In Silico Biology 12 (2012-2014) 1–7
DOI 10.3233/ISB-140461
IOS Press

Opinion

Limits of computational biology

Dennis Bray*
Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK

Received 28 July 2014
Accepted 2 September 2014

Abstract. Are we close to a complete inventory of living processes so that we might expect in the near future to reproduce every essential aspect necessary for life? Or are there mechanisms and processes in cells and organisms that are presently inaccessible to us? Here I argue that a close examination of a particularly well-understood system—that of *Escherichia coli* chemotaxis—shows we are still a long way from a complete description. There is a level of molecular uncertainty, particularly that responsible for fine-tuning and adaptation to myriad external conditions, which we presently cannot resolve or reproduce on a computer. Moreover, the same uncertainty exists for *any* process in *any* organism and is especially pronounced and important in higher animals such as humans. Embryonic development, tissue homeostasis, immune recognition, memory formation, and survival in the real world, all depend on vast numbers of subtle variations in cell chemistry most of which are presently unknown or only poorly characterized. Overcoming these limitations will require us to not only accumulate large quantities of highly detailed data but also develop new computational methods able to recapitulate the massively parallel processing of living cells.

The expectation that, with enough details, a model will miraculously spring to life ... is the stuff of fiction. Jeremy Gunawardena 2012.

Mathematical and computational models—together with the experiments on which they are based — are framed under severely restricted conditions. Biology is so complicated that any investigator or theoretician has to eliminate as many variables as possible. Thus, one might specify the composition of the culture medium; fix the ATP concentration; assume that an embryo is at such and such stage with exactly *this* number of cells; hold temperature and pH constant; and so on. There is no other way to proceed: you have to isolate the process of interest—separate it from extraneous factors—in order to find out how it works. But the real world is not like this. Living creatures are subjected to wind and rain, heat and cold, flood and drought, feast and famine. Dangers are everywhere, externally from the physical world and predators and internally from viruses and parasites. Consequently every organism, in order to survive, has acquired the ability to morph its molecular makeup and change into a myriad of slightly different forms. It has 'learnt' in an evolutionary sense to recognize salient features of its surroundings and how to respond to them.

We know that this plasticity arises at multiple levels, from the selective inhibition or activation of genes to the alternate splicing of RNA molecules and the chemical modification of protein molecules. Taken together these mechanisms create an enormous pool of variant macromolecules and most importantly of protein molecules (sometimes referred to as 'mod forms' or 'proteoforms') [1]. Acting like a buffer, or filter, between the information in DNA (the genotype) and the structure or behaviour of the organism (the phenotype), this filter is highly modifiable. It changes with past history and surroundings and in this way equips a cell and organism to deal with the plethora of possible conditions it encounters. But how much of this universe of chemical forms do we know and understand? And if (as I will argue) there are huge areas outside our

*Corresponding author: Dennis Bray, Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3DY, UK. E-mail: db10009@cam.ac.uk.

present knowledge, how does this affect our ability to build biological simulations on a computer?

1. E. coli chemotaxis

To start with a particularly well-understood biological mechanism, consider that in which bacteria such as *Escherichia coli* smell and swim towards distant sources of food [2, 3]. It is sometimes said that we know everything about *E. coli* chemotaxis but this is not true. There are first of all fundamental areas of uncertainty, such as how flagellar motors are driven by a flux of protons or how clusters of receptors in the bacterial membrane parse incoming information. These fascinating, important topics are the focus of much on-going research, but they are specific to a particular phenomenon and so outside the scope of this article. However, in addition to these specific issues there are also more general limitations of a kind that could apply to any cellular process and therefore serve as a general benchmark of our present level of understanding.

It is well established that *E. coli* detects substances of interest in its environment by means of a simple circuit composed of half a dozen or so well-known proteins (receptors and other signalling proteins) together with flagella and motors [4]. Identification of these components came initially from genetic screens in which mutagenized bacteria were tested for their ability to swim towards distant attractants or away from distant repellents. Tests of this kind provide a powerful selection (since mutants are simply left behind) and—given the added simplicity of a haploid organism—they allow the principal genes to be mapped. Three categories of mutant were found to occur repeatedly in screens: cells that are non-motile; cells that can swim but are unable to respond to any chemical signal of any kind; and cells that are selectively 'blind' only to particular chemicals. An inspired set of physiological, biochemical, and morphological studies then led to the identification of the protein products of these core genes and the mechanisms by which they work. This in turn led to the widely-accepted canonical pathway of *E. coli* chemotaxis, which has been incorporated in a wide variety of computer models of varying sophistication.

However, the short list of components used in such studies is far from complete. In a high throughput screen performed a few years ago dozens of previously uncharacterized genes affecting motility or chemotaxis were detected [5]. Some gene products, for example, modulate the efficiency of swimming when conditions favour biofilm formation. Moreover, the screens used in this study were still restricted in scope and did not test the ability of the cells to chemotax to dozens of potential attractants (or repellents). They were also relatively coarse and would not have identified mutants that were simply reduced in chemotactic efficiency.

Then there is the question of gene expression. It is known that the intracellular concentrations of different components of the chemotaxis machinery vary according to conditions. Growth at high culture density for example changes the ratio of two different chemotaxis receptors Tar and Tsr and hence modifies responses to temperature and pH [6]. But the extent of these changes in gene expression and the conditions under which they occur are largely unexplored.

Then again, take protein posttranslational modification. The *E. coli* system adapts to ambient concentrations of attractant by adding methyl groups to the membrane receptors. The diffusing protein CheY is modified in a different way: being phosphorylated or dephosphorylated as attractant concentrations fall and rise. This same protein is also acetylated according to the metabolic state of the cell. The enzymes performing these and other posttranslational modifications are themselves regulated in activity and sensitive to cellular conditions. Moreover the products of modification are highly variable (each receptor dimer has at least eight sites of potential methylation, for example) and interact with other proteins. These include not only proteins of the chemotaxis pathway but also components of other pathways in the cell such as those involved in the uptake of glucose, or energy production, or control of cell division. Since the interacting proteins are almost all modified and exist in multiple forms we have an explosion of distinct chemical combinations that would be impossible to resolve experimentally.

We also have to remember that a cell is organized in space and subject to numerous physical constraints. What it does depends on *where* molecular components are positioned and *how* they respond to external conditions. For example, protein components of the motors driving the bacterial flagella have been found to diffuse in and out of the surrounding membrane. In contrast to machines made by humans, these bacterial machines are capable of subtle changes in their makeup, for example according to the load they experience. A motor attached to a short, newly made, easy-to-turn flagellum typically contains just one or two copies of the force-generating protein MotB. But as the flagellum grows longer, or if the cell is experimentally exposed to a viscous medium, the motor has to work harder. Under these conditions it can acquire up to 11 MotB units [7]. Evidently if we wanted an exact description of the performance of the cell we would have to include the instantaneous composition of each of its motor together with its position and mechanical properties.

The last black box is protein function. Even if we were given the identity and position of each and every atom of a protein molecule, we might still be in the dark regarding its biological role. Gene ontogeny, a major topic in bioinformatics, draws on an impressive toolbox of analytical procedures such as binding assays, co-expression profiles, cellular locations, and structural homologies. But these tests are all imprecise and prone to error. Even the genome of *Escherichia coli*—tiny by comparison to most organisms and closely examined for half a century—contains genes of unknown function. According to recent estimates, just over half of *E. coli* genes have an experimentally validated function; another third or so have an 'imputed' function (guessed to be part of the DNA replication machinery, for example, or something to do with membrane transport); while the remaining cohort is designated as 'uncharacterised'. That is, we have no idea what they do.

Nor is there is anything to stop any of these proteins playing more than one role. Take as an example the histone-like H-NS protein, an important regulator of gene expression with a major role in suppressing transcription of spurious RNAs. This is located in the *E. coli* nucleoid and seems to be a bona fide DNA-associated protein. But—surprise, surprise— this very same protein also turns up in the flagellar motor as part of the spokes that link the rotor with the stator [8]. Apparently, binding of N-HS stabilizes the motor and promotes its ability to rotate, thereby working in opposition to proteins such as YcgR that inhibit motility. Numerous other examples of multifunctional proteins have been documented, but just how widespread the phenomenon is, no one can tell.

The problem, ultimately, is that the function of a protein is determined by its selective affinity for other molecules and there is no sure-fire way to screen all possible targets. Suppose I told you that two proteins, A and B, coexist in the same cell. How could you ever prove that A and B never interact under any circumstances? You might point to yeast two-hybrid assays, affinity purification schemes, coexpression profiles and so on and tell me that they show no evidence of binding. But I could then respond: 'Well it depends on a particular modification of protein A', or 'Binding occurs only if B is in contact with a third protein', or 'You have to stress the cell first by raising the temperature'. It seems beyond reason that one could ever rigorously exclude every possible interaction and hence identify every potential physiological function.

2. Too much detail?

So there are things about *E. coli* chemotaxis we do not know. Some are specific to this particular phenomenon, such as the mechanism by which the flagellar motors rotate and switch direction. But others are more general. Even for this relatively simple process we cannot enumerate all of the genes that affect its operation. Nor can we say for every conceivable environmental condition, how these genes are regulated. We have little hope of identifying all the combinations of protein modification present in the cell at any instant of time. Nor can we say exactly what these proteoforms do within the cell.

But does it matter? If we know the basic mechanisms by which a cell operates then surely anything more is unnecessary detail? In principle, you might say, we could reproduce on a computer everything necessary to simulate bacteria swimming in a gradient of nutrient. In fact, by extension, since we have a good grasp of most vital processes occurring in these cells, we should be able to reproduce the entire life cycle and predict phenotype from genotype. Something of the kind has been claimed for the simple organism *Mycoplasma genitalium* [9]. The problem with such a high-handed approach is that it ducks the question of why an efflorescence of molecular types is there in the first place. If they have no function then why have they evolved? It is true that molecular processes are inherently stochastic and that this inevitably produces a large amount of meaningless noise. But noise does not explain changes that produce specific effects and require specific interactions between macromolecules. These functional variations must have produced a selective advantage. Many will be subtle in effect and enhance the 'efficiency', or 'reliability', or 'versatility' of fundamental cell processes, which in a laboratory or on a computer may seem like secondary concerns. But for a living creature under natural conditions they can be critical for survival. Organisms compete over a vast landscape of possibilities and this demands innumerable subtle modifications in performance. Even a 1% reduction in reproductive capacity will be enough – taken over many generations—to lead to extinction.

As an illustration, imagine an *E coli* mutant in which the response of motor composition to changes in flagellar load, mentioned above, is defective. In such a mutant, every motor in each cell will always have the maximum number of MotB units. From what we know, we would expect these modified bacteria to swim normally and respond to gradients of food molecules in an adequate fashion. However, every time a mutant cell made a new flagellum (an event that occurs continually as growth proceeds) there would be a period during which the new motor spun furiously, burning energy in a futile manner. In an artificial broth, this hyperactive strain would survive and be barely distinguishable from the wild type cells. But in the wild it would exhaust its food sources sooner and be outcompeted by other cells having better management of their resources. Thus, a mutant defective in the ability to adjust motor torque ... or one that was unable to acetylate CheY, or respond to high-density cultures, and so on ... would be in a sense a "sick" cell. It would survive at some level but would be outcompeted by normal healthy cells.

3. Multicellular organisms

What now happens if we move from bacteria to other systems, especially those in multicellular organisms? Quite obviously, the extent of molecular variation within any eukaryotic cell is much larger than in any bacterium and its potential significance far greater. All of the sources of uncertainty described above for *E. coli* chemotaxis are also found in eukaryotic organisms ... in spades! There are more genes—about 20,000 in humans as compared to 4,000 in *E. coli*—and these genes are further diversified by alternative RNA spicing. Splicing multiplies the number of distinct protein products many times over and can, in some cases, produce hundreds or even thousands of different mRNAs from an individual gene; the *Drosophila* DSCAM and mammalian neurexin are notable examples (both, incidentally, involved in neuronal specificity). Which of many spliced variants is expressed in any particular cell depends on local signals that, for the most part, are poorly understood. An impressive array of sophisticated controls determines how much of each gene product is made. And recall that after it is made, each protein is subject to a plethora of potential modifications that exceeds in variety and frequency anything encountered in prokaryotic organisms.

Our knowledge of the molecular events in a eukaryotic cell is therefore superficial at best. Any systematic attempt to itemize the sources of molecular variation in, for example, a MAP kinase cascade, or the circuit controlling mitosis, or a protein secretion pathway, quickly runs into a morass of incomplete data. In most cases we do not even have a list of principal components, while subtleties such as interactions with other cell processes, controls over efficiency or speed, responses to metabolic state, temperature, and a myriad quantitative considerations are barely considered. Just as we saw in *E. coli* chemotaxis but to a far greater extent, efficiency, adaptation, fitness, fine-tuning to the environment, and other niceties will be almost entirely absent from our simulations. Our simulated cell will be like a robot designed to perform a certain function in one well-defined situation but incapable of adapting to multiple environmental conditions.

Crosstalk between cells is particularly evident in a growing embryo where sequences of cell-cell interactions determine the type and location of the many different tissues and organs. These interactions are based on the same subtle modifications of gene expression and protein composition we have been discussing, so it is legitimate to ask how accurately they can be modelled. We can illustrate this question by considering the crucially important process of the development of somites, future segments of the vertebrate body. In 1997, a group led by Olivier Pourquié proposed a mechanism in which these regularly-spaced anatomical features are produced by a clock-like oscillation of gene expression at the tail of the developing embryo [10]. According to this hypothesis, the oscillations cease as cells are left behind by the dividing tissue and become locked in a particular phase, which then determines the future position in the somatic or nonsomatic tissue.

More recent studies add weight to this notion, with particular reference to transcription factors, *her1* and *her7*, which do indeed oscillate in level of expression at an appropriate rate. Current models of this process show how these oscillations could be created through a mechanism of autoregulation created by negative feedback and transcriptional delays. Moreover, formation of mixed dimers of *her1*, *her7* and a third transcription factor *her6* can account in a simple way for the outcome of various genetic manipulations [11]. Other detailed simulations include such features as the degradation of the transcription factors, stochastic variations in protein numbers, and coordination of the intrinsic cycles between neighbouring cells [12].

Taken together these computational studies represent a major advance in our understanding of a fundamental event in the formation of a vertebrate embryo. But one should be aware that the models on the table are essentially phenomenological and only loosely tied to real molecules. Parameters such as rates of synthesis, affinities of binding, rates of transcription and so on, are given in relative terms and for the most part adjusted to give the desired outcome. Candidate molecules have almost never been isolated in biochemically pure form or their activities measured *in vitro* (in contrast to the situation with *E. coli* chemotaxis). Nor does anyone claim that a handful of molecules including a few transcription factors and one or two surface receptors could provide anything like a complete description of the phenomenon. Such a fundamental development mechanism, on which so much depends, will engage many thousands of massively interconnected circuits, acting as switches, integrators, oscillators, coincidence detectors, and so on. Each circuit will be the target of multiple fine-tuning mechanisms of the kind already mentioned, about which we know virtually nothing.

4. Verisimilitude

In the award-winning animated movie Shrek, the eponymous hero—an ogre with a large and conspicuously ugly green face—displays an astonishing range of human emotions. As the plot unfolds, his face expresses grief, anger, hubris, ecstasy, contrition, amusement, embarrassment, determination, self-doubt, and so on—often segueing from one emotion to the next in a trice. The apparent reality of Shrek's feelings is a tribute to both the artistic abilities of his creators and also the techniques available to contemporary animators. One of the latter, known as *facial capture*, employs small reflective beads attached to salient points on an actors face. The tiny relative motions of the beads as the actor portrays different emotions are captured on a head-mounted camera and later projected onto a graphical image of Shrek's face. What the viewer sees in the final movie is therefore a balletic display of pixels on a two-dimensional surface, encoded by long linear sequences of digital signals. Despite our anthropomorphic interpretations there is no face, no underlying neuronal circuitry, no emotion; it is all in the eye of the beholder.

Something similar applies to cell biology simulations, particularly those involving graphical displays. Dynamic simulations of diffusing macromolecules, growing and shrinking microtubules, crawling cells, folding embryos on the computer screen and so on, are seductively convincing. They behave exactly as we expect, so it is tempting to take them at face value and forget they are just pixels on a screen. It is easy to treat the program as if it were a live experiment and make observations just as though you were watching a real specimen. The results can be valuable and informative; indeed this is one of the main reasons to build a model in the first place. But there are also dangers.

As we saw above, our understanding of the embryonic process of somatogenesis is superficial at best and limited to certain key molecular events (which are themselves still hypothetical). Since there is nothing special about somatogenesis in this regard, we must be in a similar state of ignorance regarding a long list of other embryonic processes. Moreover, molecular plasticity is not restricted to development but essential also for the maintenance of adult tissues. Muscle, skin, fat, and other tissues are continually modified by conditions and changed by the different physiological constraints imposed by exercise, diet, disease and so on. Even a single small structure such as a cilium on the cell surface is the product of many hundreds of different protein molecules, most of which are subjected to a plethora of modifications. Multiply a cilium's volume by a factor of 10^{12} or so to match to the volume of a tissue and you will see that we have only scratched the surface of the underlying complexity.

5. Does it matter?

The introduction of computers to biological research has revolutionized our ability to analyse and understand living systems. Contemporary simulations based on mathematical models of a host of different systems provide us with rigorous quantitative tests. They can tell us whether a hypothetical mechanism could work in the manner proposed and reveal flaws in current thinking—even point the way to improved models and novel insights. But computer programs only recapitulate logical processes thought to occur in organisms; they are not substitutes for the living tissue itself. Any features not explicitly included in the simulation code—which might include spatial dimensions, physical parameters, mechanical constraints, temperature dependence, nutritional status—will not appear by magic. Nor can computer programs incorporate the jaw-dropping biochemical complexity of living systems, most of which is presently uncharacterized.

Whether this matters, and how much, is debatable. One view is to say that these are minor considerations, since we are not concerned with minutia of cell individuality. If we have a good grasp of the essential mechanism of *E. coli* chemotaxis, for example, or somatic development then this is sufficient. Should we need to know more about, say, the modulation of motor torque or the interplay between transcription factors in presomitic mesoderm we can always dig more deeply. We can collect more data and refine our models step-by-step. In this way our computer simulations will become increasingly inclusive and accurate, rather as a molecular structure becomes more detailed through x-ray crystallography of increasingly resolution. These simulations will never be *final*, since there will always be unresolved questions and molecular uncertainties. But they could be as accurate as we wish.

But there is also a possibility that the astonishing variety of chemical forms within living cells is actually important, has functions that we only dimly appreciate. Living tissues are the product of myriad cells working together and this cooperation requires continual fine adjustments. Each cell has to be almost infinitely adaptable so it performs in the desired manner, rather like individuals in a human society. If cells lack this ability to make subtle adjustments (as they do in a typical computer simulation) we can expect a progressive accumulation of errors and the eventual catastrophic breakdown of the entire organism.

There are also biological systems—most notably the immune system and the nervous system—whose function seems to depend on a seemingly inexhaustible supply of molecular heterogeneity. Thus the process of neuronal specificity and learning requires that every individual nerve cell in the mammalian CNS is essentially distinct. We usually interpret this plasticity in electrical terms, arising from action potentials, electrotonic spread, synaptic delays, and so on. But the electrical events are produced by protein channels, pumps, and receptors embedded in the nerve cell membranes, and these proteins are themselves subject to the full panoply of variation—alternative splicing, post translational modification, spatial differentiation. It is at least possible that the primary substrate for long-term memory formation is the post-translational modification of synaptic proteins [13].

6. Future prospects

Although I have argued that computational models of living organisms have intrinsic limitations, this is not meant as a counsel of despair. Nothing in biology is outside chemistry and physics, so it should be possible, eventually, to overcome present barriers and develop simulations that are much closer to reality. Evidently this will require the collection of an enormous quantity of data at an unprecedented level of detail. We may also have to develop novel hardware and software approaches so as to better handle the myriad modifications and interactions that characterize living matter. As an example of the kind of qualitative change that will be needed, consider the novel “brain chips” currently under development to simulate the neuronal activity of the brain [14]. Here the canonical sequential architecture proposed by von Neumann is replaced by massively-parallel networks, which may be digital, analogue, or hybrid in character. A similar approach could be applied to intracellular reactions and would then allow programmers to produce a far more detailed and integrated picture of a cell.

If we look to a future in which these challenges have been met then we should see a transformation in our expectations. Computer-based cells will now be able to detect and respond correctly to a whole constellation of external conditions, sending and receiving hugely complex sets of messages enabling them to cooperate and work together to form tissues and organs. Eventually it should be possible to recapitulate embryonic development with such detail that we include individual variations in cells. We should then be able to recognize the origins of defects in tissues, understand the consequences these have for the organism, and most importantly identify the steps necessary to correct these defects. In other words, medicine itself will come within a rigorous analytic framework, and our health and longevity will be under our control to an unprecedented degree.

References

[1] S., Prabakaran, et al., Post-translational modification: Nature’s escape from genetic imprisonment and the basis for dynamic information encoding, *WIREs Syst Biol Med* (2012), 1–19.

[2] H.C. Berg, *E. coli In Motion*. Biological and Medical Physics Biomedical Engineering, ed. E. Greenbaum, New York: Springer-Verlag, 2004, 133.

[3] Y. Tu, Quantitative modeling of bacterial chemotaxis: Signal amplification and accurate adaptation, *Annu Rev Biophys* **42** (2013), 337–359.

[4] D. Bray, M.D. Levin and K. Lipkow, The chemotactic behavior of computer-based surrogate bacteria, *Curr Biol* **17** (2007), 12–19.

[5] T. Baba, et al., Construction of Escherichia coli K-12 in-frame single-gene knockout mutants: The Keio collection, *Mol Systems Biol* (2006).

[6] M. Demir, et al., Effects of population density and chemical environment on the behavior of Escherichia coli in shallow temperature gradients, *Physical Biology* **8** (2011), 1–8.

[7] J. Yuan, et al., Adaptation at the output of the chemotaxis signalling pathway, *Nature* **484** (2012), 233–237.

[8] P.K. Carlquist and D.F. Blair, Adjusting the spokes of the flagellar motor with the DNA-binding protein H_NS, *J. Bacteriol* **193**(21) (2011), 5914–5922.

[9] J.R. Karr, et al., A whole-cell computational model predicts phenotype from genotype, *Cell* **150**(2) (2012), 389–401.

[10] I. Palmeirim, et al., Avian hairy gene expression identifies a molecular clock linked to vertebrate segmentation and somitogenesis, *Cell* **91** (1997), 639–648.

[11] C. Schröter, et al., Topology and dynamics of the zebrafish segmentation clock core circuit, *PLoS Biology* **10**(7) (2012), e1001364.

[12] A. Ay, et al., Short-lived Her proteins drive robust synchronized oscillatons in the zebrafish segmentation clock, *Development* **140** (2013), 3244–3253.

[13] A. Routtenberg, Long-lasting memory from evanescent networks, *Eur J Pharmacol* **585** (2008), 60–63.

[14] R.F. Service, The brain chip, *Science* **345** (2014), 614–616.