European Journal of Surgical Oncology 2000; **26**: 805–809
doi:10.1053/ejso.2000.1008, available online at http://www.idealibrary.com on IDEAL®

EDUCATIONAL SECTION

Modelling in surgical oncology—Part III: Massive data sets and complex systems

D. A. Rew

Consultant Surgeon and Honorary Senior Lecturer, Royal South Hants Cancer Centre, Southampton University Hospitals, Southampton, UK

Human tumours are complex and unstable biological systems. New intellectual and mathematical approaches together with massive computing power are transforming our capacity to model and investigate such complexity. Computers also allow massive data sets to be collated and analysed. Such sets include the medical and epidemiological records of entire populations; the entire genetic code of the human being and of other species, including parasites and disease vectors; and the genotype of each and every individual.

Massive data sets take us into new dimensions of complexity for which simple linear mathematics are insufficient. The analysis of the grades of complexity which determine protein and cell construction, cell to cell interactions within tissues and organs, the morphogenesis of entire organisms and population interactions with disease vectors require the sophisticated mathematical tools of non-linear analysis, neural networks, chaos and complexity theory. The capacity for closer representations of reality through powerful computational models also allows us to look afresh at the generalizations of conventional statistics.

Within this computational cauldron, we may also find help in the better understanding of oncogenesis and cancer therapy. This paper, the third in our series on modelling in tumour biology, considers the breadth of opportunity and challenge at the interface between cell biology and biomathematics.

Key words: **modelling; complex systems; cancer; bioinformatics.**

Introduction

Human tumours are complex, unstable populations of billions of cells, each comprising billions of molecules and over 100 000 variably expressed genes.[1] To understand and control cancer by means other than the knife, we must tease out and model the most subtle disorders and instabilities from this complexity. Raw information from biological systems grows exponentially.[2] In population biology, we can catalogue the millions of identifiable species and track their evolutionary relationships. We may document each of the six billion or more human beings alive now. We may document and analyse molecular composition and interactions within cells. We may model the characteristics of the billions of cells that comprise normal and dysfunctional tissues and tumours. We may study the DNA base sequence of any organism, including those which have already been characterized, such as the fruit fly and various human pathogens. We are achieving a much closer representation of the biological world in digits.

Until 50 years ago, the need for huge computing capacity was not generally apparent. We seemed able to describe the world to our satisfaction in terms of linear mathematics and models, amplified with such exuberant mathematical models as the theories of Einstein. Computers and computer-controlled instruments have thrown up whole new fields of opportunity for research, from the universal to the microscopic scale. Modern machines possess a massive capacity for information storage, management and retrieval for the digital description of the physical and biological world. They allow us to convert massive volumes of raw data into knowledge and understanding. This computer power comes in the form of potent personal and desk top machines, and powerful supercomputers, and through connectivity across the Internet which gives each individual a window on the computing world. Most exciting are internet enabled software tools which allow large numbers of personal computers to cooperate and share capacity on huge computing tasks.[3]

Digital computers have thus created a virtuous spiral of advancement in biology. Computer driven instrumentation allows us to create vast catalogues of the natural world, from the astronomical to the genetic. This paper considers the nature and the significance of massive data sets in biology and clinical practice, and the tools for the study of complexity which may help us to make best use of the data in clinical practice.

Tools for the analysis of complex dynamic systems

The modelling of systems which change rapidly in time and space, and which are immensely dynamic even when seemingly static, defies simple inferential techniques of data collection. Linearity implies simple relationships between cause and effect. Mathematicians have applied a theory of complexity to the description of the form and behaviour of biological systems.

Complex adaptive systems are non-linear modelling tools which can be used to analyse complexity in many physical, economic or biological systems in terms of their components and rules of behaviour. Such tools should be predictive and mathematically rigorous. Complex adaptive systems have seven basic elements.[4,5] There are four properties: aggregation, non-linearity, flow of information and diversity; and three mechanisms: identity tags, internal models and building blocks. These systems evolve. Their component parts are relatively independent but are interactive. They can change strategies and change to maximize the 'fitness' of the system. Both Chaos Theory and Complexity Theory describe the properties of non-linear systems: Complexity Theory is related to Chaos Theory.[6] Chaotic systems are reactive, while complex systems are proactive and comprise mechanisms which allow them to adapt to their environment and improve the global properties of the system.

Biological systems comprise vast numbers of connected elements, and we do not usually know the logic of coupling of these elements. Many of the elements are in fixed, inactive, or conserved states. The other elements are free to change in simple, complex or chaotic ways. Boolean networks are an important component of complex adaptive systems. They find particular applications in computer science. Boolean networks are made up of binary (on–off) variables, such as AND and OR, TRUE or FALSE, IF or THEN, STOP and GO, each of which can be represented by binary data (0 or 1). The behaviour of each variable is governed or specified by a logical switching rule, or a Boolean function. Such behaviours may be ordered, chaotic or complex. Among the many possible states of stability and instability in molecular biology, simple or chaotic rules dictate possible behaviours of a system, reducing the number of options, and making certain outcomes more likely. These complex systems do not occupy single states, but there are a range of possible states that the system can occupy. Complex systems may also change in response to their environment, through mutation or adaptation, thus influencing the properties of the entire system.

Kaufmann[7] describes the modelling of complex systems with the concept of the 'fitness landscape'. This is a mathematical or graphical plot which describes the variability in behaviour of a system as its structure is altered. A smooth landscape will be very stable, while a rugged landscape will describe considerable variability. Within this fitness landscape may be found regions which describe a stable, steady state. Of the many ways that a biological process may be conducted by a range of molecules, some combinations or structures will be more efficient or stable than others. Sequential evolutionary changes in molecular stucture have tended to select out those conformations of proteins and groups of proteins which are most efficient. Through frequent gene mutation, proteins and functional combinations of proteins have evolved towards optima of form and function. This process is repeated across all molecular functions in the cell, including those self-serving processes which optimize genetic stability and molecular repair.

Stable and unstable systems

Stable systems

Complex biological states are underpinned by stability of molecular structure and function down the generations and across species. The DNA–RNA–protein sequence appears to be the universal template for eukaryotic cells, and the molecular regulatory mechanisms work consistently through embryogenesis and differentiation to reproduce tissue, organ and body structure across billions of individuals. The laws of physics and nature dictate against the coalescence of molecules and the continuity of life. The Second Law of Thermodynamics dictates that systems tend to maximum disorder. The progression and syntheses of biology are thus a rearguard action against the trends in the physical universe. The forces and processes which order genetic material and expression, cell, tissue and organism structure must thus be very stable in a universe of physical dysequilibrium. We need to understand how disordered systems represented by random mixtures of organic molecules may 'crystallize' to ordered systems from a huge range of possibilities, such that unstable or unreliable molecules have been eliminated from the construction set.

We may deduce that stable molecular states which provide optimal fault free function in data transmission and in all aspects of biological function have been heavily selected out by evolution. We can see evidence of stability in key molecular regulatory processes; in the control of the expression of the genome; in the morphogenesis of individual organisms from the linear genetic code of the gametes; and in the evolution of species by stepwise change in genes.

Stability and instability are both dynamic concepts with a time element which may extend from fractions of a microsecond to generations. Stability may be defined as the attainment of a steady state. No biological structure is fixed and immutable in space and time. The steady state may include patterns of cyclicity, or even cross generations. Stability allows integrity and consistency of form and function in cell biology, from generation to generation, and from individual to individual.

Genetic programmes exhibit great order, but we do not yet understand the controls and timing of gene expression, or of the proteins in the cell and organism over short and long periods. Spontaneous order is clearly critical to gene regulatory systems. The simultaneous and sequential function of complex patterns of large numbers of genes must be coordinated through self-organizing behaviours working to thermodynamic laws. Genomic function appears to comprise organized and evolutionary conserved functional subsystems and regulatory cascades of genes.[8–10]

Unstable systems

Cancer is defined by instability in cellular processes. Instability is defined as dysequilibrium from the steady state. It is crucial in biology. Over short time scales, it may involve the accretion or loss of components of a system. At the cellular level, this may include the gain or loss of cells within a diseased tissue or tumour. Over long time scales, it allows for change, adaptation and evolutionary speciation. Nothing in the biological world is static. Everything is dynamic at every timescale, from molecular processes over nanoseconds to physiological changes over decades. The static model of the cell describes a structure composed of various subunits (nucleus, cell membrane, organelles) with solid boundaries. That model disguises the dynamic complexity of cell biology. Cells are unstable in time and space. A dynamic model of the cell embraces fluidity and instability at all levels. This manifests in molecular vibration and mobility, synthesis and degradation, interaction and catalysts over huge numbers of moieties with millions of iterations per second, in the nucleus, cytoplasm, organelles and membranes.

The genetic code as an information technology

DNA is an information technology.[2] The human genome is a massive data set, whose raw information content has been documented by the Human Genome Project (HGP).[11,12] The information framework is the genetic code, with its four base elements, akin to binary code with two elements; its triplet combinations, akin to the eight bit elements of digital computing; its genes, akin to data set files; and its chromosomes akin to data file directories.

Technology now exists to catalogue each individual human genotype using microarrays of gene probes on a silicon chip. Genetic databases may assist in disease prediction, but may also have profound consequences for insurance, confidentiality and forensic science. We may use components of these data sets to understand the processes of disease, including human malignancy, but the interpretation of genetic data will require detailed studies of the biology and health of numerous individuals. Attention has recently been drawn to this issue in the press by debate on the commercial use of the health records of the Icelandic population, a relatively homogenous genetic pool of some 200 000 individuals in which to correlate clinical and HGP data.

Modelling higher orders of biological complexity

The raw sequence data of DNA does not tell us how genes are organised into functional modules, or how this one dimensional linear data set coordinates three-dimensional structure at the organelle, cell, tissue, organ, whole body and multi-organism level, let alone its creation of the conditions for higher intellectual function. The information template of the linear genetic code does not contain a three-dimensional map of the cell, tissue, organ or body. All structure and behaviour must derive from spontaneous interactions of linear amino acid sequences which have been optimized according to the laws of thermodynamics. Molecules self-organize structure and behaviour. Linear instructions translate into morphology as each molecule imposes order on the next element. Thus, a protein product of a gene will always assume a particular three-dimensional structure in given conditions. Similarly, specific molecular interactions, such as gene switching or enzyme action, will always proceed in a given way in given thermodynamic conditions. Thus, a bilipid membrane will form spontaneously when component lipids are synthesized and come together.

Modelling the complexity of molecular and cell biology should thus be easier because life exists within stable molecular and physical states which are far from random. Cell function emerges from the cooperation of numerous individual molecules, and the autocatalysis of molecules and polymers under thermodynamic laws. This cooperative function of individual components cannot be mapped out precisely within the genome. Spontaneous order may emerge in complex dynamic systems containing large numbers of elements, for example membrane function, intracellular transport systems, and cell skeleton assembly, such that the cell is a factory comprising automatic interactions of subassemblies of molecular machines.[8] Many of these subassemblies appear to have emerged early in evolution and to have been conserved and adapted thereafter. One example is the Nucleotide Binding Domain (NBD), an energy dependent engine for cell functions in organisms as diverse as bacteria and human cells.[9] Darwinian evolution thus appears to work through the selection of such ensembles of gene systems as well as through individual mutations. Stability models make genomic evolution of the huge number of possible combinations much more explanable and ‘realistic’.

Tumours as complex adaptive systems

Gene regulatory subsystems which are usually in balance may be tipped to disorder by slight changes. Changes in the control of critical cell functions in tumour genesis may be due to the most subtle mutational changes in key regulatory genes, such as p53, producing cascades of dysfunctional changes at the interface of the various genetic components.

Complex adaptive systems may thus be a tool to unravel the behaviour of tumours. The complexity and interrelationships of signalling pathways defy simple models. Schwab and Pienta[4] have considered how many of the observed disorders in tumour structure and function result from disturbances of the signal transduction pathways which correlate environmental events with instructions to the nucleus to transcribe genes. They apply complex adaptive modelling to the ras gene transduction pathway and its growth factor signalling pathway which can change gene transcription or cytoskeletal structure. The ras protein is a membrane bound receptor which appears to recognize a variety of cell signals (tags), and to produce a variety of responses, according to locally prevailing cell properties. Mutation in the ras gene thus produces a dysfunctional protein with consequences for both signal response and global cell architecture. The cmyc and bc12 are other well characterized families of genes which can influence both cell proliferation and apoptosis and thus exhibit compex behaviour according to prevailing conditions.

The three major structural components of tissues, the extracellular matrix (ECM), the cytoskeleton and the nuclear matrix, exhibit the properties of complex adaptive systems which are altered in neoplasia.[4] The ECM influences cell shape, function and differentiation through cell adhesion molecules, such as integrins and E-cadherin for example. Metastatic cells in particular exhibit altered responsiveness to the ECM, as illustrated by their migratory and invasive tendencies. The cytoskeletal forms are also highly modified by the adjacent environment and in turn modify cell structure and function, as, for example, in signal transduction. The nuclear matrix may be cell type specific and influence cell proliferation, as through the organization of cell cycle proteins.[13] Thus, small mutational changes in signalling genes may produce dramatic shifts in form, function and differentiation from within the nucleus.

The control of three dimensional form in tissues and tumours

Form and shape are the key consequences of all the cooperating and interdependent but autonomous processes of biological self-organization. Morphogenesis must be inherently robust. The range of outcomes must be reproducible and relatively limited for a given combination of developmental gene expressions. Such rhythmic phenomena cannot be defined in DNA and yet are reproducible. Thus, for example, a pattern determining colour or organ shape and size does not interfere with the control of other systems such as cell cycle regulation. It has been hypothesized that specific morphogens and gradients of growth factors might determine form and impose coordinates for development and positional information during embryogenesis and tissue homeostasis later in life.[14] This search may thus be unproductive. It may be that even complex structures are manifestations of spontaneous order. Ordered pattern formation may emerge from simple general laws, such as those imposed by geometric positions of cells relative to each other in two and three dimensions, the internal polarity of cells and asymmetric distribution of molecules and skeletal elements.

In 1952, Turing[15] showed how simple mathematical models could describe the spontaneous generation of complex spatial patterns such as standing waves in organisms. Many changes in morphology, such as those demonstrated by experiments with homeotic mutants in the fruit fly Drosophila and the nematode *C. elegans*, can be modelled by simple binary codes and developmental decisions.

Modelling differentiation and metaplasia

The precise control of differentiation is essential to the emergence of normal tissues and organs and to the development of multicellular organisms. The reproducibility of form down and across generations testifies to the power and reliability of those genetic and molecular mechanisms which determine morphology during embryogenesis, postnatal growth and maturity. Some pathways of differentiation are reversible, while others are irreversible. Some forms which we see in tumours may be a true reversal in the differentiation pathways, while others may be a true manifestation of disorder during tumour growth. Patterns of differentiation are also likely to be stable and highly conserved. One example of this is seen in the way that many of the stages of the early embryo resemble those of other species. Embryogenesis and differentiation unfold along simple branching pathways. Mature cell types are discretely different and do not ‘intergrade’, or merge across a spectrum. This suggests that certain morphologies represent peaks of functional stability rather than representing a continuum of change in the spectrum of cell types. Cell types may thus represent combinatorial systems of gene development and function, or patterns of stable ‘attractor’ systems.

Some of the progressive modifications in cell development, morphogenesis and differentiation may lead to dramatic switches in form and function, similar to the bifurcation behaviours of Chaos Theory.[6] This suggests other possible explanations for malignant behaviour based on simple and subtle changes in cell regulation. According to the thermodynamic model, a cell type or a stem cell may be defined as a recurrent pattern of gene activity (a chaotic attractor). Mathematical modelling of complexity in the genome and ensembles of gene regulatory systems then allows predictions to be made about phenomena such as metaplasia and the ability of cell types to differentiate to a few other cell types. For example, human epithelial tissues are believed to comprise 14 subsets of mesodermal and endodermal origin.[7] These tissues are differentiated but display metaplastic potential which may be differentiated by relatively simple control switches. Thus, the study of patterns of metaplasia allows us to study processes of interchange between epithelia, as for example in endometriosis, squamous metaplasia in the bladder in bilharzia, and glandular metaplasia in Barrett’s oesophagus.

How much of malignant transformation is new acquisition, and how much is pre-existing suppressed patterns of instability within the genome? The genome may harbour unused patterns of gene expression for redundant or unwanted cell types including the malignant phenotype. Metaplasia would then represent one deviation from the usual maturation pathways. There may well be uncommon developmental opportunities or unused pathways within the genome which may be activated in specific circumstances. Because patterns of metaplasia are also consistent and highly ordered, it seems likely that the patterns are intrinsic and reactivated by triggers such as environmental damage or mutation.

Instability in tumours

The balance between stability and instability, or between order and disorder, is poised both in normal and abnormal cell systems. The states of perturbation permissible to tumours may also be far from random. The characterization of such instabilities may help us understand, define, predict and disrupt the behaviour of tumours. To survive and prosper, tumour cells must retain most of the functional controls of normal cells, including those of cell maturation, proliferation and programmed cell death. The switch to malignant behaviour often reflects very subtle changes in normal regulation. A single sequence mutation can significantly change the electrodynamic conformation of the protein and hence its function, as for example in the haemoglobinopathies and mutant forms of p53. Single mutations in regulatory genes may cause changes throughout systems of interrelated genes.[16]

All cells in solid tissues and tumours occupy a unique position in time and space in relation to others, and thus influence the behaviour of neighbouring cells. Tumour cells may be influenced by stromal interactions, such as during angiogenesis. Kinzler and Vogelstein[17] have modelled the function of mutations in stromal cells which support tumour development. They refer to such facilitating mutations as landscaper defects. Thus, in the spectrum of colorectal cancers, sporadic mutations may underly the mucosa–adenoma–carcinoma sequence. Gatekeeper defective genes such as FAP increase genetic instability and hence the tumour rate in Familial Adenomatous Polyposis syndrome. Caretaker gene defects in the DNA repair system, [18] as in Hereditary Non-Polyposis Colorectal Cancer, substantially increase the cancer risk, while landscaper defects produce a lesser risk, as in hamartomas of the colon. The nature of the gene instability thus has a demonstrable bearing on tissue and tumour morphology.

The complex relationships between genetic components and subcellular systems suggests that seemingly unrelated perturbations may be used to change the behaviour of cells. Thus, for example, therapeutic efforts targeted on the cytoskeleton may disrupt many key elements of cell regulation and coordination, such as spindle formation at mitosis. Modulation of cell function by subtle techniques may be as effective as strategies of cell destruction using potent, toxic anticancer drugs. The complexities of tumour form may lead to better therapies, because the genetic defects unique to each tumour also define specific molecular targets. The challenge is to amplify these subtle functional and structural defects to create more highly selective therapies in a form which can be delivered to the tumours.

Conclusions

We now have mathematical tools and concepts with which to model the complexity in tissue and tumour biology, and to explain the interdependence and the autonomy of biological processes within normal and malignant cells. We are now in a position to supercede simplistic, linear thinking of cause and effect in the aetiology of tumours with a greater appreciation of the exquisite subtlety of natural order and disorder. This may produce new and unusual strategies to manipulate the behaviour of those tumours which defy surgical control.

Many aspects of the information explosion and the new biological models which they generate have yet to impact upon the practice and craft of surgical oncology. They can nevertheless have a profound effect upon the world in which we work, and our intellectual approach to the problems of cancer. They provide enormous opportunities for progress in surgical oncology.

References

1. Rew DA. Heterogeneity, biodiversity and bioperversity in solid neoplasms. *Eur J Surg Oncol* 1996; **22**: 469–73.
2. Anon. Science and technology: drowning in data. *The Economist* 1999; **351** 8125: 131–2.
3. Butler D. Computing 2010: from black holes to biology. *Nature* 1999; **402** (Suppl.): C67–C70.
4. Schwab ED, Pienta KJ. Explaining aberrations of cell structure and cell signalling in cancer using complex adaptive systems. *Advances in Molecular and Cell Biology* 1997; **24**: 207–47.
5. Holland J. *Hidden Order: How Adaptation Builds Complexity.* New York: Addison Wesley, 1995.
6. Rew DA. Chaos, tumour biology and non-linear dynamics. *Eur J Surg Oncol* 1999; **25**: 86–9.
7. Kauffman SA. *The Origins of Order and Disorder: Self Organisation and Self Selection in Evolution.* New York: Oxford University Press, 1993.
8. Alberts B. The cell as a collection of protein machines. *Cell* 1998; **92**: 291–4.
9. Welsh MJ, Robertson AD, Ostedgaard LS. The ABC of a Versatile Engine. *Nature* 1998; **396**: 623–4.
10. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. Impacts: from molecular to modular cell biology. *Nature* 1999; **402** (Suppl.): C47–C52.
11. Little P. The book of genes. *Nature* 1999; **402**: 467–8.
12. Collins FS, Jegalian KG. Deciphering the code of life. *Sci Am* 1999; **12**: 50–5.
13. Fey EG, Penman S. Nuclear matrix proteins reflect cell type of origin in cultured human cells. *Proc Natl Acad Sci USA* 1988; **85**: 121–5.
14. D'Arcy Thompson. *On Growth and Form.* (2nd edn). Cambridge University Press, 1966.
15. Turing AM. The chemical basis of morphogenesis. *Philos Trans Roy Soc London B* 1952; **237**: 37.
16. Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers: a review article. *Nature* 1998; **396**: 643–9.
17. Kinzler KW, Vogelstein B. Landscaping the cancer terrain. *Science* 1998; **280**: 1036–7.
18. Wood RD. DNA repair in eukaryotes. *Ann Rev Biochem* 1996; **65**: 135–67.
19. Malins DC, Polissar NL, Schaefer S, Su Y, Vinson M. A unified theory of carcinogenesis based on order–disorder transitions in DNA structure as studied in the human ovary and breast. *Proc Natl Acad Sci USA* 1998; **95**: 7637–42.