Some Brief Thoughts on Philosophical Questions in Systems Biology

(or, my reaction to Sydney Brenner and other critics)

Is Systems Biology useless because it attempts to tackle ill-posed 'inverse problems'?

Jacques Hadamard, over a century ago, set out the conditions he believed were required for a mathematical problem to be well-posed [6]. In essence, the three requirements of a solution to such a problem were existence, uniqueness and smoothness. Sydney Brenner points out that the problems Systems Biology purports to solve are of an inverse nature (i.e., they seek to go from behaviour/function/output to mechanism/structure/model, rather than the other way round), and are thus ill-posed in the Hadamardian sense [2]. He says that because in inverse problems one has to overcome an information gap (the observed data does not in general uniquely specify the underlying system), these problems can only be solved by including some sort of prior information or assumptions. Others like Lewis Wolpert have also raised similar doubts about the efficacy of Systems Biology approaches.

This argument seems to involve an implicit (and in my view, erroneous) assumption that if a problem is ill-posed, no useful solution or model can be obtained by attempting to solve it. It seems to me that nature is proficient at solving inverse problems. Evolution itself is solving a kind of inverse problem, that of constructing organisms which best fit the prevailing environmental circumstances. All of human and animal cognitive learning involves the solution of inverse problems. One of the most dramatic examples is human language acquisition. It is generally accepted that the amount and type of linguistic input a child receives is insufficient to precisely specify a given grammar: this is often referred to as the "poverty of stimulus" and the consequent "paradox of language acquisition". It has been proven mathematically that any formal language that has hierarchical structure capable of infinite recursion is unlearnable from positive evidence alone [4]. And yet, we learn; indeed, as studies of cultures such as the Kaluli of Papua New Guinea and the Warlpiri of Australia have shown, infants are even capable of acquiring language without ever being directly spoken to [12]. Linguists have varying theories for how this happens. We may be hard-wired to search within some sort of restricted class of grammars (e.g., Chomsky's universal grammar hypothesis [3]). We may simply be able to come to good heuristic approximations such that we all have slightly differing grammars which overlap sufficiently to allow for communication. There is little consensus on the cognitive algorithms or mechanisms, but what is clear is that children do solve the inverse problem of learning language, very efficiently and very usefully.

This question also seems to touch upon a broader philosophical issue, which is the nature of scientific reasoning and the role of induction. It seems to me that the forward/inverse problem distinction closely mirrors the distinction between deductive logic and inductive inference. So perhaps there is a parallel between Brenner's rejection of the utility of solving inverse problems and Popper's rejection of inductive reasoning as a justifiable scientific method [14]. Nevertheless, despite its logical unsoundness, inductive generalisation of one form or another has surely been at the heart of all scientific progress. The problem of induction has long intrigued philosophers; the classical treatment was by Hume, who advocated a practical approach based on common sense, recognising the inevitability of inductive reasoning in our daily lives [9]. In recent decades, work in statistical learning theory has for the first time allowed us to give a mathematical underpinning to Hume's notion of 'common sense'. Approaches such as Bayesian inference allow us to quantify uncertainty and construct a reasonable framework for reliable reasoning [7]. Popper's 'falsifiability' criterion has been controversial, but work by Vapnik and Chervonenkis suggests that a modified version of it can be resurrected in a statistical learning

theory framework, with the notion of VC dimension serving as an analogue for Popper's 'degree of falsifiability' [8,16]. Thus it would appear that in practice, a scientific method founded on falsifiability cannot really be separated from induction.

At a more practical level, one may ask: "OK, so nature can do all sorts of wonderful things, we can come up with philosophical justifications and so on, but in Systems Biology, it's essentially computers that need to be able to solve inverse problems. So how do we know that our technology and algorithms are good enough to do that?". I think advances in Computer Science, in particular in Machine Learning, provide considerable grounds for optimism. Brenner himself suggests the sorts of conditions needed to be able to solve inverse problems: you need some extra information, in the form of *a priori* assumptions [2]. Machine Learning methods allow us to do this via the process of regularisation, a technique developed initially by Tychonoff precisely in order to deal with ill-posed or inverse problems [15]. Essentially regularisation just means invoking some sort of smoothness or simplicity assumption, i.e., penalising overly complex models. It can be seen as an implementation of Occam's razor; in a Bayesian context, it corresponds to specifying a prior distribution over model parameters. Regularisation, combined with heuristic optimisation methods, actually make it feasible for computers to search over infinitely large spaces (e.g., of possible molecular structures) and come up with good models. Of course, such methods do not always work well, but there are already a large number of success stories from a wide range of scientific domains.

None of this is to belittle the importance of detailed experimental investigation in order to build models in the 'forward' direction. However, such work is generally expensive and time-consuming; at the very least, Systems Biology approaches can be useful in focusing experimental efforts along potentially fruitful directions, a point Brenner seems not to acknowledge. To me, Systems Biology is fundamentally about the symbiotic cyclic interplay between the forward and inverse problems; computational models undoubtedly need to be continuously refined based on experiment, but the models can in turn help to make the use of limited experimental resources more efficient.

Is all the high-throughput data Systems Biologists like to use too noisy to tell us anything meaningful?

Noise certainly is a problem. However, as we saw in the previous section, inverse problems already involve the inference of *missing information*, even if the data is entirely reliable. So the addition of noise does not fundamentally change the situation; it just adds an extra source of uncertainty to what was already an ill-posed problem. This uncertainty has to be dealt with within a statistical framework, and the approaches outlined above provide a way of doing so; any reasonable machine learning method allows for the incorporation of an error term to account for noise in the observations. The wider area of statistical signal processing has long been concerned with the problem of extracting signal from noise, so this is a well-studied issue that is certainly not unique to biology. Of course there is a certain threshold of noise beyond which the data will become useless, but surely not all high-throughput data should be discarded just because we expect it to be noisy.

Brenner also makes the related point that not everything we are measuring may be of significance; biologically, there are likely to be many 'don't-care conditions', i.e., variables whose values may fluctuate substantially because they are not functionally or evolutionarily constrained [2]. Again, this is not something that need be a problem: it corresponds to the *feature selection* problem which has been extensively studied in the Machine Learning community. One can use statistical concepts like entropy to quantify the degree of uncertainty in one's observations of any given variable, and a large number of methods exist to rank and select variables based on their information value for the system and circumstances being studied.

Does Systems Biology seek to avoid questions of causality? Does it seek to release us from having to think by somehow automating science?

I'm not sure I understand this criticism entirely, but presumably it stems from the 'black-box' nature of many computational models; they may be able to reproduce the data we observe, but it is not always easy to understand how they do so and what correspondence that bears to reality. This is certainly true in some cases, although not always; for instance, certain machine learning approaches like Inductive Logic Programming [11] have been designed specifically to produce interpretable, rule-based models, and have been successfully applied to a number of biological problems such as drug design [10]. More generally, however, I don't think the outputs of the large-scale computational modelling methods used in Systems Biology should be necessarily seen as ends in themselves. As I have said earlier, it is the iterative cycle of modelling and experiment that leads us to more and more refined understanding of the system being studied.

On the broader issue of causality, this has of course been one of the most vexed notions in philosophy at least since Aristotle. Whilst I certainly don't have any novel insights into the fundamental questions, I do think that we should be careful not to subscribe to a sort of 'causality fundamentalism': the idea that we can find a proximate cause for every phenomenon we observe. Perhaps our notion of causality tends to be too simplistic at times; this seems to be something that has been ingrained into our psychology for evolutionary reasons, and it is probably one of the factors underlying the invention of religion [5]. There is a large body of literature on emergent phenomena in complex systems [1], and while not all of this work may be of the highest standard, I don't think one can dismiss it entirely, as Brenner seems to do [2]. It would appear that causality can in fact be quite complicated at times. I like Denis Noble's conception of The Music of Life [13]: the vision of a biological system as a kind of orchestra, producing beautiful music as a result of all the parts functioning in perfect harmony. However, the biological orchestra has no conductor: it somehow manages to conduct itself.

References

1. Bedau, M.A.: Weak emergence. In: Tomberlin, J. (ed.) Philosophical Perspectives: Mind, Causation, and World, vol. 11, pp. 375–399. Blackwell (1997)
2. de Chadarevian, S.: Interview with Sydney Brenner. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 40(1), 65–71 (March 2009), http://dx.doi.org/10.1016/j.shpsc.2008.12.008
3. Chomsky, N.: Aspects of the Theory of Syntax. MIT Press, Cambridge, MA (1965)
4. Gold, E.M.: Language identification in the limit. Information and Control 10(5), 447 – 474 (1967), http://www.sciencedirect.com/science/article/B7MFM-4DX4964-C/2/cc2c59f85d26a52e92aad38d37790439
5. Guthrie, S.: A cognitive theory of religion. Current Anthropology 21(2), 181 (1980)
6. Hadamard, J.: Sur les problèmes aux derivées partielles et leur signification physique. Princeton University Bulletin 13, 49–52 (1902)
7. Harman, G., Kulkarni, S.: Reliable Reasoning: Induction and Statistical Learning Theory. MIT Press, Cambridge, MA (2007)
8. Harman, G., Kulkarni, S.: Statistical learning theory as a framework for the philosophy of induction. In: Bandyopadhyay, P., Forster, M. (eds.) Philosophy of Statistics. Elsevier (to appear)
9. Hume, D.: An Enquiry Concerning Human Understanding (1748)
10. King, R.D., Muggleton, S., Lewis, R.A., Sternberg, M.J.: Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. Proceedings of the National Academy of Sciences of the United States of America 89(23), 11322–11326 (1992), http://www.pnas.org/content/89/23/11322.abstract
11. Muggleton, S., de Raedt, L.: Inductive logic programming: Theory and methods. The Journal of Logic Programming 19-20(Supplement 1), 629 – 679 (1994), http://www.sciencedirect.com/science/article/B6V0J-45GMW00-14/2/f4275705a38350560bdcfbb82bacc061

12. Naigles, L.R.: Form is easy, meaning is hard: resolving a paradox in early child language. Cognition 86(2), 157 – 199 (2002), http://www.sciencedirect.com/science/article/B6T24-472JJ3V-1/2/2fe95f0b5cb2e32ea54ad9e40242ac4a
13. Noble, D.: The Music of Life: Biology Beyond Genes. Oxford University Press (2006)
14. Popper, K., Miller, D.: A proof of the impossibility of inductive probability. Nature 302, 687–688 (1983)
15. Tychonoff, A.N.: On the stability of inverse problems [in Russian]. Doklady Akademii Nauk SSSR 39(5), 195–198 (1943)
16. Vapnik, V.N.: The Nature of Statistical Learning Theory (2nd ed.). Springer-Verlag New York (2000)