anti-cancer drugs. Potent cytokines and monoclonal antibodies directed against cell surface-associated structures are already prominent within a radically revised pharmaceutical armamentarium in areas including cancer, autoimmunity, allergy and transplantation. DNA research is therefore crucial to a new generation of immunologists, from those striving towards the development of novel vaccines to those seeking to understand and control autoimmune diseases, allergy and transplant tolerance. □

doi:10.1038/nature01409

1. Ehrlich, P. The Croonian Lecture. On immunity with special reference to cell life. *Proc. R. Soc. Lond. B* **66,** 424–448 (1900).
2. Landsteiner, K. *The Specificity of Serological Reactions* (Harvard Univ. Press, Boston, 1945).
3. Breinl, F. & Haurowitz, F. Untersuchungen des Präzipitates aus Hämoglobin und anti-Hämoglobin-Serum und Bemerkungen über die Natur der Antikörper. *Hoppe-Seyler's Z. Physiol. Chem.* **192,** 45–57 (1930).
4. Jerne, N. K. The natural selection theory of antibody formation. *Proc. Natl Acad. Sci. USA* **41,** 849–857 (1955).
5. Talmage, D. W. Allergy and immunology. *Annu. Rev. Med.* **8,** 239–256 (1957).
6. Burnet, F. M. A modification of Jerne's theory of antibody production using the concept of clonal selection. *Aust. J. Sci.* **20,** 67–69 (1957).
7. Nossal, G. J. V. & Lederberg, J. Antibody production by single cells. *Nature* **181,** 1419–1420 (1958).
8. Edelman, G. M. & Gall, W. E. The antibody problem. *Annu. Rev. Biochem.* **38,** 415–466 (1969).
9. Hilschmann, N. & Craig, L. C. Amino acid sequence studies with Bence-Jones proteins. *Proc. Natl Acad. Sci. USA* **53,** 1403–1409 (1965).
10. Dreyer, W. J. & Bennett, J. C. The molecular basis of antibody formation: a paradox. *Proc. Natl Acad. Sci. USA* **54,** 864–869 (1965).
11. Wu, T. T. & Kabat, E. A. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.* **132,** 211–250 (1970).
12. Hozumi, N. & Tonegawa, S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc. Natl Acad. Sci. USA* **73,** 3628–3632 (1976).
13. Bernard, O., Hozumi, N. & Tonegawa, S. Sequences of mouse immunoglobulin light chain genes before and after somatic changes. *Cell* **15,** 1133–1144 (1978).
14. Kocks, C. & Rajewsky, K. Stable expression and somatic hypermutation of antibody V regions in B-cell developmental pathways. *Annu. Rev. Immunol.* **7,** 537–559 (1989).
15. Hedrick, S. M., Cohen, D. I., Nielsen, E. A. & Davis, M. M. Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. *Nature* **308,** 149–153 (1984).
16. Yanagi, Y. *et al.* A human T cell-specific cDNA clone encodes a protein having extensive homology to immunoglobulin chains. *Nature* **308,** 145–149 (1984).
17. Zinkernagel, R. M. & Doherty, P. C. Restriction of *in vitro* T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. *Nature* **248,** 701–702 (1974).
18. Bernard, O., Cory, S., Gerondakis, S., Webb, E. & Adams, J. M. Sequence of the murine and human cellular myc oncogenes and two modes of myc transcription resulting from chromosome translocation in B lymphoid tumours. *EMBO J.* **2,** 2375–2383 (1983).
19. Vaux, D. L., Cory, S. & Adams, J. M. *Bcl*-2 gene promotes haemopoietic cell survival and cooperates with c-*myc* to immortalize pre-B cells. *Nature* **335,** 440–442 (1988).
20. Donnelly, J. J., Ulmer, J. B., Shiver, J. W. & Liu, M. A. DNA vaccines. *Annu. Rev. Immunol.* **15,** 617–648 (1997).
21. Krieg, A. M. CpG motifs in bacterial DNA and their immune effects. *Annu. Rev. Immunol.* **20,** 709–760 (2002).
22. Burnet, F. M. & Fenner, F. J. *The Production of Antibodies* (Macmillan, Melbourne, 1949).
23. Billingham, R. F., Brent, L. & Medawar, P. B. Actively acquired tolerance of foreign cells. *Nature* **172,** 603–606 (1953).
24. Nossal, G. J. V. & Pike, B. L. Clonal anergy: persistence in tolerant mice of antigen-binding B lymphocytes incapable of responding to antigen or mitogen. *Proc. Natl Acad. Sci. USA* **77,** 1602–1606 (1980).

DIFFERENCE AND REPETITION (DETAIL) BY SUZANNE ANKER (2000)
COURTESY: UNIVERSAL CONCEPTS UNLIMITED, NYC.

The digital code of DNA

Leroy Hood* & David Galas†

**Institute for Systems Biology, 4225 Roosevelt Way NE, Seattle, Washington 98105, USA (e-mail: lhood@systemsbiology.org)*
†Keck Graduate Institute of Applied Sciences, 535 Watson Drive, Claremont, California 91711, USA (e-mail: david_galas@kgi.edu)

The discovery of the structure of DNA transformed biology profoundly, catalysing the sequencing of the human genome and engendering a new view of biology as an information science. Two features of DNA structure account for much of its remarkable impact on science: its digital nature and its complementarity, whereby one strand of the helix binds perfectly with its partner. DNA has two types of digital information — the genes that encode proteins, which are the molecular machines of life, and the gene regulatory networks that specify the behaviour of the genes.

"Any living cell carries with it the experiences of a billion years of experimentation by its ancestors." Max Delbruck, 1949.

The discovery of the double helix in 1953 immediately raised questions about how biological information is encoded in DNA[1]. A remarkable feature of the structure is that DNA can accommodate almost any sequence of base pairs — any combination of the bases adenine (A), cytosine (C), guanine (G) and thymine (T) — and, hence any digital message or information. During the following decade it was discovered that each gene encodes a complementary RNA transcript, called messenger RNA (mRNA)[2], made up of A, C, G and uracil (U), instead of T. The four bases of the DNA and RNA alphabets are related to the 20 amino acids of the protein alphabet by a triplet code — each three letters (or 'codons') in a gene encodes one amino acid[3]. For example, AGT encodes the amino acid serine. The dictionary of DNA letters that make up the amino acids is called the genetic code[4]. There are 64 different triplets or codons, 61 of which encode an amino acid (different triplets can encode the same amino acid), and three of which are used for 'punctuation' in that they signal the termination of the growing protein chain.

The molecular complementary of the double helix — whereby each base on one strand of DNA pairs with its complementary base on the partner strand (A with T, and C with G) — has profound implications for biology. As implied by James Watson and Francis Crick in their landmark paper[1], base pairing suggests a template-copying mechanism that accounts for the fidelity in copying of genetic material during DNA replication (see article in this issue by Alberts, page 431). It also underpins the synthesis of mRNA from the DNA template, as well as processes of repairing damaged DNA (discussed by Friedberg, page 436).

Tools to modify DNA

The enzymes that function in cells to copy, cut and join DNA molecules were also exploited as key tools for revolutionary new techniques in molecular biology, including the cloning of genes and expression of their proteins, and mapping the location of genes on chromosomes. The ability to recreate the process of DNA replication artificially in the laboratory led to the development of two techniques

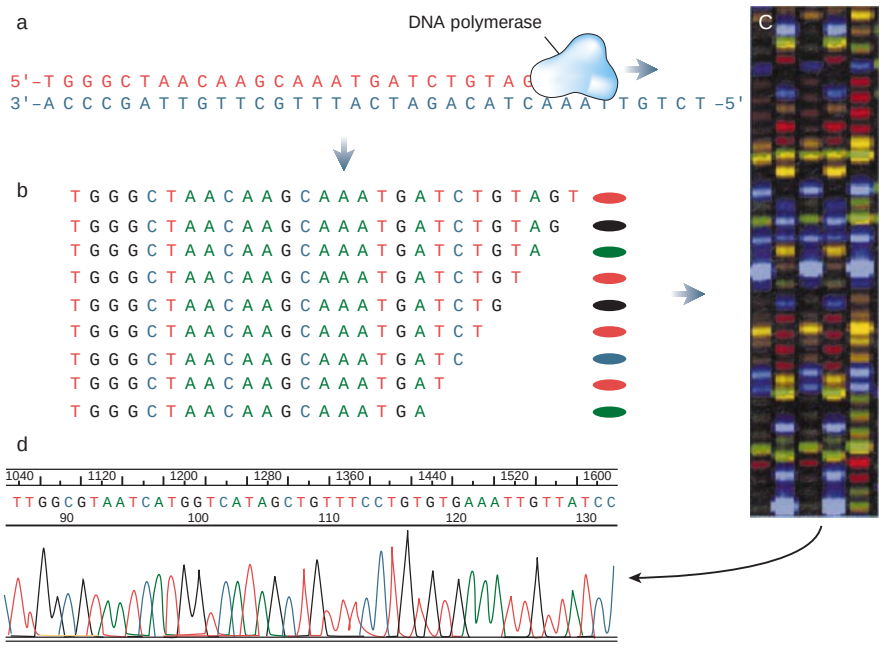

Figure 1 How to sequence DNA. **a**, DNA polymerase copies a strand of DNA. **b**, The insertion of a terminator base into the growing strand halts the copying process. This is a random event that results in a series of fragments of different lengths, depending on the base at which the copying stopped. The fragments are separated by size by running them through a gel matrix, with the shortest fragments at the bottom and largest at the top. **c**, The terminators are labelled with different fluorescent dyes, so each fragment will fluoresce a particular colour depending on whether it ends with an A, C, G, or T base. **d**, The sequence is 'read' by a computer. It generates a 'sequence trace', as shown here, with the coloured peaks corresponding to fluorescent bands read from the bottom to the top of one lane of the gel. The computer translates these fluorescent signals to DNA sequence, as illustrated across the top of the plot. Image adapted from ref. 20.

that transformed biology: a manual DNA sequencing method in 1975 and, in 1985, the discovery of the polymerase chain reaction (PCR), whereby DNA sequences could be amplified a millionfold or more[5].

Although sequencing and PCR transformed the science of biology, they also had wide applications for medicine and forensics. The detection of variation in DNA sequence from one individual to the next — so-called 'polymorphisms' — forms the basis of DNA 'finger-printing' of individuals. Forensics uses these fingerprints to deal with paternity disputes, as well as criminal cases such as rape. The finding that many specific DNA polymorphisms are associated with disease or disease susceptibility has brought DNA diagnostics to medicine and opened the pathway to truly predictive medicine, where the risks of disease can be identified in advance of symptoms (see article in this issue by Bell, page 414).

Automated DNA sequencing

The first efforts to sequence DNA, pioneered by Walter Gilbert[6] and Fred Sanger[7] in the 1970s, decoded stretches of DNA a few hundred bases long. When the first complete genome was sequenced over a period of about one year in 1977–78 — that of a viral genome of about 5,000 bases[8] — it became clear that DNA sequence data could provide unique insights into the structure and function of genes, as well as genome organization. It was this potential to generate vast amounts of information about an organism from its genetic code that inspired efforts towards the automation of DNA sequencing (Fig. 1).

The combination of technical wizardry and intensive automation in the decade that followed launched the 'genomic era'. A series of new instruments enabled novel approaches to biological analysis[9–11]. The first sequencing machine — invented by Leroy Hood, Lloyd Smith and Mike Hunkapiller in 1986 (ref. 12) — was automated in data acquisition, but still required substantial manual attention and the sequencing rate was low, roughly 250 bases per day. Over the next ten years, the development of automated DNA sequencing accelerated, rapidly passing through three distinct stages: the prototype sequencing machine (1986); a robust instrument that could be used routinely in a standard laboratory (1989); and finally, a machine that formed part of an integrated factory-like production line where DNA sample preparation and sequencing were all fully automated (1998). The advances in sequencing capacity have been striking — the latest sequencing machines are able to decode approximately 1.5 million bases over 24 hours — 6,000 times the throughput of the prototype.

The goals of high-throughput biological instrumentation are to increase throughput, enhance the quality of the data, and greatly reduce the cost of per unit information acquired. To reach these goals in the future, the miniaturization, automation, parallelization and integration of successive procedures will propel DNA sequencing technology into the realm of microfluidics and microelectronics, and eventually into the area of nanotechnology. With single-DNA-molecule sequencing, we foresee a time when the entire genome of an individual could be sequenced in a single day at a cost of less than $US10,000 (compared with the US$50 million or more it would cost today). This will readily enable the decoding of the genomic sequence of virtually any organism on the planet and provide unparalleled access to the foundations of biology and the study of human genetic variability.

The Human Genome Project

The breathtaking speed at which automated DNA sequencing developed was largely stimulated by the throughput demands of the Human Genome Project (HGP), which officially started in 1990 following discussions and studies on feasibility and technology that began in earnest in 1985. The objectives of the HGP were to generate a finished sequence in 15 years[13], but a draft of the human genome sequence was available in 2001. Two versions of the draft were generated and published in 2001, one by the publicly funded International Human Genome Sequencing Consortium[14], and another by the biotechnology company Celera[15] (Box 1). In the process of developing the tools and methodology to be able to sequence and assemble the 3 billion bases of the human genome, a range of plant, animal and microbial genomes was sequenced and many more are currently being decoded. As genome sequences become available, different areas of biology are being transformed — for example, the discipline of microbiology has changed significantly with the completion of more than 100 bacterial genome sequences over the past decade.

The HGP profoundly influenced biology in two respects. First, it illustrated the concept of 'discovery science' — the idea that all the elements of the system (that is, the complete genome sequence and

the entire RNA and protein output encoded by the genome) can be defined, archived in a database, and made available to facilitate hypothesis-driven science and global analyses. Second, to succeed, the HGP pushed the development of efficient large-scale DNA sequencing and, simultaneously, drove the creation of high-throughput tools (for example, DNA arrays and mass spectrometry) for the analysis of other types of related biological information, such as mRNAs, proteins and molecular interactions.

The digital nature of biological information

The value of having an entire genome sequence is that one can initiate the study of a biological system with a precisely definable digital core of information for that organism — a fully delineated genetic source code. The challenge, then, is in deciphering what information is encoded within the digital code. The genome encodes two main types of digital information — the genes that encode the protein and RNA molecular machines of life, and the regulatory networks that specify how these genes are expressed in time, space and amplitude.

It is the evolution of the regulatory networks and not the genes themselves that play the critical role in making organisms different from one another. The digital information in genomes operates across three diverse time spans: evolution (tens to millions of years), development (hours to tens of years), and physiology (milliseconds to weeks). Development is the elaboration of an organism from a single cell (the fertilized egg) to an adult (for humans this is 10^{14} cells of thousands of different types). Physiology is the triggering of specific functional programmes (for example, the immune response) by environmental cues. Regulatory networks are crucial in each of these aspects of biology.

Regulatory networks are composed of two main types of components: transcription factors and the DNA sites to which they bind in the control regions of genes, such as promoters, enhancers and silencers. The control regions of individual genes serve as information processors to integrate the information inherent in the concentrations of different transcription factors into signals that mediate gene expression. The collection of the transcription factors and their cognate DNA-binding sites in the control regions of genes that carry out a particular developmental or physiological function constitute these regulatory networks (Fig. 2).

Because most 'higher' organisms or eukaryotes (organisms that contain their DNA in a cellular compartment called the nucleus), such as yeast, flies and humans, have predominantly the same families of genes, it is the reorganization of DNA-binding sites in the control regions of genes that mediate the changes in the developmental programmes that distinguish one species from another. Thus, the regulatory networks are uniquely specified by their DNA-binding sites and, accordingly, are basically digital in nature.

One thing that is striking about digital regulatory networks is that they can change significantly in short periods of evolutionary time. This is reflected, for example, in the huge diversity of the body plans, controlled by gene regulatory networks, that emerged over perhaps 10–30 million years during the Cambrian explosion of metazoan organisms (about 550 million years ago). Likewise, remarkable changes occurred to the regulatory networks driving the development of the human brain during its divergence from its common ancestor with chimpanzees about 6 million years ago.

Biology has evolved several different types of informational hierarchies. First, a regulatory hierarchy is a gene network that defines the relationships of a set of transcription factors, their DNA-binding sites and the downstream peripheral genes that collectively control a particular aspect of development. A model of development in the sea urchin represents a striking example[16] (Fig. 2). Second, an evolutionary hierarchy defines an order set of relationships, arising from DNA duplication. For example, a single gene may be duplicated to generate a multi-gene family, and a multi-gene family may be duplicated to create a supergene family. Third, molecular machines may be assembled into structural hierarchies by an ordered assembly process. One example of this is the basic transcription apparatus that involves the step-by-step recruitment of factors and enzymes that will ultimately drive the specific expression of a given gene. A second example is provided by the ribosome, the complex that translates RNA into protein, which is assembled from more than 50 different proteins and a few RNA molecules. Finally, an informational hierarchy depicts the flow of information from a gene to environment: gene → RNA → protein → protein interactions → protein complexes → networks of protein complexes in a cell → tissues or organs → individual organisms → populations → ecosystems. At each successively higher level in the informational hierarchy, information can be added or altered for any given element (for example, by alternative RNA splicing or protein modification).

Box 1
Sequencing the human genome

The first complete drafts of the human genome sequence were published in 2001 by the International Human Genome Sequencing Consortium (IHGSC), a publicly funded effort, and Celera, a biotechnology company, using different approaches. Both efforts used a random or shotgun approach where the original DNA to be sequenced was randomly broken into overlapping fragments that were then cloned, and 500 base pairs (bp) were 'read' from one or both ends of the clones.

For the draft genome sequences, each base was read six to ten times to optimize the accuracy of the sequence. The stretches of DNA sequence were read by a computer and assembled into a complete sequence. The IHGSC effort randomly cleaved DNA into ~200,000-bp fragments and generated a map of these fragments across the 24 different human chromosomes; it then used the shotgun approach to sequence the pre-ordered fragments clone by clone. In contrast, Celera randomly fragmented the entire genome into three sizes of fragments (approximately 2,000, 10,000 and 200,000 bp), sequenced both ends of the clones and then used the end sequences to assemble the entire genome sequence, without the aid of a map.

Celera's 1998 announcement that it would sequence the human genome within three years was greeted with considerable scepticism, but it succeeded in producing a draft sequence and considerably accelerating the public effort. The efforts of both groups benefited science by producing draft genome sequences considerably earlier than expected.

Although minor differences were noted between the two drafts, the overall conclusions concerning gene numbers, repeated sequences and chromosomal organization were remarkably similar. For example, both groups identified 30,000–35,000 genes, far fewer than the 100,000 expected from an earlier (admittedly 'back of the envelope') calculation.

Systems approaches to biology

Humans start life as a single cell — the fertilized egg — and develop into an adult with trillions of cells and thousands of cell types. This process uses two types of biological information: the digital information of the genome, and environmental information, such as metabolite concentrations, secreted or cell-surface signals from other cells or chemical gradients. Environmental information is of two distinct types: deterministic information where the consequences of the signals are essentially predetermined, and stochastic information where chance dictates the outcome.

Random, or stochastic, signals can generate significant noise in biological systems, but it is only in special cases that noise is converted into signals. For example, stochastic events govern many of the genetic mechanisms responsible for generating antibody diversity. In the immune response, those B cells that produce antibodies that bind

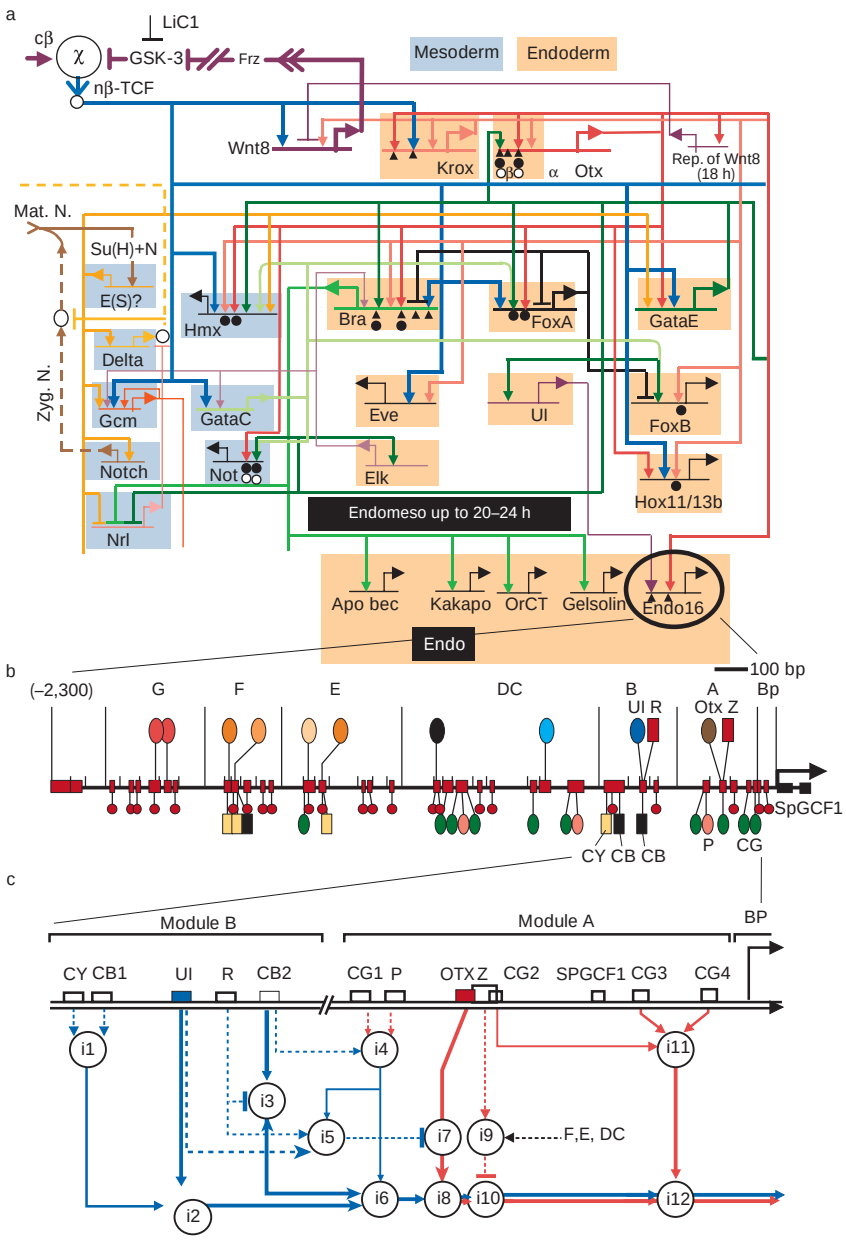

Figure 2 A gene regulatory network involved in sea urchin development[16]. **a**, Part of the network of transcription factors and their interactions with the control regions of other transcription factors. Genes are indicated by horizontal lines; arrowheads indicate activation; '⊥' symbols indicate gene repression. **b**, An enlargement of the promoter region of a gene, called *endo 16*, that helps modulate the development of the endoderm. It contains 34 binding sites (rectangles) for 13 different transcription factors and cofactors (illustrated as rectangles or lollipops, respectively). Six modules (A–G) of transcription factors and binding sites carry out discrete functions to developmentally regulate *endo 16*. **c**, Diagram depicting the logical structures of the A and B control circuits during sea urchin development.

tightly to the antigen (that is, those having high affinities) undergo an expansion in number that is proportional to the strength of the antibody affinity (see article in this issue by Nossal, page 440). Hence, the signal (high affinity) is distinguished from the noise (low affinity). Moreover, high levels of mutation in the B cells causes specific diversification of antibody genes in the presence of antigen and permits the affinity to increase even more. The cells carrying the higher-affinity antibody genes are then preferentially selected for survival and proliferation.

The key question is what and how much signal emerges from the noise. Analysis of stochastic events and the differentiation between signal and noise will be a future challenge for contemporary biology. The immune response has been studied for more than 100 years, yet we still have only a partial understanding of its systems properties, such as the immune response and tolerance (the unresponsiveness to one's own cells). This is because until recently immunologists have been able to study this complex system only one gene or one protein at a time.

The systems approach permits the study of all elements in a system in response to genetic (digital) or environmental perturbations. Global quantitative analyses of biological information from different levels each provide new insights into the operation of the system; hence, information at as many levels as possible must be captured, integrated, and ultimately, modelled mathematically. The model should explain the properties of the system and establish a framework that allows us to redesign the system in a rational way to generate new emergent properties.

Several systems have been explored successfully. The utilization of the sugar galactose in yeast has been analysed using genetic perturbations (inactivation of genes) and four levels of information were gathered — RNA and protein concentrations as well as protein–protein and protein–DNA interactions[17]. Using an iterative and integrative systems approach, new insights into the regulation of galactose use were gained. Moreover, the relationships of the galactose regulatory network to other modules in the yeast cell were also delineated. Likewise, systems approaches to early embryonic

development in the sea urchin have delineated a regulatory network that has significant predictive power[16] (Fig. 2). Finally, systems approaches to metabolism in an archaeal halobacterium (an organism thriving in up to five-molar salt solutions, such as the Dead Sea) have revealed new insights into the inter-relationships among several modules controlling energy production in the cell[18].

The study of cellular and organismal biology using the systems approach is at its very beginning. It will require integrated teams of scientists from across disciplines — biologists, chemists, computer scientists, engineers, mathematicians and physicists. New methods for acquiring and analysing high-throughput biological data are needed. A powerful computational infrastructure must be leveraged to generate more effective approaches to the capture, storage, analysis, integration, graphical display and mathematic formulation of biological complexity. New technologies must be integrated with each other. Finally, hypothesis-driven and discovery science must be integrated. In short, both new science and technology must emerge for the systems biology approach to realize its promise. A cultural shift in the biological sciences is needed, and the education and training of the next generation of biologists will require significant reform.

Gordon Moore, the founder of Intel, predicted that the number of transistors that could be placed on a computer chip would double every 18 months. It has for more than 30 years. This exponential growth has been a driver for the explosive growth of information technology. Likewise, the amount of DNA sequence information available to the scientific community is following a similar, perhaps even steeper, exponential increase. The critical issue is how sequence information can be converted into knowledge of the organism and how biology will change as a result. We believe that a systems approach to biology is the key. It is clear, however, that this approach poses significant challenges, both scientific and cultural[19]. The discovery of DNA structure started us on this journey, the end of which will be the grand unification of the biological sciences in the emerging, information-based view of biology. □

doi:10.1038/nature01410

1. Watson, J. D. & Crick, F. H. C. A structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (1953).
2. Brenner, S., Jacob, F. & Meselson, M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190,** 576–581 (1961).
3. Crick, F. H. C., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General nature of the genetic code for proteins. *Nature* **192,** 1227–1232 (1961).
4. Nirenberg, M. W. & Matthaei, J. H. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polynucleotides. *Proc. Natl Acad. Sci. USA* **47,** 1588–1602 (1961).
5. Saiki, R. K. *et al.* Enzymatic amplification of β-globin sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230,** 1350–1354 (1985).
6. Maxam, A. M. & Gilbert, W. A new method of sequencing DNA. *Proc. Natl Acad. Sci. USA* **74,** 560–564 (1977).
7. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 444–448 (1975).
8. Sanger, F. *et al.* Nucleotide sequence of bacteriophage ϕX174. *Nature* **265,** 678–695 (1977).
9. Hunkapiller, M. W. & Hood, L. New protein sequenator with increased sensitivity. *Science* **207,** 523–525 (1980).
10. Horvath, S. J., Firca, J. R., Hunkapiller, T., Hunkapiller M. W. & Hood. L. An automated DNA synthesizer employing deoxynucleoside 3′ phosphoramidites. *Methods Enzymol.* **154,** 314–326 (1987).
11. Kent, S. B. H., Hood, L. E., Beilan, H., Meister S. & Geiser, T. High yield chemical synthesis of biologically active peptides on an automated peptide synthesizer of novel design. *Peptides* **5,** 185–188 (1984).
12. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321,** 674–679 (1986).
13. Collins, F. & Galas, D. J. A new five-year plan for the US Human Genome Project. *Science* **262,** 43–46 (1993).
14. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
15. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
16. Davidson, E. H. *et al.* A genomic regulatory network for development. *Science* **295,** 1669–1678 (2002).
17. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292,** 929–933 (2001).
18. Baliga, N. S. *et al.* Coordinate regulators of energy transduction modules in the *Halobacterium* sp. analyzed by a global systems approach. *Proc. Natl Acad. Sci. USA* **99,** 14913–14918 (2002).
19. Aderem A. & Hood, L. Immunology in the post-genomic era. *Nature Immunol.* **2,** 1–3 (2001).
20. Dennis, C. & Gallagher, R. (eds) *The Human Genome* (Palgrave, Basingstoke, 2001).

Controlling the double helix

Gary Felsenfeld* & Mark Groudine†

**Laboratory of Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Building 5, Room 212, Bethesda, Maryland 20892-0540, USA (e-mail: gary.felsenfeld@nih.gov)*
†Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, and Department of Radiation Oncology, University of Washington School of Medicine, Seattle, Washington 98195, USA (e-mail: markg@fhcrc.org)

Chromatin is the complex of DNA and proteins in which the genetic material is packaged inside the cells of organisms with nuclei. Chromatin structure is dynamic and exerts profound control over gene expression and other fundamental cellular processes. Changes in its structure can be inherited by the next generation, independent of the DNA sequence itself.

Genes were first shown to be made of DNA only nine years before the structure of DNA was discovered (ref. 1; and see article in this issue by McCarty, page 406). Although revolutionary, the idea that genetic information was protein-free ultimately proved too simple. DNA in organisms with nuclei is in fact coated with at least an equal mass of protein, forming a complex called chromatin, which controls gene activity and the inheritance of traits.

'Higher' organisms, such as yeast and humans, are eukaryotes; that is, they package their DNA inside cells in a separate compartment called the nucleus. In dividing cells, the chromatin complex of DNA and protein can be seen as individual compact chromosomes; in non-dividing cells, chromatin appears to be distributed throughout the nucleus and organized into 'condensed' regions (heterochromatin) and more open 'euchromatin' (see article in this issue by Ball, page 421). In contrast, prokaryotes, such as bacteria, lack nuclei.

The evolution of chromatin

The principal protein components of chromatin are proteins called histones (Fig. 1). Core histones are among the most highly conserved eukaryotic proteins known, suggesting that the fundamental structure of chromatin evolved in a common ancestor of eukaryotes. Moreover, histone equivalents and a simplified chromatin structure have also been found in single-cell organisms from the kingdom Archaeabacteria[2,3].

Because there is more DNA in a eukaryote than in a prokaryote, it was naturally first assumed that the purpose of histones was to compress the DNA to fit within the nucleus. But subsequent research has dramatically revised the view that histones emerged as an afterthought, forced on eukaryotic DNA as a consequence of large genome size and the constraints of the nucleus.

It was known that different genes are active in different tissues, and the distinction of heterochromatin and euchromatin suggested that differences in chromatin structure were associated with differences in gene expression. This led to the early supposition that the histones were also repressor proteins designed to shut off unwanted expression. The available evidence, although rudimentary, does indeed suggest that archaeal histones are not merely packaging factors, but function to regulate gene expression[2–5]. They