

A FRAMEWORK FOR THE INTEGRATION OF INFORMATION RETRIEVAL AND PARSE TREE
DATABASE WITH APPLICATIONS IN THE GENOMICS DOMAIN

by

Luis Babaji Ng Tari

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

ARIZONA STATE UNIVERSITY

December 2009

UMI Number: 3391866

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.

UMI[®]

Dissertation Publishing

UMI 3391866

Copyright 2010 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.

ProQuest[®]

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

A FRAMEWORK FOR THE INTEGRATION OF INFORMATION RETRIEVAL AND PARSE TREE
DATABASE WITH APPLICATIONS IN THE GENOMICS DOMAIN

by

Luis Babaji Ng Tari

has been approved

August 2009

Graduate Supervisory Committee:

Chitta Baral, Chair
Yi Chen
Hasan Davulcu
Seungchan Kim
Huan Liu

ACCEPTED BY THE GRADUATE COLLEGE

ABSTRACT

With the ever increasing number of biomedical articles, keeping up with new information has become a big challenge for biomedical researchers. Much of the information biologists need resides in semi-structured biomedical text articles, making it difficult for researchers to realize the full benefits of these findings. Information retrieval (IR) and information extraction (IE) have been the central technologies for seeking information from large corpora of unstructured text. Advances in these technologies can have a direct impact to the research methodologies for research areas such as biomedical research.

While the fields of IR and IE have matured in the past decade, current technologies still have yet to fulfill the promise of supporting biomedical research. In particular, traditional IE technologies adopt a 'black-box' approach, in which biologists have no means in expressing their extraction needs. In addition, typical automated IE technologies rely on manually curated data to learn syntactic patterns for extraction. However, curation of such data is known to be labor-intensive, limiting the applicability of IE in the biomedical domain. While there have been successes in utilizing linguistic structures for IE, linguistic structures have yet to be adopted in the current technologies for IR. Syntactic parsing over large corpus of text is known to be computationally expensive, and this is not ideal for IR, which is expected to respond to users in a timely manner. However, the lack of usage of linguistic structures leads to suboptimal performance for certain queries in the biomedical domain.

In this thesis, these issues in IR and IE are tackled by proposing a novel framework called *IR+PTQL*. The core idea of the framework is to model and store the syntactic and semantic information of the text corpora in a specialized database called the parse tree database. Extraction is then expressed in the form of database queries. A core component is the automated query generation that generates extraction patterns without training data. The evaluation results demonstrate that the query generation component contributes positively to the performance of IR and IE. The applicability of the framework is illustrated with various applications in the genomics domain.

PREVIEW

This dissertation is dedicated to my parents Madeu Babaji Tari and Maria Fatima Ng Tari, and my sisters
Ana Tari Ashizawa and Mila Tari Wasson, who have been very supportive with my PhD degree.

ACKNOWLEDGMENTS

I would like to express my gratitude to my thesis adviser Dr. Chitta Baral for his guidance and patience throughout my degree. Dr. Baral's creativity and inspiration in research serves as a role model in how good research should be done. I would also like to thank the committee members for their time in supervising this dissertation. In particular, I want to thank Dr. Yi Chen, Dr. Graciela Gonzalez and Dr. Seungchan Kim for their collaboration in the research.

I am also indebted to the lab members, particularly the two postdocs Dr. Phan Huy Tu and Dr. Jörg Hakenberg for their advises and contribution in this research, as well as Saadat Anwar, Robert Leaman, Shanshan Liang and Võ Hà Nguyễn for their effort in this research. I am fortunate to be surrounded by lots of good friends who have been with me through the good and rough times. Lastly, my words are not enough to express my gratitude to my family for their support, understanding and patience in this long journey.

TABLE OF CONTENTS

	Page
LIST OF TABLES	xi
LIST OF FIGURES	xviii
CHAPTER 1 INTRODUCTION	1
1.1. Information retrieval	1
1.2. Information extraction	3
1.3. DB+IR integration	6
1.4. Overview	8
1.5. Specific research contributions	8
1.5.1. Generic extraction of information from text	9
1.5.2. Automated generation of linguistic queries for information retrieval and extraction	10
1.5.3. Combining knowledge acquisition with logical reasoning for synthesis of biological pathways	11
1.6. Summary	12
1.7. Outline of the dissertation	13
CHAPTER 2 BACKGROUND AND FOUNDATION	14
2.1. Background	14
2.1.1. Answer Set Programming	14
2.1.2. Link Grammar	15
2.1.3. IntEx: a protein-protein interaction extractor	17
2.1.4. Phoenix: extraction based on constituent trees	18
2.2. Foundation	19
2.2.1. Parse tree database	20
2.2.2. TEQL: Text Extraction Query Language	20
2.2.3. Labeling scheme	26

	Page
2.2.4. Query Evaluation	28
CHAPTER 3 INFORMATION EXTRACTION USING DATABASE QUERIES	31
3.1. Introduction	31
3.2. System Architecture	33
3.2.1. Parse tree database and inverted index	34
3.2.2. PTQL: Parse Tree Query Language	39
3.2.3. Optimization of PTQL query evaluation with IR	43
3.3. Query generation	45
3.4. Results	47
3.4.1. Extraction performance for PTQL	48
3.4.2. Time performance for PTQL	49
3.5. Related work	51
3.6. Conclusion	52
CHAPTER 4 AUTOMATED QUERY GENERATION WITH THE IR+PTQL FRAMEWORK	54
4.1. IR+PTQL framework	54
4.1.1. Overview	54
4.1.2. IR+PTQL query language	56
4.1.3. Query evaluation for IR+PTQL queries	58
4.2. Training data driven query generation	62
4.2.1. Method	63
4.2.2. Datasets and results	65
4.2.3. Related work	65
4.2.4. Conclusions	68
4.3. Pseudo-relevance feedback driven query generation	69
4.3.1. Method	73

	Page
4.3.1.1. Sentence retrieval	74
4.3.1.2. LCA finder	76
4.3.1.3. String encoding generation	76
4.3.1.4. Clustering	77
4.3.1.5. PTQL generation	79
4.3.2. Using pseudo-relevance query generation for information retrieval	80
4.3.3. Using pseudo-relevance query generation for information extraction	82
4.3.4. Experimental Results	82
4.3.4.1. Experimental settings for IR	82
4.3.4.2. Evaluation results for IR	83
4.3.4.3. Experimental settings for IE	86
4.3.4.4. Evaluation results for IE	87
4.3.4.5. Time performance	90
4.3.4.6. Analysis of the effects of the parameters	90
4.3.5. Related work	93
4.3.6. Conclusions	96
CHAPTER 5 APPLICATIONS OF THE IR+PTQL FRAMEWORK	97
5.1. Querying parse tree database of Medline text to synthesize user-specific biomolecular networks	97
5.1.1. Method	99
5.1.2. PTQL ^{LITE} : a simplified parse tree query language for users	101
5.1.3. Synthesis of various biomolecular networks	103
5.1.3.1. Drug-enzyme relationship networks	105
5.1.3.2. Gene-disease relationship networks	107
5.1.4. Keyword-based queries with pseudo-relevance query generation	108
5.1.5. Conclusion	111

	Page
5.2. Synthesis of pharmacokinetic pathways through knowledge acquisition and reasoning	113
5.2.1. Introduction	114
5.2.2. Pharmacokinetics	117
5.2.3. Methods	119
5.2.3.1. Fundamentals behind our approach	119
5.2.3.2. Fact and interaction extraction from knowledge bases	122
5.2.3.3. Automated text extraction of facts and interactions	124
5.2.3.4. Ordering of interactions through reasoning	126
5.2.4. Synthesis of pharmacokinetic pathways	130
5.2.4.1. Repaglinide pharmacokinetic pathway	131
5.2.4.2. Pravastatin pharmacokinetic pathway	133
5.2.5. Evaluation and analysis	134
5.2.6. Conclusion	140
CHAPTER 6 CONCLUSIONS	143
6.1. Future work	144
6.1.1. NetSynthesis with keyword queries	144
6.1.2. Ranking of interactions	144
6.1.3. Generalized inference of pathways	145
6.1.4. Extending IR+PTQL framework for development purposes	145
6.2. Summary	146
APPENDIX A PHARMACOKINETIC PATHWAYS	147
A.1. Atorvastatin pharamcokinetic pathway	148
A.2. Clopidogrel pharamcokinetic pathway	154
A.3. Desipramine pharamcokinetic pathway	158
A.4. Erlotinib pharamcokinetic pathway	162

	Page
A.5. Fluoxetine pharamcokinetic pathway	165
A.6. Fluvastatin pharamcokinetic pathway	170
A.7. Gefitinib pharamcokinetic pathway	175
A.8. Ifosfamide pharamcokinetic pathway	180
A.9. Irinotecan pharamcokinetic pathway	185
A.10.Lovastatin pharamcokinetic pathway	193
A.11.Nateglinide pharamcokinetic pathway	198
A.12.Nicotine pharamcokinetic pathway	202
A.13.Omeprazole pharamcokinetic pathway	210
A.14.Phenytoin pharamcokinetic pathway	217
A.15.Pravastatin pharamcokinetic pathway	226
A.16.Repaglinide pharamcokinetic pathway	232
A.17.Rosuvastatin pharamcokinetic pathway	237
A.18.Simvastatin pharamcokinetic pathway	242
A.19.Tamoxifen pharamcokinetic pathway	248
A.20. Warfarin pharamcokinetic pathway	256
APPENDIX B PTQL QUERIES FOR PHARMACOKINETIC PATHWAYS	264
B.1. Drug-enzyme metabolic relations	265
B.2. Expression of genes in liver/intestine	269
B.3. Drug transporters responsible for drug elimination	273
B.4. Drug transporters and the corresponding drugs	274
B.5. Metabolites and the corresponding drugs	274
APPENDIX C ANSPROLOG PROGRAM FOR PHARMACOKINETIC PATHWAYS	278
C.1. AnsProlog program	279
REFERENCES	286

LIST OF TABLES

Table	Page
1. Relational Representation of a Parse Tree	27
2. Relational Representation of a Linkage	27
3. Axes in PTQL Queries and their Translated Conditions in SQL Queries	29
4. Relational representation of the Constituent table for a Sentence	37
5. Relational Representation of the Linkage Table for a sentence	37
6. Relational Representation of the Bioentities Table for a Sentence	38
7. Examples of PTQL Queries and their Meaning	42
8. Performance of Various Approaches on the BioCreative 2 IPS Test Data	48
9. Number of Link Paths per Event Class and Pair of Arguments	66
10. PTQL Queries per Argument Pair with the Highest Support	67
11. Official Results for the BioNLP'09 Shared Task Task 2	68
12. Performance Comparison based on Document Mean Average Precision between the Top-performing IR Models only and the IR Models with our Query Generation Method	83
13. Topics in which the Query Generation Method contributes Positively and Negatively to the Document Retrieval Performance	84
14. Performance Comparison of Individual Topics between TF-IDF only and our Method using TF-IDF together with our Query Generation Method with $\alpha=0.75$	85
15. Sample PTQL Queries generated by our Query Generation Method	86
16. Keyword-based Queries for the Extraction of Gene-Drug Interactions, Protein-Protein Interactions and Gene-Disease Associations	87
17. Classes and their Corresponding Lexical Variants as Instances of the Classes	87
18. Sample PTQL Queries generated by our Query Generation Method	88
19. Precision and Recall for Gene-Drug Metabolic Relations between the Cooccurrences Method and our Query Generation Method	89

Table	Page
20. Precision, Recall and F-Measure for each kind of Gene-Drug Metabolic Relations between the Cooccurrences Method and our Query Generation Method	90
21. A Comparison of Precision, Recall and F-Measure between the Cooccurrences Method and our Query Generation Method among Various Extraction	91
22. Support Evidences that are Extracted Incorrectly by the Query [DRUG] _ metabolized by [GENE]	104
23. A Partial List of Gene-Drug Relations generated by our Approach using the Pattern [DRUG] _ metabolized by [GENE]	106
24. A List of Correct Drug-Enzyme Inhibitions and the Corresponding Support Evidences . . .	107
25. A List of Gene-Disease Associations and the Corresponding Support Evidences	108
26. Number of Relations Extracted for the Query [DRUG] and (metabolize or metabolizes or metabolized or metabolised or metabolism) and [GENE] for Different Degrees of m , the Maximum Number of Descendants to include in the m -th Level String Encodings	111
27. String Encodings and Samples of the Corresponding Sentences Retrieved by using the Query (glycan or glycans) and (modification or modifications or modify or modifies or modified)	112
28. A List of Predicates used in Representing the Pharmacokinetics Domain	122
29. A List of Fluents that describe the Properties of the World in the Pharmacokinetics Domain .	122
30. A List of Actions that can take place in the Pharmacokinetics Domain	123
31. Logic Facts and Evidence Sentences Extracted by our PTQL Framework	126
32. The Logical Representation of the Pharmacokinetic Pathway of Repaglinide Generated by our System	132
33. Evidence Sentences for the Metabolites of Pravastatin Extracted by our PTQL Extraction Framework	134

Table	Page
34. Coverage of each of the Sources in the Pharmacokinetic Pathways for the 20 Manually Annotated Pharmacokinetic Pathways in PharmGKB	135
35. Precision and Recall for PTQL Extraction of Enzymes and Transporters, as well as Metabolites	135
36. Extracted Enzymes for Repaglinide, their Evidences and the Normalized Names	137
37. Extracted Metabolites for the Repaglinide Pharmacokinetic Pathway and their Evidences	138
38. The Logical Representation of the Pharmacokinetic Pathway of Pravastatin Generated by our System	142
39. The Extracted Logic Facts for Atorvastatin Pharmacokinetic Pathway	150
40. The Logic Representation of the Atorvastatin Pharmacokinetic Pathway	151
41. Extracted Proteins for the Drug-Protein Interactions for Atorvastatin, their Evidences and the Normalized Names	152
42. (Continued from Table 41) Extracted Proteins for the Drug-Protein Interactions for Atorvastatin, their Evidences and the Normalized Names	153
43. The Extracted Logic Facts for Clopidogrel Pharmacokinetic Pathway	156
44. The Logic Representation of the Clopidogrel Pharmacokinetic Pathway	157
45. Extracted Proteins for the Drug-Protein Interactions for Clopidogrel, their Evidences and the Normalized Names	157
46. The Extracted Logic Facts for Desipramine Pharmacokinetic Pathway	158
47. The Logic Representation of the Desipramine Pharmacokinetic Pathway	160
48. Extracted Proteins for the Drug-Protein Interactions for Desipramine, their Evidences and the Normalized Names	161
49. The Extracted Logic Facts for Erlotinib Pharmacokinetic Pathway	163
50. The Logic Representation of the Erlotinib Pharmacokinetic Pathway	164
51. Extracted Proteins for the Drug-Protein Interactions for Erlotinib, their Evidences and the Normalized Names	164

Table	Page
52. The Extracted Logic Facts for Fluoxetine Pharmacokinetic Pathway	167
53. The Logic Representation of the Fluoxetine Pharmacokinetic Pathway	168
54. Extracted Proteins for the Drug-Protein Interactions for Fluoxetine, their Evidences and the Normalized Names	169
55. The Extracted Logic Facts for Fluvastatin Pharmacokinetic Pathway	172
56. The Logic Representation of the Fluvastatin Pharmacokinetic Pathway	173
57. Extracted Proteins for the Drug-Protein Interactions for Fluvastatin, their Evidences and the Normalized Names	174
58. The Extracted Logic Facts for Gefitinib Pharmacokinetic Pathway	177
59. The Logic Representation of the Gefitinib Pharmacokinetic Pathway	178
60. Extracted Proteins for the Drug-Protein Interactions for Gefitinib, their Evidences and the Normalized Names	179
61. The Extracted Logic Facts for Ifosfamide Pharmacokinetic Pathway	182
62. The Logic Representation of the Ifosfamide Pharmacokinetic Pathway	183
63. Extracted Proteins for the Drug-Protein Interactions for Ifosfamide, their Evidences and the Normalized Names	184
64. The Extracted Logic Facts for Irinotecan Pharmacokinetic Pathway	187
65. The Extracted Logic Facts for Irinotecan Pharmacokinetic Pathway (Continued from Table 64)	188
66. The Logic Representation of the Irinotecan Pharmacokinetic Pathway	189
67. The Logic Representation of the Irinotecan Pharmacokinetic Pathway (Continued from Table 66)	190
68. Extracted Proteins for the Drug-Protein Interactions for Irinotecan, their Evidences and the Normalized Names	191
69. (Continued from Table 68) Extracted Proteins for the Drug-Protein Interactions for Irinotecan, their Evidences and the Normalized Names	192

Table	Page
70. The Extracted Logic Facts for Lovastatin Pharmacokinetic Pathway	195
71. The Logic Representation of the Lovastatin Pharmacokinetic Pathway	196
72. Extracted Proteins for the Drug-Protein Interactions for Lovastatin, their Evidences and the Normalized Names	197
73. The Extracted Logic Facts for Nateglinide Pharmacokinetic Pathway	200
74. The Logic Representation of the Nateglinide Pharmacokinetic Pathway	201
75. Extracted Proteins for the Drug-Protein Interactions for Nateglinide, their Evidences and the Normalized Names	201
76. The Extracted Logic Facts for Nicotine Pharmacokinetic Pathway	204
77. The Extracted Logic Facts for Nicotine Pharmacokinetic Pathway (Continued from Table 76)	205
78. The Logic Representation of the Nicotine Pharmacokinetic Pathway	206
79. The Logic Representation of the Nicotine Pharmacokinetic Pathway (Continued from Table 78)	207
80. Extracted Proteins for the Drug-Protein Interactions for Nicotine, their Evidences and the Normalized Names	208
81. (Continued from Table 80) Extracted proteins for the Drug-Protein Interactions for Nicotine, their Evidences and the Normalized Names	209
82. The Extracted Logic Facts for Omeprazole Pharmacokinetic Pathway	212
83. The Logic Representation of the Omeprazole Pharmacokinetic Pathway	213
84. Extracted Proteins for the Drug-Protein Interactions for Omeprazole, their Evidences and the Normalized Names	214
85. (Continued from Table 84) Extracted Proteins for the Drug-Protein Interactions for Omepra- zole, their Evidences and the Normalized Names	215
86. (Continued from Table 85) Extracted Proteins for the Drug-Protein Interactions for Omepra- zole, their Evidences and the Normalized Names	216
87. The Extracted Logic Facts for Phenytoin Pharmacokinetic Pathway	219

Table	Page
88. The Extracted Logic Facts for Phenytoin Pharmacokinetic Pathway (Continued from Table 87)	220
89. The Logic Representation of the Phenytoin Pharmacokinetic Pathway	221
90. The Logic Representation of the Phenytoin Pharmacokinetic Pathway (Continued from Table 89)	222
91. Extracted Proteins for the Drug-Protein Interactions for Phenytoin, their Evidences and the Normalized Names	223
92. (Continued from Table 91) Extracted Proteins for the Drug-Protein Interactions for Phenytoin, their Evidences and the Normalized Names	224
93. (Continued from Table 92) Extracted Proteins for the Drug-Protein Interactions for Phenytoin, their Evidences and the Normalized Names	225
94. The Extracted Logic Facts for Pravastatin Pharmacokinetic Pathway	229
95. The Logic Representation of the Pravastatin Pharmacokinetic Pathway	230
96. Extracted Proteins for the Drug-Protein interactions for Pravastatin, their Evidences and the Normalized Names	231
97. The Extracted Logic Facts for Repaglinide Pharmacokinetic Pathway	234
98. The Logic Representation of the Repaglinide Pharmacokinetic Pathway	235
99. Extracted Proteins for the Drug-Protein Interactions for Repaglinide, their Evidences and the Normalized Names	236
100. The Extracted Logic Facts for Rosuvastatin Pharmacokinetic Pathway	239
101. The Logic Representation of the Rosuvastatin Pharmacokinetic Pathway	240
102. Extracted Proteins for the Drug-Protein interactions for Rosuvastatin, their Evidences and the Normalized Names	241
103. The Extracted Logic Facts for Simvastatin Pharmacokinetic Pathway	244
104. The Logic Representation of the Simvastatin Pharmacokinetic Pathway	245

Table	Page
105. Extracted Proteins for the Drug-Protein Interactions for Simvastatin, their Evidences and the Normalized Names	246
106. (Continued from Table 105) Extracted Proteins for the Drug-Protein Interactions for Simvastatin, their Evidences and the Normalized Names	247
107. The Extracted Logic Facts for Tamoxifen Pharmacokinetic Pathway	250
108. The Extracted Logic Facts for Tamoxifen Pharmacokinetic Pathway (Continued from Table 107)	251
109. The Logic Representation of the Tamoxifen Pharmacokinetic Pathway	252
110. The Logic Representation of the Tamoxifen Pharmacokinetic Pathway (Continued from Table 109)	253
111. Extracted Proteins for the Drug-Protein Interactions for Tamoxifen, their Evidences and the Normalized Names	254
112. (Continued from Table 111) Extracted Proteins for the Drug-Protein Interactions for Tamoxifen, their Evidences and the Normalized Names	255
113. The Extracted Logic Facts for Warfarin Pharmacokinetic Pathway	258
114. The Extracted Logic Facts for Warfarin Pharmacokinetic Pathway (Continued from Table 113)	259
115. The Logic Representation of the warfarin Pharmacokinetic Pathway	260
116. The Logic Representation of the Warfarin Pharmacokinetic Pathway (Continued from Table 115)	261
117. Extracted Proteins for the Drug-Protein Interactions for Warfarin, their Evidences and the Normalized Names	262
118. (Continued) Extracted Proteins for the Drug-Protein Interactions for Warfarin, their Evidences and the Normalized Names	263

LIST OF FIGURES

Figure	Page
1. A workflow of text processing modules that takes a paragraph of text as input to perform interaction extraction.	4
2. An Overview of the IR+PTQL Framework	9
3. Linkage of a Sample Sentence	16
4. Constituent Tree of a Sample Sentence	16
5. Linkage of a sample complex sentence	18
6. Database Schema of the Parse Tree Database	21
7. An Example of a Parse Tree for a Document	22
8. TEQL Grammar	24
9. An Example of a Translated SQL Query	30
10. System Architecture of the PTQL Framework	34
11. Database Schema of the Parse Tree Database	35
12. An example of processing a paragraph of text in our framework	36
13. An Extended Inverted Index	39
14. A workflow to illustrate our text processor that stores the intermediate output of each text processing module in the initial phase. When a revised or new processing module such as a named entity recognizer is deployed (denoted as <i>NER'</i>) due to improvement of the module or setting a new extraction goal, only <i>NER'</i> is executed without the reprocessing of the other text processing modules.	40
15. PTQL Grammar	41
16. An Overview of the Training Set Driven Query Generation	46
17. Examples to Illustrate the Process of Pattern Generalization	47
18. Time Distribution for PTQL Execution on BioCreative 2 IPS Testing Corpus	50
19. A System Overview of the IR+PTQL Framework	55

Figure	Page
20. Grammar for IR Queries	57
21. IR+PTQL Grammar	58
22. Linkage of a Sample Sentence	64
23. Linkages of three sentences that express the concept phosphorylation of p53.	69
24. An Overview of the Pseudo-Relevance Driven Query Generation	73
25. An Illustration of the m -th Level String Encoding	76
26. Parts of the Constituent Trees for Sample Sentences	78
27. Precision of our Query Generation Approach using Various Configurations	92
28. Recall of our Query Generation Approach using Various Configurations	93
29. Precision and Recall of our Query Generation Approach using Various Configurations for Gene-Drug Metabolic Relations	94
30. A System Overview of NetSynthesis and its Interactions with the IR+PTQL Framework . . .	101
31. EBNF Grammar for PTQL ^{LITE} Queries	101
32. A Screenshot of our NetSynthesis Prototype	113
33. Pharmacokinetic Pathway of Fluvastatin	117
34. A Network Representation of the Drug-Protein Interactions for the Fluvastatin Pathway . . .	118
35. An Overview of the System Architecture for Pathway Synthesis	130
36. The Manually Curated Pharmacokinetic Pathway of the Drug repaglinide from PharmGKB .	133
37. The Pharmacokinetic Pathway of Repaglinide Synthesized by our System	134
38. Pharmacokinetic Pathway of Atorvastatin	148
39. Synthesized Version of the Pharmacokinetic Pathway of Atorvastatin	149
40. Pharmacokinetic Pathway of Clopidogrel	154
41. Synthesized Version of the Pharmacokinetic Pathway of Clopidogrel	155
42. Pharmacokinetic Pathway of Desipramine	159
43. Synthesized Version of the Pharmacokinetic Pathway of Desipramine	159

Figure	Page
44. Pharmacokinetic Pathway of Erlotinib	162
45. Synthesized Version of the Pharmacokinetic Pathway of Erlotinib	163
46. Pharmacokinetic Pathway of Fluoxetine	165
47. Synthesized Version of the Pharmacokinetic Pathway of Fluoxetine	166
48. Pharmacokinetic Pathway of Fluvastatin	170
49. Synthesized Version of the Pharmacokinetic Pathway of Fluvastatin	171
50. Pharmacokinetic Pathway of Gefitinib	175
51. Synthesized Version of the Pharmacokinetic Pathway of Gefitinib	176
52. Pharmacokinetic Pathway of Ifosfamide	180
53. Synthesized Version of the Pharmacokinetic Pathway of Ifosfamide	181
54. Pharmacokinetic Pathway of Irinotecan	185
55. Synthesized Version of the Pharmacokinetic Pathway of Irinotecan	186
56. Pharmacokinetic Pathway of Lovastatin	193
57. Synthesized Version of the Pharmacokinetic Pathway of Lovastatin	194
58. Pharmacokinetic Pathway of Nateglinide	198
59. Synthesized Version of the Pharmacokinetic Pathway of Nateglinide	199
60. Pharmacokinetic Pathway of Nicotine	202
61. Synthesized Version of the Pharmacokinetic Pathway of Nicotine	203
62. Pharmacokinetic Pathway of Omeprazole	210
63. Synthesized Version of the Pharmacokinetic Pathway of Omeprazole	211
64. Pharmacokinetic Pathway of Phenytoin	217
65. Synthesized Version of the Pharmacokinetic Pathway of Phenytoin	218
66. Pharmacokinetic Pathway of Pravastatin	227
67. Synthesized Version of the Pharmacokinetic Pathway of Pravastatin	228
68. Pharmacokinetic Pathway of Repaglinide	232

Figure	Page
69. Synthesized Version of the Pharmacokinetic Pathway of Repaglinide	233
70. Pharmacokinetic Pathway of Rosuvastatin	237
71. Synthesized Version of the Pharmacokinetic Pathway of Rosuvastatin	238
72. Pharmacokinetic Pathway of Simvastatin	242
73. Synthesized Version of the Pharmacokinetic Pathway of Simvastatin	243
74. Pharmacokinetic Pathway of Tamoxifen	248
75. Synthesized Version of the Pharmacokinetic Pathway of Tamoxifen	249
76. Pharmacokinetic Pathway of Warfarin	256
77. Synthesized Version of the Pharmacokinetic Pathway of Warfarin	257

PREVIEW

1. INTRODUCTION

Finding information from the literature is a necessary process in scientific discovery for biologists. However, biologists face the problem of information overload with the increasing number of published articles. From 1994 to 2004, close to 3 million biomedical articles were published by US and European researchers. This publication rate has resulted in approximately 18 million publications in PubMed, which serves as a repository of biomedical articles. This implies that biologists consume most of their time in finding relevant information from articles rather than focusing on their efforts in developing hypotheses for research. There is an urgent need to reduce the information burden of biologists so that they can focus and speed up the process of scientific discovery.

1.1. Information retrieval

Information retrieval (IR) is an active research area that studies the problem of handling information conveyed in large amount of unstructured natural language text. Web search is a well-recognized form of information retrieval, which allows users to seek information from the web by expressing their search interest with keywords. In this thesis, we focus on the retrieval of information from a collection of text articles, which is sometimes referred as *document retrieval*, particularly biomedical articles in the form of abstracts or full-text articles.

A typical IR system (or IR engine) is composed of the *indexing* and *retrieval* components. The indexing component tokenizes each of the words in the document collection to build an inverted index for efficient document retrieval. Common terms that appear with very high frequencies, such as the word “*the*”, are seen to have little value to retrieval. These terms are known as *stopwords*. It is a common practice to discard stopwords in documents from being indexed. The retrieval component fetches relevant documents and ranks the documents according to their relevance with respect to the query. There are various *retrieval models* in deciding the relevance of a document. These include boolean model, vector space model, language model and probabilistic model [1]. The main idea behind these models is that documents are modeled as a *bag of words*, so that ordering of words in a document are not captured by such retrieval models. Word frequency is the main factor in computing the relevance of documents. This means that in the bag-of-words approach,

document with the sentence “*Mary is quicker than Tom*” is treated as the same as document with the sentence “*Tom is quicker than Mary.*” [1]

While natural language processing (NLP) has been viewed as a critical part of IR [2–4], the role of NLP in IR has been limited to *stemming*. Stemming is a method to find the inflectional forms of the words. For instance, “*be*” is the inflectional form of the word “*is*”. By applying stemming to the process of indexing and retrieval, it is largely seen as a way to increase the recall of the system, with possible decrease in precision. Whether stemming has a positive impact to IR has been inconclusive for years [5], until recent publications show that applying stemming with certain conditions results in a significant impact to the performance of IR [6, 7].

Another fundamental component to increase the recall of IR systems is *query expansion*. Basic query expansion considers synonyms or acronyms of terms, which can be obtained through resources such as WordNet or through automated extraction [1]. Effective expansion of queries is particularly important for biomedical applications, due to the wide variety of ways that can be used to express a concept in the biological domain. *Pseudo-relevance feedback* is one of the query expansion techniques that has been shown to be effective to the performance of retrieval [8]. The idea behind pseudo-relevance feedback is that an initial query is used to retrieve documents, and frequently occurring terms are selected from the top- k documents. These terms are then used to augment the initial query, and the enhanced query is applied to perform another retrieval of documents. However, a recent study [9] shows that using frequently occurring keywords in relevant documents do not necessarily improve the performance of retrieval. It is necessary to consider the context of the frequently occurring keywords that are used in query expansion.

To evaluate the performance of IR, it is important to consider both precision and the rank of results. Ranking is critical as thousands of results may be retrieved, and it is ideal to present the relevant results to the users as highly ranked. The mean average precision (MAP) is a popular measure for the evaluation of ranked lists of results. MAP incorporates aspects of both precision and recall. Let n be the number of retrieved documents for a particular query, and $rank(i)$ be the i -th document in the ranked list of documents.