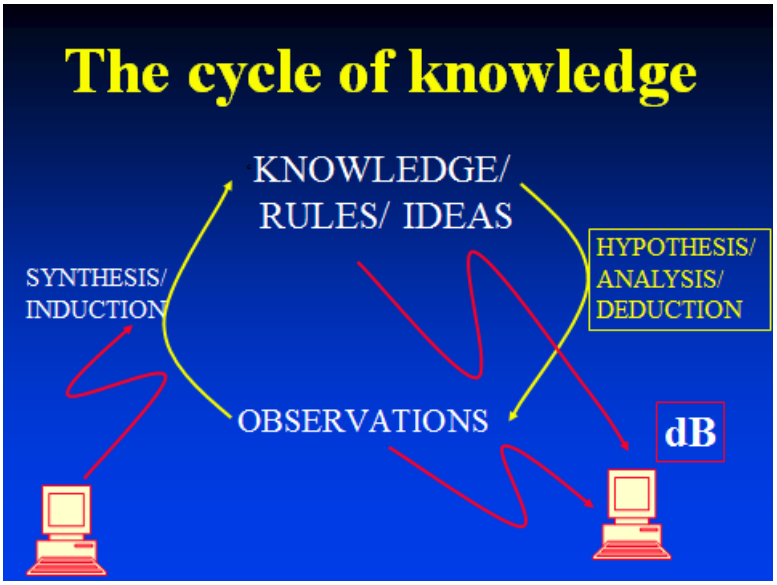
Elements of computational intelligence and their applications in bio(techno)logy

Douglas B. Kell **@dbkell**
School of Chemistry and Manchester Institute of Biotechnology, The University of Manchester, MANCHESTER M1 7DN, UK. **dbk@manchester.ac.uk** **http://dbkgroup.org/publications**

It is reasonable to recognise that most scientific problems are best considered in terms of an (iterative) search between the worlds of ideas (hypotheses, abstract thoughts) and of observations ('data') [1] (and Figure at right). We recognise that in general terms this search occurs on a rugged (but not pathological) landscape, in which 'good' solutions (in terms of either the 'best' experiment to do next, hence of improving one's understanding) are a tiny fraction of the possible ones to test [2; 3]. However, the state of the art of machine intelligence as applied to consumer goods (mostly based on 'deep' neural methods of various kinds (e.g. [4-6])) far exceeds that typically being applied to scientific problems; this provides many opportunities.



Following this, the background (and foreground) knowledge of most scientific problems can be structured as tables of examples and properties (objects and variables), and decomposed, depending on the availability of the objective function(s), into supervised (learning/optimisation [7]), unsupervised (clustering/ similarity [8]), or semi-supervised problems. Supervised methods have a trade-off between potency and intelligibility; our experience is that evolutionary computing (GA and GP) [9-11], random forests [12], and both shallow [13; 14] and deep [6; 15-17] neural systems cover the space adequately. Given the usual availability of many more 'unsupervised' (domain) examples than those with output data, we anticipate considerable growth in the use of variants of semi-supervised learning (much as has been beneficial in object recognition [18]). Feature extraction lies at the heart of computational intelligence [19]. Thus, the biggest issue for computational intelligence remains the representation of ideas and physical objects; thus the various means of representing chemical structures as 'fingerprints' can give very different results, even in simple clustering analyses (e.g. [20-23]).

We recognise that a useful trend is towards the entire ('closed loop') automation of scientific discovery [9; 24]. Synbio-based directed protein evolution [25] coupled to deep learning is a realistic short-term goal, *en route* to a more general intelligence that might help human beings to solve most biological (and other) problems much more efficiently.

[1] Kell, D. B. & Oliver, S. G. (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* 26, 99-105.
[2] Kell, D. B. (2012). Scientific discovery as a combinatorial optimisation problem: how best to navigate the landscape of possible experiments? *Bioessays* 34, 236-244.
[3] Kell, D. B. & Lurie-Luke, E. (2015). The virtue of innovation: innovation through the lenses of biological evolution. *J R Soc Interface* 12, 20141183.

[4] Fogel, D. B. (2002). *Blondie42: playing at the edge of AI*. Morgan Kaufmann, San Francisco.
[5] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484-9.
[6] Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*. MIT Press, Boston.
[7] Handl, J., Kell, D. B. & Knowles, J. (2007). Multiobjective optimization in bioinformatics and computational biology. *IEEE Trans Comput Biol Bioinformatics* 4, 279-292.
[8] Handl, J., Knowles, J. & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21, 3201-3212.
[9] O'Hagan, S., Dunn, W. B., Brown, M., Knowles, J. D. & Kell, D. B. (2005). Closed-loop, multiobjective optimisation of analytical instrumentation: gas-chromatography-time-of-flight mass spectrometry of the metabolomes of human serum and of yeast fermentations. *Anal Chem* 77, 290-303.
[10] Kell, D. B. (2002). Genotype:phenotype mapping: genes as computer programs. *Trends Genet.* 18, 555-559.
[11] O'Hagan, S. & Kell, D. B. (2015). Software review: The KNIME workflow environment and its applications in Genetic Programming and machine learning. *Genetic Progr Evol Mach* 16, 387-391.
[12] Knight, C. G., Platt, M., Rowe, W., Wedge, D. C., Khan, F., Day, P., McShea, A., Knowles, J. & Kell, D. B. (2009). Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape. *Nucleic Acids Res* 37, e6.
[13] Goodacre, R., Trew, S., Wrigley-Jones, C., Saunders, G., Neal, M. J., Porter, N. & Kell, D. B. (1995). Rapid and quantitative analysis of metabolites in fermentor broths using pyrolysis mass spectrometry with supervised learning: application to the screening of *Penicillium chryosgenum* fermentations for the overproduction of penicillins. *Anal Chim Acta* 313, 25-43.
[14] Goodacre, R., Timmins, É. M., Burton, R., Kaderbhai, N., Woodward, A. M., Kell, D. B. & Rooney, P. J. (1998). Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. *Microbiology UK* 144, 1157-1170.
[15] LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature* 521, 436-44.
[16] Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw* 61, 85-117.
[17] Gawehn, E., Hiss, J. A. & Schneider, G. (2016). Deep learning in drug discovery. *Mol Inform* 35, 3-14.
[18] Hinton, G. E. & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504-7.
[19] Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction, 2nd edition*. Springer-Verlag, Berlin.
[20] O'Hagan, S., Swainston, N., Handl, J. & Kell, D. B. (2015). A 'rule of 0.5′ for the metabolite-likeness of approved pharmaceutical drugs. *Metabolomics* 11, 323-339.
[21] O'Hagan, S. & Kell, D. B. (2015). Understanding the foundations of the structural similarities between marketed drugs and endogenous human metabolites. *Front Pharmacol* 6, 105.
[22] O'Hagan, S. & Kell, D. B. (2016). MetMaxStruct: a Tversky-similarity-based strategy for analysing the (sub)structural similarities of drugs and endogenous metabolites. *Front Pharmacol* 7, 266.
[23] O'Hagan, S. & Kell, D. B. (2017). Consensus rank orderings of molecular fingerprints illustrate the 'most genuine' similarities between marketed drugs and small endogenous human metabolites, but highlight exogenous natural products as the most important 'natural' drug transporter substrates. bioRxiv version. *bioRxiv*, 110437.
[24] King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., Kell, D. B. & Oliver, S. G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* 427, 247-252.
[25] Currin, A., Swainston, N., Day, P. J. & Kell, D. B. (2015). Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem Soc Rev* 44, 1172-1239.