

Chapter 19

THREE FUNDAMENTALS OF THE BIOLOGICAL GENETIC ALGORITHM

A Biologist's View of Biological Inspiration vs. Biology Envy

Stephen Freeland

Department of Biology, UMBC, Catonsville, MD 21250 email: freeland@umbc.edu

Abstract: Evolutionary computing began by lifting ideas from biological evolutionary theory into computer science, and continues to look toward new biological research findings for inspiration. However, an over enthusiastic 'biology envy' can only be to the detriment of both disciplines by masking the broader potential for two-way intellectual traffic of shared insights and analogizing from one another. Three fundamental features of biological evolution illustrate the potential range of intellectual flow between the two communities: particulate genes carry some subtle consequences for biological evolution that have not yet translated mainstream EC; the adaptive properties of the genetic code illustrate how both communities can contribute to a common understanding of appropriate evolutionary abstractions; finally, EC exploration of representational language seems pre-adapted to help biologists understand why life evolved a dichotomy of genotype and phenotype.

Key words: particulate genes; genetic code; phenotype; genotype; biology envy

1. Introduction

At present, we know of only two explanations that can account for complex (i.e. highly improbable) 'design' wherever it is found: conscious intelligence and natural selection. Biological explanations are necessarily limited to using the latter to explain what has spontaneously emerged, *ex nihilo*, during 4 billion years: any appeal to Intelligent Design steps beyond the remit of science. By contrast, evolutionary computing (EC) inherently relies on a degree of intelligent design to define a system of rules within which evolution takes place. Each community faces a different challenge:

the biological theorist is ultimately charged with explaining in detail why one 'run' of the algorithm turned out precisely as it did. The EC researcher is charged with designing rules that, once initiated, auto-generate adaptive solutions without compromising the spontaneity of the adaptive process.

These challenges are qualitatively different, and there is no explicit need for EC to emulate biology. In these proceedings, Ostrowski and Reynolds provide an excellent example of how a conscious rejection of biological analogy need not impede the creation of an efficient EC search algorithm. That said, the sophisticated adaptation evident within the biosphere is a natural inspiration for the intelligent design component of EC systems. Indeed, EC not only began by transposing fundamental concepts of natural selection into computer science (Holland, 1992) but continues to look towards biology for new ideas (e.g. Banzhaf, these proceedings). Thus, elsewhere in this volume, Spector's thought-provoking essay argues that further 'borrowing' from biology is both inevitable, as EC techniques continue to grow in complexity and desirable, as biological research continues to change our perception of the evolutionary process. As an evolutionary biologist, I do not disagree with this conclusion but extend it to include Goldberg's sentiment, expressed at the workshop, that over zealous borrowing could transform biological inspiration into a blind "biology envy" (Notably, biologists themselves have long decided that an analogous blind "physics envy" is unhelpful to their research, (Yates, 1978)). In this context, I introduce three fundamental features of biological evolution:

1. particulate genes and population genetics
2. the adaptive genetic code
3. the dichotomy of genotype and phenotype

Each phenomenon is chosen to represent a different point in the spectrum of possible relationships between computing and biological evolutionary theory. The first is chosen to ask whether current EC has fully transferred the basics of biological evolution. The second demonstrates how both biological and computational evolutionary theorists can contribute to common understanding of evolutionary abstractions. The third is chosen to illustrate a question of biological evolution that EC seems better suited to tackle than biology. Taken together, their relevance is not to showcase what biology can offer EC, but rather to explore the difference between usefully borrowing from (or feeding into) biological evolutionary theory, and the indiscriminate one-way intellectual traffic endemic to biology envy.

In principle, distinguishing biological inspiration from biology envy is straightforward. It is the difference between importing biological phenomena into EC because they improve the evolutionary heuristic (e.g. genetic drift), as opposed to importing phenomena simply because they have emerged during 4 billion years of evolution (e.g. walruses). However, this distinction

also inherently implies equivalency: to qualify as inspiration, a 'useful' feature of biological evolution must also translate, functionally, into EC.

A plausible case in point is recombination. Exchange of genetic material between individuals appears near-ubiquitous in biological systems, and has long been regarded as an adaptive feature that increases genetic variation and thereby the power of natural selection (e.g. Fisher, 1930). Within EC, the introduction of recombination has proved far more controversial: effects analogous to reductions in biological fitness caused by epistasis (the different phenotypic effects of a gene depending on its genomic context) often result, enough to elicit the comment during this workshop that:

Our biological friends assure us that recombination is not a problem, but our experiences tell us otherwise. (*Goldberg, pers. comm.*)

This observation from EC may stimulate biologists to re-evaluate the perceived status of recombination as an adaptive phenomenon. However a different possibility is that the communities' observations are correct, but that biological systems somehow limit the potential for epistatic chaos with features that have not been incorporated into mainstream EC. In other words, recombination in the two systems may not be equivalent. One suggestion for such a feature relates to the population genetics of particulate genes.

2. Particulate Genes and Population Genetics

Today it is widely recognized that Darwin's theory of natural selection, as published, would not have worked. The problem lay not with what he added to our understanding, but rather in what he failed to address:

The laws governing inheritance are quite unknown...Noone can say why the child often reverts in certain characters to the grandmother or grandfather. (Darwin, 1859)

Mainstream thinking of the time viewed the genetic essence of phenotype as a liquid that blended whenever male and female met to reproduce. It took the world's first professor of engineering, Fleeming Jenkin (1867), to point out the mathematical consequence of blending inheritance: a novel advantageous mutation arising in a sexually reproducing organism would dilute itself out of existence during the early stages of its spread through any population comprising more than a few individuals. This is a simple consequence of biparental inheritance. Mendel's theory of particulate genes (Mendel, 1866) replaced this flawed, analogue concept of blending inheritance with a digital system in which the advantageous version (allele) of a gene is either present or absent and biparental inheritance produces

diploidy. Thus natural selection merely alters the proportions of alleles in a population, and an advantageous mutation can be selected into fixation (presence within 100% of individuals) without any loss in its fitness. Though much has been written about the Neo-Darwinian Synthesis that ensued from combining Mendelian genetics with Darwinian theory (e.g. Mayr and Provine, 1998), it largely amounts to biologists' gradual acceptance that the particulate nature of genes alone provided a solid foundation to build detailed, quantitative predictions about evolution. Indeed, derision of mathematical models of genes in populations as "bean bag genetics" (Mayr, 1963) overlooks the scope of logical deductions that follow from particulate genetics. They extend far beyond testable explanations for adaptive phenomena and into deeper, abstract concepts of biological evolution. For example, particulate genes introduce stochasticity into evolution. Because genes are either present or absent from any given genome, the genetic makeup of each new individual in a sexually reproducing population is a probabilistic outcome of which particular alleles it inherits from each parent. Unless offspring are infinite in number, their allele frequencies will not accurately mirror those of the parental generation, but instead will show some sampling error (genetic drift). The magnitude of this sampling error is inversely proportional to the size of a population. Wright (1932) noted that because real populations fluctuate in size, temporary reductions can briefly relax selection, potentially allowing gene pools to diffuse far enough away from local optima to find new destinations when population size recovers and selection reasserts itself. In effect, particulate genes in finite populations improve the evolutionary heuristic from a simple hill climbing algorithm to something closer to simulated annealing under a fluctuating temperature.

One final property of particulate genes operating in sexual populations is worthy of mention. In the large populations where natural selection works most effectively, any novel advantageous mutation that arises will only reach fixation over the course of multiple generations. During this spread, recombination and diploidy together ensure that the allele will temporarily find itself in many different genetic contexts. Classical population genetics (e.g. Fisher, 1930) and experimental EC systems (e.g. O'Reilly 1999) have focused on whether and how this context promotes selective pressure for gene linkage into "co-adapted gene complexes". A simpler observation is that a novel, advantageous allele's potential for negative epistatic effects is integral to its micro-evolutionary success. Probability will favor the fixation of alleles that are good 'team players' (i.e. reliably imbue their advantage regardless of genetic background, (see Dawkins, 1989). Many mainstream EC methods simplify the population genetics of new mutations (e.g. into tournaments), to expedite the adaptive process. This preserves non-blending inheritance and even genetic drift, but it is not clear that it incorporates

basic difference between biology and EC contribute anything to our understanding of why recombination seems to play such different roles in the two systems?

3. The Adaptive Code Book

Molecular biology's Central Dogma connects genes to phenotype by stating that DNA is transcribed into RNA, which is then translated into protein.

The terms transcription and translation are quite literal: RNA is a chemical sister language to DNA. Both are polymers formed from an alphabet of four chemical letters (nucleotides), and transcription is nothing more than a process of complementing DNA, letter by letter, into RNA. It is the next step, translation, that profoundly influences biological evolution.

Proteins are also linear polymers of chemical letters, but they are drawn from a qualitatively different alphabet (amino acids) comprising 20 elements. Clearly no one-to-one mapping could produce a genetic code for translating nucleotides unambiguously into amino acids, and by 1966 it was known that the combinatorial set of possible nucleotide triplets forms a dictionary of 'codons' that each translate into a single amino acid meaning (Frisch, 1966). The initial surprise for evolutionary theory was to discover that something as fundamental as the code-book for life would exhibit a high degree of redundancy (an alphabet of 4 RNA letters permits $4 \times 4 \times 4 = 64$ possible codons that map to one of only 20 amino acid meanings, Crick, 1966). Early interpretation fuelled arguments for Non-Darwinian evolution (King and Jukes, 1969): genetic variations that make no difference to the protein they encode must be invisible to selection and therefore governed solely by drift (Kimura, 1968). More recently, both computing (e.g. Banzhaf, 1994; Kargupta, 2001) and biological (e.g. Kauffman, 1993; Huynen *et al.*, 1996) evolutionary theory have started to place this coding neutrality in the bigger picture of the adaptive heuristic. Essentially, findings appear to mirror Wright's early arguments on the importance of genetic drift: redundancy in the code adds selectively neutral dimensions to the fitness landscape that renders adaptive algorithms more effective by increasing the connectedness of local optima.

At present, an analogous reinterpretation is underway for a different adaptive feature of the genetic code: the observation that biochemically similar amino acids are assigned to codons that differ by only a single nucleotide. Early speculations (e.g. Woese, 1965) that natural selection organized the genetic code so as to minimize the phenotypic impact of

mutations have gained considerable evidential support as computer simulation enables exploration of theoretical codes that nature passed over (comprehensively reviewed in Freeland *et al.*, 2003). However, it seems likely that once again this phenomenon has more subtle effects in the broader context of the adaptive heuristic. We have suggested that an 'error minimizing code' may in fact maximize the probability that a random

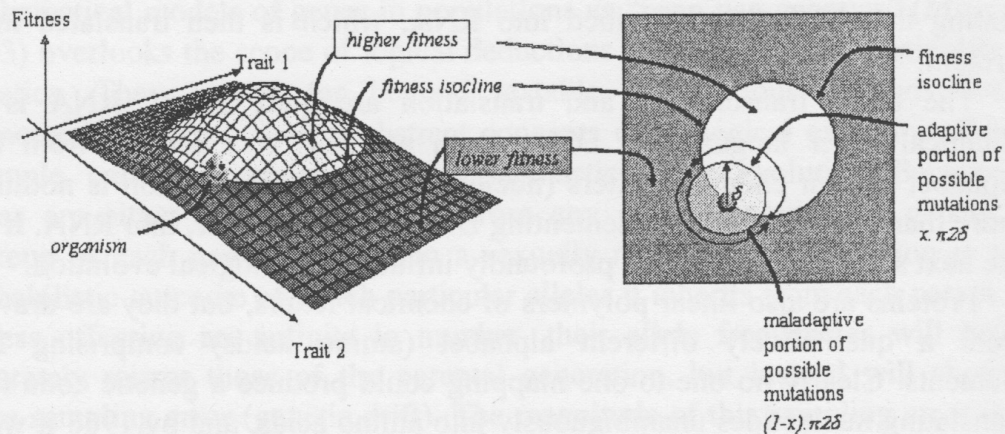


Figure 19.1. The fitness landscape for an organism of 2 phenotypic traits: (a) for any organism, we may define an isocline that connects all trait combinations of equal fitness; (b) (the fitness landscape from above): a random mutation of magnitude δ that has tradeoff effects on both traits defines a circle of radius δ around the organism. The probability that this mutation will improve fitness (i.e. that the organism will move within the white area) is inversely proportional to its magnitude (δ) (from Freeland *et al.*, 2002)

mutation produces an increase in fitness (Freeland, 2002) According to Fisher's (1930) Geometric Theory of gradualism (Fig. 19.1). Preliminary tests for this phenomenon reveal an even simpler influence: the error minimizing code smoothes the fitness landscape where a random genetic code would render it rugged (unpublished data). By clustering biochemically similar amino acids within mutational reach of one another it ensures that any selection towards a specific amino acid property (e.g. hydrophobicity) will be towards an interconnected region of the fitness landscape rather than to an isolated local optimum.

4. The Genotype/Phenotype Dichotomy

Implicit to the concept of an adaptive genetic code is a deeper question that remains largely unanswered by biology: why does all known life use two qualitatively different polymers, nucleic acids and proteins, with the

associated need for translation? Current theories for the origin of this dichotomy focus on the discovery that RNA can act both as a genetic storage medium, and as a catalytic molecule. Within the most highly conserved core of metabolism, all known organisms are found to use RNA molecules in roles we normally attribute to proteins (White, 1976). Moreover, *in vitro* experiments have begun to demonstrate a unexpectedly broad potential for RNA-enzyme ('ribozyme') catalysis (reviewed in Doudna *et al.* 2002). Put together, these observations have produced the RNA World hypothesis (Gesteland *et al.*, 1999) in which primordial, self-replicating RNA molecules eventually evolved metabolic protein clothing.

However, this answer to *how* the dichotomy evolved has largely eclipsed the question of *why* RNA evolved a qualitatively different representation for phenotype. A typical biological answer would be that the larger alphabet size of amino acids unleashed a greater catalytic diversity for the replicators (e.g. Szathmary and Maynard Smith, 1995), with an associated increase in metabolic sophistication that optimized self-replication. Interestingly, we know that nucleic acids are not limited to the 4 chemical letters we see today: natural metabolically active RNA's utilize a vast repertoire of post-transcriptional modifications (reviewed in Grosjean and Benne, 1998) and synthetic chemistry has demonstrated that multiple additional nucleotide letters can be added to the genetic alphabet even with today's cellular machinery (Picirilli *et al.* 1990). Furthermore, an increasing body of indirect evidence suggests that the protein alphabet itself underwent exactly the sort of evolutionary expansion early in life's history (Ronneberg *et al.* 2001).

Given the ubiquity of nucleic acid genotype and protein phenotype within life, biology is hard-pressed to assess the significance of evolving this 'representational language'. The choice of phrase is deliberate: clearly the EC community is far ahead of biology in formalizing the concept of representational language, and exploring what it means. Biology will gain when evolutionary programmers place our system within their findings, illustrating the potential for biological inspiration *from* EC.

5. CONCLUDING REMARKS

Three fundamental features of the biological genetic algorithm are: particulate genes, the adaptive genetic code and the dichotomy of genotype and phenotype. For particulate genes, I ask whether EC can gain from biology by considering the detailed dynamics by which a novel, advantageous allele invades a wild-type population. This naïve question serves to illustrate an important, general criterion for curtailing biology envy: does the EC representation of a biological phenomenon achieve true

functional equivalency? It is unclear whether this question gets routinely and rigorously assessed when novel biological phenomena are brought into EC.

On a different front, research into adaptive features of the genetic code illustrates how EC and biological evolutionary research can and do contribute to a common understanding of the general evolutionary dynamic. Neutrality, redundancy and fitness-landscape smoothing are abstractions that apply unambiguously to both communities.

For the dichotomy of genotype and phenotype, I argue that biology has so far failed to even recognize a question that the EC community is actively addressing: how and why does representational language influence the evolutionary dynamic? A luxury of EC is to bend fundamental rules of evolution beyond anything biologically plausible, and thus to answer questions where biologists have assumptions. Wherever conceptual equivalency can be demonstrated, EC will penetrate biological evolutionary theory not by borrowing from it, but by escaping biological assumptions long enough to tell us why they exist.

Acknowledgements

Funding for this work came from Templeton Foundation grant 938-COS27, and from UMBC. I am indebted to Rick Riolo and Bill Worzel for inviting me to this workshop, to Lee Spector, Wolfgang Banzhaf, Duncan MacLean and all the workshop attendees for their warm introduction to the EC community and to Anjeanette Agro for persistent inspiration in a complex world.

References

- Banzhaf, W. (1994). Genotype-phenotype mapping and neutral variations ---a case study in Genetic Programming. In *Parallel Problem Solving From Nature III*: 322-332. Davidor, Y., Schwefel, H.-P. and Männer, R., Springer: Berlin.
- Banzhaf, W. (*these proceedings*). "Artificial Regulatory Networks and Genetic Programming."
- Crick, F. H. C. (1966). Codon-anticodon Pairing: The Wobble Hypothesis. *Journal of Molecular Biology* 19: 548-555.
- Darwin, C. R. (1859). *On the Origin of Species by means of Natural Selection*. John Murray.
- Dawkins, R. (1989). *The Selfish Gene*. 2nd edition, Oxford University Press; 84-86
- Doudna, J. A. and Cech, T. R. (2002). The Chemical Repertoire of Natural Ribozymes. *Nature* 418: 222-228
- Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press.
- Freeland, S. J. (2002). The Genetic Code: An Adaptation for Adapting? *Journal of Genetic Programming and Evolvable Machines* 3(2): 113-127

- Freeland, S. J., Wu, T., and Keulmann, N. (2003). "The Case for an Error Minimizing Genetic Code." In *Origins of Life and Evolution of the Biosphere*. (in press)
- Frisch, L. (ed.) (1966). The Genetic Code. In *Proceedings of Cold Spring Harbor Symposia on Quantitative Biology*, pp. 1-747.
- Gesteland, R. F., Cech, T., and Atkins, J. F. (eds.) (1999). *The RNA World*. 2nd edition, Cold Spring Harbor Laboratory monograph series 37.
- Grosjean, H. and Benne, R. (1998). *Modification and Editing of RNA*. American Society for Microbiology.
- Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems*. The MIT Press.
- Huynen, M., Stadler, P., and Fontana, W. (1996). Smoothness Within Ruggedness: The Role of Neutrality in Adaptation. *Proc. Natl. Acad. Sci.* 93: 397-401.
- Jenkin, F. (1867). (Review of) The Origin of Species. *North British Review* 46: 277-318.
- Kargupta, H. A. (2001). Striking Property of Genetic Code-like Transformations. *Complex Systems* 13: 1-32.
- Kauffman, S. A. (1993). *The Origins of Order - Self-Organization and Selection in Evolution*. New York: Oxford University Press.
- Kimura, M. (1968). Genetic Variability Maintained in a Finite Population due to Mutational Production of Neutral and Nearly Neutral Isoalleles. *Genetical Research* 11: 247-269.
- King, J. L. and Jukes, T. H. (1969). Non-Darwinian Evolution. *Science* 164: 788-798.
- Maynard, Smith J. and Szathmáry, E. (1995). *The Major Transitions in Evolution*. W.H. Freeman and Co.
- Mayr, E. (1963). *Animal Species and Evolution*. Harvard University Press.
- Mayr, E. and Provine, W. B. (1998). *The Evolutionary Synthesis: Perspectives on the Unification of Biology*. Harvard University Press.
- Mendel, G. (1866). "Versuche über Pflanzen-Hybriden" Verhandlungen des naturforschenden Vereines, Abhandlungen Brünn 4:3-47. (English translation: available at <http://www.mendelweb.org/MWGerText.html>)
- O'Reilly, U-M. (1999). Foundations of Genetic Programming: Effective Population Size, Linking and Mixing.
- Ostrowski D. and Reynolds R.G. (these proceedings). "Using Software Engineering Knowledge to Drive Genetic Programming Design Using Cultural Algorithms."
- Picirilli, J. A., Krauch, T., Moroney, S. E., and Benner, S. (1990). Enzymatic Incorporation of a New Base Pair into DNA and RNA Extends the Genetic Alphabet. *Nature* 343: 33-37.
- Ronneberg, T. A., Landweber, L. F., and Freeland, S. J. (2001). Testing a Biosynthetic Theory of Genetic Code Evolution: Fact or Artefact? *Proc. Natl. Acad. Sci.* 97(25): 13690-13695.
- Spector, L. (these proceedings). "An Essay Concerning Human Understanding of Genetic Programming."
- White, H. B. (1976). Coenzymes as Fossils of an Earlier Metabolic State. *Journal of Molecular Evolution* 7.
- Woese, C. R. (1965). On the Evolution of the Genetic Code. *Proc. Natl. Acad. Sci.* 54: 1546-1552.
- Wright, S. (1932). The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution. In *Proceedings of the 6th International Congress of Genetics* 1, 356-366.
- Yates, F. E. (1978). Is Physics Envy the Curse of Biology? *American Journal of Physiology* 234(1).