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

Genetic programming applied to gene function identification.

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 how to obtain the gene function 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 gene description and its function is used to obtain the discriminate function and then application of this function to undetermined genes to determine their function.

1. *Introduction.*

As stated in the literature (see Brunel), the human genome project is officially completed. Out of all the genes sequenced, approximately 40% of these genes code for a protein that has an unknown function.

Knowing protein structure can lead to the deduction of its function, which can then be verified. Current experimental techniques of protein structure determination can be costly and time consuming. The requirement for a protein to be crystallized to undergo X-ray crystallization complicates matters when a protein cannot form crystals. Not all institutions will have a NMR (Nuclear Magnetic Resonance) facilities, which does not require a protein to be crystallized. However, 3-D NMR works best with small proteins or protein fragments. Therefore, some proteins cannot be structurally characterized using current experimental methods.

This paper proposes mining the information available to extract the discriminate function that model the features that characterize the function and apply it to unknown genes, obtaining their function without experimental techniques.

2. *The modelling 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 discriminate 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.

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 gene/protein function" – see KDD reference.

A filter reads the gene data and stores it in a memory record with the previous treatment for each field:

- ESSENTIAL / NON-ESSENTIAL / AMBIGUOUS-ESSENTIAL: attribute value 1.0 to the respective flag float variable, 0.7 to ambiguous or 0.5 to both if unknown.
- CLASS, COMPLEX, PHENOTYPE, MOTIF: attribute value 1.0 to the respective flag float variable, or 0.5 to both if unknown.
- CHROMOSOME: attribute 1.0 value to the respective flag float variable.

The interacting information feeds a relation record with pointers to genes involved with a specific function. The first element is the same gene into both positions and correlation factor 1.0 and relation PHYSICAL. The following entries contain the pointers to each gene information, the type of interaction (PHYSICAL (1), GENETIC (2) and Genetic-Physical(3)) and its correlation factor, totalising 5219 records.

The information used to training the algorithm is if the gene acts (NO-YES) into the functions: CELL GROWTH CELL DIVISION AND DNA SYNTHESIS (0), CELL RESCUE DEFENSE CELL DEATH AND AGEING (1), CELLULAR BIOGENESIS (proteins are not localized to the corresponding organelle) (2), CELLULAR COMMUNICATION / SIGNAL TRANSDUCTION (3), CELLULAR ORGANIZATION (proteins are localized to the corresponding organelle) (4), CELLULAR TRANSPORT AND TRANSPORT MECHANISMS (5), ENERGY (6), IONIC HOMEOSTASIS (7), METABOLISM (8), PROTEIN DESTINATION (9), PROTEIN SYNTHESIS (10), TRANSCRIPTION (11), TRANSPORT FACILITATION (12), TRANSPOSABLE ELEMENTS VIRAL AND PLASMID PROTEINS (13). The numbers indicate the sequence and are used into the tables.

Terminals are defined with gene features and random number between 0 and 1.

The functions are:

- AND: the product of both probabilities.
- OR: the average of both probabilities.
- NOT: 1-probability.
- XOR: if only one input >=0.5, the output is this value, otherwise 0.
- NOR : NOT OR
- NAND: NOT AND
- XNOR: if both >=0.5 or both <0.5 output 1.0, otherwise 0.

Fitness function evaluates accuracy of mathematical relation (discriminate function) between interactions, gene information, and gene function. Further discriminate function is ready to be applied to obtain unknown genes function.

4. *Gene function obtained by discriminate function.*

The GP parameters are 30 individuals and 300.000 generations, 60% crossover probability and 20% mutation probability.

Genetic Programming Training

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

$$Fitness = \frac{OK\_NO}{OK\_NO + NOK\_NO} + \frac{OK\_YES}{OK\_YES + NOK\_YES}$$

where OK is the right prediction for gene action and NOK is the wrong predictions. NO and YES inform if the gene act into the function.
Table 1 presents training results for each different function and the discriminate function for each function is listed in annex 1.

Function	Actual	Prediction		Case hit the mark (%)	Total hit the mark		Fitness	Generation of Solution
		No	Yes		#	%		
0	No	3033	329	90	4074	78	1.4624	192,943
	Yes	816	1041	56				
1	No	4069	724	84	4331	82	1.4640	99,227
	Yes	164	262	61				
2	No	3545	1348	72	3772	72	1.4207	13,659
	Yes	99	227	69				
3	No	4361	548	88	4539	86	1.4626	62,048
	Yes	132	178	57				
4	No	127	25	83	2246	43	1.2536	112,197
	Yes	2948	2119	41				
5	No	3783	144	96	4339	83	1.3937	97,303
	Yes	736	556	43				
6	No	4714	325	93	4817	92	1.5077	48,458
	Yes	77	103	57				
7	No	4752	351	93	4809	92	1.4226	61,931
	Yes	59	57	49				
8	No	2124	2256	48	2777	53	1.2635	69,753
	Yes	186	653	77				
9	No	2409	1862	56	3060	58	1.2506	38,473
	Yes	297	651	68				
10	No	4869	48	99	5109	97	1.7849	8919
	Yes	62	240	79				
11	No	3117	239	92	4272	81	1.5484	30,100
	Yes	708	1155	61				
12	No	4651	412	91	4771	91	1.6879	52,828
	Yes	36	120	76				
13	No							
	Yes	0	0					

Test dataset results.

The test dataset prediction, with information do not be used during training, are presented in table 2.

Function	Actual	Prediction		Case hit the mark (%)	Total hit the mark
		No	Yes		%
0	No	218	46	82	80
	Yes	30	87	74	
1	No	266	82	76	74
	Yes	15	18	54	
2	No	173	182	48	51
	Yes	3	23	88	
3	No	281	92	75	75
	Yes	3	5	62	
4	No	12	6	66	57
	Yes	155	208	57	
5	No	280	19	93	84
	Yes	41	41	50	
6	No	328	37	89	88
	Yes	8	8	50	
7	No	315	58	84	83
	Yes	4	6	60	
8	No	107	210	33	42
	Yes	9	55	85	
9	No	74	214	25	38
	Yes	19	74	79	
10	No	345	10	97	96
	Yes	3	23	88	
11	No	212	49	81	81
	Yes	21	99	82	
12	No	284	76	78	77
	Yes	8	13	61	
13	No	380	0	100	99
	Yes	1	0	0	

5. *Future works and conclusion.*

Through the use of GP we show that it is possible to find a discriminate function that predict the gene action into some function, without experimental equipment. The results accuracy depends how many information available for each function, and what the available information of interactions and genes.

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.

References.

BRUNEL "An introduction to Protein Prediction Servers" http://http1.brunel.ac.uk:8080/depts/bl/project/biocomp/mak_fan/intro.htm

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

ROC Receiver Operating Characteristic (ROC) curve analysis http://www.medcalc.be/roccman.html

Werner,J.C.; "Active noise control in ducts using genetic algorithm" PhD. Thesis- São Paulo University- São Paulo-Brazil-1999.