On the complete determination of biological systems

Douglas W. Selinger, Matthew A. Wright and George M. Church

Harvard Medical School, Department of Genetics, 200 Longwood Ave, Boston, MA, 02115, USA.

The nascent field of systems biology ambitiously proposes to integrate information from large-scale biology projects to create computational models that are, in some sense, complete. However, the details of what would constitute a complete systems-level model of an organism are far from clear. To provide a framework for this difficult question it is useful to define a model as a set of rules that maps a set of inputs (e.g. descriptions of the cell's environment) to a set of outputs (e.g. the concentrations of all its RNAs and proteins). We show how the properties of a model affect the required experimental sampling and estimate the number of experiments needed to 'complete' a particular model. Based on these estimates, we suggest that the complete determination of a biological system is a concrete, achievable goal.

Scientific investigation has long been a technology-limited endeavor: from Aristotle's passive observations, to Galileo's experimental probing, to our own elaborately contrived and controlled micro-dissections of nature. New technologies, in the form of systematic, quantitative, large-scale experiments with machine-readable outputs have recently unleashed a torrent of data onto the biological community, resulting in abundant speculation about the future of post-genomic biology.

With new tools, naturally come new goals. Classical molecular methods forced us to focus our gaze on small numbers of molecules at a time, so we laboriously built up descriptions in human language, pictures and the occasional video clip. The overarching goal of biology, if there was one, was to compile a large number of systems that are interesting (those that define a general rule, break one or appeal to us as idiosyncratic human beings) or applicable (those that contribute to the engineering, reverse-engineering or modification of a system). The defining feature of this 'compilation strategy' is that it is more a process than a goal – it specifies no endpoint other than continual accumulation.

Completion in biology

Long the goal of physicists searching for a 'theory of everything', completion has now become a pervasive idea in biology, raising the question of where it rightfully applies and whether it constitutes a new sort of goal for biological inquiry. The proliferation of the '-ome' suffix attests to widespread acceptance that biology is rife with things to be completed, whether it's the genome, transcriptome or proteome. Genome projects and large-scale experiments have already yielded important advances in medicine, biotechnology and basic research. Moreover, systems level descriptions promise predictions for cell, organ and organism behavior.

There seem to be two distinct levels of completion. The first, and conceptually simpler of the two, is 'parts list completion', defined as the fraction of observed to total predicted parts. This is well underway in the various 'ome' projects. The second, more ambitious and less well-defined level of completion, is at the level of 'systems biology', the study of how the parts work together to form a functioning biological system [1,2]. There is no clear correspondence between these two levels of completion – a nearly complete parts list could lead to an inaccurate description of the system if the missing parts, essential genes for example, were crucial for system function.

But how can we know when a systems level description is complete? Whereas crystallographers can state an R_{free} to describe the extent of agreement between a structural model and the data from which it was derived, biologists still lack a coherent framework for deciding the extent to which a model of a biological system is consistent with experimental data. Such a framework would be useful for setting systems biology goals, assessing progress and identifying areas in need of further investigation.

A model for modeling

We can think of a systems level description as a formal mathematical construct, or model. Thus, consideration of the properties of a model is necessary to understand in what sense a systems level description might be considered complete. A model can be defined as a set of rules that maps a set of inputs (see Fig. 1, blue area; e.g. descriptions of the cell's environment), to a set of outputs (see Fig. 1, yellow area; e.g. the concentration of all of its RNAs). Large-scale experimental sampling of input–output pairs (Fig. 1, yellow-red dots) such as condition–transcriptome pairs, can be used to derive these rules [3].

To specify a particular model we must decide on its basic properties (Table 1). First, we must decide on the inputs and outputs. This choice will depend, for example, on whether we are interested in predicting transcriptomes from temperature and pH, or in predicting successive molecular states. Second, we must decide on the range of values the inputs can assume and finally, we must decide

Corresponding author: George M. Church (church@arep.med.harvard.edu).

http://tibtec.trends.com 0167-7799/03/$ - see front matter doi:10.1016/S0167-7799(03)00113-6

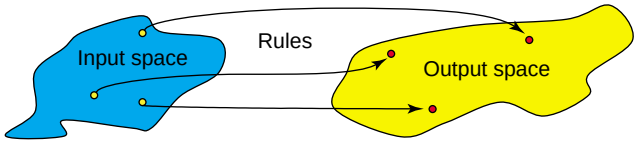


Fig. 1. A general scheme for modeling as an exercise in mapping input space (blue area; for example all possible environments in which a cell can live), to output space (yellow area; e.g. all possible cellular responses). The yellow-red dot pairs represent measured input–output pairs, which, in large numbers, can be used to derive rules (arrows) to predict outputs for novel inputs. Examples of possible inputs and outputs are given below.

on what level of accuracy and precision we require in our predictions. For instance, if we are predicting relative RNA levels, do we need predictions such as, 'upregulation by a factor of 3.3 ± 0.1' or would a predicted factor of 3 ± 1 allow us to reach the same biological conclusions? Once we have made these three decisions, we must choose a rule type that will allow us to realize the model, that is, one that will allow us to map our chosen input space to our chosen output space with the desired level of accuracy and precision.

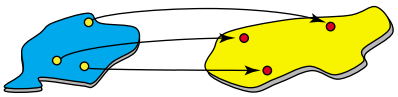
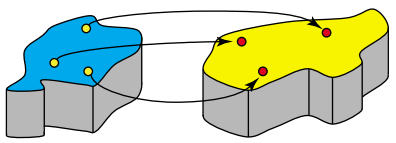
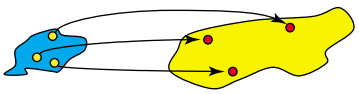
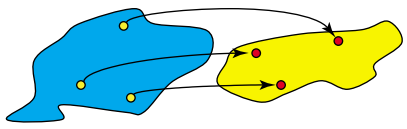
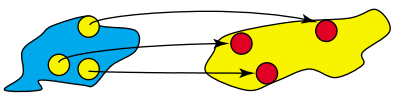
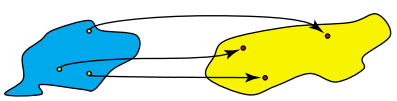
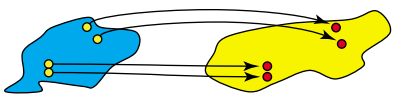
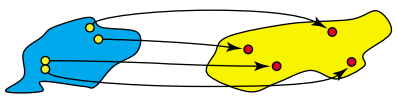
From these basic model properties we can determine how many measurements, at least to the order of magnitude, it would take to populate the space of all possible inputs (e.g. conditions) with enough measured outputs (e.g. transcriptomes, proteomes) to make prediction feasible. In other words, we can establish how many measurements are needed to adequately sample input space to allow the rule parameters to be determined. A similar issue has been addressed in the field of supervised learning by Probably Approximately Correct (PAC) theory [4], which gives the probability that a given number of measurements will generate rules of a given accuracy. Of course, once a model has been generated, its actual accuracy must be assessed by additional experiments that were not used to infer the rules.

We can readily determine how the properties of a model affect the number of measurements required to derive its rules (Table 1). A larger number of inputs and outputs will require more individual measurements per input–output pair, that is, per sample (Row I). A larger range of input values might require a greater number of samples (Row II). A higher desired accuracy generally will require more samples and increased accuracy per individual measurement (Row III). Finally, a more complex rule type will probably require more samples (Row IV). For example, if nearby points in input space do not map to nearby points in output space then we must sample the space more densely. It should be noted that because the output space is simply a function of the input space we can focus exclusively on the properties of the input space and rule type when considering the required experimental sampling density.

Model types

The choice of a model type is a crucial part of any completion effort because it determines the type of rules that need to be discovered and the number and type of

Table 1

Property	Minimizing (*maximizing*) case	Minimizing case	Maximizing case
I. Number of inputs and outputs	Low (*high*) level of model detail, less (*more*) comprehensive model		
II. Range of inputs	Can live in few (*many*) environments		
III. Accuracy and precision	Predictions useful at low (*only at high*) level of accuracy		
IV. Rule type	Similar inputs give similar (*different*) outputs, requires simple (*complex*) rule types.		

Four model properties that contribute to the number of measurements needed for complete determination by either influencing the effective size of the spaces to be sampled, or the necessary sampling density. Cases that maximize and minimize measurements are given, with corresponding schematic representations based on Figure 1. Row I, with more inputs and/or outputs each sampling of an input–output pair requires more individual measurements, represented by the depth of the spaces. Row II, the size of the input space contracts or expands depending on the range of the input values. Row III, larger input-output dots in the minimizing case indicate that each measurement or prediction effectively covers more input and/or output space because, with low accuracy requirements, nearby points can be considered equivalent. Row IV, when nearby points in input space map to nearby points in output space, simpler rules with fewer parameters will need to be determined, requiring fewer measurements. When nearby points in input space do not map to nearby points in output space, the model will require more complex rules with more parameters, requiring more measurements.

Table 2

Model	Scope	Applicable Rules	Model Outputs	# of Outputs	Examples of Outputs
Atomic	Cell *c* at time *t*	Physics	Atomic positions and momentums	10^9–10^{17}	^{12}C position and momentum
Molecular	Cell *c* at time *t*	Chemistry	Small molecule positions and momentums	10^8–10^{16}	Glucose position and momentum
Biomolecular (discrete)	Cell *c* at time *t*	Molecular mechanics	Macromolecule positions and momentums	10^5–10^{12}	Hexokinase position and momentum
Biomolecular (statistical)	Biochemically equivalent cells	Chemical kinetics and thermodynamics described by differential equations	Macromolecule concentrations, compartments	10^3–10^6	Hexokinase concentration in cytoplasm
Biomolecular (steady-state)	Genetically equivalent cells, similar growth conditions, steady state	Flux balance, physical and chemical constraints	Molecular fluxes	10^2–10^4	Flux of glucose to glucose-6P
Boolean	Genetically equivalent cells	Genetic and metabolic 'circuits'	Regulons, pathways	10^2–10^4	Glycolysis 'on', gluconeogenesis 'off'

Examples of hypothetical levels at which a systems biology project can be completed, listed from most detailed (top) to least (bottom). The number of outputs is estimated for single cells. The details of these calculations can be found at http://arep.med.harvard.edu/completion. We might soon be able to collect complete datasets for some classes of biomolecules at the level of macromolecular concentrations.

measurements that need to be made. There are a host of issues, discussed in several reviews [5–8], which must be considered when planning a modeling strategy. Table 2 gives examples of model types organized by the level of detail of their predictions, that is, their outputs. On one end of the spectrum we can imagine atomic level, or even subatomic level descriptions of a complete cell, which set an upper bound on detail. Towards the lower end of the detail spectrum we have Boolean models, which consist of logical statements such as, 'if the *lac* repressor is bound to the operator then the *lac* operon is off'.

As we move from more to less detailed models we make certain trade-offs. The more detailed models make fewer assumptions and are therefore potentially more accurate for the systems they describe. However, they tend to be more problematic with regard to computability and measurement, and are therefore difficult to apply to large systems. Furthermore, computational predictions at too high a level of detail might not be useful for human understanding of the biological phenomena under study. As we enhance our ability to make large numbers of measurements, we might be able to generate enough input–output pairs to allow the complete determination of more and more detailed model types.

Practical application

Now let's consider specific examples of projects we might wish to complete. A useful model for many biological purposes is one in which the resulting expression level of each gene can be predicted using the input levels of all of the genes. Such a model would predict the effects of overexpression, genetic knockouts, or even various environmental stimuli, provided that the effects of those stimuli on individual genes are known. In fact, historically, much of genetic research has been devoted to finding small parts of such a model. Specifically, we consider a discrete transcriptional network model that maps all N genes as inputs to all N genes as outputs, in which the genes can take on three levels of expression (low, medium and high) and each gene has, at most, K direct regulators (Table 3). We consider this model for a range of organisms: *Mycoplasma pneumoniae*, *Escherichia coli* and *Homo sapiens*.

A very simple cell, such as *M. pneumoniae*, lives in an exquisitely controlled environment within its human host and has a minimal number of genes (low N) that seem to lack any transcriptional regulation (low K) [9]. At an intermediate level is *E. coli*, which can live in many environments and consequently has more genes and requires more genetic regulation (intermediate N and K). At the upper extreme are humans, which have a large number of genes and highly complex regulatory mechanisms. Additionally, as multicellular organisms, humans have abundant intercellular communication and a large number of cell types, each potentially with its own set of transcriptional states.

In Table 3, we use formulae given by Krupa [10] to estimate the upper and lower bounds for the number of microarray experiments needed to complete the discrete transcriptional network model described above. The lower bound is related to the amount of information needed to specify the network structure and mapping functions. It assumes that each microarray experiment is maximally informative and independent from previous measurements and also assumes perfectly efficient use of experimental measurements to determine model parameters. These assumptions make it likely that this estimate is far below the actual number of measurements needed. The upper bound reflects the number of random experiments needed to complete the model with a 99% probability of

Table 3

Organism	N	K	Estimated number of microarrays	
			Lower bound	Upper bound
M. pneumoniae	688	1	10	80
E. coli	4,288	3	50	40 000
H. sapiens	50 000	4	100	700 000

Upper and lower bounds on the number of microarrays (or equivalent transcriptome-wide experiments) to complete discrete transcriptional network models for various organisms, calculated according to Krupa [10]. N represents the number of nodes (genes in this example). K represents the maximum number of regulatory connections per node. The expression level of each gene is categorized as high, medium, or low ($\xi = 3$). The lower bound (information-theoretic) is given by $\xi^K + K\log_\xi(N/K)$. The upper bound is given by $\xi^{2K}(2K(\ln N + \ln\xi) + \ln C)$, where the measurements fail to determine the model with probability 1/C. Here we set 1/C equal to 0.01. It is important to note that the upper bound estimate increases exponentially with K, making it the dominant parameter.

success and is probably a more realistic estimate. It is important to note that the upper bound estimate grows rapidly (exponentially) with the maximum number of regulatory connections (K) per gene. Fortunately, genes tend to have a low number of regulatory inputs, making an estimate based on a low K reasonable. It is also encouraging to note that the upper bound estimate grows slowly (logarithmically) with the number of genes (N), making determination feasible even for large genetic networks. Furthermore, if the structure of the network is already known, far fewer measurements will be needed.

The upper bound of 80 transcriptome experiments for *M. pneumoniae* is already feasible with current technology. Although 40 000 microarrays (or equivalent transcriptome-wide experiments) for *E. coli* and 700 000 for human might seem daunting, we should keep in mind that the initial version of the human genome required ~30–40 million sequencing reads [11,12] – a number that was not practical at the time the project was first proposed.

There are other methods for inferring rules directly from large-scale datasets and for estimating the number of measurements necessary for a given level of accuracy [3,13,14]. Additionally, current microbial models based on flux balance analysis have shown considerable progress towards a complete description of metabolism, with mappings from culture conditions and genotype (input) to growth phenotype (output) that reach accuracies >90% (106/116) [15]. Models of this type have even been shown to be predictive of the biological evolution of metabolic fluxes [16]. Additional refinements promise to further increase their accuracy [17].

Conclusion

With the advent of large-scale projects, synthesis has become an important goal in biology. Completion of a large number of genome projects and the pursued completion of other 'ome' projects raises the question of what it might mean to complete a systems biology project and what might be gained from such an effort. We have found it useful to consider this question within a framework for modeling and show how the number of experiments necessary can be related to the model properties. Furthermore, we present an example of a discrete transcriptional network model and estimate the number of experiments necessary for its completion. When viewed through the framework of modeling, the complete determination of a biological system becomes a concrete, achievable goal.

Acknowledgements

We thank Patrik D'haeseleer, Daniel Segrè, Jeremy Edwards, and many members of the Church laboratory for helpful discussions.

References

1 Kitano, H. (2001) Systems biology: toward system-level understanding of biological systems. In *Foundations of Systems Biology* (Kitano, H., ed.), pp. 1–36, The MIT Press
2 Kitano, H. (2002) Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. *Curr. Genet.* 41, 1–10
3 D'haeseleer, P. *et al.* (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707–726
4 Valiant, L.G. (1984) A theory of the learnable. *Commun. ACM* 27, 1134–1142
5 Palsson, B. (2000) The challenges of in silico biology. *Nat. Biotechnol.* 18, 1147–1150
6 Kitano, H. (2002) Systems biology: a brief overview. *Science* 295, 1662–1664
7 Palsson, B. (2002) In silico biology through 'omics'. *Nat. Biotechnol.* 20, 649–650
8 Kitano, H. (2002) Computational systems biology. *Nature* 420, 206–210
9 Razin, S. *et al.* (1998) Molecular biology and pathogenicity of mycoplasmas. *Microbiol. Mol. Biol. Rev.* 62, 1094–1156
10 Krupa, B. (2002) On the number of experiments required to find the causal structure of complex systems. *J. Theor. Biol.* 219, 257–267
11 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
12 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
13 Ideker, T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18 (Suppl 1), S233–S240
14 Akutsu, T. *et al.* (1999) Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Symp. Biocomput.* 1, 17–28
15 Covert, M.W. and Palsson, B.O. (2002) Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J. Biol. Chem.* 277, 28058–28064
16 Ibarra, R.U. *et al.* (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* 420, 186–189
17 Segrè, D. *et al.* (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* 99, 15112–15117