

The Structure of the Computational Biosciences

A Shortest Tour of Bioinformatics, Systems Biology and the OMICS

by Jotun Hein (1.6.07)

Life on Earth

The overwhelming percentage of bioscience is directed towards earth life, although many principles applied make no specific reference to this and many ideas and questions apply generally. Life beyond earth issues are clearly very interesting, but of little practical importance presently. The moment (and if) other life forms are found, question will be asked about similarities and differences in overall structure of life forms. In certain cases, like Earth/Mars, it could even legitimately be asked if the overall structure was a sign of homology? This is not pure silliness, as it is known that on very long time span there is exchange of surface material between Mars and Earth. It has been tested, in what retrospectively seems naïve, but is still a valiant and fun attempt. This was done by asking if the root of life on Earth was older than the age of the Earth (Eigen,). (The answer was “no”!).

Life on Earth has an overall architecture and has chosen certain key components. Due to evolution all instances of Life on Earth are related by one very large genealogical structure that presently is being characterized. Behind all research about Life on Earth looms the question about its origin, which clearly remains unsettled.

The Architecture and basic building blocks of Life on Earth.

Defining life has been the subject of whole books (Deamer, 1994), but for our purposes it is: A connected physical system, that has the ability to replicate identically or with slight modification and has a metabolism. Metabolism will take input molecules process them to maintain itself and sustain replication. It can be illuminating to search for examples that falls into or almost falls into this definition, but still not be life. Such examples could be fire, crystals and more difficult again – viruses. There has been quite a literature on abstract approaches to defining life starting with von Neumann's Self-reproducing automata (published posthumously, 1967), Conway's game of life (1968), Ganti's soft automata (1971, 2003), Rosen's (1991) essays on “Life Itself” and a large Artificial Life literature (1999 – special issue). These approaches are very interesting, but are quite detached from real examples of life. Given the increased ability to simulate computationally and the ability to build bio-nanostructures, mathematical definitions of self-reproducing automata will likely be of increasing practical interest.

It is very difficult to deduce any specific details from how life must be from abstract life definitions and in the search for life in the Universe, one quickly resorts to restrictions that are strongly Earth inspired: water, chirality, duplicating information molecule and more.

Common to all members on the tree of life is cellularity. (This leaves viruses out. Viruses are molecular parasites and of major interest, but have been left out in the above description. They are incomplete viewed as organisms and probably exchange genetic material with hosts and each other, that they need separate treatment in all respects.) This creates that natural division of life characteristics: sub-cellular, cellular and supra-cellular. Key to organisation of the central is what has been called “The Central Dogma”, that describes the relationship between 3 major classes of macromolecules - DNA, RNA and protein: DNA can replicate itself (with some help), it can be transcribed into RNA that can be translated into protein. DNA will be information carrier and protein will have enzymatic, structural and regulatory roles. This picture needs some refinements, but still is a decent rough description of key organisational features of cells on Earth. Sugars and lipids are on par in importance to the three mentioned classes and are central in energy storage and membranes (compartmentalisation and cellularity). (Question: Is there a natural algorithm taking the known network of a cell (metabolism, regulation, duplication,..) and getting “The Central Dogma” and possibly levels of detail lower than complete description and higher than the “The Central Dogma”? Question: What alternatives to “The Central Dogma” could be imagined?)

The features just described have amino acids, nucleotides, sugars, water, ions and some key metabolites as building blocks. Again concerning these building blocks, question remains concerning if life could have been implemented using very different sets. The fact that many organic molecules arise easily by non-biological means does not answer this as maybe alternative molecules do also arise easily. Benner () and

Szathmary () have asked more limited questions, concerning which amino acids and nucleotides could be imagined if proteins and DNA should have their present functionalities.

The genealogical structure of Life on Earth.

Any description of Earth life on covering its history will centre on the “Tree of Life” (TOL). The present knowledge of TOL is the result of centuries of work, where Linneaus’ (1790) hierarchal, but still non-evolutionary, classification was one of the first major steps. The establishment of evolution (1859) legitimized the search for global tree. For the century following Darwin, phylogenetic knowledge accumulated with focus on closely related species, as their characteristics (morphology, cellular architecture, chromosomal re-arrangements,..) are more easily comparable by the then available methods. Only in the late 1960s, did sequence data expand the scope of phylogenetic knowledge and we are now seeing a completion of this catalogization driven by genome sequencing on a massive scale. Sequence data heralded a revolution in phylogenetics for several reasons: Sequence data can be obtained on a very large scale, they are exact, quantifiable, their evolution much easier to model and finally they allow comparison of very distant organisms. The phylogenetic project will probably been completed by 2015. This is clearly a triumph, but there is also a negative corollary to this, in what is not known by then, probably will not be known ever, as it is difficult to see what information could complement complete genomes. The TOL has a few major features, that will be summarized in any modern textbook on evolution, but is otherwise characterized by the mass of information and detail. Optionally, the TOL can include extinct species. Reducing the TOL is key to summarizing it usefully. Starting with the origin of life (~3.5-3.8 BYR) via the last universal common ancestor (LUCA) (~2.7-3.2 BYR) and moving towards the present the duplications leading to the major divisions among organisms will be encountered. Prokaryotes splits out, Archeobacter splits out, metazoans appears, animals-plants-fungi separates. Alternatively, as is the perspective in Dawkins’ *Ancestors* (2004) one could start with humans and move towards the root. This creates a list and one can then ask when any key characteristic of humans were first seen on the phylogeny (primates, mammals, , metazoans,..), this would also traces a series of splits, but allow focus on humans. Finally, how does TOL look, when addressing large scale features of the tree? How well-determined is the estimated TOL? How large is the determined tree within the complete tree of extant species? How large is the tree of extant tree within the tree also including extinct species? Within any reconstructed TOL the evolutionary fate of a set nucleotides have been traced. At a given time in the reconstruction of the TOL, how many nucleotide-years has been observed. Presently, the number must be of the order 10^{20} nucleotide-years, but this number will increase as more genomes are determined. A final issue is the problem of horizontal transfer that undermines the use of phylogenies to describe the history of sequences. Just get a general feeling for the size of several of the above quantifies could be done through idealized models involving speciation, extinction and horizontal transfer.

The Origin of Life

Origin of Life research in this context tries to address: How do biological systems arise out of a non-biological system? Given a definition of life, this is should be a solid question to address.

Origin of Life research occupies a peculiar niche in science, in that it easily gets top ranking in any survey of providing the grand questions to be answered. On the other hand funding has been low and in contrast to most other fields there has been little possibility of an application of the research. As a consequence Origin of Life research has been often been conducted by individuals that had the interest and freedom to pursue this. Alternatively progress has been a bi-product of other research, such as the chemistry of self-replicating molecules, geology or astronomy. The research has been dominated by an unusual amount of speculation, which is probably the consequence or its fragmentedness, lack of an “empirical program” and (on the positive side) that it actually involves wide spanning questions that needs bold intellectual steps. But in view of many scientists, Origins of Life research has something ridiculous over it. In the longer time perspective this might be un-justified.

The reasons for this optimism is that the factors that historically have made it special in a negative sense, have disappeared: the planned space mission to bring samples back from Mars and test for life elsewhere, the large scale discovery of other planetary systems, the increased ability to simulate and design small molecular systems, the massively increased research in molecular evolution that has increased our knowledge of ancestral organism (still later than and very complex compared to the organisation of biomolecules ~3.5-3.8 BYR ago). This change in situation does not create instant success for the field, but could well do so over one or two decades.

Some of the central concepts in the description of Life on Earth – cell, metabolism, information carrying molecule, individuals, two sexes, populations, selection/neutrality, ontogeny - are so general that one suspects that they are universal. Even if they were, questions remain of how many details would also be universal.

Clearly, almost all present modelling of biological systems occur with focus within the framework created by Life on Earth.

Systems Biology, Integrative Modelling and Computational Biology.

The Biosciences has during the last 5 decades experienced stunning scientific successes. The last twenty years have seen experienced a major quantification of data and the increased use of databases, statistics and computers. This trend is often associated with the field of Bioinformatics. We are presently witnessing an even more ambitious phase (by some associated the term Systems Biology or Integrative or Multilevel Modelling). There is great variation in the interpretation of these terms, but central goals are:

- Dynamic Modelling of a Biological System at Multiple Levels.
- Obtain Data in sufficient quantity, quality and variety to be able to characterize and parametrize dynamic models.
- Make Testable and Interpretable Predictions of the behaviour of the system.

There is also widespread scepticism about the feasibility of such goals as even small biological systems are highly complex, but it is clear that event partial success of such goals would have major scientific consequences: It presents an unprecedented level of understanding and will lead to important medical advances. Biology will undergo a transformation, where models of biological systems will become ubiquitous tools for researchers, long-term goals in biology are dynamic and predictive modelling of increasingly complex systems, like cells, tissues, organs and eventually individuals. Some projects attempting this have already been launched, but success of these projects could well take decades of research.

Although systems biology only recently has become a major activity, it has an extended history. Very early descriptions of biological systems are Cannon (1932) describing blood dynamics in terms of homeostasis, Wiener (1948) using stabilizing feedbacks and Bertalanfy (1950) in his general systems theory. The concept of feedback creates a description of a system where characteristics using many components are conserved and thus can be formulated without reference to individual components. This leads to the often elusive concept of emergence (Kim, 1999). In the wake of the operon model of the gene came the first mathematical models of gene regulation (Goodwin, 1964) and this was quickly followed by models of network dynamics either in the form of the biochemical systems theory of Savageau (1969+, 1976) metabolic control analysis starting in 1965 (Higgins) and developing into a general theory early-mid 70s independently by two groups (Kacser and Burns, 1973 and Heinrich and Rapoport, 1973). These models described metabolic networks, but networks include at least regulatory networks, signal transduction networks, protein interaction networks. Modelling now also includes interacting networks (for instance regulatory and metabolic, Yeung et al., 2006) and evolution (Wiuf et al., 2006).

Although systems biology has been described as the biology of networks, it is clearly more and does cover topics such as biological structures/objects and their dynamics, biological motors, spatial organisation, form and pattern formation. These are very general concepts and most, if not all, biological systems can be described in terms of these. Biological structures/objects (ribosome, cell, horse,..) will have a certain permanence and has rules for its creation, destruction and interaction with other biological structures/objects.

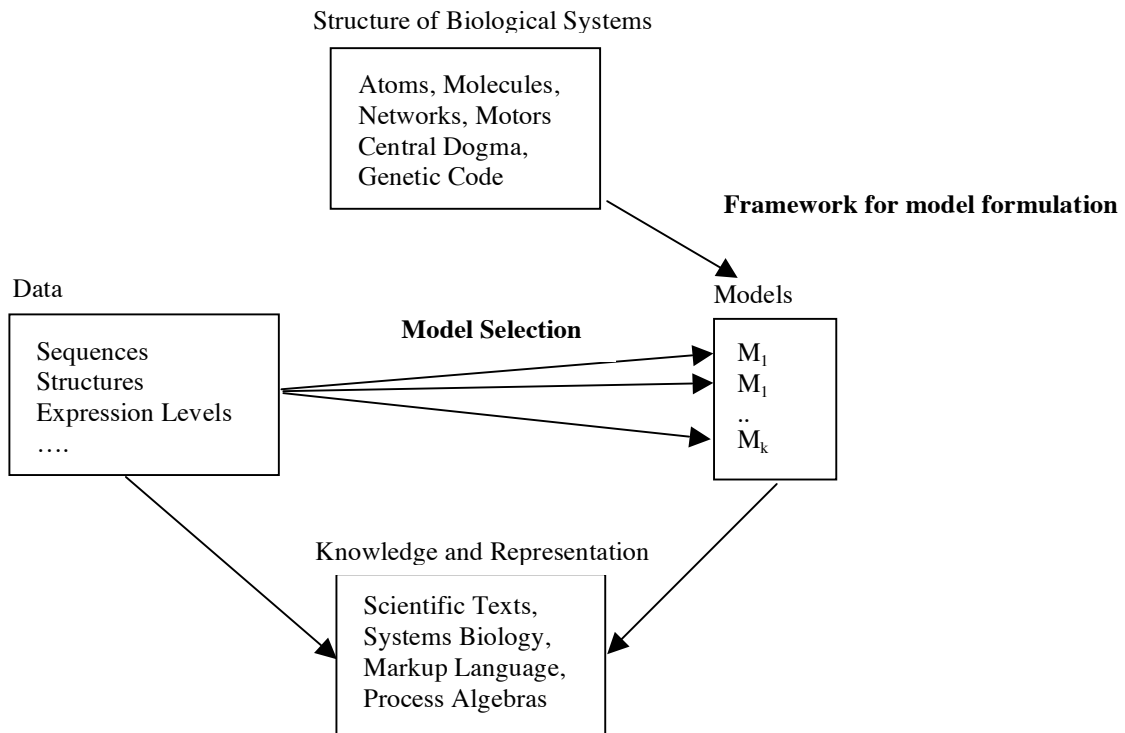
Structure determination for biological macromolecules has at least a 50 year history, but has also experienced a major revolution both in quantity and quality (Campbell, 2002, 2007), where the main recent advances covers the ability of monitor structure and dynamics *in vivo*, single molecule transitions and also much larger structures due to new microscopy techniques.

Motors describe molecular assemblies that convert chemical energy into mechanical motion such as muscle proteins, bacterial rotors or the spindle separating the chromosomes during meiosis and mitosis (Bustamente et al., 2004).

The most famous model pattern formation was proposed by Turing in 1952 and still is the focus of much research, although biology research has led to a series of contending theories (Deutsch and Dorman, 2003).

A general trend in the formulation of these models, besides diversification and specialisation, is that the first models are fully deterministic, followed by stochastic analogues and finally formulated as a statistical inference problem.

The recent rise of system biology is mostly due to the availability of high throughput data and dynamical modelling due the increased power of computers. The completion of the human genome has yielded a “parts list” that has laid down the challenge of providing an integrated understating of their function. This new area has at least 7 major components: i. the data, ii. the concepts and biological knowledge, iii. the models, iv. evolution and variation, v. inference and model selection, vi. representation of models and knowledge and finally, vii. what are the big questions relating to this new bioscience?



The structure of biological systems summarized present knowledge of the architecture of life on earth. This naturally provides as framework for the formulation of dynamic models of biological systems. Structure of biological systems clearly also influences how we represent knowledge and data. Data will in conjunction with model selection techniques choose models that are compatible with observations and discard models that are not. Both data and models are stored using different kinds of representation, from more informal scientific text to highly formalized representations, that are computer readable and readably

It is clear that a field encompassing so many facets necessitates collaboration by statisticians, computer scientists, physical chemists and biologists. This does not materialize without large-scale and targeted investment on a relative to classical bioscience.

Data and High-Throughput Technologies.

Recent years have been dominated by both the large amount of data and the emergence of techniques allowing new observations of cellular states and content. These data have fundamentally changed the Biosciences from being “Formulate Hypothesis, then get appropriate Data” to having a large component of “Get Large and Diverse Data Sets, then explore/mine its consequence for present understanding”.

This data-revolution started with sequence data, but has then spread to higher levels and more complex data types involving both dynamics and structure. First, at the gene scale, but since to genomes and

presently both to genomes characterizing a population and bulk sequencing of genomes (Tringe and Rubin, 2005). A window to the dynamics of a cell is through its level of mRNA monitored by the techniques of expression analysis and this has since been complemented by the levels and presence of other molecules, like proteins or smaller biomolecules, through the use of techniques of mass spectroscopy and metabolomics/metabonomics.

Sequence Data (Metzker, 2005). Starting in the 50s with very time consuming methods allowing the sequencing of small size proteins and followed in the late 60s with the first RNA sequencings, DNA sequencing accelerated enormously in the late 70s and has experienced an ever-increasing pace of sequence accumulation. Presently, we are experiencing large scale sequencing of genomes and partial sequencing of genomes from a population. It is clear that present technologies will improve and in 2-5 years sequencing will be so cheap that of interest will be sequenced on demand.

Expression Data (Allison et al., 2006; Elvidge, 2006). With the first investigations from 1995, these techniques measure mRNA levels absolutely or relatively to a standard, unfortunately subject to variation. The price has dropped enormously and their domain of applications have continuously expanded to include classification, network inference, regulatory signal inference and more. Additionally, precision and quality will go up.

Proteomics and Protein Interactions (Bork et al., 2004; Souchelnytskel, S.,2005; Kleppe et al.,2006). The set of all proteins (including modifications) and their pairwise affinities can be determined by mass spectroscopy and immuno-precipitation, creating an enormous set of noisy data of potential great use. The most immediate biological conclusions relate to protein complexes, signalling, degradation and modification.

Metabonomics/Metabolomics and Small Molecule Detection (Shulaev, 2006; Lindon, Holmes and Nicholson, 2006). This refer to *in vivo* techniques monitoring the levels of metabolites in cell and its main areas of application are the study of metabolism and for instance drug degradation.

Structures from Crystallography, NMR and Cryo-EM (Campbell, 2002; Chandonia and Brenner, 2006; Bravo and Aloy, 2006). Again from being very labour intensive techniques 5 decades ago, these have become high-throughput and almost all protein types will be known by the end of this decade. The 3 techniques cover different areas with crystallography giving high precision information about crystallisable structures, NMR allowing *in vivo* determination of small molecules and cryo-EM giving coarse information about large structures.

Microscopy (Glaeser et al. 2007; Frank, 2006, Santos et al, 2004; Sverlov et al. 2003; Pepperkok and Ellenberg, 2006) has over the last 5-10 years provided a long series of new techniques, like atomic force microscopy (AFM), small angle scattering techniques (SAXS) and cryo-techniques giving partial information in small, medium large scale molecules and biological structures, respectively.

Single Molecule Measurements (Kulzer and Orrit, 2004; Kou, Xie and Liu, 2005) allows the detection of movements of labelled positions on a molecule in real time and can give detailed information in the dynamics of movements of for instance parts of RNA, proteins or molecules in a membrane.

Clearly, these data sources constitute a revolution in biology and together possibly legitimizes the ambitious goals of systems biology. However, the inherent limitations must still be underlined: Many measurements are cell averages (both over cells and time as for instance expression data), possibly confined to *in vitro* conditions (like many kinds of microscopy or crystallography), extremely noisy (like expression and protein interaction data) or quite limited (like single molecule measurements), or measured under unphysiological or at least very special conditions, like most kinetic parameters. *In vivo* measurement might be measured under one set of conditions, but used in models describing another set of conditions – kinetic parameters under different pH and temperature for instance.

Relative to the ultimate goal of systems biology, these data still needs major advances in both inference and computational tools to be efficiently harvested in a translation to a dynamic model of a cell (for instance). And the question of the sufficiency of these data relative to this goal is still completely open.

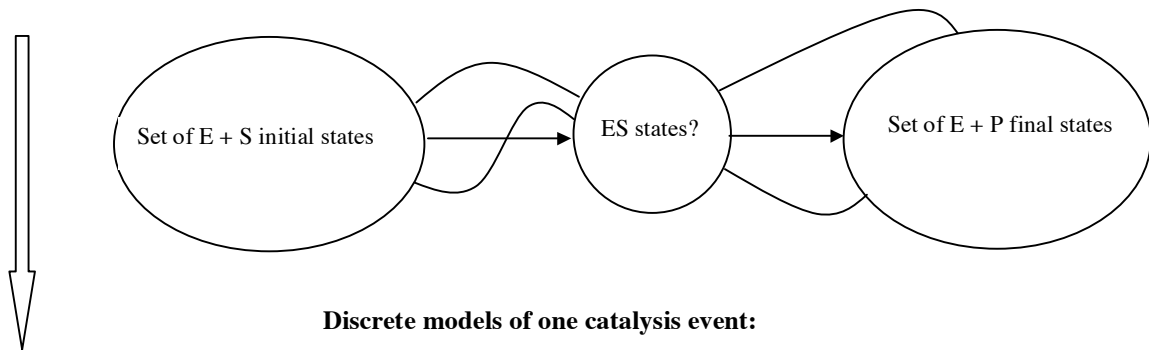
Ranking of importance of individual data types is not meaningful without reference to specific biological investigations, but quantification of different kinds of data types in a way that would be meaningful, relative the goals of systems biology is definitely of interest. The size in bits is very straightforward and would be informative, but how much does each of them constrain dynamical models of the biological system could be very interesting to know. For instance to characterize a dynamic model of a cell, what is the relative worth of proteomic and expression data?

Biological Levels, Themes and Structural Knowledge.

Not any domain of biology is suited to be addressed by the techniques from Systems Biology. At least 3 criteria must be fulfilled: i. It must be of a size and nature that can be modelled. ii. Data must be obtainable, so the system can be acceptably parameterized or determined and iii. It must be sufficiently autonomous, so modelling in itself is meaningful.

Biological concepts are frequently arranged in levels, where the concept at one level (for instance an enzyme) are autonomous in the sense that its behaviour (as for instance described by Michaelis-Mentens' constant) can be summarized with properties that does not need reference to lower levels (for instance its constituent atoms). The concepts and levels are naturally defined during biological research and is not part of an explicit meta-scientific program, although these concepts are often central to such debates (Kim, 1999; Polanyi, 1968). The levels are necessities both for modelling and for understanding. Using kinetics to describe enzyme behaviour instead of molecular dynamics allows an acceleration of more than 8 orders of magnitude in calculations. Using the concept "enzyme" allows a much lower dimensional representation that can be handled intellectually in contrast to a more complete description.

A molecular dynamics sample path involving one catalysis event:



Discrete models of one catalysis event:



The first models of enzymatic action were published in the first decades of the 20th century by Michaels, Menten, Henry, Haldane and others. These are models with a few states and interactions and lead to simple equations allowing explicit analysis of their dynamics. It is clear that the real picture is vastly more complicated and started to be analyzed in the last decade of the 20th century by molecular dynamics. A naïve MD simulation could have 10³-10⁴ atomic positions (enzyme, substrate, water and ions) and would in principle be simulated for 10⁹ steps of 10⁻¹⁵ second duration. Clever techniques such as Transition Path Sampling (Bolhuis et al, 2002) can improve significantly on this and force dynamic trajectory toward a desired end state and still allow rates to be calculated and identification of key interacting groups.

Making discrete approximations have also been applied to other complex molecular events such as protein folding (Fersht, 2004).

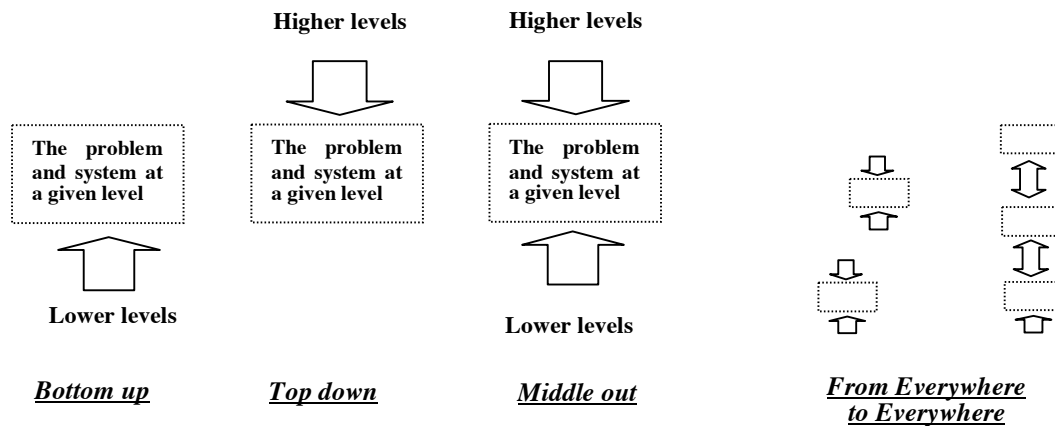
These are simple illustrations of events and objects formulated at different levels. As computational modelling become increasingly dominant in biology, stringent definitions of levels, objects and their dynamics could be increasingly necessary.

The cell has a series of components that are not nested: networks, motors, molecular assemblies and more. There are at least four major classes of sub-cellular networks: metabolic networks, regulatory networks, signal transduction network and protein interaction network (PIN).

| Level | Example(s) | Data | Modelling Techniques |
|---------------------|--|---|---|
| Atomic, Molecules | α -globin, water, cell membrane | Single molecules measurements, X-ray diffraction of crystals, NMR | Classical potentials and Newtonian Dynamics, Quantum Mechanics, |
| Molecular complexes | Ribosome, hemoglobin, | single molecule measurements | Mechanical analogues models, Continuous Time Markov Chain with finite state space |
| Molecule | | Concentration of metabolites, | ODEs (many |

| | | | |
|-----------------------------|---|---|--|
| concentrations | | fate of isotopes in different molecules, | molecules/concentrations), kinetics, |
| Metabolic Network | Citric Acid Cycle | Enzyme and metabolite concentrations and metabonomics | ODEs, Kinetic Models, Flux Analysis, |
| Regulatory Network | α -globins and their regulators | Expression data | Boolean networks, Petri Nets, ODEs |
| Signal Transduction | Mitogen-activated protein-kinase (MAPK) | Protein Interaction and Expression Data | ODEs, Continuous Time Markov Chains, |
| Protein Interaction Network | Yeast PIN | Mass Spectroscopy | No dynamics involved, i.e. a data type. |
| Motors | Flagellar Motor, | Microscopy, single molecule fluorescence | Mechanical Analogue Models |
| Cell(s) | B-Cell, zygote, E.coli, | Microscopy, expression data, proteomics,... | Integration of genetic, mechanical and network models. |
| Tissue | Cancer, | | Partial differential equation (PDEs), cellular automata. |
| Organ | Liver, lung, heart | Mechanical measurements, | Multilevel integrated modelling, including mechanics. |

Levels are partially ordered since objects in one can consist of many objects from a lower level. This has led to characterisations of the approaches to modelling. *Bottom-up* will start with thorough descriptions of low level and move to higher levels always explaining higher level objects in terms of lower level objects. This approach is also often called reductionism. *Top-down* is used in biology, when the function of an object (for instance network) is assumed known and lower level (for instance a specific enzyme in the network) is then explained by serving a purpose at the higher level. Sydney Brenner and Denis Noble (Noble, 2006) often describes much practical modelling as *middle-out* in the sense that one starts at the level of interest and refer of lower levels for reductionist explanations and to higher levels to establish purpose.



In systems biology and the analysis of throughput data, there often is a lack of initial driving question and an emphasis on global modelling where top-down and bottom-down approaches will be applied simultaneously at many levels in an approach that could rather be called from everywhere to everywhere.

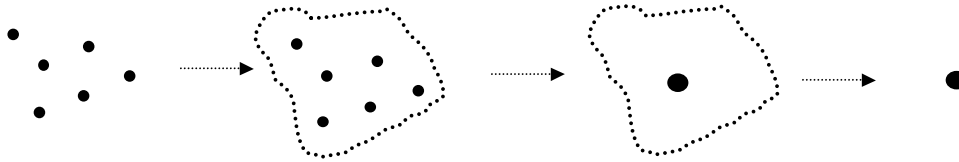
Physical-chemical, Stochastic and Dynamical Modelling Techniques.

Due to the size of biological systems modelling is done at many levels: Quantum Mechanical for small sets of atoms involved for instance enzymatic reactions, Classical Force Fields for motions and interactions of molecules involving hundreds to many thousands of atoms, Stochastic Chemical Kinetics for reactions involving tens to a thousand molecules and Deterministic Kinetics for larger number of molecules. Kinetic Models are often combined to integrate network models used to describe a regulation, metabolic pathways and more. Diffusion Models describes transport and thermal noise in biological systems.

Biological Systems are physical systems, and physics have a hierarchy of models. Berendsen (2007) describes one eleven-level hierarchy starting with relativistic quantum dynamics and ending with steady flow fluid dynamics. There is not easy identification between biological hierarchies and physical

hierarchies, although they do intersect and some biological objects can be identified with naturally defined physical objects. Biological concepts are partially historical constructs, where certain entities have been elevated to be “biological object” and its behaviour described and predicted by useful rules and equations.

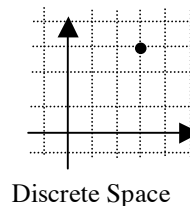
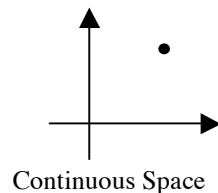
Defining or recognizing Biological Objects



For computational and conceptual reasons it is convenient to identify a series of objects (for instance atoms or molecules that can be view as a unit with properties that can be defined from the component. Often this is straight forward – a molecule is a set of atoms bound by covalent bonds or a concentration is the number of a specific type of molecule per volume. In other cases it can be harder and the quality of modelling is at risk as when coarse graining in MD simulations.

Since all biological objects are physical a reduction is possible in principle, but can be computationally prohibitive or even detrimental as it would erase central qualities in its biological definition. In modelling biological objects a series of essential decisions have to be made:

Space can enter models in a series of ways. It can be absent as in Ordinary Differential Equations (ODEs) that assume space homogeneity, which can be a reasonable assumption for a small prokaryote, but increasingly less so for larger volumes. Heterogeneity can then either enter by modelling space by a Euclidian space and operating with Partial Differential Equations (PDEs), but heterogeneity can also be modelled discretely as in cellular automata or by having a few compartments like nucleus, cytoplasm or mitochondria, between which an object can be transported.



Time can also be represented either continuously or discretely, with equations being indexed by either the real numbers or integers. This can be used to keep a space homogeneity assumption, where transport over distance is then appearing as a time-delay.



Models can be *deterministic*, where there is one dynamic path from a given set of initial conditions or stochastic, where the state of system will be described in terms of distributions. *Stochastic* models will dependent on space-time representations, different classes of models will result from more tractable discrete time – discrete state Markov chains or for instance much harder, both conceptually and computationally, Stochastic Partial Differential Equations. In general, stochasticity makes models harder, but clearly also more realistic and flexible in terms of handling uncertainty.

Biological systems are developing over time and are thus described by a subclass of differential equations, where one parameter can be singled out as time and are called dynamical systems. A few distinctions are of importance in understanding such systems: Ordinary versus Partial Differential Equations (ODEs/PDEs), where partial involves derivatives of several parameters and ordinary only of a single. These equations do not involve distributions, but have been generalized to have stochastic analogues.

These have widespread use in economics, but so far little in systems biology. Given the probabilistic nature of many biological phenomena and the centrality of diffusion in many systems, their relevance is clear.

As systems biology typically involves many levels, models can be mixed and typically are. Useful models arise in a trade-off between tractability and realism, so experience or testing can allow certain aspects to be simplified to for instance Boolean networks, if the underlying model is sufficiently switch like. The complexity of integrated models is no virtue and the simplicity of 1-level models are often attempted to be regained, by observing that the dynamics of different levels occur at very different scales, where fast dynamics levels can be assumed to be at equilibrium relative to boundary conditions determined by slow dynamic levels.

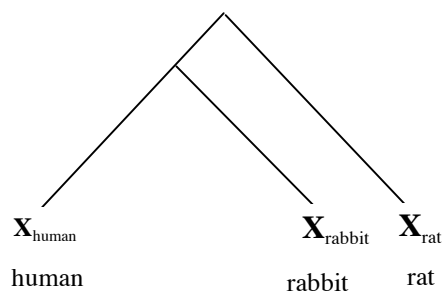
Recently (Ball et al., 2006) have also used multi-level models for enzymatic reactions, where the stochastic dynamics moves between different for instance Poisson and Diffusion Processes dependent on the number of molecules in the system.

Evolutionary and Population Variation.

Despite the emphasis on modelling, the relevance of long-term evolution, population variation and bioinformatics is as high as ever. Comparative Genomics is the outstanding success of the last 5 years, where sequence comparison is used to annotate genomes, but comparison can be applied to other levels of biology, such as protein interactions (PINs), motors, networks, forms/patterns/shapes and more. Due to the size and complexity of biological systems, population variation and association mapping is still a major contributor in proposing a few relevant genes of the 24.000 possible relevant for understanding a biological system.

Comparative Biology has recently experienced a major boom in the form of comparative genomics, where the interpretation of a genome is strongly augmented by the comparison with other genomes (Boffelli, Nobrega and Rubin, 2004, Abel Ureta-Vidal, Laurence Ettwiller, Ewan Birney, 2003). This trend is now moving into networks (Sharan and Ideker, 2006), but comparison is ubiquitous in biology and as data and models accumulate at higher biological levels and structures, the application of evolutionary modelling will expand accordingly. Optimal use of comparison will have following components:

- A set of species/biological systems with known phylogeny/relationship.
- A set of observations within the species on some homologous feature (sequence, network, heart,..)
- An evolutionary model of how this feature evolves over time on the branches of the phylogeny.



The benefit of such an analysis is the following:

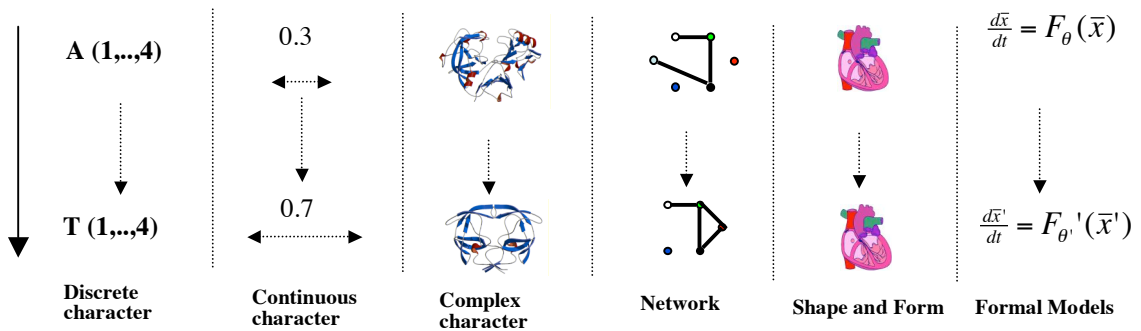
- A rational combination of information of data obtained from different species. For instance given measurements in dog and rabbit, what is the error when this is transferred to humans?
- An ability to state which information would be most valuable next. How far away evolutionary should a chosen model organism be? Which indirect measurements would be of greatest value to gain information about a parameter/data that cannot be observed directly?
- An understanding of the evolutionary process. Although this is not the primary aim of experimentalists, it has more hands-on advantages than most realize and could address fundamental questions like: Does key kinetic parameters co-evolve with known external factors like habitat or size? Which components are under positive or purifying selection? Have features of

system evolved at a constant pace or in sudden spurts? What were the properties of ancestral systems?

Evolution and *Homology* are of course of interest in them selves, but is also useful for purely functional purposes. If the two structures/molecules are homologous (derived from common ancestor), then it is legitimate to compare them and knowledge about properties and functions about one has a higher probability of the being true of the other than in the absence of homology. This is used on a large scale in sequence comparison, but is useful in general. The observation of evolution is even more useful as it allows if observation of selection and the ability to distinguish between functionality and non-functionality. These concepts are especially simple in the case of comparative genomics, where a deceleration of evolution indicates conservation of function and is used for finding genes, regulatory signal and more. An acceleration can mark a change for functional and phenotypic reasons and can thus delineate factors underlying a trait. A neutral rate of evolution is often taken as indicator of lack of importance, which again is useful information. Homology modelling of protein structures (Xiang, 2006) is good example of how evolution is used structurally: Experimental knowledge of structure of one protein sequence is transferred to a homologous sequence, where no experimental knowledge is available. Although mainly used for protein structure prediction, this concept is generally applicable. An interesting extension of this is what could be called non-homology modelling (Bystroff, 1996), where evolution is used to generate a set of variants and predictions are then biased towards these.

In modelling evolution a triad of problems must be addressed: Firstly, how to represent to object of study? Normally, this is clear from the scientific context, but can span a variety of objects from sequences, structures, graphs, processes and shapes. These problems are very well studied for sequences, less so for graphs, structures and processes, and phylogenetic shape comparison is a new field. The problems are easier in the beginning of this list grows increasingly harder. To make comparison meaningful, there must be a uniform format for representation across species, which for instance can be a problem for structures that have been determined to different levels of precision. For shapes present representation are often not suited for comparison – frequently a shape are covered by cubes with interaction among neighbors and how this is transformed into another representation in another species is not clear.

Secondly, how to describe the evolutionary change? For sequences and discrete characters in general, this is extremely well developed, based in event of individual elements (nucleotides, amino acids,..) described by a continuous time markov chain and independence of the evolution of different elements. Quantities described by continuous quantities (height, concentration, certain parameters) are often assumed to evolve by Brownian Motion or a general Diffusion Process. Models exist for networks, but are less investigated for structures, processes and shapes.



This is a non-exhaustive list of objects that can evolve. Complex characters would cover a long series that cannot be simply described. RNA and protein structures would be complex, but it easy to list others. Molecular movements and behavior are other examples. The mathematical models themselves could be subject to the same reasoning and clearly the models used to described a phenomena in humans should not be formulated independent of the models describing the homologous phenomena in mouse. A similar consideration can be made to diseases and dysfunction. As we move from simple to complex, the modelling becomes harder and have more parameters and less data and thus the inference weaker. However, complex modelling is indispensable and also more biological relevant.

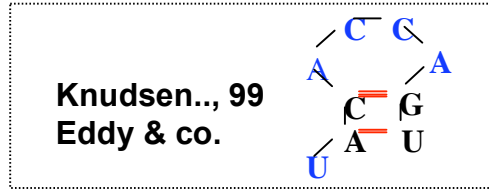
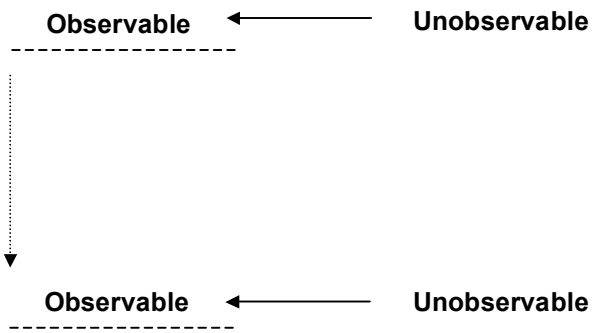
Pathways: There are key classes of networks in biology: Metabolic, Regulatory, Signalling and Protein Interaction Networks. The concept of network is so general (set of objects with pairwise relationships) that it will appear everywhere in science. The different kinds of network in biology describe different kinds of dynamics and the reasonable model of evolution will vary from class to class. Mathematical modelling of networks is a well-explored area (Dorogovtsev and Mendes, 2003), but evolutionary models are only now being undertaken (Wiuf et al., 2006). A network will typically have a discrete component – nodes and edges- and possibly a continuous component that labels edges (for instance flux) or nodes (for instance concentration). The discrete component are associated different kinds of models. For instance, metabolic pathways will have reactions (edges) deleted/added. Protein Interaction Networks will have nodes duplicated and simultaneously the edges of the duplicated node copied as well. A variety of models exists describing network change and are presently being explored.

Kinetic, Continuous and Physiological Parameters: The present setting is the natural for the use of parameters from multiple species. The problem of parameterisation of Heart Models has great similarity the phylogenetic analysis of quantitative characters that has been studied for decades in evolutionary biology, but in the last few decades this field have been in the shadow of phylogenetic analysis based on molecular sequences. In the present problem, we want to make statements about the parameters of the human Heart. The full model need k parameters, the parameters of the models – $X(t)$ – is evolving according to a Brownian motion, i.e. $X(t_1)-X(t_2) \sim N(0,(t_1-t_2)\Sigma)$, where Σ is a covariance matrix. Two additional factors complicate the problem: i. Each observation is subject to error, which again could be modelled by adding a k -dimensional normally distributed variable Y to each observed heart. ii. The physiological parameters have not been determined under identical conditions, but under varying pH, temperature etc. This adds a regression problem to the basic modelling.

Shape and Form: These concepts are ubiquitous in biology and elucidating the underlying genetic mechanisms for this is a major area of research. Patterns have been described by a series of contending models (Deutsch and Dorman, 1994) and recently models describing the evolution of the mechanism have been explored (Demidova, 2006). Statistical modelling of form is an established field of statistics (Dryden and Mardia, 2004) that views form as a random variable. Phylogenetic model of form is a new concept, but would be highly relevant for an analysis of how gross structures of the heart evolves over time. Three-dimensional physiological/mechanical models (Smith and Hunter, 2004) are well explored, but further decisions would have to be made concerning how uniquely defined equivalent (read homologous) positions in models from different species.

Mathematical Models: Models tends to be developed with the aim of describing one physical systems, while in biology the situation arises that multiple models are developed to homologous systems. In principle such models should be develop simultaneously and be appropriately (read: phylogenetically) coupled. This is fully feasible and is to some extent already being done: Potentially, a model in one species and be viewed as a general model with one set of parameter values, while another species has another set of parameter values, which would correspond the Brownian motion discussed above. In the networks the edges would correspond to reactions that could be added and deleted over evolutionary time. A network can often be translated into a dynamic model, so this can also be viewed as a model evolving over time.

The above list describe a series of objects that could be modelled individually, but it clear that combined modelling of two or more can have major advantages if the dependence between the phenomena can be modelled properly. This idea has been exploited on a large scale in comparative genomics in finding RNA and protein genes. In most applications one of the objects is sequences, but eventually other setups could be imagined. Additionally, in present applications is also assumed that the non-sequence object evolves very slowly compare to the sequences and thus can be assumed constant for the evolutionary time period under consideration. The first application of this idea, was Goldman, Thorne and Jones that in 1996 tried to predict protein secondary by making an evolutionary model of protein evolution where an amino acid evolved differently dependent on which secondary structure it was part of. The same idea was used by Knudsen and Hein (1999) to predict RNA secondary structure and from 2002 onwards it was used on a large scale to predict genes in genomes (Meyer and Durbin, 2002; Pedersen and Hein, 2003; Siepel and Haussler, 2003; Brent et al.)



$$P(\text{Sequence}|\text{Structure})P(\text{Structure}) =$$

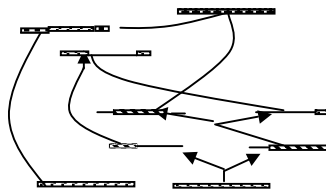
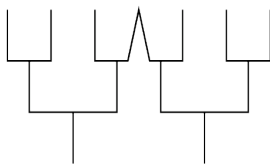
$$P(\text{Structure}|\text{Sequence})P(\text{Sequence})$$

The generic situation of co-modelling is illustrated on the left and it has so far only been applied to when one object was sequences and the other was a structure the sequence could undertake: We can obtain sequence data on a large scale, while the structures they participate in are harder to obtain. Fortunately, the probability of a sequence and how it evolves depends on the structure it participates in, so observing many homologous variants can give information on the unobserved structure. This is especially efficient in the case of RNA secondary structure as illustrated top right: The nucleotides that basepair evolves in coordination, while the ones that do not evolves independently. So observing correlation implies physical basepairing. To turn this into a proper analysis tool a model of sequence evolution dependent on structure and a distribution on structures must be given.

Lastly, which genealogical structure is appropriate? There are three major genealogical structures: Pedigrees, Phylogenies and the Ancestral Recombination Graph (ARG), where only the latter isn't universally known. The ARG describes the relationship of homologous sequences subject to recombination, where each small segment will be related by a phylogeny, while different segments can have different phylogenies that must obey certain constraints. Variants of these exist to cover phenomena like selfing and gene conversions. Radically different structures exists like phylogenetic networks, but does not represent an actual representation of history, but must rather be viewed as tools in exploratory data analysis. In models of cultural transmission (Cavalli-Sforza and Feldman, 1982) it can be meaningful of an "object" to have multiple parents. For instance, if an opinion is adopted by taking the consensus of a set of opinion of class mates.

In many contexts, most clearly in population genetics, it is of interest and natural to have a probability measure on the genealogical structures. A famous example of such a measure is "The Coalescent" (Kingman, 1982) that arises naturally in population genetics models of relationships of alleles from a population. This can be extended to probabilities/densities of ARGs when the alleles are subject to recombination (Hudson, 1982).

Such probability measures can also provide the basis of inference, when the relationship of the objects cannot be observed.



The pedigree to the left is defined by each individual having two parents of opposite sex, being born further back in time. The Ancestral Recombination Graph (ARG) to the left describes the relationship of homologous sequences subject to recombination in a population. The ARG should be embedded in a pedigree, but this aspect is typically ignored. Each sequence in the ARG is carried by an individual in a pedigree. Each individual carries two sequences. When a recombination happens the two ancestors must reside in two distinct individuals. Pedigree and ARG generally related to within population analysis, although ARG like structures are used to described long term evolution in the presence of horizontal transfers. For system biology purposed the phylogeny will almost always be the relevant genealogical structure.

In the illustration of these genealogical structures, time is at the bottom with the observations and going upwards is going back in time. A last issue (potentially difficult) is what to place at the points furthest back in time (root in a phylogeny). These ancestors/roots are unobservable, but assumptions must be made about what they are. The ideal solution is to draw them from a distribution defined by a biological process. For a phylogeny for instance, the root sequence would be drawn from the equilibrium distribution of the substitution process.

Above it modelling was described as a one-level endeavour, but can be extremely advantageous to do simultaneously if dependencies can be properly described. In such cases, observation at one level can yield information on evolution at other levels. The standard success story is genome annotation, where observation on sequence evolution (a low level) can give information on gene and RNA structure (a higher level). The final destination of this trend would be a comparative systems biology.

Population Variation Modelling (Hein, Schierup and Wiuf, 2005) at the molecular level can be viewed as molecular evolution at smaller time scale. Any major evolutionary innovation must as such be explicable as the cumulative effect of events at the population level, whether it happened in large or small individual step. Variation studies from a population (for instance human) clear is of major interest in itself as it allows investigation of key evolutionary factors and the specific historical path evolution has taken. However, it has proven a major tool in dissecting factors contributing to traits such as disease susceptibility, intelligence of athletic performance. Association mapping locates loci that are responsible for such traits by co-occurrence with specific genetic variants in the population at large and pedigree performs a similar task by observing co-segregation in a pedigree. Such genetic analysis narrows the potential analysis of the genetic basis of traits from 24,000 genes to a handful in successful cases and again underlines the fundamental difficulty in a purely functional/mechanistic analysis based only on experimental observations of an organism.

A fact of life on earth, not normally considered by modelling, is that all organisms have an ontogeny, ie a development over time. This must clearly be a major constraint on the genetic programs of any real organism. It is conceivable to have organisms designed that didn't have an ontogeny although this hasn't been done.

Model Inference and Validation.

Especially low copy number of regulatory molecules and diffusion has lead to the importance of stochastic models and this trend will increase in coming years. Besides stochasticity, recent research has also started to focus on model integration and models with stochasticity at different levels, such as stochastic number of molecules or fluctuating concentration.

The actual line of reasoning underlying science has been the subject matter of the philosophy of science for more than many decades and will not be discussed here. Ideally, *experiments* can be defined as a designed set of investigations that allows the differentiation between competing hypotheses. In reality the situation can be very complex. But a major novelty in present high through-put technologies is that experiments now are supplemented by a data generation not part of designed hypothesis testing. However, experiments have provided knowledge and framework, within which almost all modelling will take place. Examples of this would be the central dogma and the roles of the major classes of molecules – DNA, RNA, proteins.

Model Selection (Burnham and Anderson, 2002; Ripley, 2004) is the key statistical procedure for acquiring scientific knowledge and central to statistical systems biology. Two major contending frameworks for statistical inference: Likelihood theory and Bayesian statistics. The former creates parameterized statistical models of observations. One set of parameters is true and statements about the true parameters are major aims. Bayesian statistics view the hidden parameters as stochastic variables in

themselves. This latter assumption creates a very clear interpretation for inference, but is also a source of criticism. From technical and computational standpoints, the two frameworks lead to very similar approaches and algorithms. Model selection will often involve two aspects: (i) continuous parameter estimation - for instance what are the kinetic parameters in a network model and (ii) structural aspect – for instance different underlying networks could be used to explain the same data. (i) above is well studied and hypothesis testing are often formulated as restricting parameters to different subspaces in the same overall space. With increasingly complex models (ii) is becoming important. One approach to (ii) is to embed different models in even more general models. A second approach is based on Akaike's Information Criteria $AIC = -2\log(\text{likelihood}) + 2p$ (or similar criteria). Such criteria make non-nested models comparable.

Non-physical models. Major emphasis in systems biology clearly is on modelling that reflects actual physical dynamics, but a large set of models does not attempt this. An example of this could be graphical models that make statements about dependencies of quantities of interest. Such models are of great use in describing the relationship of genes in expression data. Such models can be used as stepping stones towards dynamic models, since they would make statements about which classes of molecules that would interact and thus also which kinds of dynamic models should be considered. Clearly, dynamic models are superior in principle, but other approaches can be more realistic and useful at present. This distinction does not influence the considerations underlying model selection.

Deterministic Inference. As stated, many models are completely deterministic and such models should not be associated statistical inference. Given observations as a suitable number of time points, competing hypotheses could be distinguished. In reality any observation will be associated error or noise. This is often solved by turning deterministic models into stochastic by adding a stochastic error term. Alternatively, dynamic trajectories can be selected by maximum fit to data. Explicit modelling of stochastic dynamics and measurement error is clearly superior, but often prohibitive in practice.

Data Mining (Hand et al., 2000; Hand, 2006) has in face of the data from the new high throughput technologies proven extremely valuable. Data mining is a collective term for a series of statistical techniques that uses weak assumptions about independence or the underlying structure of the data and typically no modelling of the underlying biological process or evolution. Different types of classification – supervised and unsupervised – using techniques such as neural networks, vector support machines, k-means – are examples of methods exploited to investigate large data sets from throughput technologies.

Given the ambitious goals of systems biology, a major issue is clearly if the data can distinguish models at a sufficiently high level. This cannot be answered in general, but only for specific situations. However, despite the large quantities of data, predictive modelling of even a simple system based on for instance time series of expression data can often be totally unrealistic at present. There is a constant growth in both quantity and precision of data, so this could one day change.

Formal Representation of Models and Knowledge.

The widespread use of models in biology creates the need for standards allowing models and knowledge to be exchanged, compared and integrated in a reproducible fashion. Additionally, automatic model generation from different kinds of specifications (for instance ODEs from graph descriptions of interactions) are of increased importance. Sets of tools are presently Ontologies, Systems Biology Markup Language and Process Algebras. A novelty in this field within the last 5 years that also complements the traditional representation of knowledge in articles, books, lectures, etc. – is that of text mining.

Ontologies (Bodenreicher and Stevens, 2006; Bard and Rhee, 2004; Gkoutos et al., 2004) are structured vocabularies of concepts. These are typically organized as a tree or directed acyclic graph reflecting that terms are naturally related by inclusion. A famous example in molecular biology is Gene Ontology (GO). A major function of ontologies is to have a consensus and exchangeable vocabulary in describing for instance the function of gene products.

Systems Biology Markup Language (Hucka et al., 2003; Finney et al., 2006) is a set of conventions for how to formulate models to be exchanged and allowing computer parsing. It is based on a more general set of conventions called extensible Markup Language (XML) for . First versions were publically available in 2003 and have since seen version 2.0 with extended flexibility. At present it is directed towards cellular biochemical models and does not place restrictions on which languages the model is formulated in (MatLab, MATHEMATICA,...). There are related Markup Languages, like for instance CellML devised by

the groups creating the multilevel heart models and automated methods of translation between these formalisms have been created.

Formal computer science (Calder et al., 2006; Baeten, 2005; Phillips et al., 2006; Aceto, 2004, Kwiatkowska et al., 2006) has created disciplines devoted to the description and analysis of processes, such as Process Algebras, π -calculus and Petri Nets. Given the rise in biology of processes as a supplement to data/knowledge the relevance of this is obvious. PA has many biology-realistic properties such as interaction among processes and refinement of descriptions. Researchers such as Cardelli and Calder have designed automatic model generation taking Process Algebra descriptions and generating ODE models for simple biological systems. PA could be the natural framework for further developments in Systems Biology, but also presently lacks flexibility to incorporate several continuous features necessary in biological models, such as space and concentration.

Text Mining (Raychaudhuri, 2006) is useful to extract information in a coarse manner from large bodies of text that could not be read by a single researcher. The underlying principles are simple. Words in articles can be tabulated according to frequency, contextuality, combinations, information content and more. This very efficiently allows linking genes with genes, biological objects with sets of properties and much more. The widespread use of text mining shows that in the large body of articles information is encoded in a way that is not detectable by simple sequential reading of a smaller set of articles. Text mining is progressing fast and with the increased use of ontologies and associated well-defined terms, text mining will increasingly acquire abilities closer to traditional reading.

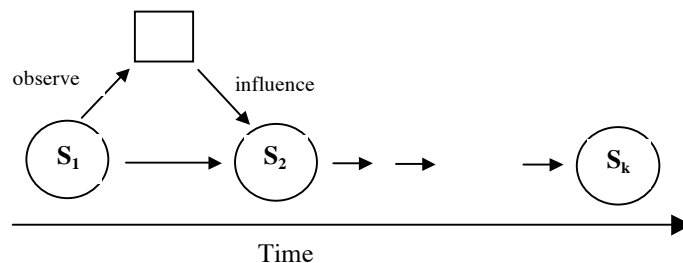
Discussion.

The emergence of predictive modelling will most likely change the biosciences in the coming decade, were it will become a standard tool for all researchers. It will also pose a series of challenging questions.

Quality Measures. Predictive modelling by definition is to predict, but exactly how and what would be a fair standard for success? Exact prediction of for instance the state of a cell of a long time interval is probably not feasible and possibly even of limited value. More restricted measures are of more use at present. Key examples are: i. That the proposed dynamic systems displayed behaviour that has been observed in experiments. ii. Stability of the proposed dynamic system. Although a central feature of modelling is quantitative and predictions in deterministic models can be made arbitrarily precise, the behaviour of a model is often evaluated on purely qualitative aspects.

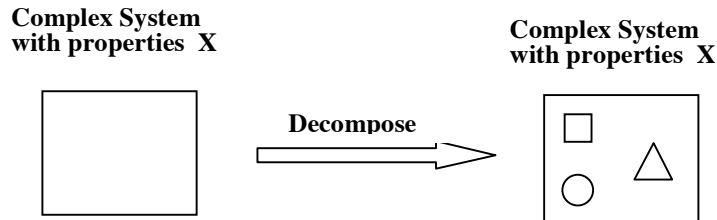
Systems, Control and Pathology. The very concept of system with an associated dynamics invites the concept of control – what to modify in parameters/input/environment to achieve a desired end – and as a special case how does a system malfunction and how to rectify it. This is already being done, but would be much more widespread in the future. Present examples could be the use of metabolic control theory to predict where to optimize yield in for instance yeast of a desired product or the use Noble's Heart models to predict the effect of different drugs. Metabolic Control Analysis have been applied for suggesting which enzyme should be targeted by drug design to maximally disrupt a pathway of a pathogen (Bakker et al., 1999). Drug delivery can be optimized by having a realistic model of drug transport, metabolism and effect and viewing dosage as a continuous time control problem (Drews et al., 2006, Danhof et al., 2007).

Control and Feedback Systems (Åstrom and Murray, 2007) have been generalized in a great variety of directions like stochastic and quantum systems and the nature of control also depends on what is observable in the systems and what can be influenced. Since most of medical science aims for control or alteration of the behaviour of a system, Control Theory could get achieve central importance in biology, but at present dynamical systems descriptions are too crude to use this theory in most cases. Comparison of control mechanism observed in biological systems with designed systems would be very interesting.



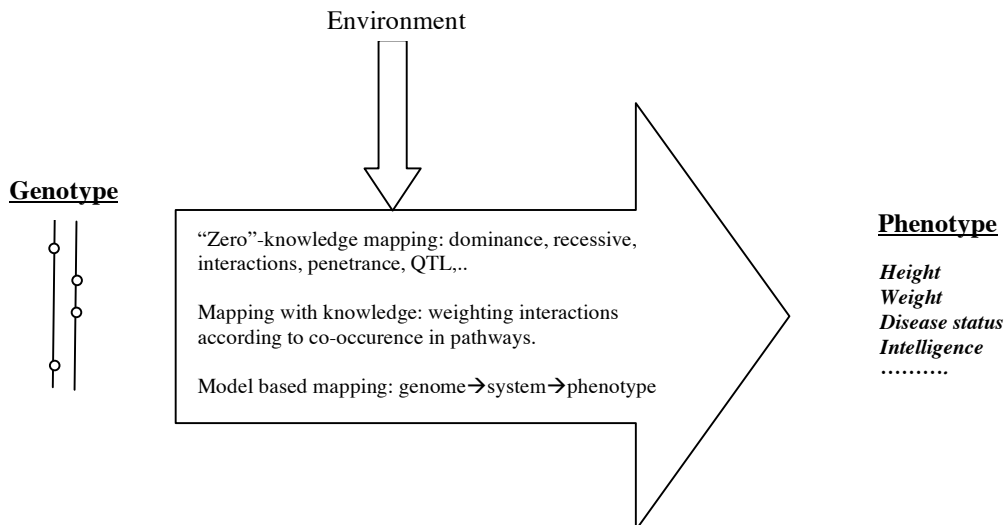
The specific nature of control can take a variety of forms dependent on what of the system can be observed – this could be all internal states of the system or some output function. The most common form of control in designed systems is called PID (proportional integral derivative control), that attempts to minimize error presently over a time period covering both the past and the predicted future.

Understanding and Theory Simplification. Models of biological systems can be increasingly complex and poses problems in themselves. Models with hundreds to thousands of parameters and equations can be hard to grasp in themselves and methods to extract overall behaviour in a understandable way will be needed. Metabolic Control Theory is one such construct, where the effect of individual enzymes and metabolites can be assigned an effect although it is embedded in a web of interactions. It is an open question or assumption that biological systems can be decomposed in an understandable way. One concept to legitimize this assumption is *modularity* that should be favoured evolutionary, where biological systems are composed of modules with distinguishable functions or roles. The concept of function is easily handled by biologists, but is fiendishly difficult to define exactly and is normally couched in teleological terminology – “the purpose of X is to accomplish Y”.



Modularity implies that an entity can be understood in terms of components with well defined goals. If biological systems are organized like that it is very convenient, but it is also an empirical fact and it could be possible to design systems to could only be described *in toto*. It has been argued (Kashtand and Alon, 2005) that evolution creates modularity.

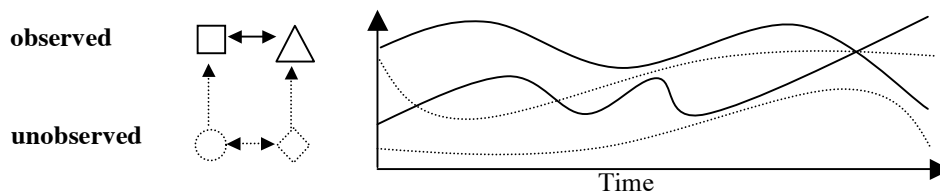
Genotype to Phenotype Functions have been described as yet another Holy Grail of biology. Such functions are inherent in all genetic mapping investigations, where genetic variation has been observed together with a phenotype. The genetic terms recessive, dominant, penetrance, quantitative character, epistasis and more describe some aspect of the general genotype to phenotype function. Based on a few assumptions about this function, statements can then be made about the position in the genome of the genetic determinants (mapping). The present genotype-to-phenotype functions are based on minimal biological knowledge. An exception is Omholt et al. (2001) that formulated network models exhibiting classical genetic concepts like epistasis etc. Many attempts of systems biology can be seen as attempts to create genotype-to-phenotype functions based on functional knowledge. Ideally, one should take a standard genome with a standard model of the organism and predict the result of a change in the genome.



Many mapping models assume a genotype→phenotype function that is so simple that users of the method can totally overlook the hidden assumption. It might just state “if disease allele is present, then sick”. But this can be refined to incorporate standard genetic concepts, but still not assume anything about the underlying mechanism (Zero knowledge above). A step will incorporate some external functional information, either in which regions of the genome to care about or if considering interactions, weight these according to external knowledge (Mapping with knowledge). In the presence of complete models of either an individual or a subsystem, where phenotypic consequences are predictable, then much more complex genotype→phenotype functions would be possible (Model based mapping). In association mapping a sample from a population is taken and the combined genotype→phenotype function and relationship of genomes/individuals must be translated into a parameterized distribution of the observation. Parameters would include population structure and parameters of the genotype→phenotype function such as number of causative SNPs and their position. In pedigree mapping the relationship of the individuals is known, but the actual path back through the individuals are still stochastic. In both cases distributions are defined and inference is then possible for quantities of interest.

As modelling of biological systems progress, questions relating to its contribution questions about the nature of its contribution will be relevant. Integrative modelling might seem like an ambitious undertaking, but very technical in nature – to convert knowledge/data about a biological system into a global dynamic description. This would leave major and conceptual discoveries out, but not remove its practical value. If this will be so, remains to be seen. It could well be that increasingly ambitious modelling could reveal hidden components and lead to fundamental discoveries.

Hidden variables in this sense is already used in a series of recent publications (Beal et al. (2005), Pournara and Wernish (2007), Sabatti and James (2006), Sanguinetti, Lawrence and Rattray (2006)) using a variety of techniques. Especially Gaussian Processes (Rasmussen and Williams, 2006) provides a natural framework for this. An obvious class of model with hidden variables is dynamic modeling of expression data, when RNA levels are observed, but regulatory proteins are not.



In this illustration the systems is composed of 4 variables, but only 2 can be observed. This situation is generic to much of systems biology, where only partial information about a system can be obtained. Real biological systems can be dominated by many hidden variables but with weak interactions with observables. The reverse can also be the case, as is often seen in expression data analysis, that much has been observed that doesn't influence the system of interest. An assessment of how much needs to be observed to allow dynamic modelling is clearly a prerequisite for defining realistic projects.

Massive modelling will also pose new interesting questions. A large-scale dynamic trajectory of a biological system does not necessarily contain the standard biological concepts and objects, such as the Central Dogma or nucleosome. Can these be recovered in an automated fashion? Much systems biology research will focus on subsystems – how easy is it to integrate local knowledge to a global model?

References:

- Aceto, L (2003). "Some of My Favorite Results in Classic Process Algebra." Bulletin of the EATCS, volume 81, pp. 89-108
- Adams, Rohlf, and Slice 2004. "Geometric morphometrics: ten years of progress following the 'revolution'" Italian Journal of Zoology, 71(1):5-1
- Alberghina, L and H.V. Westerhoff (eds.) (2005) "Systems Biology: Definitions and Perspectives" Springer**
- Allison, DB, X. Cui, GP Page and M. Sabripour (2006) "Microarray Data: From Disarray to Consensus and Consolidation" Nature Genetics. 7.55-65
- Alon, U. (2006) "An Introduction to Systems Biology: Design Principles of Biological Circuits" CRC Press.*
- Arkin, A., HH McAdams (1997) **Stochastic mechanisms in gene expression.**PNAS 94(3):814-9.
- Baeten, JCM (2005) "A brief history of process algebra" Theoretical Computer Science Volume 335 , Issue 2-3.131 - 146
- Albert-Laslo Barabasi et al. (2004) "Network Biology: Understanding the cell's functional Organisation" Nature Reviews: Genetics vol 5 Feb 2004**
- Barbara M. Bakker[§], Paul A. M. Michels, Fred R. Opperdoes, and Hans V. Westerhoff[§] "What Controls Glycolysis in Bloodstream Form *Trypanosoma brucei*?" J Biol Chem, Vol. 274, Issue 21, 14551-14559, May 21, 1999**
- Ball, K., TG Kurtz, L. Popovic and G. Rempala (2006) "Asymptotic analysis of multi-scale approximations to reaction networks" Annals of Applied Probability (in press)
- Bard, JBL and SY Rhee (2004) "Ontologies in Biology: Design, Applications and Future Challenges" Nature Reviews Genetics 5.213-22
- Basner, JE and SD Schwartz (2005) "How Enzyme Dynamics Helps Catalyze a Reaction in Atomic Detail: A Transition Path Sampling Study" JACS 127.13822-31
- Beal, MJ, F Falciani, Z Ghahramani, C Rangel and DL Wild (2005) "A Bayesian approach to reconstructing genetic regulatory networks with hidden factors" Bioinformatics 21.3.349-56.
- Berendsen, H. (2007) "Simulating the Physical World: Hierarchical Modelling from Quantum Mechanics to Fluid Dynamics" Cambridge University Press.*
- Blanchette, M, B. Schwikowski and M. Tompa (2002) "**Algorithms for Phylogenetic Footprinting**" J. Comp. Biol. 9.2.211-
- Bodenreicher, O. and R. Stevens (2006) "Bio-ontologies: current trends and future directions" Briefings in Bioinformatics 7.3.256-74.
- Bolhuis, PG, -D Chandler, C Dellago and PL Geissler "TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark" Annual Review of Physical Chemistry Vol. 53: 291-318
- Burnham and Anderson (2002) "Model Selection and Multimodel Inference" 2nd edition Wiley
- Bystroff, C., Simons, K. T., Han, K. F., Baker, D. (1996). "**Local sequence-structure correlations in proteins**" Curr Opin Biotechnol 7, 417-21.
- Carlos Bustamante., Yann R. Chemla., Nancy R. Forde. and, David Izhaky (2004) "**MECHANICAL PROCESSES IN BIOCHEMISTRY**" Annual Review of Biochemistry. Volume 73, Page 705-748, 2004
- M. Calder, S. Gilmore and J. Hillston. "Modelling the influence of RKIP on the ERK signalling pathway using the stochastic process algebra PEPA" Transactions on Computational Systems Biology VII, vol. 4230, pp. 1-23, Springer, 2006.
- Campbell, ID (2002) "The march of structural biology" Nature Molecular Cell Biology Review 3.5.377-82.
- Campbell, ID (2007) "The Croonian lecture 2006: Structure of the living cell" Phil. Trans. R. Soc. B.
- Cavalli-Sforza and Feldman (1982) "Cultural Transmission and Evolution" Princeton
- Chandonia, J-M and SE Brenner (2006) "The Impact of Structural Genomics: Expectations and Outcomes" Science 311.347-51.
- Cornish-Bowden, A (1995) "Fundamentals of Enzyme Kinetics" Portland Press

Danhof M, D deJongh, ECM De Lange, OD Pasqua, BA Ploeger and RA Voskuyl (2007) "Mechanism-Based Pharmacokinetic-Pharmacodynamic Modeling: Biophase Distribution, Receptor Theory, and Dynamical Systems Analysis" *Annu.Rev.Pharmacol.Toxicol.* 47.357-400.

Deutsch, A. and Dormann, S. (2003). Cellular Automaton Modeling of Biological Pattern Formation. Birkhauser.

Drews, FA, N Syroid, J Agutter, DL Strayer and DR Westenskow (2006) "Drug Delivery as a Control Task: Improving Performance in a Common Anesthetic Task" *Human Factors* 48.1. 85-94

Dryden and Mardia (1998) "Shape analysis" Wiley

Dllage, C. and PG Bolhuis (2007) "Transition Path Sampling Simulations of Biological Systems" *Top. Curr Chem* 268.291-317.

P.D'haeseleer, Liang & Somogyi (2000) Genetic network inference: from co expression clustering to reverse engineering. *Bioinformatics* 16.8.707-726

DiVenture, B., C.Lemerle, K.Michalodimitrakis and L. Serrano (2006) "From in vivo to in silico and back" *Nature* 443.527-534.

Dorogovtsev, SN and JFF Mendes (2003) "Evolution of Networks: From Biological Nets to the Internet and WWW" OUP

Elvidge, G. (2006) "Microarray expression technology: From start to finish" *Pharmacogenomics* 7.1.123-34.

Fall, Marland, Wagner, Tyson (eds.) "Computational Cell Biology" Springer (2002)

Fell, D. (1997): "Understanding the control of metabolism." Portland Press: London.

Feldman, M and LL Cavalli-Sforza (1982) "Cultural Transmission and Evolution" Princeton

Fersht, A. (2004) "Structure and Mechanism in Protein Science" Freeman

Felsenstein, J. (2004) "Inferring Phylogenies" (chapt. 23 + 24) Sinauer.

Finney, A., M.Hucka and N. Le Novere (2006) "Systems Biology Markup Language (SBML) Level 2: Structures and Facilities for Model Definitions" www.sbml.org/specifications/sbml-level-1-v2.pdf

Frank, J. (2006) "Three-Dimensional Electron Microscopy of Macromolecular Assemblies" Oxford University Press

Garcia-Viloca, M, J.Gao, M. Karplus and DG Truhlar (2004) "How Enzymes Work: Analysis of Modern Rate Theory and Computer Simulations" *Science* 303.186-195.

Gardner, S. (2005) "Ontologies and semantic data integration" *Drug Discovery Today* 10.14,1001-7

Gardiner, CW (2004) "Handbook of Stochastic Methods: For Physics, Chemistry and the Natural Sciences" Springer

Geremia, JM (2003) "An introduction to control theory from classical to quantum applications"

Gillespie, D (1977) "Exact Stochastic Simulation of Coupled Chemical Reactions" *The Journal of Physical Chemistry*, Vol. 81, No. 25, pp. 2340-2361

Gillespie, D (2007) "Stochastic Simulation of Chemical Kinetics" *Ann.Rev.Chem.Phys.*58.35-55

G.V. Gkoutos, E.C.J. Green, A.M. Mallon, J.M. Hancock, and D. Davidson (2004) "Building Mouse Phenotype Ontologies" *Pacific Symposium on Biocomputing* 9.178-189.

Glaeser, R., Wah, C. and Frank, J. (2007) "Electron Crystallography of Biological Macromolecules" Oxford University Press

Goldbeter, A. (1996). *Biochemical oscillations and cellular rhythms: The molecular bases of periodic and chaotic behaviour.* 2nd edition. Cambridge University Press: Cambridge.

Hand, DJ, G.Blunt, MG Kelly and NM Adams (2000) "Data Mining for Fun and for Profit" *Statistical Science* 15.2.111-131

Hand, DJ (2006) "Classifier Technology and the Illusion of Progress" *Statistical Science* 21. no. 1 (2006), 1-14

Hein, J., Schierup, M. & Wiuf, C. (2004) "Gene Genealogies, Variation and Evolution" Oxford University Press

Heinrich, R. and Schuster, S. (1996). "The Regulation of Cellular Systems" Chapman and Hall: New York.

Heinrich and Rapoport (1973) "Linear Theory of Enzymatic Chains: its application for the analysis of the crossover theorem and of the glycolysis of human erythrocytes" *Acta Biol Med germ.* 31.479-494"

- Higgins, (1965) "Dynamics and Control in cellular reactions" in Chance, B, Estabrook, RW and Williamson, JR (eds.) Control of Energy Metabolism, Academic Press. P13-46.
- M.W. Hirsch and S.O. Smale (1974) "Differential Equations, Dynamical Systems and Linear Algebra" (Academic Press)
- Hudson, RR (1983) " " Theor. Pop. Biol.
- Huson, DH and D. Bryant (2006) "Application of Phylogenetic Networks in Evolutionary Studies" Mol. Biol. Evol., 23(2):254-267.
- John, F. (1982) "Partial Differential Equations" 4th Ed. Springer
- Kacser H and Burns JA. "The control of flux." Symp Soc Exp Biol. 1973;27:65-104.
- Kashtan N and U Alon (2005) "Spontaneous evolution of modularity and network motifs" PNAS vol. 102.39.13773-8
- Kaufmann, S.(1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol. Mar;22(3):437-67.
- Keller, E. F. (2003). "Making sense of life: Explaining biological development with models, metaphors, and machines" Harvard University Press: Cambridge, Massachusetts.
- Kim, Jaegwon (1999). "Making Sense of Emergence," *Philosophical Studies*, 95, pp. 3-36.
- Kingman, JFC (1982) "The Coalescent"
- Kitano, H. (2002) "Foundations of Systems Biology" MIT
- Kitano H (2007) "**A robustness-based approach to systems-oriented drug design.**" 6(3):202-210. Nat. Rev. Drug Disc.
- Kleppe, R., E.Kjarland and F.Selheim (2006) "Proteomic and Computational Methods in Systems Modeling of Cellular Signaling" Current Pharmaceutical Biotechnology 7.135-145.
- Klipp, Herwig, Kowald, Wierling and Lehrach (2005) "Systems Biology in Practice" Wiley**
- Kou, SC, XS Xie and Jun S. Liu (2005) "Bayesian analysis of single-molecule experimental data" Applied Statistics. 54.3.469-506
- Kulzer, F. and M. Orrit (2004) "Single-Molecule Optics" Annual Review Physical Chemistry 55.585-611.
- Kwiatkowska M, J Heath and E Gaffney (2006) "Simulation and Verification for Computational Modelling of Signalling Pathways" Proc. Winter Simulation Conference.
- Leach, A. (2001) "Molecular Modelling: Principles and Applications" Prentice Hall
- Lindon, JC, E.Holmes and JK Nicholson "Metabonomics Techniques and Applications to Pharmaceutical Research and Development" Pharmaceutical Research 23.6.1075-88
- Liu, JS (2001) "Monte Carlo Strategies in Scientific Computing" Springer
- Maini, PK (2004) Using mathematical models to help understand biological pattern formation. C R Biol. 2004 327(3):225-34.
- Metzker, M. (2005) "Emerging Technologies in DNA Sequencing" Genome Research 15.1767-76.
- Mishra B, Daruwala RS, Zhou Y, Ugel N, Policriti A, Antoniotti M, Paxia S, Rejali M, Rudra A, Cherepinsky V, Silver N, Casey W, Piazza C, Simeoni M, Barbano P, Spivak M, Feng J, Gill O, Venkatesh M, Cheng F, Sun B, Ioniata I, Anantharaman T, Hubbard EJ, Pnueli A, Harel D, Chandru V, Hariharan R, Wigler M, Park F, Lin SC, Lazebnik Y, Winkler F, Cantor CR, Carbone A, Gromov M. "A sense of life: computational and experimental investigations with models of biochemical and evolutionary processes." OMICS. 2003 Fall;7(3):253-68
- Murray, J. (2002) "Mathematical Biology I-II" Springer
- Noble, D. (2006) "The Music of Life" OUP**
- Omholt SW, Plahte E, Oyehaug L, Xiang K. Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. Genetics. 2000 Jun;155(2):969-80.
- Pournara I and L Wernish (2007) "Factor analysis for gene regulatory networks and transcription factor activity profiles" BMC Bioinformatics 8.61.1-20
- Rasmussen CE and CKI Williams (2006) "Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)" MIT Press

- Raychaudhuri, S. (2006) "Computational Text Analysis for Functional Genomics and Bioinformatics" Oxford University Press.
- Pepperkok, R. and J. Ellenberg (2006) "High-throughput fluorescence microscopy for systems biology" *Nature Molecular Cell Biology Review* 7.9:690-6.
- Andrew Phillips, Luca Cardelli and Giuseppe Castagna. (2006) "A Graphical Representation for Biological Processes in the Stochastic pi-calculus" *Transactions in Computational Systems Biology (TCSB)*, 4230:123–152
- Ripley, B. D. (2004) Selecting amongst large classes of models. In *Methods and Models in Statistics* eds N. Adams, M. Crowder, D. J. Hand and D. Stephens, Imperial College Press, pp. 155-170.
- Sabatti C and GM James (2006) "Bayesian sparse hidden components analysis for transcription regulation networks" *Bioinformatics* 22.2.739-46
- Sanguinetti G, ND Lawrence and M Rattray (2006) "Probabilistic Inference of transcription factor concentrations and gene-specific regulatory activities" *Bioinformatics* 22.22.2775-81.
- Santos, NC and ARB Castanho (2004) "An overview of the biophysical applications of atomic force microscopy" *Biophysical Chemistry* 107. 133-149
- Savageau, M. (1976) "Biochemical Systems Theory" Addison-Wesley
- Shulaev, V. (2006) "Metabonomics technology and bioinformatics" *Briefings in Bioinformatics* 7.2.128-139
- Somogyi, R., Sniegowski, C.A., (1996) "Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation" *Complexity* 1(6):45-63
- Spasic, I. et al. (2005) "Text mining and ontologies in biomedicine: Making sense of raw text" *Briefings in Bioinformatics* 6.3.239-51.
- Svergos, DI and MHJ Koch (2003) "Small-angle scattering studies of biological macromolecules in solution" *Rep. Prog. Phys.* 1735-1782.
- Schaffner, K. (1993) "Discovery and Explanation in Biology and Medicine" University of Chicago Press.
- Sharan R, Ideker T. (2006) "Modeling cellular machinery through biological network comparison" *Nat Biotechnol.* 4:427-33.
- Sontag, E. (2004) "Some new directions in control theory inspired by systems biology" *Systems Biology* 1: 9-18.
- Sontag, E. (2005) "Molecular Systems Biology and Control" *European J. of Control* 11: 396-435.
- Souchelnyskel, S. (2005) " Bridging Proteomics and systems biology: What are the roads to be travelled?" *Proteomics* 5.4:123-37.
- Szallasi, Z, Jorg Stelling, and Vipul Periwal (2006) "System Modelling in Cellular Biology: From Concepts to Nuts and Bolts" MIT**
- Thagaard, P. (2000) "How Scientists Explain Disease" Princeton University Press
- Tringe, SG and Rubin, EM (2005) "Metagenomics: DNA Sequencing of Environmental Samples" *Nature Review Genetics* 6.11.805-14
- Turing, A. M. (1952). "The chemical basis of morphogenesis" *Philosophical Transaction of the Royal Society of London Series B* 237, 37-72.
- Abel Ureta-Vidal, Laurence Ettwiller, Ewan Birney (2003) "Comparative genomics: genome-wide analysis in metazoan eukaryotes" *Nature Reviews Genetics* 4, 251 – 262
- Van Kampen, NG. (1992) "Stochastic Processes in Physics and Chemistry" (North-Holland)
- Wilkinson, D. (2006) "Stochastic Modelling for Systems Biology" Chapman Hall/CRC Press
- Wiuf C, Brameier M, Hagberg O, Stumpf MP. (2006) "A likelihood approach to analysis of network data" *PNAS* 103(20):7566-70
- Wiener, N. (1948) "Cybernetics" MIT
- Xiang Z. (2006) "Advances in homology protein structure modeling" *Curr Protein Pept Sci.* 2006 Jun;7(3):217-27.
- Yeang CH, Vingron M. (2006) "A joint model of regulatory and metabolic networks" *BMC Bioinformatics.* 2006 Jul 4;7:332.

Yu, AC (2006) "Methods in biomedical ontology" Journal of Biomedical Informatics 39.252-66

Åstrom KJ and RM Murray (2007) "Feedback Systems - An Introduction for Scientists and Engineers"