Making sense of the metabolome using evolutionary computation: seeing the wood with the trees

Royston Goodacre*
School of Chemistry, The University of Manchester, PO Box 88, Sackville Street, Manchester M60 1QD, UK

Received 14 May 2004; Accepted 1 January 2004

Abstract

One should perhaps start off by asking the question, ‘But what wood is it we want to see?’ There are so many trees that make up the wood; within a post-genomics context, genes, transcripts, proteins, and metabolites are the more tangible ones. Rather than studying these components in isolation, a more holistic approach is to unravel the interactions between the myriad of subcellular components and this is vital to systems biology. Moreover, this will help define the phenotype of the organism under investigation. Metabolomics is complementary to transcriptomics and proteomics, and despite the immense metabolite diversity observed in plants, metabolomics has been embraced by the plant community and in particular for studying metabolic networks. Whilst post-genomic science is producing vast data torrents, it is well known that data do not equal knowledge and so the extraction of the most meaningful parts of these data is key to the generation of useful new knowledge. A metabolomics experiment is guaranteed to generate thousands of data points (e.g. samples multiplied by the levels of particular metabolites) of which only a handful might be needed to describe the problem adequately. Evolutionary computational-based methods such as genetic algorithms and genetic programming are ideal strategies for mining such high-dimensional data to generate useful relationships, rules, and predictions. This article describes these techniques and highlights their usefulness within metabolomics.

Key words: Evolutionary computation, metabolomics, subcellular components, systems biology.

Introduction

Whole genome sequence projects continue to remind us that our knowledge is really rather scant when it comes to putting some function to orphan genes. The recent completion of the human genome sequence (The International Human Genome Mapping Consortium, 2001; Venter et al., 2001) and the plant genome sequences of Arabidopsis thaliana (Arabidopsis Genome Initiative, 2000) and Oryza sativa (rice) (Goff et al., 2002; Yu et al., 2002), amongst others (http://wit.integratedgenomics.com/GOLD/) have accelerated demand for determining the biochemical function of orphan genes and for validating them as molecular targets for therapeutic intervention. The search for biomarkers (real or surrogate) that can serve as indicators of disease progression or response to therapeutic intervention has also increased (Harrigan and Goodacre, 2003). Functional analyses (Fig. 1) have thus emphasized analyses at the level of gene expression (transcriptomics), protein translation (proteomics; including post-translational modifications), and the metabolite network (metabolomics), with a view within a systems biology approach of defining the phenotype and bridging the genotype-to-phenotype gap (Fiehn, 2002).

The ‘metabolome’ is the quantitative complement of all the low-molecular-weight molecules present in cells in a particular physiological or developmental state (Oliver et al., 1998), and whilst it is complementary to transcriptomics and proteomics it does have special advantages, particularly with respect to the paradigm shift from thinking of linear metabolic pathways to more global integrated metabolic networks and neighbourhoods (Barabási and Oltvai, 2004; Kell, 2004), since it is obviously better to measure these networks directly. This is no easy task and despite the immense metabolite diversity observed in plants metabolomics, up to 200 000 different metabolites in the plant kingdom (Fiehn, 2002), it is highly encouraging to see...
that metabolomics has been readily embraced by plant biologists.

Generally there are two sorts of data that are generated by a metabolomics experiment. The first is a list of metabolites with their levels, usually obtained via some hyphenated technique involving chromatography, and the second type of data are fingerprint-like traces from an FT-IR or NMR analysis. Irrespective of the type of data generated, these metabolomic strategies generate large amounts of data, and it is obvious (Fiehn, 2001; Mendes, 2002) that current informatic approaches need to adapt and grow in order to make the most of these data. In particular, good robust databases (Hardy and Fuell, 2003; Mendes, 2002), very good data, excellent visualization methods (Li et al., 2003), and even better algorithms (Goodacre and Kell, 2003), are needed with which to turn these data into knowledge.

Consider a hypothetical experiment where data on a modest 250 metabolites have been collected using GC-MS from two plant populations; control plants, and test plants exposed to drought. So that any natural variation in plant growth stages is accounted for 100 plants of each are measured. Once the data have been collected it is necessary to be able to discriminate reproducibly between the two different populations (or classes), and this has classically been achieved by using supervised learning algorithms such as discriminant analysis, partial least squares or artificial neural networks (Fig. 2). Each of those metabolites could be used in the construction of a discriminatory model.
to differentiate between the plants, however, in addition to important metabolites many will be non-relevant and it is known from the statistical literature that better (i.e. more robust) predictions can often be obtained when only the most relevant input variables are considered (Seasholtz and Kowalski, 1993), i.e. that ‘parsimonious’ models tend to generalize better.

An exhaustive search of whether a metabolite is used or not in the model would be computationally prohibitive since the possible permutations (to use or not use 250 metabolites, i.e. two choices 250 times) is $2^{250}$ or $1.8 \times 10^{75}$. This number is so big that if a computer could check 10 million orderings every second it would still take $>3 \times 10^{62}$ years to check them all! And the lifetime of the Universe is after all only $\sim 10^{17}$ seconds (Barrow and Silk, 1995). Thus, even though the way to solve this problem is known, it cannot be done. These problems are NP-complete (Garey and Johnson, 1979); that is to say, as the number of variables in the search space goes up linearly, the number of possible solutions, and hence the time to solve the problem, is a polynomial of the number of variables. Therefore, to solve these problems requires time which is exponential in the problem size and so to find the global optimum requires exhaustive searching and this is computationally impossible. Thus route A in Fig. 3 is unfeasible as no algorithm can do this and an alternative strategy needs to be found. The premise is that a ‘good’ solution is acceptable and so an alternative method is needed to search the huge spaces of possible solutions. Importantly, however, if the search space is large but the solution space is small, i.e. the problem can be solved with just a small number of variables, the effective search space becomes much narrower. Thus the number of permutations of five variables from 250 metabolites is $250!/[250!–5!5!]$ or just $9.4 \times 10^{11}$. Evolutionary computational-based methods offer such an approach. These are classified as heuristic algorithms which are considered to work ‘reasonably well’; that is to say a ‘good’ solution is acceptable, but for many cases there is no proof that they are faster or that they find the best solution.

**Evolutionary computation**

It is commonly accepted that the origin of life started during the early Archaean period when biogenic signatures in metasediments from Akilia Island, West Greenland were discovered which dated from $\sim 3.85$ billion years ago (Mojzsis et al., 1996). This proto-organism (or precellular stage) positioned at the root of the tree of life, gave rise to the immense species diversity that is present on the planet today. This rich diversity has been created from DNA mutation, DNA crossover (including horizontal gene transfer), and the survival of the fittest, which has led to a plethora of different organisms occupying different ecological niches. What nature has achieved in $4 \times 10^9$ years can be mimicked *in silico* and this forms the basis of evolutionary computation.

In evolutionary computation a population of individuals, each representing the parameters of the problem to be optimized as a string of numbers or binary digits, undergoes a process analogous to evolution in order to derive an optimal or near-optimal (a good) solution (Fig. 3B). The parameters stored by each individual are used to assign it a *fitness*, a single numerical value indicating how well the solution using that set of parameters performs. New individuals are generated from members of the current population by processes analogous to asexual and sexual reproduction (Fig. 4), these populations can then be ‘pruned’ based on the concepts of Darwinian selection (Bäck et al., 1997; Darwin, 1859). This process might enrich the population for false positives or negatives and so to avoid this during, as well as after, evolution an independent test set is used to validate the model.

![Fig. 3](https://academic.oup.com/jxb/article-abstract/56/410/245/483996)

**Fig. 3.** (A) The complex problem which it is hoped to solve but cannot, and (B) the strategy used by evolutionary computing methods.
Evolutionary algorithms are particularly popular inductive reasoning and optimization methods (Corne et al., 1999; Michalewicz and Fogel, 2000) and include genetic algorithms (GAs; Goldberg, 1989; Holland, 1992; Mitchell, 1995), evolution strategies (Beyer, 2001; Schwefel, 1995), evolutionary programming (Fogel, 2000; Fogel and Fogel, 1996), genetic programming (GP; Banzhaf et al., 1998; Koza, 1992, 1994; Koza et al., 1999), and genomic computing (GC; Kell, 2002; Kell et al., 2001) and because the models are in English and by penalizing complex expressions may be made comparatively simple. Evolutionary computational-based algorithms are thus explanatory supervised learning techniques (Fig. 2) (Beavis et al., 2000; Goodacre and Kell, 2003; Kell and King, 2000; Mitchell, 1997) where answers to questions of biological interest are sought, such as ‘What metabolites have I measured in my metabolome that makes plants exposed to drought different from the same isogenic plants that have been adequately watered?’

The overall evolutionary procedure employed by GAs and GP is depicted in Fig. 4. However, whilst there are many analogies between how GAs and GP operate, when it comes to using these methods for variable (or metabolite) selection, each method should be considered separately. What follows is a brief description of the salient features of GAs and GP.

Genetic algorithms (GAs)

In a GA the population of individuals, each representing the parameters of the problem to be optimized, are encoded as a string of numbers or binary digits (Fig. 5A). In the GA representation each individual in the population contains a string of ‘1’s and ‘0’s, the number of which would be the total number of metabolites to choose from. Each input variable represented by a ‘1’ is selected to be used in the model, whilst each ‘0’ is not selected (Broadhurst et al., 1997; Horchner and Kalivas, 1995). Other GA variants based on the selection of spectral windows for FT-IR and Raman spectroscopy are also popular (Leardi et al., 2002; Roger and Bellon-Maurel, 2000; Taylor et al., 1998; Williams and Paradkar, 1997). In genetic terms each variable is called a gene and a set of variables is called a chromosome. For example, as depicted in Fig. 5A representing the selection or otherwise of seven metabolites, one possible chromosome would be 1101010, which can be translated as selection or otherwise of seven metabolites, one possible chromosome would be 1101010, which can be translated as selection or otherwise of seven metabolites.

The output from the model would be used to assign it a fitness to each of the individuals in the population. For example, to differentiate a control plant from one exposed to drought, the fitness would be a measure of how well the model predicted the two groups. New individuals are generated from members of the current population by processes analogous to asexual and sexual reproduction (Fig. 6).

Asexual reproduction would simply be cloning the chromosome into the next population, whilst mutation is performed by randomly selecting a parent with a probability related to its fitness, then randomly changing one or more of the parameters it encodes. The new individual then replaces a less-fit member of the population, if one exists. Sexual reproduction, or crossover, is achieved by selecting two parents with a frequency related to their fitnesses, and generating two new individuals by copying parameters from one parent, and switching to the other parent after a randomly-selected point; alternatively, as depicted in Fig. 6C a double crossover could be implemented. The two new individuals then replace less-fit members of the population as before. The above procedure is repeated, with the overall fitness of the population improving at each generation, until an acceptably-fit individual is produced.
However, whilst GAs are very successful search algorithms for tackling NP-hard problems, the disadvantage is that, with the GA variable selection approach, the relationship between one variable and another is not evident, only whether they contribute to a model or not. Therefore, a richer language is needed.

**Genetic programming (GP)**

A GP is an application of the GA approach to derive mathematical equations, logical rules or program functions automatically (Banzhaf et al., 1998; Koza, 1992, 1994; Koza et al., 1999; Langdon, 1998). Rather than representing the solution to the problem as a string of parameters, as in a conventional GA, a GP usually (cf. Banzhaf et al., 1998) uses a tree structure. The leaves of the tree, or terminals, represent input variables or numerical constants. Their values are passed to nodes, at the junctions of branches in the tree, which perform some numerical or program operation before passing on the result further towards the root of the tree (Fig. 5B).

As with GAs an initial (commonly random) population of individuals, each encoding a function or expression, is generated and their fitness to produce the desired output is assessed. In the second population three reproduction strategies are adopted (see Fig. 7 for pictorial details). (1) **Cloning**, some of the original individuals are allowed to survive unmodified. (2) New individuals are generated by **mutation** where one or more random changes to a single parent individual are introduced. This can be when a node is randomly chosen, and modified either by giving it a different operator with the same number of arguments, or it may be replaced by a new random subtree. Terminals can be mutated by slightly perturbing their numerical values, or randomly choosing an input variable. (3) Alternatively, new children are generated by **crossover** where random rearrangement of functional components between two or more parent individuals takes place. Two parents are chosen with a probability related to their fitness. A node is randomly chosen on each parent tree, and the selected sub-trees are then swapped. At each reproduction stage, because of the use of these trees to encode mathematical equations, the new trees are still syntactically correct. The fitness of the new individuals in population 2 is assessed and the best individuals from the total population become the parents of the next generation. An individual’s fitness is usually assessed as the root mean squared (RMS) error of the difference between expected values and the GP’s estimated values for the training set. In order to reduce ‘bloat’, a phenomenon in which the GP function trees get so huge that it lacks explanatory power (Langdon and Poli, 1998; Podgorelec and Kokol, 2000), penalties to the number of nodes and depth of the tree in the individual’s function tree can be applied. This overall process is repeated until either the desired result is achieved or the rate of improvement in the population becomes zero.

![Fig. 7. The GP reproduction processes. Three strategies are used including (A) cloning, (B) mutation, and (C) crossover events.](https://academic.oup.com/jxb/article-abstract/56/410/245/483996)

**Application of genetic search algorithms to metabolome analysis**

GAs and GPs are very efficient search algorithms and can be used to produce models that allow the deconvolution of metabolome data in chemical terms. Detailed below are three published plant metabolome examples illustrating this.

**Metabolic fingerprinting in salt-stressed tomatoes**

Samples from Edkawy tomato fruit grown hydroponically under both high- and low-salt conditions were analysed using FT-IR, with the aim of identifying biochemical features linked to salinity in the growth environment. Examination of the GP-derived trees showed that there were a small number of spectral regions that were consistently being used. In particular, the spectral region containing absorbances potentially due to a cyanide/nitrile functional group was identified as discriminatory (Johnson et al., 2000).

More recently (Johnson et al., 2003), the same authors applied FT-IR to fingerprint Edkawy and Simge F1 tomato varieties. It was observed that the exposure of the plants to salinity significantly reduced the relative growth rate of Simge F1 but had no significant effect on Edkawy. By contrast, whilst with both tomato varieties there was little effect on total fruit number, the salt treatment had significantly reduced the mean fruit fresh weight and size class in both Edkawy and Simge F1. In this study, rather than using GP, GA was used as a variable selection method prior to discriminant MLR and this approach was able to classify accurately between control and salt-treated fruit. It was encouraging that this different genetic search algorithm on two tomato varieties also identified a cyanide/nitrile functional group as being discriminatory.
It is known that cyanide is formed in plants during ethylene biosynthesis, and that ethylene production is enhanced in plants under stress conditions. Therefore, it may be proposed that plants grown under saline conditions may have enhanced levels of cyanide as a result of enhanced ethylene biosynthesis. Thus inductive reasoning via GP and GA has allowed the significance of a pathway turned on under tomatoes exposed to salinity to be highlighted as potentially important. This pathway can now be subjected to conventional biochemical analysis.

Analysis of defence in tobacco plants

Within functional genomics the potential power of evolutionary methods has been shown for the analysis of metabolites from transgenic tobacco plants (Kell et al., 2001). Tobacco is a model organism for the study of salicylate biology in plant defence, but despite a considerable amount of research, little is known regarding its synthesis, catabolism, and mode of action. Six-week-old control plants and a transgenic expressing a bacterial gene encoding the enzyme salicylate hydroxylase (SH-L), which is known to block salicylic acid accumulation in transgenic tobacco (Darby et al., 2000), were inoculated with tobacco mosaic virus and leaf samples were analysed by HPLC. GP analysis of these metabolome profiles identified three peaks as highly discriminatory for detecting the presence of the SH-L genotype in the transgenic. One of the peaks was indeed salicylate, but the other two were unknown and are now the subject of further investigation.

Analysis of the photoperiodic floral induction in Pharbitis nil

Metabolic fingerprints were obtained from unfractionated P. nil leaf sap samples by direct infusion into an electrospray ionization mass spectrometer using flow-injection (Vaidyanathan et al., 2002). Analyses took less than 30 s per sample and yielded complex mass spectra. Various chemometric methods, including cluster analysis, ANNs, and GP, could discriminate the metabolic fingerprints of plants subjected to different photoperiod treatments. A GP was evolved to discriminate plants 1 week after a short day exposure from all the other plants (24 h and 48 h after SD exposure and controls) and generated rules predominantly involving m/z 520, 229, and 143. Although these m/z values represented very minor peaks in the sap spectra, when they were plotted as a pseudo-3D plot (data not shown), 1-week plants could indeed be separated from the others. Whilst the identity of these peaks has not yet been established, thus ESIMS with GP has suggested which analytes are potential metabolite markers and thus this approach has generated new directions of research and potentially new knowledge (Goodacre et al., 2003).

Metabolic fingerprinting of a fermentation model

The ability to control the industrial bioprocess is paramount for product yield optimization, and it is imperative, therefore, that the concentration of the fermentation product (the determinand) is assessed accurately. Whilst real industrial bioprocesses have been modelled in the past (McGovern et al., 2002, 1999) and GA and GP used for spectral interpretation (McGovern et al., 2002), for the present purpose a simple model of a bioprocess that has previously been investigated using mixtures of ampicillin and E. coli as a model system will be investigated (Winson et al., 1997).

Sample preparation and metabolic fingerprint generation

The bacterium used was E. coli HB101 (Maniatis et al., 1982); this is ampicillin-sensitive, indicating that any spectral features observed are not due, for instance, to β-lactamase activity. The mixtures were prepared as described previously (Winson et al., 1997). The strain was grown in 4.0 l liquid medium: glucose (BDH) 10.0 g; peptone (LabM) 5.0 g; beef extract (LabM) 3.0 g; per litre water, for 16 h at 37 °C in a shaker. After growth, the cultures were harvested by centrifugation, washed, and resuspended in physiological saline (0.9% NaCl). Ampicillin (desiccated D-[ ]- aminobenzylpenicillin sodium salt, >=98% (titration), Sigma) was prepared in the bacterial suspensions to give concentration ranges of 0–5000 μg ml⁻¹ in 250 μg ml⁻¹ steps (0–13.46 mM) in 40 mg ml⁻¹ E. coli (dry weight), corresponding to ~1–2×10⁷ cells ml⁻¹.

Aliquots (20 μl) of the above samples were evenly applied to the wells of an aluminium plate (measuring 10×10 cm) in triplicate and dried at 50 °C for 30 min. The FT-IR instrument used was a Bruker IFS28 FT-IR spectrometer (Bruker Spectrospin Ltd., Banner Lane, Coventry, UK) equipped with an MCT (mercury–cadmium–telluride) detector (cooled with liquid N2) and a motorized stage of a reflectance accessory, onto which the Al plate was loaded. Spectra were collected over the wavenumber range 4000 cm⁻¹ to 600 cm⁻¹. Spectra were acquired as described previously (Goodacre et al., 2000; Timmins et al., 1998) at a rate of 20 s⁻¹, the spectral resolution used was 4 cm⁻¹, and to improve signal-to-noise ratio 256 spectra were co-added and averaged.

For the IR map the biomass from Escherichia coli HB101 was applied evenly to the surface of a flat 7×7 cm Al plate at a concentration of ~200 μg cm⁻² (dry weight); a β-lactam ring was then drawn with ~100 μl of a 5 mg ml⁻¹ solution of ampicillin. Data were acquired at a resolution of 1×1 mm (therefore these maps are 71×71 pixels ×882 wavenumbers). Spectra were acquired as above, but for speed (since 5041 spectra were collected) the number of co-adds was only 16.
Quantifying the level of ampicillin

Initially PLS was used to predict the level of ampicillin in the mixtures. The training data were those triplicate spectra with ampicillin concentrations 0, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000 μg ml⁻¹, whilst the test data were those mixtures containing 250, 750, 1250, 1750, 2250, 2750, 3250, 3750, 4250, and 4750 μg ml⁻¹ ampicillin. The model was test set validated and whilst 10 latent variables gave the lowest RMS error (168 μg ml⁻¹), the most parsimonious model used four latent variables with an acceptable RMS error of 288 μg ml⁻¹. The results for this are shown in Fig. 8A, where it is clear that PLS was then able to predict the concentration of this secondary metabolite accurately. PLS is a linear regression analysis and it is possible to inspect the regression coefficients from the PLS model (Fig. 8B), whilst many areas of the 882 inputs were selected as positively contributing to the model (and hence correlated with ampicillin), it is clear that a vibration at 1767 cm⁻¹ was dominant.

The same training and test sets were analysed by GP using the Genomic Computing software Gmax-bio™ (Aber Genomic Computing, Aberystwyth, UK) which runs under Microsoft Windows NT on an IBM-compatible PC. An introduction to Gmax-bio™ is given elsewhere (Kell et al., 2001). The default parameter settings for population size (1000), mutation and recombination rates were used throughout, and the fitness was assessed using OLS (ordinary least squares). The operators that were used were the arithmetical ones +, −, +, ×, asterisk, log(x), 10^x, the hyperbolic Tanh(x), and numerical inputs of 0.1, 1, 3, 5, and random. Ten GPs were run and the GP results were equivalent to the PLS predictions (data not shown). The frequency of the number of times each input (wavenumber) was used for the 10 evolved populations was calculated and plotted against the wavenumber of the infrared light (Fig. 9) where it was clear that a single dominating area of the spectra was chosen which was a vibration at 1767 cm⁻¹, this peak was also clearly evident in the E. coli+5000 μg ml⁻¹ ampicillin mix, but was absent from E. coli alone (Fig. 9).

Spectral interpretation

That the vibration at 1767 cm⁻¹ was selected as characteristic of ampicillin is highly encouraging since it is known (Winson et al., 1997) that this corresponds to the constrained carbonyl bond in the β-lactam ring of penicillins in general (the structure of ampicillin is shown in Fig. 9). To highlight the usefulness of this spectral deconvolution an FT-IR image was collected from an E. coli background with an image of the β-lactam ring drawn with ampicillin and slices of this hypercube were taken and colour maps generated (Fig. 10). The first image shows the level of

\[ \text{Absorbance (arbitrary)} \]

\[ \text{Wavenumber (cm}^{-1}) \]

\[ \text{Summed frequency plot from GP analysis of the number of times each input (wavenumber) was used for the 10 evolved populations. Also shown are the normalized FT-IR spectra from E. coli (blue trace) and E. coli+5000 μg.ml}^{-1} \text{ (red trace), and the structure of ampicillin.} \]
protein on the plate by simple integration under the Amide I band at 1662 cm$^{-1}$, and since the level of protein is constant the picture can essentially not be seen. By contrast, integrating under the molecular vibration from the constrained carbonyl on the $\beta$-lactam ring at 1767 cm$^{-1}$ allows the hidden chemical image to be clearly seen.

**General conclusions**

In the early stages of functional genomics programmes there is a scenario where current knowledge is minute, that is there are no ideas about the role of an orphan ORF and there are few if any hypotheses to test (Brent, 2000; Kell and King, 2000). However, experiments can be designed based, for example, on gene knockouts and controlled over-expression and the effect on the phenotype of the organism observed. Alternatively, an isogenic organism might be exposed to different abiotic and biotic stresses to assess how it adapts to these new environments.

Metabolomics is one ‘omics approach with which data floods can be generated from these genetic manipulations and environmental stimuli (as indeed are transcriptomics and proteomics, and the same general conclusions given here apply equally to these methods), however, deconvolution of these data in terms of which metabolites are of key importance to the genetic manipulations and environmental stimuli is essential to generate new knowledge. Evolutionary algorithms can be considered to be rule induction methods that are entirely data-driven and are thus especially appropriate for problems that are data-rich but hypothesis/information-poor. As described above rule induction can be used to generate rules and hence hypotheses from suitable examples. Of course, these new theories will not necessarily be correct, but by testing them new knowledge will be generated which will lead to an increased understanding of the function of the orphan gene, or how an organism responds to different environmental conditions.

**Acknowledgements**

I am indebted to the UK BBSRC (Engineering and Biological Systems Committee), the UK EPSRC and the Royal Society of Chemistry for supporting our research within metabolomics.

**References**


