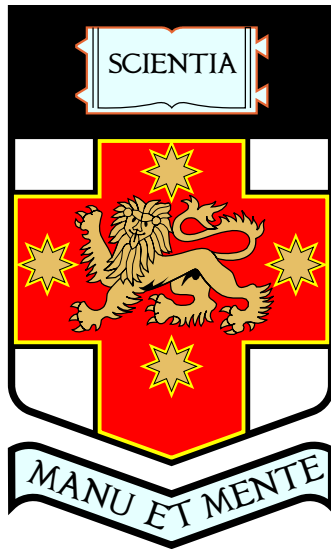


A Flexible Representation for Genetic Programming: Lessons from Natural Language Processing



A thesis submitted to the
School of Information Technology and Electrical Engineering
University College
University of New South Wales
Australian Defence Force Academy
for the degree of Doctor of Philosophy

By
Nguyen Xuan Hoai
December 2004

© Copyright 2004 by Nguyen Xuan Hoai

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgment is made in the thesis. Any contribution made to the research by colleagues, with whom I have worked at UNSW or elsewhere, during my candidature, is fully acknowledged.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Nguyen Xuan Hoai

Abstract

This thesis principally addresses some problems in genetic programming (GP) and grammar-guided genetic programming (GGGP) arising from the lack of operators able to make small and bounded changes on both genotype and phenotype space. It proposes a new and flexible representation for genetic programming, using a state-of-the-art formalism from natural language processing, Tree Adjoining Grammars (TAGs). It demonstrates that the new TAG-based representation possesses two important properties: non-fixed arity and locality. The former facilitates the design of new operators, including some which are bio-inspired, and others able to make small and bounded changes. The latter ensures that bounded changes in genotype space are reflected in bounded changes in phenotype space.

With these two properties, the thesis shows how some well-known difficulties in standard GP and GGGP tree-based representations can be solved in the new representation. These difficulties have been previously attributed to the tree-based nature of the representations; since TAG representation is also tree-based, it has enabled a more precise delineation of the causes of the difficulties.

Building on the new representation, a new grammar guided GP system known as TAG3P has been developed, and shown to be competitive with other GP and GGGP systems.

A new schema theorem, explaining the behaviour of TAG3P on syntactically constrained domains, is derived.

Finally, the thesis proposes a new method for understanding performance differences between GP representations requiring different ways to bound the search space, eliminating the effects of the bounds through multi-objective approaches.

Acknowledgments

The first person I would like to thank is my principal supervisor, Dr Robert (Bob) Ian McKay for introducing me to the field of genetic programming. Bob has been the ideal mentor that I was looking for. He taught me how to love research and how not to be a research-alcoholic. His genius has been a constant source of help. His encouragement and constructive criticism have been the triggers for much of the research work in this thesis.

I wish also to thank my co-supervisors Dr Daryl Essam and Dr Hussein Aly Abbass. It is a pleasure and luxury for me to work with such brilliant people during my candidature. The seemingly never-ending, but always useful, discussions between us on the matters related to the work presented in this thesis will be in my mind forever.

I would like to express my gratitude to Professor Jason Daida for sending some of the figures in his papers on the problem of structural difficulty in GP and for also allowing me to reproduce them in chapter 8 of this thesis. I must also thank Dr Zitzler for promptly replying to my questions related to the implementation of his SPEA2 algorithm.

During my candidature, I have had opportunities to attend a number of top academic conferences in the field, such as: CEC, EuroGP and GECCO. I wish to thank the school of IT&EE, University of New South Wales at the Australian Defence Force Academy, for providing me the necessary funding to go and present my research papers at those conferences. While attending those conferences, I was lucky to have had a number of useful and interesting discussions with a number of researchers in the field. The exchange of ideas and information with them was

very important to my research. For that reason, I wish to thank those people who were generous to me by giving information, suggestions, and discussions related to my work and research interests. I would like to mention some of their names in no particular order: Bill Langdon, Michael O’Neil, Man Leung Wong, Una-O’Reilly, Peter Whigham, Vic Ciesielski, Ivan Tanev, Karl Lehre, Sung-Bae Cho, Andras Kornai. Also, when I asked for people who have done some work related to grammar-guided genetic programming over the GP mailing list, a number of respondents stood out and sent me information about their work. Thank you.

During my period of three and a half years at ADFA, I shared my room with Shan Yin, a nice and brilliant Chinese PhD student, whom I share interest in research, modern Chinese history and politics. I wish to thank him for sharing with me many interesting discussions both related and not related to research. Hoang Tuan Hao, my former student, must also be thanked for helping me to draw some of the figures in this thesis.

I am also indebted to Professor Nguyen Xuan Huy at the Vietnamese Institution of Information Technology and Professor Pham The Long, Vice Rector of Vietnamese Military Technical Academy for their care, support, and encouragement when I was a university student in Vietnam.

Last, but most important, I would like to dedicate this work to my family, my wife, Nguyen Thi Thu Huyen for having gone side by side with me through all the ups and downs in life for the last five years, and to my parents, Nguyen Tai and Nguyen Thi Lai, for working so hard to bring up their son and for never losing the hope, that one day, he will become one of the best in his generation.

Contents

Certificate of Originality	i
Abstract	ii
Acknowledgments	iii
1 Introduction	1
1.1 Motivation	1
1.2 Statement of the Thesis	3
1.3 Outline of Dissertation	4
2 Related Work	7
2.1 Evolutionary Algorithms	7
2.2 The Genetic Algorithm	8
2.3 Genetic Programming	10
2.3.1 Standard tree-based GP	11
2.3.2 Some Problems with Tree-Based Genetic Programming . .	15
2.3.3 GP with Linear Representation	16
2.4 Grammar Guided Genetic Programming	20
2.4.1 Some Early Systems	22
2.4.2 Grammar Guided Genetic Programming with Tree-Based Representation	23
2.4.3 Grammar-Guided Genetic Programming with Linear Rep- resentation	27

2.4.4	Some Problems with Grammar-Guided Genetic Programming	30
2.4.5	Applications of Grammar-Guided Genetic Programming	31
2.5	Conclusion	31
3	TAG-based Representation	33
3.1	Tree Adjoining Grammars	34
3.1.1	Definitions of Tree Adjoining Grammars	35
3.1.2	Derivation Trees in Tree Adjoining Grammars	39
3.1.3	Some Properties of TAGs	44
3.2	Tree Adjoining Grammar Based Representation for Genetic Programming	48
3.2.1	Representation in Evolutionary Algorithms	48
3.2.2	TAG-based representation	51
3.2.3	A Working Example of TAG-based representation	54
3.2.4	Why TAG-based representation ?	55
3.3	Conclusion	57
4	TAG3P System	58
4.1	Introduction	58
4.2	The Components of a Tree Adjoining Grammar Guided Genetic Programming (TAG3P)	59
4.2.1	Program Representation	60
4.2.2	Initialisation Procedure	62
4.2.3	Fitness Evaluation	65
4.2.4	Main Genetic Operators	65
4.2.5	Other Operators	72
4.2.6	Parameters	76
4.3	Some Information on TAG3P Implementation	77
4.4	Conclusion	77

5	TAG3P: Preliminary Comparisons	79
5.1	Test Problems	79
5.1.1	Simple Symbolic Regression Problem	80
5.1.2	6-Multiplexer Problem	81
5.1.3	Symbolic Integration Problem	81
5.1.4	Symbolic Differentiation Problem	82
5.2	Experiment Setup	82
5.3	Results and Discussion	85
5.4	Some further analyses on TAG3P	89
5.5	Conclusion	94
6	Some Operators	95
6.1	Insertion and Deletion Operators	96
6.1.1	Description of Insertion and Deletion Operators	97
6.1.2	Experiments	99
6.1.3	Insertion and Deletion Conclusion	113
6.2	Relocation Operator	113
6.2.1	Description of the Relocation Operator	115
6.2.2	Experiments	117
6.2.3	Relocation Operator Conclusion	126
6.3	Duplication Operator	126
6.3.1	Description of Duplication Operator	128
6.3.2	Experiments	129
6.3.3	Duplication Operator Conclusion	136
6.4	Conclusion and Future Work	137
7	A Schema Theory for TAG3P	139
7.1	Schema Theory in GP	140
7.2	Schema Theory in GGGP	141
7.3	Schema Definition in TAG3P	142
7.4	A Schema Theorem for TAG3P	144

7.4.1	Expected Number of Individuals Matching a Schema after Selection	145
7.4.2	Schema Disruption due to Genetic Operators	145
7.4.3	A Schema Theorem for TAG3P	146
7.5	Conclusion	148
8	Structural Difficulty in GP	149
8.1	Structural Difficulty in Genetic Programming	150
8.2	Insertion and Deletion Operators and the Structural Difficulty in Genetic Programming	154
8.3	Experiments	155
8.3.1	Experiment Setup	157
8.3.2	Results and Discussion	157
8.4	Conclusion	160
9	Fitness Landscape Study	162
9.1	Fitness Landscape Study in EAs	163
9.2	Problems with GP and GGGP in Fitness Landscape Study	164
9.3	TAG-based Representation and Fitness Landscape Study	166
9.4	Experiments	167
9.4.1	Experiment Setup	171
9.4.2	Results and Discussion	172
9.5	Conclusion	177
10	EMO comparison	179
10.1	Difficulties in Making Meaningful Comparison between Genetic Programming Systems	180
10.2	Multi-objective Evolutionary Optimization	181
10.3	The Use of Multi-objective Techniques for Comparisons	182
10.4	Experiments and Results	184
10.4.1	Experiment Design	184

10.4.2	Results	185
10.4.3	Discussion of Results	186
10.5	Conclusion	192
11	Conclusions	194
11.1	Contributions of this thesis	195
11.2	Future Work	197
11.2.1	Future Work on TAG-based Representation	197
11.2.2	Future Work on TAG3P	199
A	Some More Information on Grammars	201
A.1	Context Free Grammars	201
A.2	Attribute Grammars	203
A.3	Definite Clause Grammars (DCGs)	203
A.4	Definite Clause Translation Grammars	204
A.5	Schabes's Lexicalisation Algorithm	204
A.6	The Grammars for Some Problems in the Thesis	206
A.6.1	Some Grammars for the Problems in Chapter 6	206
A.6.2	Some Grammars for the Problems in Chapter 9	208
B	Schema Theory	211
B.1	Introduction	211
B.2	Holland's Schema Theorem for GAs	211
B.3	Fixed Shape and Size Schema Theorem for GP by Poli and Langdon	213
B.4	Whigham's Schema Theorem for CFG-GP	215
C	Techniques for Fitness Landscape Analysis	217
C.1	Correlation Analysis of Fitness Landscapes	217
C.2	Information Content Measures for Fitness Landscape	219
D	SPEA2	222
E	Some Supplementary Figures	225

List of Figures

2.1	The basic flow chart for evolutionary algorithms (EAs) [Fog1995].	8
2.2	Representation for individuals in standard GA	8
2.3	Basic genetic operators in standard GA	9
2.4	Lisp-expression tree representation in standard GP, where $F =$ (<i>AND, OR, NOT</i>) and $T = (a, b, c)$	12
2.5	Crossover operator in standard GP	14
2.6	Mutation operator in standard GP	14
2.7	Genotype and Phenotype in GEP, where AND=0, OR=1, NOT=2, a=3, b=4, c=5.	18
2.8	Genotype to phenotype map in GEP is not causal. The top-left corner is an genotype and the top-right corner is its corresponding phenotype. Just by one mutation taken in the first gene (bottom- left corner), the phenotype has changed completely (bottom-right corner)	19
2.9	Strongly Typed GP	20
2.10	Derivation tree as programs in GGGP	24
2.11	Crossover operator in GGGP	25
2.12	Mutation operator in GGGP	26
2.13	An example of GE genotype-phenotype mapping	28
3.1	A simple TAG for some English sentences	36
3.2	Adjunction Operation.	37
3.3	Substitution	37

3.4	Weir's TAG derivation tree	40
3.5	Schabes-Shielber TAG derivation tree	41
3.6	Joshi-Schabes TAG derivation tree	42
3.7	Example of the new form of TAG derivation tree.	44
3.8	Another simple TAG for some English sentences	47
3.9	An example of ELD in TAGs	48
3.10	An example of FRD in TAGs	48
3.11	An example of the conflict between tree alignment metric and the intuitive sense of the similarity between two trees. The left and the right trees are very similar in the intuitive sense but far from each other using tree alignment metric	52
3.12	The elementary trees for G_{lex} . L1 is a lexicon that can be substituted with any lexeme in (p,n,s,t,l,o,co,chla,r1,r2)	55
3.13	Translation from a genotype to a phenotype. The first four parts are the intermediate steps. The final genotype and final phenotype are given at the bottom of the figure	56
4.1	Mapping Process	61
4.2	Crossover in TAG3P	68
4.3	Subtree mutation in TAG3P	71
4.4	Node replacement in TAG3P	74
4.5	Truncation in TAG3P	75
5.1	A 6-multiplexer and a solution	81
5.2	TAG elementary trees for symbolic regression, symbolic integration, and symbolic differentiation problems	83
5.3	TAG elementary trees for the 6-multiplexer problem	84
5.4	Cumulative Frequencies for Symbolic Regression Problem	87
5.5	Cumulative Frequencies for 6-Multiplexer Problem	88
5.6	Cumulative Frequencies for Symbolic Integration Problem	88
5.7	Cumulative Frequencies for Symbolic Differentiation Problem	89

5.8	Average Fitness of the population	90
5.9	Average fitness of the best in the population	90
5.10	Evolution of size in the runs for the symbolic regression problem .	92
5.11	Evolution of size in the runs for the 6-Multiplexer problem	93
6.1	Insertion Operator	98
6.2	Deletion Operator	99
6.3	Cumulative Frequencies for ORDER Problems	104
6.4	Cumulative Frequencies for MAJORITY Problems	105
6.5	Cumulative Frequencies for symbolic regression Problem	105
6.6	Cumulative Frequencies for 6-Multiplexer Problem	106
6.7	Cumulative Frequencies for ORDER Problems	108
6.8	Cumulative Frequencies for MAJORITY Problems	109
6.9	Cumulative Frequencies for symbolic regression Problem	109
6.10	Cumulative Frequencies for 6-Multiplexer Problem	110
6.11	Relocation Operator	116
6.12	Cumulative frequencies for four problems	121
6.13	Cumulative Frequencies for four problems	123
6.14	Duplication Operator	129
6.15	Cumulative frequencies for X4-X9	132
6.16	Cumulative frequencies for X4-X9	134
7.1	An example for Whigham’s schema, where a schema can match more than once in an individual	142
7.2	An example for a schema on TAG-based representation	143
8.1	Four regions in the space of tree structures. Reprinted with per- mission from [Dai2003b]	151
8.2	The “horizontal and vertical cuts”. Reprinted with permission from [Dai2003b]	152

8.3	Proportion of Success for GP on the “Horizontal cut”. Reprinted with permission from [Dai2003b] The lower half of the figure shows which part of the structure space the problem instances belong to. It can be seen that the hard-to-find instances are those in region II, and III	153
8.4	Proportion of Success for GP on the “Vertical cut”. Reprinted with permission from [Dai2003b]. The lower half of the figure shows which part of the structure space the problem instances belong to. It can be seen that the hard-to-find instances are those in region II, and III	154
8.5	Elementary trees of G_{lex} for LID problem	156
8.6	Results of TAG-HILL on the “horizontal cut”	158
8.7	Results of TAG-HILL on the “vertical cut”	159
8.8	Average number of fitness evaluations for the ”horizontal cut” . . .	160
8.9	Average number of fitness evaluations for the ”vertical cut”	161
9.1	Graph of functions $F_1(X3)$ to $F_4(X6)$	174
10.1	Cumulative Frequencies (POPSIZE=500)	187
10.2	Average First Fitness (POPSIZE=500)	187
10.3	Average Second Fitness (POPSIZE=500)	188
10.4	Tree Size Frequencies for Symbolic Regression Problem, POPSIZE=500	190
10.5	Tree Size Frequencies for 6-multiplexer Problem, POPSIZE=500 . .	191
A.1	An example of CFG derivation sequences and trees	203
A.2	Elementary trees for the grammar of the ORDER and MAJORITY problems	207
A.3	Elementary trees for the grammar of the SEXTIC and QUINTIC problems	207
A.4	Elementary trees for the grammar of the TRIGO problems	208
A.5	Elementary trees for the grammar of the TWOBBOX problem . . .	208

A.6	TAG elementary trees for $G1_{lex}$	209
A.7	TAG elementary trees for $G2_{lex}$	209
A.8	TAG elementary trees for $G3_{lex}$	210
A.9	TAG elementary trees for $G4_{lex}$	210
B.1	An example of a fixed-shape and -size schema in GP expression tree representation	214
B.2	An example for a CFG-GP schema. The Schema is $F \Rightarrow A A$. . .	216
E.1	Cumulative Frequencies (POPSIZE=250)	225
E.2	Cumulative Frequencies (POPSIZE=1000)	226
E.3	Average First Fitness (POPSIZE=250)	226
E.4	Average First Fitness (POPSIZE=1000)	226
E.5	Average Second Fitness (POPSIZE=250)	227
E.6	Average Second Fitness (POPSIZE=1000)	227
E.7	Tree Size Frequencies for Symbolic Regression Problem, POP- SIZE=250	228
E.8	Tree Size Frequencies for 6-multiplexer Problem, POPSIZE=250 .	229
E.9	Tree Size Frequencies for Symbolic Regression Problem, POP- SIZE=1000	230
E.10	Tree Size Frequencies for 6-multiplexer Problem, POPSIZE=1000	231