Published in Molecular Biology Reports 29, 237-241 (2002)

Metabolomics and machine learning: explanatory analysis of complex metabolome data using genetic programming to produce simple, robust rules.

Douglas B. Kell

Institute of Biological Sciences, Cledwyn Building, University of Wales, Aberystwyth SY23 3DD, UK. dbk@aber.ac.uk http://qbab.aber.ac.uk http://www.abergc.com.

From Sept 1st, 2002: Dept Chemistry, UMIST, PO Box 88, Sackville St, MANCHESTER M60 1QD.

Introduction

There is a clear trend in post-genomic studies [1-5] to understand gene function [6, 7], pharmaceutical mode of action [8], cytotoxicity [9, 10] and the like by expression profiling at the level of the transcriptome [11-13], the proteome [14-17] and the metabolome [1, 18-28]. Our interest is focused on the latter [7, 29-37].

The result of these expression profiling studies is likely to be values for concentration of hundreds or thousands of molecules. Finding useful rules to 'explain' e.g. the differences between healthy and diseased individual is a combinatorial optimization problem [38, 39] of high dimensionality. Conventional analyses merely look at the differences between individuals, but the biggest differences may not be relevant to the higher-order trait of interest; this is of course well-known in MCA where large changes in enzyme concentration may cause negligible changes in flux through pathways of which they are a part [40-42]. We consider that finding the most interesting and significant differences from such data is in fact best cast as a (more or less standard) machine learning problem [7, 29, 34, 36, 37].

Of the numerous methods available (e.g. [43-47]) we have found that a variant of genetic programming [48-54], which we call genomic computing [29, 34, 35, 55], allows one to evolve simple rules that are highly discriminatory and have great explanatory power, i.e. not only do the rules provide the correct answers but the answers are intelligible and provide the nonlinear mapping directly from the important 'input' variables to the trait of interest.

Results and discussion

We shall give three metabolomic examples in which highly complex datasets, which could not be deconvoluted successfully in their original form, succumbed to genomic computing such that we can simply describe which segments of metabolism best explain differences between organisms of different types and thus are most appropriate for detailed study. The examples are:

1. A study of plant defence metabolites [56] which led to the discovery of two important new candidates [34];
2. A study aimed at finding the most important metabolic differences between cultivars of olive [29, 57];
3. A study aimed at establishing targets for therapeutic intervention following the genetic induction of muscular dystrophy (for data see [58]).

We discuss case number 1 in detail. This was a 'transgene discovery' problem in which we measured a series of metabolites via HPLC and used these as the inputs to a Genetic Program designed to find a rule which would tell from the metabolome data whether the transgene of interest was present or absent. The experiment was also aimed at investigating the biosynthesis and function of salicylic acid in plant defense by the expression of a salicylate hydroxylase enzyme to block accumulation [56].

Salicylic acid has been known for many years to play a key role in defence mechanisms in many plants and is associated specifically with the hypersensitive response (HR) and the phenomenon of Systemic Acquired Resistance (SAR; [59-62]. A bacterial gene encoding the enzyme salicylate hydroxylase (SH-L) expressed from the CaMV 35S promoter has provided a useful tool to block SA accumulation in transgenic tobacco [56, 60, 61]. Six-week old transgenic tobacco plants (35S-SH-L) and control plants (Samsum NN) were inoculated with Tobacco Mosaic Virus (TMV) at a temperature (32ºC) non-permissive for the hypersensitive response [61, 63]. Under these conditions the TMV can replicate and spread without inducing lesion formation. Following a shift to a permissive temperature (24ºC) the HR is induced synchronously, with cell death visible after 8 hours. Leaf tissue from TMV-inoculated, temperature-shifted plants was sampled at different time points (0-24h), flash frozen in liquid N_2, extracted in 90% methanol, dried, partitioned with dichloromethane and then analysed by HPLC using standard procedures [56]. A total of 48 peaks (V1 – V48) from the HPLC traces were digitized and integrated using standard software provided with the instrument, and a total of 36 samples studied.

The metabolite peak values were used as inputs to the Genomic Computing software Gmax-bio (Aber Genomic Computing, Unit 8, Science Park, Aberystwyth SY23 3AH, UK), with the presence or absence of SH-L in the genotype being encoded 1 or 0.

One of many rules which evolved could be written as follows:
SCORE = Sqrt((V37/V24)) + Sqrt(V30/(V24+V42)); Probability that plant contains the transgene = 1 / (1 + Exp(-(-8.046777 + SCORE * 1.872833))).
This rule had an accuracy of more than 95%. A power of genomic computing is that it ranks variables in order of their utility in successful rules. The top 3 variables are peaks 24, 30 and 42, and peak 24 is indeed salicylate (though the computational analysis was done single-blind so this was not known to the author). The other two variables, previously unheralded in this field, are now under active study as major new components of the plant defence response. Thus the GP discovered not only what differences there were but which were important to the biological pathway of interest, and turned metabolomic data into biochemical knowledge.

1. Oliver, D. J., Nikolau, B. & Wurtele, E. S. (2002) Functional genomics: high-throughput mRNA, protein, and metabolite analyses, *Metab Eng. 4*, 98-106.
2. Delneri, D., Brancia, F. L. & Oliver, S. G. (2001) Towards a truly integrative biology through the functional genomics of yeast, *Curr. Op. Biotechnol. 12*, 87-91.
3. Paton, N. W., Khan, S. A., Hayes, A., Moussouni, F., Brass, A., Eilbeck, K., Goble, C. A., Hubbard, S. J. & Oliver, S. G. (2000) Conceptual modelling of genomic information, *Bioinformatics. 16*, 548-557.
4. Brent, R. (2000) Genomic biology, *Cell. 100*, 169-183.
5. Brent, R. (1999) Functional genomics: Learning to think about gene expression data, *Current Biology. 9*, R338-R341.
6. Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H. Y., He, Y. D. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. & Friend, S. H. (2000) Functional discovery via a compendium of expression profiles, *Cell. 102*, 109-126.

7. Kell, D. B. & King, R. D. (2000) On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning., *Trends Biotechnol. 18*, 93-98.
8. Aranibar, N., Singh, B. K., Stockton, G. W. & Ott, K.-H. (2001) Automated mode-of-action detection by metabolic profiling, *Biochem. Biophys. Res. Commun. 286*, 150-155.
9. Pennie, W. D., Tugwood, J. D., Oliver, G. J. & Kimber, I. (2000) The principles and practice of toxigenomics: applications and opportunities, *Toxicol Sci. 54*, 277-83.
10. Nuwaysir, E. F., Bittner, M., Trent, J., Barrett, J. C. & Afshari, C. A. (1999) Microarrays and toxicology: The advent of toxicogenomics, *Molecular Carcinogenesis. 24*, 153-159.
11. Deyholos, M. K. & Galbraith, D. W. (2001) High-density microarrays for gene expression analysis, *Cytometry. 43*, 229-238.
12. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D. & Brown, P. O. (2000) Genomic expression programs in the response of yeast cells to environmental changes, *Molecular Biology of the Cell. 11*, 4241-4257.
13. DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science. 278*, 680-686.
14. Kodadek, T. (2001) Protein microarrays: prospects and problems, *Chem. Biol. 8*, 105-115.
15. Mann, M., Hendrickson, R. C. & Pandey, A. (2001) Analysis of proteins and proteomes by mass spectrometry, *Annu. Rev. Biochem. 70*, 437-473.
16. Link, A. J. (1999) *2-D proteome analysis protocols*, Humana Press, Totowa, NJ.
17. Wilkins, M. R., Williams, K. L., Appel, R. D. & Hochstrasser, D. F. (1997) *Proteome research: new frontiers in functional genomics*, Springer, Berlin.
18. Fiehn, O. (2002) Metabolomics: the link between genotypes and phenotypes, *Plant Mol. Biol. 48*, 155-171.
19. Weckwerth, W. & Fiehn, O. (2002) Can we discover novel pathways using metabolomic analysis?, *Curr Opin Biotechnol. 13*, 156-60.
20. ter Kuile, B. H. & Westerhoff, H. V. (2001) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway, *FEBS Letters. 500*, 169-171.
21. Fiehn, O., Kloska, S. & Altmann, T. (2001) Integrated studies on plant biology using multiparallel techniques, *Curr. Opin. Biotechnol. 12*, 82-6.
22. Tomita, M. (2001) Whole-cell simulation: a grand challenge of the 21st century, *Trends Biotechnol. 19*, 205-10.
23. Kose, F., Weckwerth, W., Linke, T. & Fiehn, O. (2001) Visualizing plant metabolomic correlation networks using clique-metabolite matrices, *Bioinformatics. 17*, 1198-208.
24. Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R. N. & Willmitzer, L. (2000) Metabolite profiling for plant functional genomics, *Nature Biotechnol. 18*, 1157-1161.
25. Trethewey, R. N. (2001) Gene discovery via metabolic profiling, *Curr. Op. Biotechnol. 12*, 135-138.
26. Trethewey, R. N., Krotzky, A. J. & Willmitzer, L. (1999) Metabolic profiling: a Rosetta Stone for genomics?, *Curr. Op. Plant Biol. 2*, 83-85.
27. Lindon, J. C., Nicholson, J. K., Holmes, E. & Everett, J. R. (2000) Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids, *Concepts in Magnetic Resonance. 12*, 289-320.
28. Nicholson, J. K., Lindon, J. C. & Holmes, E. (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data, *Xenobiotica. 29*, 1181-1189.
29. Kell, D. B., Giansante, L. & Bianchi, G. (2002) Metabolic profiling and machine learning: the development by genomic computing of simple explanatory rules for the discrimination of Italian olive oils from their metabolic profiles, *Eur. J. Lipid Sci. Technol. submitted*.

30. Raamsdonk, L. M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M., Berden, J. A., Brindle, K. M., Kell, D. B., Rowland, J. J., Westerhoff, H. V., van Dam, K. & Oliver, S. G. (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations, *Nature Biotechnol. 19*, 45-50.
31. Johnson, H. E., Gilbert, R. J., Winson, M. K., Goodacre, R., Smith, A. R., Rowland, J. J., Hall, M. A. & Kell, D. B. (2000) Explanatory analysis of the metabolome using genetic programming of simple, interpretable rules, *Genetic Progr. Evolvable Machines. 1*, 243-258.
32. Gilbert, R. J., Johnson, H. E., Winson, M. K., Rowland, J. J., Goodacre, R., Smith, A. R., Hall, M. A. & Kell, D. B. (1999) Genetic programming as an analytical tool for metabolome data. in *Late-breaking papers of EuroGP-99* (Langdon, W. B., Poli, R., Nodin, P. & Fogarty, T., eds) pp. 23-33, Software Engineering, CWI, Amsterdam.
33. Kell, D. B. & Mendes, P. (2000) Snapshots of systems: metabolic control analysis and biotechnology in the post-genomic era in *Technological and Medical Implications of Metabolic Control Analysis* (Cornish-Bowden, A. & Cárdenas, M. L., eds) pp. 3-25 (and see http://qbab.aber.ac.uk/dbk/mca99.htm), Kluwer Academic Publishers, Dordrecht.
34. Kell, D. B., Darby, R. M. & Draper, J. (2001) Genomic computing: explanatory analysis of plant expression profiling data using machine learning, *Plant Physiol. 126*, 943-951.
35. Kell, D. B. (2002) Defence against the flood: a solution to the data mining and predictive modelling challenges of today, *Bioinformatics World (part of Scientific Computing News). Issue 1*, 16-18.
36. McGovern, A. C., Broadhurst, D., Taylor, J., Gilbert, R. J., Kaderbhai, N., Winson, M. K., Small, D. A. P., Rowland, J. J., Kell, D. B. & Goodacre, R. (2002) Monitoring of complex industrial bioprocesses for metabolite concentrations using modern spectroscopies and machine learning: application to gibberellic acid production., *Biotechnol. Bioengi.. 78*, 527-528.
37. Ellis, D. I., Broadhurst, D., Kell, D. B., Rowland, J. J. & Goodacre, R. (2002) Rapid and quantitative detection of the microbial spoilage of meat using Fourier-Transform infrared spectroscopy and machine learning, *Appl. Env. Microbiol. 68*, 2822-2888.
38. Papadimitrou, C. H. & Steiglitz, K. (1998) *Combinatorial optimization: algorithms and complexity*, Dover Publications, Mineola, NY.
39. Cook, W. J., Cunningham, W. H., Pulleyblank, W. R. & Schrijver, A. (1998) *Combinatorial Optimization*, Wiley-Interscience, New York.
40. Fell, D. A. (1998) Increasing the flux in metabolic pathways: A metabolic control analysis perspective, *Biotechnol. Bioeng. 58*, 121-124.
41. Cornish-Bowden, A., Hofmeyr, J.-H. S. & Cárdenas, M. L. (1995) Strategies for manipulating metabolic fluxes in biotechnology, *Bioorg. Chem. 23*, 439-449.
42. Westerhoff, H. V. & Kell, D. B. (1996) What BioTechnologists knew all along...?, *J. Theoret. Biol. 182*, 411-420.
43. Mitchell, T. M. (1997) *Machine learning*, McGraw Hill, New York.
44. Michie, D., Spiegelhalter, D. J. & Taylor, C. C. (1994) *Machine learning: neural and statistical classification*, Ellis Horwood, Chichester.
45. Quinlan, J. R. (1993) *C4.5: programs for machine learning*, Morgan Kaufmann, San Mateo, CA.
46. Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The elements of statistical learning: data mining, inference and prediction*, Springer-Verlag, Berlin.
47. Duda, R. O., Hart, P. E. & Stork, D. E. (2001) *Pattern classification, 2nd ed.*, John Wiley, London.
48. Langdon, W. B. & Poli, R. (2001) *Foundations of genetic programming*, Springer-Verlag, Berlin.
49. Koza, J. R., Bennett, F. H., Keane, M. A. & Andre, D. (1999) *Genetic Programming III : Darwinian Invention and Problem Solving*, Morgan Kaufmann, San Francisco.
50. Banzhaf, W., Nordin, P., Keller, R. E. & Francone, F. D. (1998) *Genetic programming: an introduction*, Morgan Kaufmann, San Francisco.

51. Koza, J. R. (1994) *Genetic programming II: automatic discovery of reusable programs*, MIT Press, Cambridge, Mass.
52. Koza, J. R. (1992) *Genetic programming: on the programming of computers by means of natural selection*, MIT Press, Cambridge, Mass.
53. Cramer, N. L. (1985). A representation for the adaptive generation of simple sequential programs. Paper presented at the *Int. Conf. Genetic Algorithms and their Applications*.
54. Langdon, W. B. (1998) *Genetic programming and data structures: genetic programming + data structures = automatic programming!*, Kluwer, Boston.
55. Gilbert, R. J., Rowland, J. J. & Kell, D. B. (2000) Genomic computing: explanatory modelling for functional genomics in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000)* (Whitley, D., Goldberg, D., Cantú-Paz, E., Spector, L., Parmee, I. & Beyer, H.-G., eds) pp. 551-557, Morgan Kaufmann, Las Vegas.
56. Bi, Y. M., Kenton, P., Mur, L., Darby, R. & Draper, J. (1995) Hydrogen peroxide does not function downstream of salicylic acid in the induction of PR protein expression, *Plant Journal. 8*, 235-245.
57. Bianchi, G., Giansanti, L., Shaw, A. D. & Kell, D. B. (2001) Chemometric criteria for the characterisation of Italian Protected Denomination of Origin (DOP) olive oils from their metabolic profiles, *Eur. J. Lipid Sci. Technol. 103*, 141-150.
58. Griffin, J. L., Williams, H. J., Sang, E., Clarke, K., Rae, C. & Nicholson, J. K. (2001) Metabolic profiling of genetic disorders: A multitissue H-1 nuclear magnetic resonance spectroscopic and pattern recognition study into dystrophic tissue, *Anal. Biochem. 293*, 16-21.
59. Draper, J. (1997) Salicylate, superoxide synthesis and cell suicide in plant defence, *Trends in Plant Science. 2*, 162-165.
60. Mur, L. A. J., Naylor, G., Warner, S. A. J., Sugars, J. M., White, R. F. & Draper, J. (1996) Salicylic acid potentiates defence gene expression in tissue exhibiting acquired resistance to pathogen attack, *Plant J. 9*, 559-571.
61. Mur, L. A. J., Bi, Y. M., Darby, R. M., Firek, S. & Draper, J. (1997) Compromising early salicylic acid accumulation delays the hypersensitive response and increases viral dispersal during lesion establishment in TMV-infected tobacco, *Plant J. 12*, 1113-1126.
62. Mur, L. A. J., Brown, I. R., Darby, R. M., Bestwick, C. S., Bi, Y. M., Mansfield, J. W. & Draper, J. (2000) A loss of resistance to avirulent bacterial pathogens in tobacco is associated with the attenuation of a salicylic acid- potentiated oxidative burst, *Plant J. 23*, 609-621.
63. Roberts, M. R., Warner, S. A. J., Darby, R., Lim, E. K., Draper, J. & Bowles, D. J. (1999) Differential regulation of a glucosyl transferase gene homologue during defence responses in tobacco, *J. Exp. Bot. 50*, 407-410.