Acquiring a Debt Worth Repaying :
why Evolutionary Computation should gain yet further inspiration from Biology

Leo Caves [1], John Clark [2], and Susan Stepney [2]

[1] Department of Biology, University of York, York YO10 5DD, UK
lsdc1@york.ac.uk

[2] Department of Computer Science, University of York, York YO10 5DD, UK
{jac,susan}@cs.york.ac.uk

Abstract. We claim that the alleged "debt" owed by Genetic and Evolutionary Computation (GEC) to Biology is exaggerated. However, we do believe that Biology is potentially an extraordinarily rich source of inspiration, and that a highly fruitful way forward is to genuinely plunder it, to be more greatly inspired than at present. We would simply like to increase the debt: to *acquire a debt worth repaying*.

1 What Debt?

> *"The field of Genetic and Evolutionary Computation (GEC) has greatly benefited by borrowing ideas from the biological sciences. Recently, it has become clear that GEC can help solve biological problems, and thereby repay its debt."* – BioGEC call for papers, 2004

We argue that far from being in a position to "repay its debt", the bio-inspired computer science community still has much to learn from Biology.

2 The power of Genetic and Evolutionary Computation

Genetic and Evolutionary Computation (GEC) has shown great promise in bioinformatics: in fitting (relatively) simple models to large quantities of biological data. GEC tends to be better for static problems – providing a fit to a given data set – rather than more dynamic problems – trying to model complex time-varying data [1].

The search approaches taken tend to be rather inflexible, and build in quite a lot of modelling assumptions. For example, the choice of particular primitive functions used in a GP fit to data has a very strong influence on the models that result.

Recent innovative advances in intrinsic hardware evolution could potentially be carried over into biology. Intrinsic GEC could be used to evolve parameters for biological experiments (we stop short of suggesting, for biological *organisms*, but maybe we should not be so coy?), with the evaluation of the population being performed *in vivo*. An interesting recent development is that of the Robot Scientist

Project which incorporates logic processes to suggest and refute hypotheses concerning the role of metabolic enzymes in a fully automatic system incorporating a robot for performing actual cell culturing experiments [14].

GEC problems tend to be posed in terms of search over static "fitness landscapes". Some work has been done on tracking changes to fitness landscapes in general, e.g. [5], and co-evolution in particular (GECCO typically has a strand dedicated to co-evolution) , but this rather static feel pervades.

More could be done to use GEC techniques examine models (evolved or otherwise) for stability, *etc*: more work on co-evolution of solutions and test case data; on evolution of parameters to stress modelling assumptions; on evolution of sets of diverse models to explore a problem space, etc.

3 The weaknesses of GEC

Many problems in biology nowadays are being cast in terms of *dynamical systems based on complex networks of interactions* – here the currency is network topology, phase spaces, attractors, trajectories – with short term changes in a system being characterised as its trajectory between attractors, longer term changes as changing phase space parameters altering the position, number and class of attractors, and even longer term changes modifying the dimension of the phase space itself [2]. Metabolic and regulatory networks [12], developmental processes [11], organism movement, and learning processes [13], all have models couched in these terms. Classic GEC, with its different conceptual basis of search over fitness landscapes, does not appear well-suited to this important class of problems.

4 Important questions in biology

Rather than looking for the biological "nails" we can attack with our shiny GEC "hammer", however, let us take a moment to step back and ask what the biologists think are the important questions and difficult problems in Biology today.

There is a trend in Biology to become more holistic, or integrative, in acknowledgement of (and response to) the limits of reductionism. With all the components (genes, proteins, metabolites) we can start to understand the system ("bottom-up"), or acknowledge the system (organism, process or pathology), and then discover the subsystems and their coupling ("top-down") [15]

Parts of Biology are undergoing a culture change to a "data rich" environment (e.g. the "discovery science" of 'omics), much as astronomy and particle physics have already undergone. Can biology learn anything from those domains? Or is the complexity of its models and systems so much greater that the situation is qualitatively different?

What are the grand challenges in Biology? At the molecular level, a robust solution to the Protein Folding problem (the prediction of the three-dimensional structure of a protein based on its amino-acid sequence alone) is something of a holy

grail in biophysics [4]. Folding is a test-bed for myriad search and optimisation methods and is a key driver in high performance computing developments [3]. Large multidisciplinary efforts are being targeted at whole cell modelling [e.g 8, 10]. Such projects exemplify the full interplay of simulation and knowledge engineering into the framework of data-driven 'omics approaches. A significant driving force for these efforts is the prospect of an *in silico* platform for developing patient-specific therapeutic strategies in molecular medicine. Understanding evolution and development in terms of the regulatory networks of molecular interactions is a huge challenge and targets the basic mechanisms of these fundamental biological processes [9]. Simulating whole organs or organ systems requires models that link across multiple scales of biological organisation. Such *computational physiology* is emerging as a highly quantitative discipline [7].

5 What further biological inspiration should we seek?

The fact that there are *bio*-inspired approached being used to analyse *biological* data is possibly just coincidence – *any* search technique, however it was inspired, that works well would be useful. GEC techniques may be of interest to biologists mostly because they are couched in a veneer of biological terminology.

This is true of bio-inspired population search models in general. Whether the search be performed by an evolutionary approach, an immune system, a swarm system, or whatever, the underlying structure of the algorithms has a great deal of similarity (fittest cohort of a population being used to produce the next generation), and very little biological richness or diversity. Much effort is spent on "parameter tweaking" (eg, small changes to rates of a simplistic mutation model), and rather less on more sophisticated bio-inspiration. (There are, of course, exceptions to all these, but they are indeed the exception, rather than the rule.)

Where is the biological *inspiration*? Biological populations are much larger than their algorithmic progeny. Biological chromosomes have much more structure than digit strings. Biological search does not start from "random" populations. Evolutionary landscapes involve ecologies, not single species. Genotype and phenotype are separated by a huge developmental distance, involving complex dynamics. Biology has several levels of organisational, spatial, and temporal structure. Biological "algorithms" are *not* solving optimisation problems. Biological organisms operate within the real world, and suffer constraints (for example, and organism has to be viable at all stages during its development, not only at its "final fitness evaluation"), and exploit emergent consequences of embodiment.

This gap between the biology and its implemented inspiration may be due to a cultural difference. Biologists are interested in understanding the world *as it is (and how it came to be)*. Computer Scientists are interested in building simple and powerful models and tools – with the emphasis on *simple* (but not simplistic). Abstraction is the name of the game – but when considering biological mechanisms, too much abstraction can discard the very complexity that gives biology its power. In the race to become simple, comprehensible, and tractable, the domain may have been divested of its complexity, emergence, realism, and power [6].

In addition to the subject matter of biology, GEC (and CS in general) has much to learn from biological experimental methods. Simulation constitutes an *experiment*. Thus issues of hypothesis, experimental design, control, reproducibility, and interpretation become relevant. Never again should we see papers or presentations describing the import of a single data point output from an uncontrolled simulation.

6 In conclusion

Rather than noticing the current similarity between vocabularies, based on naïve implementations of biological "algorithms", computer science should be embracing and plundering the vast wealth of structure and diversity available in biology. Only once all this richness has moved into the artificial domain, and genuine light shed on complex problems, can computer science hope to start thinking of "repaying its debt" to biology.

References

1. J. Allen *et al*. High-Throughput Classification of Yeast Mutants for Functional Genomics using Metabolic Footprinting. *Nature Biotechnology* **21** 692–696, 2003
2. Y. Bar-Yam, I. R. Epstein. Response of complex networks to stimuli. *Proc. National Academy of Sciences* **101** 4341-4345, 2004
3. BLUEGENE: http://www.research.ibm.com/bluegene/
4. CASP: Comparative Assessment of Structure Prediction http://predictioncenter.llnl.gov/
5. H. G. Cobb and J.J. Grefenstette, J. J. (1993). Genetic algorithms for tracking changing environments. *Genetic Algorithms: Proc. Fifth Intl. Conf. (ICGA93),* San Mateo: Morgan Kaufmann, 523-530.
6. Marie E. Csete, John C. Doyle. Reverse Engineering of Biological Complexity. *Science* **295** 1664–1669, 2002
7. Edmund J. Crampin *et al*. Computational Physiology and the Physiome Project. *Experimental Physiology* **89**(1) 1–26, 2004
8. CYBERCELL: www.cybercellproject.org
9. E. H. Davidson *et al*. A genomic regulatory network for development. *Science* **295** 1669–1678, 2002
10. ECELL: www.e-cell.org
11. Brian C. Goodwin. *How the Leopard Changed Its Spots: the evolution of complexity*. Phoenix, 1994
12. Stuart A. Kauffman. *The Origins of Order: self-organization and selection in evolution*. Oxford University Press, 1993
13. J. A. Scott Kelso. *Dynamic Patterns: the self-organization of brain and behavior*. MIT Press, 1995
14. R. King *et al*. A Robot Scientist: automated hypothesis generation and experimentation for functional genomics. *Nature* **427** 247–252, 2004
15. H. Kitano. Systems Biology: a brief overview. *Science* **295** 1662–1664, 2002