Cup 2001 of The Seventh ACM SIGKDD International Conference on Knowledge discovery and data mining – USA/ San Francisco, August 26-29,2001.
http://www.cs.wisc.edu/~dpage/kddcup2001/

Genetic programming applied to pharmaceutical drugs design.

James Cunha Werner (wernerjc@sbu.ac.uk)

Terence C. Fogarty (fogarttc@sbu.ac.uk)

SCISM
South Bank University
103 Borough Road
London SE1 0AA

Abstract. This paper addresses the problem of determine if a drug is or not bioactive, from a dataset of structure and binding to a target site on a receptor. The application problem is model the binding to thrombin through a mathematical discriminate function using artificial intelligence algorithm (genetic programming-GP). The algorithm consist of a two steps processing: the training, where a set of binding description and effects is used to obtain the discriminate function and then application of this function to undetermined compounds to determine their bioactivity.

1. Introduction.

Globus described the problem of custom designed molecules. Frequently it is possible to precisely define what a molecule must do and still have significant problems designing a molecule to do the task.

It is known that some drugs fit precisely into receptor sites to block molecular processes in the body. This must be accomplished without fitting the receptor sites of the body's healthy molecular machinery. Furthermore, drug molecules must survive in the body long enough to be effective. Early drug discovery was accomplished without understanding these mechanisms, but modern drug designers often consciously create molecules with atomic precision to bind well to receptor sites.

One approach to drug design is to find molecules similar to good drugs that have fewer negative side effects. Ideally, a candidate replacement drug is sufficiently similar to have the same beneficial effect but is different enough to avoid the side effects.

Therefore, a design technique to automatically generate candidate molecules given requirements may be useful, determining what separates the active (binding) compounds from the inactive (non-binding) ones. Such a determination can then be used in the design of new compounds that not only bind, but also have all the others properties required for a drug (solubility, oral absorption, lack of side effects, appropriate duration of action, toxicity, etc).

Cup 2001 of The Seventh ACM SIGKDD International Conference on Knowledge discovery and data mining – USA/ San Francisco, August 26-29,2001.
http://www.cs.wisc.edu/~dpage/kddcup2001/

This paper proposes mining the information available to extract a discriminate function that model the boundary between activity and inactivity region and apply it to unknown compounds to find their activity.

2. The modeling algorithm.

The genetic programming (GP) algorithm (Holland 1975, Koza 1992) mimics the evolution and improvement of life through reproduction, when each individual contributes with its own genetic information to building a new individuals with greater fitness to the environment and higher chances of survival. Each 'individual' in a generation represents, with its chromosome, a feasible solution to the problem; in our case, a discriminate function to be evaluated by a fitness function.

The best individuals are continuously being selected, and crossover and mutation take place. Following a number of generations, the population converges to the solution that best represents the discrimination function.

There are two kinds of information defined for the algorithm: terminals (variable values and random numbers) and functions (mathematical functions used in the generated model).

The final GP output is a model or a function (we term discriminate function) of the process under study, with would be the binding activity, cancer diagnostic (Werner 2001), collagen & thrombosis diagnostic (Werner 2001), etc.

3. Processing description.

The software we have developed is an adaptation of LilGP (see reference), where GP is structured in a pre-compiled library, with other artificial intelligence procedures (such as genetic algorithm (GA), an adaptive algorithm (AA), neural networks (NN), and fuzzy control (FC)), integrating the model obtaining process with GP from with real time adaptation by the GA, AA, NN, or FC. This has been applied to the KDD Cup 2001 dataset "Prediction of molecular bioactivity for Drug design – Binding to Thrombin" – see KDD reference.

A filter reads the information of 1909 compounds tested for their ability to bind to a target site on thrombin, a key receptor in blood clotting. Of theses compounds 42 are active (bind well) and the others are inactive, and are described by a single feature vector comprised of a class value (A for active, I for inactive) and 139,351 binary features, which describe three-dimensional properties of the molecule.

The filter stores the pointer to positive features only and its activity and fitness function evaluates accuracy of mathematical relation (discriminate function) between features and activity. Further discriminate function is ready to be applied to obtain unknown activity.

Cup 2001 of The Seventh ACM SIGKDD International Conference on Knowledge discovery and data mining – USA/ San Francisco, August 26-29,2001.
http://www.cs.wisc.edu/~dpage/kddcup2001/

4. Binding activity obtained by discriminate function.

The GP parameters are 20 individuals and 17000 generations, 60% crossover probability and 20% mutation probability. The functions are: multiplication, division, add, and subtraction. The terminals are the binary features (BinF(x), an indexed vector of (139351*x)%x) and random numbers between 0 and 1.

4.1 Genetic Programming Training

The fitness function to evaluate the performance of each individual of GP generation is:

$$Fitness = \frac{OK}{OK + contI * NOK\_A / contA + NOK\_I} \quad (1)$$

where OK is the right activity prediction, NOK_A is the wrong prediction of active compound, NOK_I is wrong prediction of inactive compound, contI is the total number of inactive compounds and contA is the total number of active compounds. Table 1 presents training and test results, annex 1 present the discriminate function, coded in tree notation and table 2 total comparative results.

	Actual		Predicted value		% hit the mark
			I	A	
Training	I	1866	1815	51	**97**
	A	42	15	27	64
Test	I	484	203	280	**42**
	A	150	56	94	62

Table I. Thrombin dataset results.

	%hit the mark	%wrong A	%wrong I
Training	96	0.78	2.67
Test	47	44	8.83

Table 2: comparative results.

5. Future works and conclusion.

Through the use of GP we show that it is possible to find a discriminate function that predict compounds bioactivity with a specific shape, such as less Active errors with the concomitant cost of a decrease in inactive accuracy. The application of this approach would be improved during its own application, because each confirmed prediction would be included into the training dataset, closing the loop.

Cup 2001 of The Seventh ACM SIGKDD International Conference on Knowledge discovery and data mining – USA/ San Francisco, August 26-29,2001.
http://www.cs.wisc.edu/~dpage/kddcup2001/

References.

GLOBUS, A.; LAWTON,J.; WIPKE, T; "Automatic molecular design using evolutionary techniques"
http://www.nas.nasa.gov/Pubs/TechReports/NASreports/NAS-99-005/NAS-99-005.pdf
http://www.nas.nasa.gov/~globus/papers/Nanotechnology98/paper.html
HOLLAND,J.H. "Adaptation in natural and artificial systems: na introductory analysis with applications to biology, control and artificial intelligence." Cambridge: Cambridge press 1992 reedição 1975.
KDD The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining http://www.acm.org/sigs/sigkdd/kdd2001/ KDD Cup 2001 http://www.cs.wisc.edu/~dpage/kddcup2001/
KOZA,J.R. "Genetic programming: On the programming of computers by means of natural selection." Cambridge,Mass.: MIT Press, 1992.
LilGP "Genetic Algorithms Research and Applications Group (GARAGe)", Michigan State University; http://garage.cps.msu.edu/software/lil-gp/lilgp-index.html
Werner,J.C.; "Active noise control in ducts using genetic algorithm" PhD. Thesis - São Paulo University- São Paulo-Brazil-1999.
Werner,J.C.; Fogarty,T.; "Genetic programming applied to severe diseases diagnosis" Proceedings of Intelligent Data Analysis in Medicine and Pharmacology, London,UK, September 4th,2001 IDAMAP2001.
Werner,J.C.; Fogarty,T.; "Genetic programming applied to Collagen disease & thrombosis" Discovery Chalenges at PKDD2001 (Freiburg,Germany, Sep.3-7,2001) www.uncc.edu/knowledgediscovery/DataDesc.htm

Annex 1. Discriminate function for thrombin binding.

(- (/ 0.37 (- (* 0.55 (- (* 0.99 (/ (* (* (- 0.87 0.41) (* 0.57 0.88)) (BinF (+ 0.60 0.73))) (BinF (- (- 0.08 0.50) (- 0.30 0.75))))) (+ (/ (- (* 0.40 0.47) 0.57) (- (* 0.40 (BinF (+ (* (- (BinF 0.91) (BinF 0.23)) 0.57) (BinF (- (/ (* (* 0.55 (- (BinF (* 0.78 0.10)) 0.57)) 0.57) 0.76) (* 0.86 0.53)))))) 0.57)) (* (BinF (BinF (* 0.78 0.10))) (- 0.05 (- (- 0.00 0.59) (* (- (* 0.40 (BinF (* 0.78 (* (/ 0.35 0.57) (/ 0.72 0.74))))) 0.57) 0.04))))))) (/ (- (* (- (BinF 0.21) (+ 0.02 0.87)) (BinF (* 0.78 0.10))) (* 0.40 (BinF (- (- 0.65 (* 0.86 0.53)) (+ 0.91 0.45))))) (- (* 0.40 (BinF (* 0.78 (* 0.97 0.04)))) 0.57)))) (+ (- (- (/ 0.37 (+ (/ (/ (- (* 0.40 (BinF (* (/ 0.72 0.74) 0.57))) 0.57) 0.17) 0.76) (/ (- (BinF 0.91) (BinF 0.23)) 0.35))) (- (- (/ 0.35 0.57) 0.57) (- (* (/ (* (* 0.55 (- (BinF (* 0.78 0.10)) 0.57)) 0.57) 0.76) (- 0.78 0.16)) (- 0.05 (- (- 0.00 0.59) (* 0.97 0.04)))))) (/ (- (* 0.40 0.95) 0.57) (- (/ (BinF (- (BinF 0.21) (- (/ 0.35 0.40) 0.57))) (/ (+ 0.93 0.36) (* 0.42 0.82))) 0.57))) 0.45))