The Application of Genetic Programming for Drug Discovery in the Pharmaceutical Industry

EPSRC RAIS Secondment with GlaxoSmithKline

4 July 2002–30 September 2003.

W. B. Langdon and B. F. Buxton

Computer Science, University College,
Gower Street, London, WC1E 6BT, UK
W.Langdon@cs.ucl.ac.uk B.Buxton@cs.ucl.ac.uk
http://www.cs.ucl.ac.uk/staff/W.Langdon

24 November 2003

1 Background and Context for the Secondment

This Individual Grant Review (IGR) reports on EPSRC Research Assistant Industrial Secondment (RAIS) grant GR/S03546/01, which enabled Dr. W. B. Langdon to collaborate closely with GlaxoSmithKline (GSK), mostly by working at their research laboratories in Greenford, Middlesex. This RAIS project's primary purpose has been to disseminate research results from the UCL Computer Science department's EPSRC funded Faraday INTErSECT project GR/M43975 into a UK Industrial setting, specifically one of the primary collaborating partners, GSK. Dr. Langdon was the Research Assistant employed by the Faraday project from May 2000 until its successful conclusion.

The INTErSECT Faraday project was very successful with an outcome rated as "internationally leading" in research quality, "tending to internationally leading" in research planning and practice, and "tending to outstanding" overall. In particular, we had [1], by means of the efforts of the research fellow, Dr. W. B. Langdon, in close collaboration with the enthusiastic and extensive support of the industrial sponsor, Dr. S. K. Barrett, not only been able to develop a method of classifier fusion by genetic programming, but to demonstrate its potential for application to large scale problems of industrial interest.

1.1 Work completed for GSK

The genetic programming code used on the Faraday project was "industrialised". This included porting to IBM PC, SUN C++ and DEC ALPHA and features to make it easier to use such as: packaging code, documentation including powerpoint slides and worked examples, speed up, as well as debugging. Major extensions were made to support DNA chip and SNP data. 8 releases where made. The code and documentation is available via `ftp://cs.ucl.ac.uk/genetic/gp-code`.

In addition to improved documentation, we held several in house (GSK) one-on-one training sessions covering both genetic programming in general and the specific code. Mostly these were held at Greenford. However there were also telephone conferences augmented with PC network connections “net-meetings”, particularly between Greenford and Philadelphia. A meeting was held in Philadelphia.

A number of powerpoint slide packs describing GP and specific experiments were produced and distributed within GSK via the Data Exploration Sciences (DES) group internal web pages. A number of presentations were given within GSK both at Greenford and also Harlow and Stevenage, e.g. to Statistical Sciences Europe and to the Computational And Structural Sciences (CASS) cheminformatics group, as well as meetings with GSK's Research Statistics Unit. The CASS group held a blind trial in which a number of practising cheminformatics groups within GSK where invited to predict biochemical activity using their favourite machine learning technique. We were one of two external groups who were also invited. The GP evolved model was the best of 12 submitted at predicting the holdout set, marginally improving over the existing production system. The organisers are in the process of completing a write-up of the complete workshop [2] while a description of the GP specifically can be found in the CEC-2003 conference [3].

As with any research project, we have also disseminated research results via seminars (cf. Section 5.1) conference presentations and journal papers. These, plus work initiated during the Faraday project, are listed on the IGR form.

1.2 Problems Analysed

- High Throughput Screening (HTS) P450 measurement
- P450 IC50 measurements (cf. CASS blind trial)
- Human Bioavailability and Rat Bioavailability
- DNA chip gene expression datasets (public domain and GSK internal).
- Single Nucleotide Polymorphism (SNP)

2 Project Plan Review

The original plan was for a 12 month secondment.

Additional funds from GSK, some from another fully industrially funded project [4] to which Dr Langdon has contributed expert knowledge and advice, and some contributed specifically for further work on the topics studied in this RAIS project, have enabled the secondment to run for an additional 3 months duration.

The objectives as listed on the accompanying IGR form for the transfer of the genetic programming expertise to the industrial partner, in particular, for classifier fusion were retained and all met, viz:

1. The identification and pursuit of on going and new applications. See Section 1.2 above. The first two items were activities under way during the INTErSECT Faraday project [GR/M43975], the last three are new activities initiated during this RAIS project.

2. To provide feedback on the system, and use this to guide further development and modification. See Section 1.1 above, in particular the “industrialisation”, modification and enhancement of the genetic programming code, and participation in the GSK CASS blind trial and workshop. The workshop results are being written up for publication, led by the GSK staff who organised the event.

3. In the long term, to indicate areas for further industrial development and for further collaborative research.

The DNA chip gene expression analysis is at an early stage, but the work carried out (see for example, Figure 1) indicates that there is scope for much further study in this difficult and important area as the microarray chips become ever more widely used, both in academe and in industry. The results of our work have not only been published as indicated in the accompanying IGR, but have also been presented to Biology groups working in the Bloomsbury area and in particular to collaborators in two of our Wellcome funded microarray projects [5, 6] respectively on the functional genomics of pluripotent stem cells and on integrating transcriptomics and structural data to reveal protein function. There is currently only one other significant group (in the USA) pursuing the application of genetic programming to the analysis of microarray chips [7]. Similarly, the results obtained in the bioavailability work and, especially, in the SNP work indicate scope for further work. For commercial reasons the SNP results have not been published.

GSK is an international company. While the secondment was mostly based in the United Kingdom, towards the end of the project Dr. Langdon also worked with GSK's research laboratories in Philadelphia. Visiting this site in USA once and otherwise liasing electronically from Greenford. Code written in the UK has been ported to ALPHA workstations and run in Philadelphia.

GSK have also provided on-the-job training for various aspects of this work, in particular with aspects of the work requiring Microsoft expertise that were unfamiliar to Dr. Langdon who is a very experienced user of unix systems.

The Company operate a strict firewall policy aimed at controlling access both to the Internet from inside GSK and into GSK from outside. This impeded access to data and email from UCL and GSK causing operational difficulties, a side effect of which was that W. B. Langdon had to decline the offer of being editor-in-chief for GECCO-2003. (GECCO is the leading international conference in the field of evolutionary computing.) The opportunity to host GECCO-2005 in London has also been missed.

3 Research Impact and Benefits to Society

GlaxoSmithKline is one of the largest pharmaceutical companies in the world (currently second largest), employing more than 100,000 people. It is a major UK industrial company. The drugs they make are used world wide. GSK's continued success depends upon its ability to discover new drug treatments. From this flows the obvious benefits to GSK's customers and the UK economy. Nonetheless, like any high tech research, drug discovery research remains a risky hit and miss business. The fraction of drug discovery projects which result in a marketable product remains very low. One of the major causes for the most expensive failures of drugs is rejection of trial chemicals due to toxicity, poor bioavailability, side-effects etc., which are only discovered late in the day. Computer based prediction of potential problems offer the hope of directing the drug discovery search away from expensive failure by giving some (albeit not 100%) guidance in the earlier stages. This would not only save money, but by focusing research effort in more productive areas, might increase the chance of finding successful drug treatments. We have shown that genetic programming (GP) can evolve such predictors for two specific areas of concern (P450 inhibition and bioavailability).

Data mining is being used in the pharmaceutical industry in many ways. For example, monitoring the in-use effectiveness of existing treatments, looking for patterns to explain or predict untoward events and looking for new opportunities to re-use existing drugs in new ways. GP is one of a range of techniques that can be used.

There is an opportunity in bioinformatics to use data mining techniques to exploit very high dimensional datasets now being produced in increasing numbers due to the success of DNA chip and spot gene expression measurements and Single Nucleotide Polymorphism (SNP)

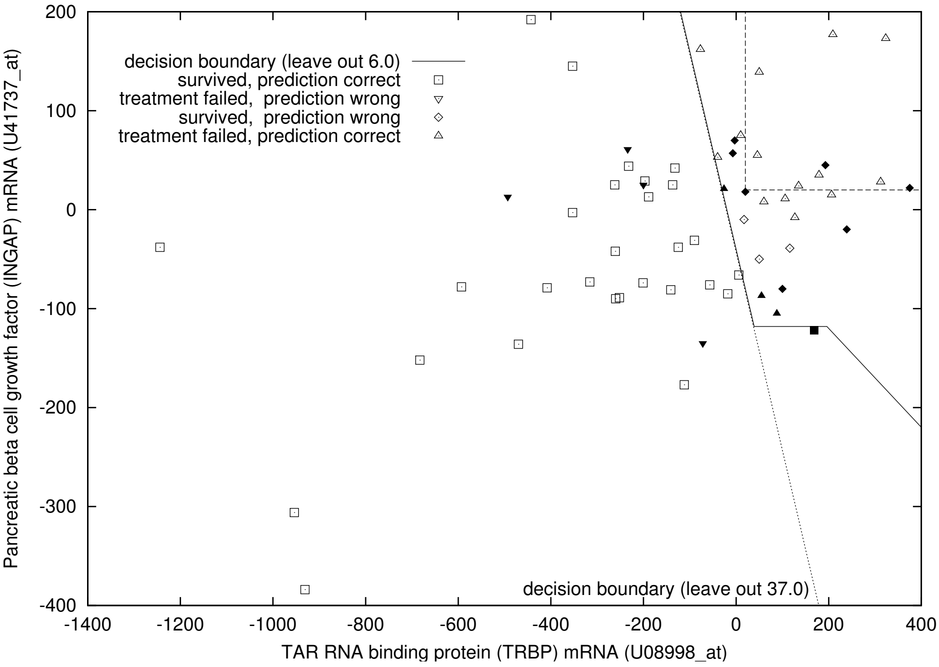

Figure 1: Expression of two genes from 7129 used to predict outcome of cancer treatment [8]. The piecewise linear decision boundary is the best of 500 random programs using just these two genes (from the first of ten runs leaving out patient record 6). Solid shapes indicate cases which are significantly harder to predict. The linear boundary produced when leaving out a different record (37) is shown dotted. It predicts survival if $2 \times \text{U08998\_at} + \text{U41737\_at} < -43$ and makes only one more error. Dashed rectangle indicates threshold (20) used by [8], only nine patient records are not affected by such thresholding.

measurements. Such high dimensionality offers a challenge to existing data mining techniques, particularly as there is often few examples, for each gene/base measured. Initial experiments (both on DNA and SNP datasets) suggest a way of using GP to perform feature selection in very high dimensional datasets and to yield predictive models.

In addition to using GP in pharmaceutical applications, fundamental research into theoretical underpinnings of GP and programming has continued. Results have cast light on existing ad hoc heuristics.

In addition, as indicated in the INTErSECT Faraday project report [1], we have, with the assistance of Dr Suran Goonatilake, one of the founders of Searchspace, been exploring the potential for setting up a company. Plans have been discussed with our GSK collaborators and discussions have taken place with several venture capitalists: MVM Limited, Atlas Venture, the Wellcome Development fund administered by Catalyst, and with Partnerships UK. The initial plans involved also Professor David Jones, head of our new Bioinformatics Unit, but feedback from the venture capitalists and other advice suggested that a medical target might be appropriate. Results of research carried out with the assistance and advice of leading biological and medical researchers within UCL and the Bloomsbury area to identify the best opportunities for such a company and selection of the optimal first target areas for commercial work support this conclusion. Chris Boshoff, Professor of cancer medicine at UCL, has thus joined the team in place of Professor Jones and a business plan is currently under consideration by two funds.

4 Explanation of Expenditure

The project was funded for 12 months by the EPSRC at a cost of £36.4k. Additionally Glaxo-SmithKline (GSK) have provided two additional sources of funding as mentioned above.

There was a gap of a few months between the end of the Faraday project and the start of the RAIS project. During this period Dr. Langdon was fully funded by an on going but separate research contract between University College and GSK. Together, these amounted to 3 months additional full time funding.

Secondly GSK provided space, full access to their research facilities and GSK staff time amounting to an indirect contribution in excess of £40,500. Additionally GSK provided Dr. Langdon with a "salary top-up" and contribution to travel and conference expenses in excess of £7000.

Finally the UK charity EvoSolve (which supports research and higher study concerning the use of advanced software technologies) provided a publication support grant of £150 (for the EvoBIO publication).

5 Further Research or Dissemination Activities

The RAIS project's major activities, publications and presentations are available via the Internet from the projects home page `http://www.cs.ucl.ac.uk/staff/W.Langdon/gsk.html`.

5.1 External Project Presentations

In addition to various presentation etc. within GSK and presenting conference papers (see list of publications on IGR form) the following talks or presentations have been delivered or are planned.

- British Computer Society, central London branch, 12 Feb 2004.
- Seminar, Kent University, 2 Feb 2004.
- Invited talk, European Workshop on Data Mining and Text Mining for Bioinformatics, Dubrovnik, Croatia, 22 September, 2003.
- Seminar, Essex University, 20 August 2003.
- BioGEC'2003 workshop, Chicago, 12 July 2003.
- Transferring Computer Science Research to Mining DNA chip Protein Expression. Talk and poster presentation at the EPSRC MIPNETS network, Liverpool 25-27 June 2003.
- Seminar Colorado State University, 19 May 2003.
- Invited talk PRCVC Prague, April 2003.
- Seminar, University of Chile, Dec 2002.
- Invited talk Hybrid Information Systems, Santiago, Dec 2002.
- Invited talk IBERAMIA, Seville, Dec 2002.
- Seminar, CWI (Dutch National Centre for mathematics and computer science research), 24 Oct 2002.
- Knowledge Discovery meets Drug Discovery, Leuven, 23 Oct 2002.

- BNAIC, Leuven, 22 Oct 2002.

Many of the slides are available online via the project world wide web page Project WWW page http://www.cs.ucl.ac.uk/staff/W.Langdon/gsk.html

References

[1] B. F. Buxton, S. B. Holden, and P. C. Treleaven. Intelligent data analysis and fusion techniques in pharmaceuticals, bioprocessing and process control, October 2002.

[2] Sandeep Modi *et. al.* *Journal of Medicinal Chemistry*, 2004. In preparation.

[3] W. B. Langdon, S. J. Barrett, and B. F. Buxton. Predicting biochemical interactions – human P450 2D6 enzyme inhibition. In *Proceedings of the 2003 Congress on Evolutionary Computation CEC2003*, pages 1340–1347, Canberra, 8-12 December 2003. IEEE Press.

[4] D. T. Jones (PI) and B. F. Buxton, November 2001. Research funded by GSK (G2558, overall value £143,984).

[5] D. H. Beach *et. al.* Functional genomics of pluripotent stem cells and their progeny, February 2002. Wellcome grant (overall value £5,119,123).

[6] C. Orengo *et. al.* Integrating transcriptomics and structural data to reveal protein functions, August 2002. Wellcome grant (overall value $\approx \pounds 800,000$).

[7] Jason H. Moore, Joel S. Parker, Nancy J. Olsen, and Thomas M. Aune. Symbolic discriminant analysis of microarray data in automimmune disease. *Genetic Epidemiology*, 23:57–69, 2002.

[8] Scott L. Pomeroy, Pablo Tamayo, Michelle Gaasenbeek, Lisa M. Sturla, Michael Angelo, Margaret E. McLaughlin, John Y. H. Kim, Liliana C. Goumnerovak, Peter M. Blackk, Ching Lau, Jeffrey C. Allen, David Zagzag, James M. Olson, Tom Curran, Cynthia Wetmore, Jaclyn A. Biegel, Tomaso Poggio, Shayan Mukherjee, Ryan Rifkin, Andrea Califanokk, Gustavo Stolovitzkykk, David N. Louis, Jill P. Mesirov, Eric S. Lander, and Todd R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, 24 January 2002.