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

1. *Introduction.*

Genes code for proteins, and these proteins tend to localize in various parts of cells and interact with one another, in order to perform crucial functions. The present study of a variety of details about the various genes of one particular type of organism. Gene names have been anonymized and a subset of the genes have been withheld for testing. The task is to predict the localizations of the proteins encoded by the genes. A gene/protein at least in this study has only one localization.

This paper proposes mining the information available to extract the discriminate function and apply it to unknown genes, obtaining their localization 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 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.

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 (a second step of the work, as described in "Future works and conclusion"). This has been applied to the KDD Cup 2001 dataset "Prediction of gene/protein localization" – 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 the location of protein into the cell: cell wall (0), cytoplasm (1), cytoskeleton (2), endosome (3), ER (4), extracellular (5), golgi (6), integral membrane (7), lipid particles (8), mitochondria (9), nucleus (10), peroxisome (11), plasma membrane (12), transport vesicles (13), and vacuole (14). 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 location. 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 300000 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:

$$Fitness = \frac{OK_Loc}{OK_Loc + NOK_Loc}$$

where OK is the right prediction of position and NOK is the wrong predictions, and Loc means the localisation under study. Table 1 presents training results for each different function.

Location	Total # loc	# right local	% right local	#total right	% total right
0	2	2	100	5219	100
1	1088	264	24	4343	83
2	458	249	54	4927	94
3	16	0	0	5203	99
4	200	59	29	5069	97
5	6	0	0	5213	99
6	294	149	50	4991	95
7	14	0	0	5205	99
8	2	2	100	5219	100
9	266	107	40	5043	96
10	2417	1252	51	3982	76
11	56	7	12	5169	99
12	234	39	16	5017	96
13	78	13	16	5151	98
14	88	13	14	5141	98

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	0	381	0	0
	Yes	0	0		
1	No	309	6	98	86
	Yes	45	21	31	
2	No	327	24	93	92
	Yes	6	24	80	
3	No	381	0	100	100
	Yes	0	0		
4	No	350	6	98	95
	Yes	13	12	31	
5	No	379	0	100	99
	Yes	2	0	0	
6	No	356	13	96	94
	Yes	7	5	41	
7	No	381	0	100	100
	Yes	0	0		
8	No	379	2	99	99
	Yes	0	0		
9	No	344	2	99	92

	Yes	25	10	28	
10	No	195	12	94	74
	Yes	84	90	51	
11	No	375	3	99	98
	Yes	2	1	33	
12	No	360	3	99	94
	Yes	17	1	5	
13	No	367	5	98	96
	Yes	8	1	11	
14	No	370	3	99	97
	Yes	8	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 localization, without experimental equipment. The results accuracy depends how many information available for each localization, 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.

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.