Genet Program Evolvable Mach (2017) 18:373–378
DOI 10.1007/s10710-017-9292-1

COMMENTARY

Evolutionary algorithms and synthetic biology for directed evolution: commentary on "on the mapping of genotype to phenotype in evolutionary algorithms" by Peter A. Whigham, Grant Dick, and James Maclaurin

Douglas B. Kell[1,2,3]

Published online: 29 March 2017

Abstract I rehearse two issues around the commentary of Whigham and colleagues. (1) There really are many more reasons than those given as to why natural evolution cannot reasonably find or select the 'optimal' individual. (2) A series of experimental molecular biology programmes, known generically as directed evolution, can use operators and selection schemes that natural evolution cannot. When developed further using the methods of synthetic biology, there are no operators or schemes for in silico evolution that cannot be applied precisely to directed evolution. The issues raised apply only to natural evolution but not to directed evolution.

Keywords Directed evolution · Synthetic biology · Navigating search spaces · Intelligent operators

1 Introduction

The basic thesis of Whigham et al. [1] is that 'molecular biology', by which they actually mean 'mechanisms of natural evolution', is a relatively poor guide to the kinds of in silico optimisation problems that are typically the targets of evolutionary algorithms. While I have no disagreement with that high-level sentiment (indeed I

This comment refers to the article available at doi:10.1007/s10710-017-9288-x.

✉ Douglas B. Kell
dbk@manchester.ac.uk

[1] School of Chemistry, The University of Manchester, 131, Princess St, Manchester, Lancs M1 7DN, UK

[2] The Manchester Institute of Biotechnology, The University of Manchester, 131, Princess St, Manchester, Lancs M1 7DN, UK

[3] Centre for Synthetic Biology of Fine and Speciality Chemicals, The University of Manchester, 131, Princess St, Manchester, Lancs M1 7DN, UK

have argued that biological genotype-phenotype mapping is entirely analogous to GP [2]!), they do omit some of the chief arguments as to why this is so. They also entirely disregard a very large literature using 'molecular biology' that is not 'natural' evolution but is aimed at producing (optimising) biological products via a set of processes typically referred to as 'directed evolution' (DE). These are therefore the points that I cover. A more extended overview of modern directed evolution may be found in [3].

2 Why natural evolution is not necessarily 'optimal' in the sense of assuming that only the fittest survive

Whigham et al. [1] listed a series of reasons outlined by Dawkins [4] as to why this is so. Natural evolution is based on the processes of (i) diversity creation within a population (some or all of whom may interact and breed), (ii) 'evaluation' of fitness(es) and (iii) selection. Each of these may contribute to the fact that, even over an arbitrary time, populations do not converge such that they are populated only by individuals of the highest possible fitness for that (nominally unchanging) environment or objective function. Consequently there are many more reasons, at different levels of detail, as to why this is the case. Some include:

At level (i) (populations and diversity):

- Populations are normally not panmictic (where all can breed with all) so global optimality is then impossible.
- The size of the search space is so much greater than any population size that no test of optimality (in the NP complete sense) is possible; even 30mers of unmodified DNA have 10^{18} possible genotypes, which as 5 μm spots each would cover 29 km^2 [5].
- When the population size is too small, some individuals who may yet contribute to optimality are simply squeezed out (especially in mutation-only organisms); this is known as Muller's ratchet (e.g. [6–8]).
- Weak mutation and strong selection (see [9–14]) is common in many ecosystems, and traps individuals and populations in local minima (or maxima) from which neutral evolution cannot help them escape.
- Given that 3/64 codons are stop codons [15], the required mutation or recombination rates to explore the entire fitness landscape in a sensible time are incompatible with the sustenance of the population, given that most mutations are deleterious [16].

At level (ii) (fitnesses):

- The linkage between genotype, fitness and phenotype may often be very weak [17], for all kinds of reasons including epistasis [3] and the nonlinear properties of biochemical networks [18, 19].

- Non-genetic factors may intervene, e.g. while reproductive fitness is normally selected for, longevity may be preferred for social reasons (grandparents help with childcare, etc.); there is not just one fitness (hence optimality).

At level (iii) (selection)

- At the population level, while selection is based on the offspring more than the parents, simple stochastic events may blur the link between selection and fitness.
- For all kinds of reasons, selection is likely to (and does [20]) favour a differentiated population.

3 Directed evolution (DE)

Directed evolution (a subset of closed loop optimisation [21]) refers to a set of biochemical procedures that are akin to natural evolution (in that they involve diversity creation and fitness-based selection). However, here the selection step is performed not by the organisms within their environment (as in natural evolution) but by an experimenter who has a specific goal, and thus selects individuals for further improvement based on those in a given generation. This is a glorified form of breeding (as in plant [22] and animal [23] breeding), but is typically done for improving individual proteins (e.g. as biocatalysts [24, 25]), where it is indeed sometimes known [17, 26, 27] as 'molecular breeding'.

Classical methods of diversity generation used random in vitro methods of mutagenesis (e.g. [24, 28, 29]), or recombination (known here as 'DNA shuffling' [30, 31]), and selection was typically done on the basis of screening via a target biochemical assay (rather than true growth rate selection, e.g. as in [32, 33]).

Nowadays, the trend is to use the methods of synthetic biology, in which the experimenter has complete (or statistical) control over precisely which amino acid is expressed at which residue (e.g. [34]), literally by synthesising the encoding DNA, coupled to sophisticated in silico sequence optimisation methods (e.g. [35]). This synbio-based DE is formally equivalent to classical DE, but is typically cast in terms of a Design-Build-Test-Learn cycle [36–38]. Knowledge of gene sequences can be exploited directly [5, 39–41], and in some cases landscapes may be searched exhaustively [42]. What we learn from this (see also [43]) is that protein fitness landscapes are rugged but not pathologically so (e.g. [3, 44]), and that the evolutionary landscape metaphor can provide insights into both innovation [45] and science more generally [46].

This brings me nicely to my last point, elaborated in detail in Table 1 of [3], where we point out that many aspects of 'classical' and synthetic biology-based directed evolution differ massively from the same ones when they are compared to natural evolution, e.g. in terms of selection (pressure), mutation rates and their randomness, evolutionary 'memory', and the degree of epistasis allowed or encouraged. Even where these are solely quantitative differences, the effective emergent properties can make them appear to be qualitatively different. In

particular, by controlling explicitly (at least statistically) the protein sequences in the population, we can combine the exploration and exploitation of fitness landscapes in DE by keeping low-fitness members in the population in a manner (arbitrarily long) that natural evolution would not be able to at any reasonable selection pressure. Notwithstanding, all these methods involve 'molecular biology'.

In conclusion, if 'molecular biology' methods are applied to synbio-based directed evolution, any operator can be used, since any subsequent genetic sequence can be based arbitrarily on what has been seen at any time before, because it can be made de novo. These operators include very high mutation rates [16], three- (or more-)way matings, matings based on information gain and the Pareto front [47], complex crossover and mutation schemes such as 'uniform' crossover [48], explicit coevolutions based on sequence alignments and 3D structures (e.g. [49]), and so on. This, indeed, is precisely why the new methods of synthetic biology are so very powerful [3, 50]. Consequently, while recognising that natural evolution may be a rather poor guide to the utility of evolutionary algorithms in optimisation [1], we would stress the field of directed evolution, especially that based on synthetic biology, provides a highly productive domain for the cross-fertilisation of evolutionary algorithms and molecular biology.

Acknowledgements I thank the Biotechnology and Biological Sciences Research Council for financial support (Grant BB/M017702/1). This is a contribution from the Manchester Centre for Synthetic Biology of Fine and Speciality Chemicals (SYNBIOCHEM).

References

1. P.A. Whigham, G. Dick, J. Maclaurin, On the mapping of genotype to phenotype in evolutionary algorithms. Genet. Progr. Evol. Mach. (2017)
2. D.B. Kell, Genotype: phenotype mapping: genes as computer programs. Trends Genet. **18**, 555–559 (2002)
3. A. Currin, N. Swainston, P.J. Day, D.B. Kell, Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. Chem. Soc. Rev. **44**, 1172–1239 (2015)
4. R. Dawkins, *The Extended Phenotype* (OUP, Oxford, 1982)
5. C.G. Knight, M. Platt, W. Rowe, D.C. Wedge, F. Khan, P. Day, A. McShea, J. Knowles, D.B. Kell, Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape. Nucl. Acids Res. **37**, e6 (2009)
6. H.J. Muller, The relation of recombination to mutational advance. Mutat. Res. **106**, 2–9 (1964)
7. A.M. Wardlaw, A.F. Agrawal, Temporal variation in selection accelerates mutational decay by Muller's ratchet. Genetics **191**, 907–916 (2012)
8. J.J. Metzger, S. Eule, Distribution of the fittest individuals and the rate of Muller's ratchet in a model with overlapping generations. PLoS Comput. Biol. **9**, e1003303 (2013)
9. H.A. Orr, The genetic theory of adaptation: a brief history. Nat. Rev. Genet. **6**, 119–127 (2005)
10. H.A. Orr, The population genetics of adaptation on correlated fitness landscapes: the block model. Evolution **60**, 1113–1124 (2006)

11. H.A. Orr, The distribution of fitness effects among beneficial mutations in Fisher's geometric model of adaptation. J. Theor. Biol. **238**, 279–285 (2006)
12. H.A. Orr, Fitness and its role in evolutionary genetics. Nat. Rev. Genet. **10**, 531–539 (2009)
13. R.L. Unckless, H.A. Orr, The population genetics of adaptation: multiple substitutions on a smooth fitness landscape. Genetics **183**, 1079–1086 (2009)
14. I.G. Szendro, J. Franke, J.A.G.M. de Visser, J. Krug, Predictability of evolution depends non-monotonically on population size. Proc. Natl. Acad. Sci. U S A **110**, 571–576 (2013)
15. L. Pritchard, D.W. Corne, D.B. Kell, J.J. Rowland, M.K. Winson, A general model of error-prone PCR. J. Theor. Biol. **234**, 497–509 (2004)
16. M.J. Oates, D.W. Corne, D.B. Kell, The bimodal feature at large population sizes and high selection pressure: implications for directed evolution, in *Recent Advances in Simulated Evolution and Learning*, ed. by K.C. Tan, M.H. Lim, X. Yao, L. Wang (World Scientific, Singapore, 2003), pp. 215–240
17. S. O'Hagan, J. Knowles, D.B. Kell, Exploiting genomic knowledge in optimising molecular breeding programmes: algorithms from evolutionary computing. PLoS ONE **7**, e48862 (2012)
18. H. Kacser, J.A. Burns, The molecular basis of dominance. Genetics **97**, 639–666 (1981)
19. D.B. Kell, H.V. Westerhoff, Metabolic control theory: its role in microbiology and biotechnology. FEMS Microbiol. Rev. **39**, 305–320 (1986)
20. D.B. Kell, M. Potgieter, E. Pretorius, Individuality, phenotypic differentiation, dormancy and 'persistence' in culturable bacterial systems: commonalities shared by environmental, laboratory, and clinical microbiology. F1000 Res. **4**, 179 (2015)
21. J. Knowles, Closed-loop evolutionary multiobjective optimization. IEEE Comput. Intell. Mag. **4**, 77–91 (2009)
22. W.G. Hill, A century of corn selection. Science **307**, 683–684 (2005)
23. J.L. Williams, The use of marker-assisted selection in animal breeding and biotechnology. Rev. Sci. Tech. **24**, 379–391 (2005)
24. N.J. Turner, Directed evolution drives the next generation of biocatalysts. Nat. Chem. Biol. **5**, 567–573 (2009)
25. U.T. Bornscheuer, G.W. Huisman, R.J. Kazlauskas, S. Lutz, J.C. Moore, K. Robins, Engineering the third wave of biocatalysis. Nature **485**, 185–194 (2012)
26. J. Minshull, W.P.C. Stemmer, Protein evolution by molecular breeding. Curr. Opin. Chem. Biol. **3**, 284–290 (1999)
27. W. Zha, S.B. Rubin-Pitel, H. Zhao, Exploiting genetic diversity by directed evolution: molecular breeding of type III polyketide synthases improves productivity. Mol. BioSyst. **4**, 246–248 (2008)
28. F.H. Arnold, G. Georgiou, *Directed Evolution Library Creation: Methods and Protocols* (Springer, Berlin, 1996)
29. J.N. Copp, P. Hanson-Manful, D.F. Ackerley, W.M. Patrick, Error-prone PCR and effective generation of gene variant libraries for directed evolution. Methods Mol. Biol. **1179**, 3–22 (2014)
30. W.P.C. Stemmer, Rapid evolution of a protein in vivo by DNA shuffling. Nature **370**, 389–391 (1994)
31. W.P.C. Stemmer, DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. Proc. Natl. Acad. Sci. **91**, 10747–10751 (1994)
32. S.W. Brown, S.G. Oliver, Isolation of ethanol-tolerant mutants of yeast by continuous selection. Eur. J. Appl. Microbiol. Biotechnol. **16**, 119–122 (1982)
33. H.M. Davey, C.L. Davey, A.M. Woodward, A.N. Edmonds, A.W. Lee, D.B. Kell, Oscillatory, stochastic and chaotic growth rate fluctuations in permittistatically-controlled yeast cultures. Biosystems **39**, 43–61 (1996)
34. A. Currin, N. Swainston, P.J. Day, D.B. Kell, SpeedyGenes: a novel approach for the efficient production of error-corrected, synthetic gene libraries. Protein Eng. Des. Sel. **27**, 273–280 (2014)
35. N. Swainston, A. Currin, P.J. Day, D.B. Kell, GeneGenie: optimised oligomer design for directed evolution. Nucl. Acids Res. **12**, W395–W400 (2014)
36. P. Carbonell, A. Currin, A.J. Jervis, N.J.W. Rattray, N. Swainston, C. Yan, E. Takano, R. Breitling, Bioinformatics for the synthetic biology of natural products: integrating across the Design-Build-Test cycle. Nat. Prod. Rep. **33**, 925–932 (2016)
37. P. Carbonell, A. Currin, M. Dunstan, D. Fellows, A. Jervis, N.J.W. Rattray, C.J. Robinson, N. Swainston, M. Vinaixa, A. Williams, C. Yan, P. Barran, R. Breitling, G.G. Chen, J.L. Faulon, C. Goble, R. Goodacre, D.B. Kell, R.L. Feuvre, J. Micklefield, N.S. Scrutton, P. Shapira, E. Takano,

N.J. Turner, SYNBIOCHEM—a SynBio foundry for the biosynthesis and sustainable production of fine and speciality chemicals. Biochem. Soc. Trans. **44**, 675–677 (2016)
38. J. Nielsen, J.D. Keasling, Engineering cellular metabolism. Cell **164**, 1185–1197 (2016)
39. C.L. Araya, D.M. Fowler, Deep mutational scanning: assessing protein function on a massive scale. Trends Biotechnol. **29**, 435–442 (2011)
40. D.M. Fowler, S. Fields, Deep mutational scanning: a new style of protein science. Nat. Methods **11**, 801–807 (2014)
41. H. Shin, B.K. Cho, Rational protein engineering guided by deep mutational scanning. Int. J. Mol. Sci. **16**, 23094–23110 (2015)
42. W. Rowe, M. Platt, D. Wedge, P.J. Day, D.B. Kell, J. Knowles, Analysis of a complete DNA-protein affinity landscape. J. R. Soc. Interface **7**, 397–408 (2010)
43. S.A. Kauffman, W.G. Macready, Search strategies for applied molecular evolution. J. Theor. Biol. **173**, 427–440 (1995)
44. P.A. Romero, F.H. Arnold, Exploring protein fitness landscapes by directed evolution. Nat. Rev. Mol. Cell Biol. **10**, 866–876 (2009)
45. D.B. Kell, E. Lurie-Luke, The virtue of innovation: innovation through the lenses of biological evolution. J. R. Soc. Interface **12**, 20141183 (2015)
46. D.B. Kell, Scientific discovery as a combinatorial optimisation problem: how best to navigate the landscape of possible experiments? BioEssays **34**, 236–244 (2012)
47. J. Knowles, ParEGO: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. IEEE Trans. Evol. Comput. **10**, 50–66 (2006)
48. G. Syswerda (1989) Uniform crossover in genetic algorithms. in *Proceedings 3rd International Conference on Genetic Algorithms*, ed. by J. Schaffer (Morgan Kaufmann, 1989), pp. 2–9
49. D.T. Jones, T. Singh, T. Kosciolek, S. Tetchner, MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics **31**, 999–1006 (2015)
50. A. Currin, K. korovin, M. Ababi, K. Roper, D.B. Kell, P.J. Day, R.D. King, Computing exponentially faster: implementing a nondeterministic universal turing machine using DNA. J. R. Soc. Interface (2017). doi:10.1098/rsif.2016.0990