (2003) 5: 123-128.

An Historical Perspective on Genomic Technologies

David J. Galas[1]* and Stephen J. McCormack[2]

[1] Keck Graduate Institute, 535 Watson Drive, Claremont, CA 91711, USA
[2] AlleCure, 28903 North Avenue Paine, Valencia, CA 91355, USA

Abstract

Genomic technologies are best defined as technologies used to manipulate and analyze genomic information. The evolution of this collective power began in earnest with the invention of DNA cloning in the 1970's and most of the technology derives from the last quarter of the 20th century. The historical impact of these technologies is clearly immense. With the genome sequence becoming available for many organisms, including humans, another new view of biology has recently emerged. This review examines the shape and texture of this recent evolution, with a particular emphasis on new technology: DNA cloning, macromolecular structure analysis (X-ray crystallography and NMR), DNA sequencing, DNA synthesis, amplification by the polymerase chain reaction, and transgenic animals (bacteria through mammals)

Introduction

Genomic technologies are as old as molecular biology. They are probably best defined as technologies used to manipulate and analyze genomic information. With this definition it is clear that the evolution of this collective power began in earnest with the invention of DNA cloning in the 1970's. While the origins of many technical advances find roots in the pre-cloning era, most of the technology derives from the last quarter of the 20th century (Galas and McCormack, 2002). As the era of "new biology" emerges with an abundance of genomic sequence information, the historical impact of these technologies is clearly immense. When we realize, however, how many times this label has been applied over the past 40 years, it is obvious that this is merely a reflection of the rapid pace of revolutionary advances, new technologies, experimental and theoretical methods and additional knowledge generated over this entire time. There was even a Nature journal in the 1960s called "Nature New Biology", which was devoted to that very new "bastard" science known as molecular biology. In fact, biology has been reinvented many times over the past 40 years, and the past 5 years have been no exception.

With the genome sequence becoming available for many organisms, including humans, another new view of biology has recently emerged. This perspective has answered many, though certainly nowhere near all, of the questions posed for years. We will examine briefly the shape and texture of this recent evolution, but before we do, it is worth taking note of a major force at work in the biological sciences – new technology. While undoubtedly true of most areas of science, it is particularly true in the biological sciences that technology drives the advance of the science. Advances in understanding the "way things work" is almost always preceded by an innovation in the technical basis for either seeing or measuring something new, or even better, doing experiments that were previously impossible. The history of biology (Mayr, 1982) is replete with examples of this effect, but perhaps a case at the dawn of European biological science makes the point well. The invention of the light microscope by van Leeuwenhoek in the 17th century revealed an entire microcosm of microbial life that was previously inaccessible. In fact, this was a world whose existence had been entirely unknown. For the first time, the small-scale structure of living things was revealed, portending a revolution in thinking about biology. Discovery has marched relentlessly in the footsteps of technological innovation in biology. The pattern has been that technology opens the door to discovery, enabling the development of newer technology, and so on. The history of this technology-discovery cycle, or 'ratchet', in biology is a long one that includes a wide range of advances. In some ways, the spectacular example of the microscope as a technical advance that drove new science is analogous to an equally spectacular technological advance underlying the biological revolution of the 20th century. While we well knew of the existence of a new world at the molecular level, the technology of DNA cloning opened the world of macromolecular information in the genomes of living things. In many ways, that single technical advance was fundamental to the ensuing biological and technological developments that ultimately led to the genomic era of the past few years. Certainly there were many other important contributing factors, but the actual opening of new worlds, like the light microscope accomplished, must be attributed to DNA cloning technology.

An Historical Perspective on Genomic Technologies

It is revealing to consider the 20th century's advances in biology from a technological point of view and ask the question: what were the key technologies that made it possible? What were the most important technical advances that enabled the transformation of the biological sciences? It may be surprising to realize that there are fewer than 10 of these advances. In fact, our particular short list of six looks like this:

- DNA cloning
- Macromolecular structure analysis (X-ray crystallography and NMR)
- DNA sequencing

*For correspondence. Email david_galas@kgi.edu.

© 2003 Caister Academic Press

- DNA synthesis
- Amplification by the polymerase chain reaction
- Transgenic animals (bacteria through mammals)

These could all be described as technologies that allow us to query the structure and function of the macromolecules of life. Indeed, one might characterize the biological revolution of the 20th century as the opening of this world of the macromolecules of life. Certainly the pivotal molecule in this world is the repository of genetic information, DNA. Thus, the past century's biological advances appropriately have culminated in vast amounts of raw information from the genomes of microbes to man. The most important technical needs to emerge from this watershed are the need to be able to efficiently analyze this sequence information – provided by the nascent discipline of bioinformatics and computational biology – and the need to fit the genomic pieces into the puzzle of biological function, understanding the functioning of complex systems of genes and macromolecules.

With the invention of efficient DNA cloning methods in the mid 1970s, biologists began to clone and study the structure and function of individual genes. This was revolutionary, for complete sequence information of a single gene and the possibility of designing and executing experiments to query aspects of the gene's function were entirely new. This era of the single gene focus lasted for almost 20 years and revealed considerable insight into biomolecular function. The characteristic genomic dimension of the era's analytical methods was a few thousand base-pairs. Indeed, in the late 1970s and early 1980s, a gene structure study of this scale was considered almost a tour-de-force. When the automated DNA sequencer emerged from Leroy Hood's laboratory and was made available to the scientific community by Applied Biosystems, the first ideas for a broader, more systematic analysis of genomic information came to the fore. The early meetings on large-scale sequencing, like that organized by Robert Sinsheimer at UC Santa Cruz in the mid-1980s, were the first serious discussions of the long-term future of biology – the need for vast amounts of sequence information, the technological challenges involved, the relation of these "big science" approaches to the predominant "hypothesis-driven" science approach, the technological and resource requirements and the need for a new perspective in both the organization of science and science funding in biology. These were the intellectual and political origins of the human genome project – really the "transformation of biology project" - that is now complete. The transformation induced by the genome project has now set the stage for an even more fundamental revolution.

If we focus on a few specific phenomena of the genome era, we can see how the technology really was driving the change in biology. In human genetics, for example, we can see from the emergence of "positional cloning" that there was clearly a proximal technological cause of the explosion in cloning of human "disease genes" during the early 1990s. Positional cloning was the process by which a genetic locus was mapped using standard pedigree analysis and statistical methods of human genetics. The gene was subsequently located using molecular technologies. The goal was the isolation of a specific gene as a DNA clone (or set of clones), and identification of a specific mutation that could be causally related to the disease phenotype in humans - genetics meets genomics. There were many of these successful projects in the 1980s and early 1990s, including the cystic fibrosis gene, the muscular dystrophy gene and others, and the process continues today in a new form. The earliest of these positional cloning successes were difficult, time consuming feats of collaborative and labor-intensive efforts. From 1990 to 1993, however, the positional cloning process underwent transformation, yielding new technologies for the human genome project. The key technology conferred the ability to clone and characterize successively larger pieces of human genomic DNA. The yeast artificial chromosome (YAC), was the first major advance (Green *et al.*, 1991). This made it clear that such things were technically possible. This invention was shortly followed by a series of new methods, notably the P1 phage cloning system (Sternberg, 1992) and the bacterial artificial chromosome (BAC), which were more efficient and simpler technologies (Mejia and Monaco, 1997). The key technical issue here is that the scale of the highest genetic mapping resolution is usually the megabase scale. Prior to the genome-project-inspired and supported technology of YACs, BACs, and P1s, the scale of cloning technologies was significantly smaller (20-40 thousand bases per clone). When efficient cloning scales - usually between 200 to 500 thousand bases up to a million at a time - reached the genetic mapping scales, a transition took place. The positional cloning of human genes now became a real technical possibility – not as a tour de force, 10-year project – but as a routine, systematic process. This technical transformation was evident in the early 1990s with the emergence of genomic-based biotechnology companies[1] , who planned to isolate key human genes involved in disease to propel the discovery and development of mechanism-based therapeutic and diagnostic products. Thus, it was again a technical advance that drove the practice of the science and was reflected in changes in the biotechnology and pharmaceutical industries. Similar arguments are evident in connection with the emergence of the human genome sequence over the past calendar year. The new approaches to the sequencing process pioneered by a company, Celera, were in stark contrast to the more slowly modified ones of the government-funded human genome project. Collectively, however, the competitive drive between both sides led to the availability of the sequence sooner than expected. This is another technology issue that is easy to miss in the heated debate about the competition between the public and private sector.

The technical challenges initially facing the genome project planners were such that it was widely assumed that "new sequencing technology" would be necessary to be able to sequence the entire genome. At the time, the focus was on methods that could avoid the electrophoresis of the nested sets of DNA fragments that resulted from the Sanger-type DNA sequencing reactions. Perhaps some

[1] Some of the companies founded in this era were: Human Genome Sciences, Millenium, Myriad, Sequana, Darwin Molecular, and Mercator.

single molecule methods, atomic force microscopy, mass spectroscopy-based methods, or other approach that could immediately increase the data acquisition rate by orders of magnitude, and decrease the costs comparably, would emerge from the technology development efforts and allow completion of the final sequencing. Yet sequencing was in fact done with the basic technology that was available in the late 1980s – nothing revolutionary was needed after all. There were some technology advances, however, that many would characterize as incremental improvements. In fact, these advances were key and included the development of automated, capillary gel electrophoresis for the Sanger reactions, the perfection of the BAC technology, the integration of automation methods and the computational analysis of sequence data, both raw and finished. Indeed, all these improvements were under development through the federally funded program as early as 1990. For example, the automated, capillary gel electrophoresis sequencer - the workhorse of the final sequence acquisition process - incorporated many of these federally funded advances (primarily by the Department of Energy's program), and was engineered into reality by private companies: Applied Biosystems, and Molecular Dynamics (now Amersham).

The publication of the first two papers describing the draft full sequence of the human genome (Venter *et al*., 2001; IGSC, 2001) is symbolic of the technology that produced it and of the beginning of a new era of biological science. It is worth reflecting on the way in which this information is used now, and the way it will be used in future. Presently, the most sought after information in the genome- the identification of the genes themselves - is still probably the most difficult to obtain, but it has instigated many changes in the way that researchers do their biological research. Since the full sequence of the genome of the yeast *Saccharomyces cereviciae* became available in 1996 (Goffeau *et al.,* 1996) we have become familiar with the use of the full genome sequence in investigations of gene expression patterns and control networks, protein-protein interaction networks and other important biological problems at the level of the full system (Ren *et al*., 2000; Laub *et al*., 2000; DeRisi *et al*., 1997). These investigations are marked by a global point of view that was simply not possible prior to this. While we still do not know what about a third of the yeast genes do, we do know that all possible protein and RNA participants in cellular function are encoded in the sequence we have; this kind of information about the whole system can transform science. As simple as it sounds, it is a fundamental shift of perspective to know that there are no other, unknown genetic components that can provide alternative explanations of experimental results. The problems under consideration include the challenges of definitively identifying the molecular components of the cell, understanding their controls, their on-off and regulatory switches for expression, their functions and interactions, and the complex interactive networks between genes that constitute the "software" of the living cell. The present form of the available sequence information of the human genome is nowhere near a complete, fully-annotated inventory of the human genes in each chromosome. Nor is the available information a single continuous and exact sequence for each chromosome. While the completeness and continuity will continue to improve, until this is achieved, there will remain significant uncertainties in inferences made from this data. The same is true of the mouse and rat genomes, which are the other mammalian organisms whose genomes are currently being sequenced. It is clear now, however, that within a very short time the genomes of all these organisms will be completely finished.

Of course, having a reliable and complete annotation of the genomic sequence – capable of query by gene type, family, gene nomenclature, and even by functional features – would be the best of all possible tools for the biological researcher. The very next major challenge of the genome program will indeed be the annotation of the finished sequence, and a significant challenge it will be, as there are no reliable technologies yet that can be used to make this process reliable and routine, much less completely automatic. Such a complete annotation will actually be feasible only when our understanding of the genome is far deeper than it is today. The complete annotation will probably take a form unlike what our present annotation efforts produce – it remains unclear how best to represent the information about each of the elements of a complex network - a gene, for example – in the one dimensional form in which the genome is represented. In every gene model constructed from the sequence, the location and structure of the sequences involved in regulation and control pose one of the most difficult annotation problems. These *cis*-regulatory regions contain information that makes them part of the network of genes in the cell, and therefore their complete annotation will be challenge to represent. At the moment, however, representation is not the immediate problem. In some cases, finding and dissecting these important sequence regions is possible using motifs known to be conserved in transcription factor binding regions. However, our ability to define and predict control regions is currently rather unreliable, much less determining which other genes these regions connect them to, directly or indirectly. At the moment, inter-specific genome comparisons are one of the most effective ways of identifying these regions, under the assumption that the regions will stand out as being conserved (Stubbs, 2002).

There has been a significant effort over the past few years to develop automated annotation tools that are powerful enough to both identify and yield reliable information about the genes (Hyatt and Uberbacher, 2002). They are most effective, however, in finding genes rather than modeling them accurately. Factors that can have strong effects on the effectiveness of such algorithms include errors in sequencing and statistical biases like base-composition. Noise in the data can sharply degrade performance, so the draft sequence, in which the error rate is higher, can be markedly inferior to finished sequence for *ab initio* prediction.

In the future, when the sequence is finished and the annotation of the genome is complete, the pertinent information will be indicated in agreed upon terms that are directly searchable by the text of the annotation rather than by sequence – for example, a gene can be found by its name, its family, by the protein domains it codes for etc.

This is called annotation-based searching. Clearly, researchers must use a combination of sequence similarity searching tools, *ab initio* methods and annotation-based searches for the foreseeable future. However, the next stage of annotation will also require the integration of independent experimental information into the gene annotation. For example, one might use pointers that link a gene to other genes by a variety of causal interactions like these: gene product is a binding partner, exerts control on the expression of that gene, produces a metabolic product that interacts with the product of that gene, participates in the same signaling pathway with the product of that gene. Only then will we be able to fully explore the properties of complex, highly interactive systems that are encoded in the genome. Judging from the past few years, it is soon likely that progress towards the annotation of model organism genomes and microbial genomes will break ground toward this new kind of biology. The availability of sequence information is paving the way to this future and we are now learning to use it to understand biology. The resulting new knowledge and the technical advances will catalyze the next transformation of biology.

One of the things it will catalyze is the formulation of complex models of the functioning of biological systems based on the information in the genome. The complex networks of interacting genes and proteins, for example, will begin to be understood on the basis of computational simulations as well as experimental inquiry (Yanai *et al*., 2002). It will become increasingly clear that the information contained in the genome and these molecular networks alike is at the heart of the functioning of living things. One way of expressing this prediction is that biology will become largely an information science. When discussing biology as an information science, one must ask: Is this just a new expression of old ideas or are there underlying principles to be asserted from this new perspective? It has been argued, for example, that biology is just an extension to a higher level of organization of physics and chemistry, albeit in systems far from equilibrium. So where does an information science come in? Biological systems represent complex, non-linear and far from equilibrium systems. They also represent systems that are stably replicated over long stretches of time from encoded information.

The key to putting order to the wealth of data in the biological sciences is likely to be through an understanding of the dynamics of complex systems. Biological systems are adaptive yet stable. Any understanding of biology must, by necessity, deal with the stability conditions for complicated nonlinear systems. Currently, these problems are too complex for traditional physico-chemical approaches. To assess the stability of a dynamical system, one can use what are called Lyapounov functions. The Gibbs entropy is an example of a Lyapounov function used to describe the stability of thermodynamic systems. Recent work suggests that is some biological applications, the Shannon information, essentially the same information measure used in communication systems, can also serve as a Lyapounov function (Lehman *et al*., 2000). This function allows the stability of the evolution of the system to be described in a direct and simple manner. In such cases, we do not need to return to thermodynamic arguments at all, but rather can use approaches from information theory and the dynamics of complex systems to characterize the stability of a living system. While we are very far from this sort of analysis at the moment, this approach may provide a theoretical foundation for discussing biology as an information science. The analysis of such fundamental issues are, perhaps, the biggest challenge in developing a quantitative, computational biology and understanding what our models mean, what constraints are in effect in biological systems and how best to shape theory to reflect experiment in biology. The computational analysis involved here can also be considered to be an important “genomic technology”.

Recent reviews (Galas and McCormack, 2002) reflect the range of technical advances and their applications that are currently transforming biology. From the application of computational methods, the robotics of modern automation methodologies, the new chemical and imaging methods that reveal chromosome structure and elucidate gene function to the emergence of evolutionary and functional information by comparative genomics, these technical approaches are the stuff of the current era of biological research. They have in common their current and future contributions to the technological drive that continues to transform biology. While the picture of the future biology of the coming decades is not within our vision, it is clear that these are among the things that will make up the new “new biology” of that future.

References

DeRisi, J.L., Iyer V.R., and Brown, P.O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-686.

Galas, D.J. and McCormack, S.J. 2002. Genomic Technologies: Present and Future. Caister Academic Press, Wymondham, UK.

Goffeau *et al*., 1996. Life with 6000 Genes. Science 274: 546-567.

Green, E.D., Riethman, H.C., Dutchik, J.E., and Olsen, M.V. 1991. Detection and characterization of chimeric yeast-artifical-chromosome clones. Genomics 11: 658-669.

Hyatt, D. and Uberbacher, E.C. 2002. Computational DNA sequence analysis and annotation. In: Genomic Technologies: Present and Future. Galas, D.J. and McCormack, S.J. (Eds). Caister Academic Press, Wymondham, UK. p. 345-374.

International Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 490: 860-914.

Laub, M.T, McAdams, H.H., Feldblyum, T, Fraser, C.M., and Shapiro, L., 2000. Global analysis of the genetic network controlling a bacterial cell cycle. Science 290: 2144-2148.

Lehman, N., Delle Donne, M., West, M., and Dewey, T. G. 2000. The genotypic landscape during *in vitro* evolution of a catalytic RNA: Implications for phenotypic buffering. J. Mol. Evol. 50: 481-490.

Mayr, E. R. 1982. The Growth of Biological Thought: Diversity, Evolution and Inheritance. Harvard University Press.

Mejia, J.E., and Monaco, A.P. 1997. Retrofitting vectors for *Escherichia coli*-based artificial chromosomes (PACs and BACs) with markers for transfection studies. Genome Research 7: 179-186.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, and S.P., Young, R.A. 2000. Genome-wide location and function of DNA binding proteins. Science 290 : 2306-2309.

Sternberg, N. 1992. Cloning high molecular weight DNA fragments by bacteriophage P1 system. Genetic Analysis: Techniques and Applications 7: 126-132.

Stubbs, L. 2002. Genome comparison techniques. In: Genomic Technologies: Present and Future. Galas, D.J. and McCormack, S.J. (Eds). Caister Academic Press, Wymondham, UK. p. 43-66.

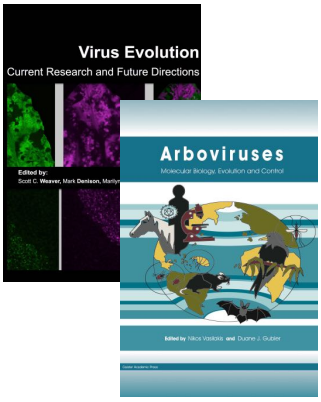
Venter, J.C. *et al.,* 2001. The sequence of the human genome. Science 291: 1304-1356.

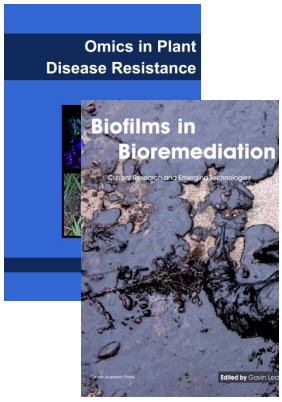
Yanai, I., Derti, A., and DeLisi, C. 2002. Beyond sequence similarity, or sequence analysis in the age of the genome. In: Genomic Technologies: Present and Future. Galas, D.J. and McCormack, S.J. (Eds). Caister Academic Press, Wymondham, UK. p. 375-410.

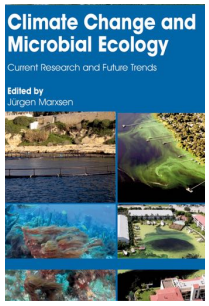
Further Reading

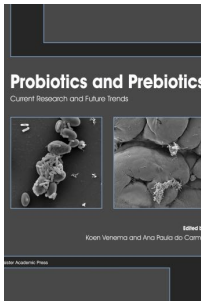
Caister Academic Press is a leading academic publisher of advanced texts in microbiology, molecular biology and medical research. Full details of all our publications at **caister.com**

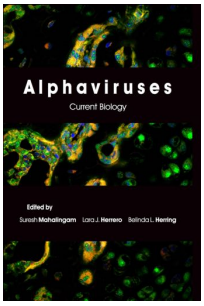
- **MALDI-TOF Mass Spectrometry in Microbiology**
  Edited by: M Kostrzewa, S Schubert (2016)
  www.caister.com/malditof
- ***Aspergillus* and *Penicillium* in the Post-genomic Era**
  Edited by: RP Vries, IB Gelber, MR Andersen (2016)
  www.caister.com/aspergillus2
- **The Bacteriocins: Current Knowledge and Future Prospects**
  Edited by: RL Dorit, SM Roy, MA Riley (2016)
  www.caister.com/bacteriocins
- **Omics in Plant Disease Resistance**
  Edited by: V Bhadauria (2016)
  www.caister.com/opdr
- **Acidophiles: Life in Extremely Acidic Environments**
  Edited by: R Quatrini, DB Johnson (2016)
  www.caister.com/acidophiles
- **Climate Change and Microbial Ecology: Current Research and Future Trends**
  Edited by: J Marxsen (2016)
  www.caister.com/climate
- **Biofilms in Bioremediation: Current Research and Emerging Technologies**
  Edited by: G Lear (2016)
  www.caister.com/biorem
- **Microalgae: Current Research and Applications**
  Edited by: MN Tsaloglou (2016)
  www.caister.com/microalgae
- **Gas Plasma Sterilization in Microbiology: Theory, Applications, Pitfalls and New Perspectives**
  Edited by: H Shintani, A Sakudo (2016)
  www.caister.com/gasplasma
- **Virus Evolution: Current Research and Future Directions**
  Edited by: SC Weaver, M Denison, M Roossinck, et al. (2016)
  www.caister.com/virusevol
- **Arboviruses: Molecular Biology, Evolution and Control**
  Edited by: N Vasilakis, DJ Gubler (2016)
  www.caister.com/arbo
- ***Shigella*: Molecular and Cellular Biology**
  Edited by: WD Picking, WL Picking (2016)
  www.caister.com/shigella
- **Aquatic Biofilms: Ecology, Water Quality and Wastewater Treatment**
  Edited by: AM Romaní, H Guasch, MD Balaguer (2016)
  www.caister.com/aquaticbiofilms
- **Alphaviruses: Current Biology**
  Edited by: S Mahalingam, L Herrero, B Herring (2016)
  www.caister.com/alpha
- **Thermophilic Microorganisms**
  Edited by: F Li (2015)
  www.caister.com/thermophile



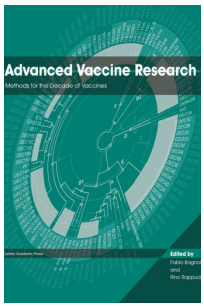


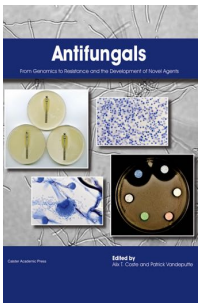












- **Flow Cytometry in Microbiology: Technology and Applications**
  Edited by: MG Wilkinson (2015)
  www.caister.com/flow
- **Probiotics and Prebiotics: Current Research and Future Trends**
  Edited by: K Venema, AP Carmo (2015)
  www.caister.com/probiotics
- **Epigenetics: Current Research and Emerging Trends**
  Edited by: BP Chadwick (2015)
  www.caister.com/epigenetics2015
- ***Corynebacterium glutamicum*: From Systems Biology to Biotechnological Applications**
  Edited by: A Burkovski (2015)
  www.caister.com/cory2
- **Advanced Vaccine Research Methods for the Decade of Vaccines**
  Edited by: F Bagnoli, R Rappuoli (2015)
  www.caister.com/vaccines
- **Antifungals: From Genomics to Resistance and the Development of Novel Agents**
  Edited by: AT Coste, P Vandeputte (2015)
  www.caister.com/antifungals
- **Bacteria-Plant Interactions: Advanced Research and Future Trends**
  Edited by: J Murillo, BA Vinatzer, RW Jackson, et al. (2015)
  www.caister.com/bacteria-plant
- ***Aeromonas***
  Edited by: J Graf (2015)
  www.caister.com/aeromonas
- **Antibiotics: Current Innovations and Future Trends**
  Edited by: S Sánchez, AL Demain (2015)
  www.caister.com/antibiotics
- ***Leishmania*: Current Biology and Control**
  Edited by: S Adak, R Datta (2015)
  www.caister.com/leish2
- ***Acanthamoeba*: Biology and Pathogenesis (2nd edition)**
  Author: NA Khan (2015)
  www.caister.com/acanthamoeba2
- **Microarrays: Current Technology, Innovations and Applications**
  Edited by: Z He (2014)
  www.caister.com/microarrays2
- **Metagenomics of the Microbial Nitrogen Cycle: Theory, Methods and Applications**
  Edited by: D Marco (2014)
  www.caister.com/n2

Order from caister.com/order