

TGAT TT TGAAGA AGAGCCCCAGAGTCAGGTCGT

180

200

210



BIOINFORMATICS

Managing
Scientific Data

Zoé Lacroix
Terence Critchlow

Contents

Preface	xix
---------	-----

1	Introduction	1
	<i>Zoé Lacroix and Terence Critchlow</i>	
1.1	Overview	1
1.2	Problem and Scope	2
1.3	Biological Data Integration	4
1.4	Developing a Biological Data Integration System	7
1.4.1	Specifications	7
1.4.2	Translating Specifications into a Technical Approach	8
1.4.3	Development Process	9
1.4.4	Evaluation of the System	9
	References	10

2	Challenges Faced in the Integration of Biological Information	11
	<i>Su Yun Chung and John C. Wooley</i>	
2.1	The Life Science Discovery Process	12
2.2	An Information Integration Environment for Life Science Discovery	14
2.3	The Nature of Biological Data	15
2.3.1	Diversity	15
2.3.2	Variability	17
2.4	Data Sources in Life Science	17
2.4.1	Biological Databases Are Autonomous	18
2.4.2	Biological Databases Are Heterogeneous in Data Formats	18

2.4.3	Biological Data Sources Are Dynamic	18
2.4.4	Computational Analysis Tools Require Specific Input/Output Formats and Broad Domain Knowledge	19
2.5	Challenges in Information Integration	19
2.5.1	Data Integration	21
2.5.2	Meta-Data Specification	24
2.5.3	Data Provenance and Data Accuracy	25
2.5.4	Ontology	27
2.5.5	Web Presentations	30
	Conclusion	31
	References	32

3 A Practitioner's Guide to Data Management and Data Integration in Bioinformatics

35

Barbara A. Eckman

3.1	Introduction	35
3.2	Data Management in Bioinformatics	36
3.2.1	Data Management Basics	36
3.2.2	Two Popular Data Management Strategies and Their Limitations	39
3.2.3	Traditional Database Management	41
3.3	Dimensions Describing the Space of Integration Solutions	45
3.3.1	A Motivating Use Case for Integration	45
3.3.2	Browsing vs. Querying	46
3.3.3	Syntactic vs. Semantic Integration	48
3.3.4	Warehouse vs. Federation	49
3.3.5	Declarative vs. Procedural Access	49
3.3.6	Generic vs. Hard-Coded	49
3.3.7	Relational vs. Non-Relational Data Model	50
3.4	Use Cases of Integration Solutions	50
3.4.1	Browsing-Driven Solutions	50
3.4.2	Data Warehousing Solutions	52
3.4.3	Federated Database Systems Approach	54
3.4.4	Semantic Data Integration	58
3.5	Strengths and Weaknesses of the Various Approaches to Integration	60
3.5.1	Browsing and Querying: Strengths and Weaknesses	61
3.5.2	Warehousing and Federation: Strengths and Weaknesses	62
3.5.3	Procedural Code and Declarative Query Language: Strengths and Weaknesses	63

3.5.4	Generic and Hard-Coded Approaches: Strengths and Weaknesses	63
3.5.5	Relational and Non-Relational Data Models: Strengths and Weaknesses	64
3.5.6	Conclusion: A Hybrid Approach to Integration Is Ideal	64
3.6	Tough Problems in Bioinformatics Integration	65
3.6.1	Semantic Query Planning Over Web Data Sources	65
3.6.2	Schema Management	67
3.7	Summary	69
	Acknowledgments	70
	References	70

4 Issues to Address While Designing a Biological Information System 75

Zoé Lacroix

4.1	Legacy	78
4.1.1	Biological Data	78
4.1.2	Biological Tools and Workflows	79
4.2	A Domain in Constant Evolution	80
4.2.1	Traditional Database Management and Changes	80
4.2.2	Data Fusion	82
4.2.3	Fully Structured vs. Semi-Structured	82
4.2.4	Scientific Object Identity	84
4.2.5	Concepts and Ontologies	85
4.3	Biological Queries	86
4.3.1	Searching and Mining	87
4.3.2	Browsing	89
4.3.3	Semantics of Queries	90
4.3.4	Tool-Driven vs. Data-Driven Integration	91
4.4	Query Processing	92
4.4.1	Biological Resources	92
4.4.2	Query Planning	94
4.4.3	Query Optimization	95
4.5	Visualization	98
4.5.1	Multimedia Data	99
4.5.2	Browsing Scientific Objects	100
4.6	Conclusion	101
	Acknowledgments	102
	References	102

5	SRS: An Integration Platform for Databanks and Analysis Tools in Bioinformatics	109
	<i>Thure Etzold, Howard Harris, and Simon Beaulah</i>	
5.1	Integrating Flat File Databanks	112
5.1.1	The SRS Token Server	113
5.1.2	Subentry Libraries	116
5.2	Integration of XML Databases	116
5.2.1	What Makes XML Unique?	118
5.2.2	How Are XML Databanks Integrated into SRS?	120
5.2.3	Overview of XML Support Features	121
5.2.4	How Does SRS Meet the Challenges of XML?	122
5.3	Integrating Relational Databases	124
5.3.1	Whole Schema Integration	124
5.3.2	Capturing the Relational Schema	125
5.3.3	Selecting a Hub Table	126
5.3.4	Generation of SQL	127
5.3.5	Restricting Access to Parts of the Schema	128
5.3.6	Query Performance to Relational Databases	128
5.3.7	Viewing Entries from a Relational Databank	128
5.3.8	Summary	129
5.4	The SRS Query Language	129
5.4.1	SRS Fields	130
5.5	Linking Databanks	130
5.5.1	Constructing Links	131
5.5.2	The Link Operators	132
5.6	The Object Loader	133
5.6.1	Creating Complex and Nested Objects	134
5.6.2	Support for Loading from XML Databanks	135
5.6.3	Using Links to Create Composite Structures	136
5.6.4	Exporting Objects to XML	136
5.7	Scientific Analysis Tools	137
5.7.1	Processing of Input and Output	138
5.7.2	Batch Queues	139
5.8	Interfaces to SRS	139
5.8.1	The Web Interface	139
5.8.2	SRS Objects	140
5.8.3	SOAP and Web Services	141
5.9	Automated Server Maintenance with SRS Prisma	141
5.10	Conclusion	143
	References	144

6	The Kleisli Query System as a Backbone for Bioinformatics Data Integration and Analysis	147
	<i>Jing Chen, Su Yun Chung, and Limsoon Wong</i>	
6.1	Motivating Example	149
6.2	Approach	151
6.3	Data Model and Representation	153
6.4	Query Capability	158
6.5	Warehousing Capability	163
6.6	Data Sources	165
6.7	Optimizations	167
6.7.1	Monadic Optimizations	169
6.7.2	Context-Sensitive Optimizations	171
6.7.3	Relational Optimizations	174
6.8	User Interfaces	175
6.8.1	Programming Language Interface	175
6.8.2	Graphical Interface	179
6.9	Other Data Integration Technologies	179
6.9.1	SRS	179
6.9.2	DiscoveryLink	181
6.9.3	Object-Protocol Model (OPM)	182
6.10	Conclusions	183
	References	184

7	Complex Query Formulation Over Diverse Information Sources in TAMBIS	189
	<i>Robert Stevens, Carole Goble, Norman W. Paton, Sean Bechhofer, Gary Ng, Patricia Baker, and Andy Brass</i>	
7.1	The Ontology	192
7.2	The User Interface	195
7.2.1	Exploring the Ontology	195
7.2.2	Constructing Queries	197
7.2.3	The Role of Reasoning in Query Formulation	202
7.3	The Query Processor	205
7.3.1	The Sources and Services Model	206
7.3.2	The Query Planner	208
7.3.3	The Wrappers	211
7.4	Related Work	213

7.4.1	Information Integration in Bioinformatics	213
7.4.2	Knowledge Based Information Integration	215
7.4.3	Biological Ontologies	216
7.5	Current and Future Developments in TAMBIS	217
7.5.1	Summary	219
	Acknowledgments	220
	References	220

8 The Information Integration System K2 225

Val Tannen, Susan B. Davidson, and Scott Harker

8.1	Approach	229
8.2	Data Model and Languages	232
8.3	An Example	235
8.4	Internal Language	239
8.5	Data Sources	240
8.6	Query Optimization	242
8.7	User Interfaces	243
8.8	Scalability	244
8.9	Impact	245
8.10	Summary	246
	Acknowledgments	247
	References	247

9 P/FDM Mediator for a Bioinformatics Database Federation 249

Graham J. L. Kemp and Peter M. D. Gray

9.1	Approach	250
9.1.1	Alternative Architectures for Integrating Databases	250
9.1.2	The Functional Data Model	252
9.1.3	Schemas in the Federation	254
9.1.4	Mediator Architecture	257
9.1.5	Example	261
9.1.6	Query Capabilities	264
9.1.7	Data Sources	265
9.2	Analysis	266
9.2.1	Optimization	267
9.2.2	User Interfaces	268
9.2.3	Scalability	271

- 9.3 Conclusions 272
- Acknowledgment 272
- References 272

10 Integration Challenges in Gene Expression Data Management 277

*Victor M. Markowitz, John Campbell, I-Min A. Chen,
Anthony Kosky, Krishna Palaniappan,
and Thodoros Topaloglou*

- 10.1 Gene Expression Data Management: Background 278
 - 10.1.1 Gene Expression Data Spaces 278
 - 10.1.2 Standards: Benefits and Limitations 281
- 10.2 The GeneExpress System 282
 - 10.2.1 GeneExpress System Components 283
 - 10.2.2 GeneExpress Deployment and Update Issues 283
- 10.3 Managing Gene Expression Data: Integration Challenges 285
 - 10.3.1 Gene Expression Data: Array Versions 285
 - 10.3.2 Gene Expression Data: Algorithms and Normalization 286
 - 10.3.3 Gene Expression Data: Variability 287
 - 10.3.4 Sample Data 288
 - 10.3.5 Gene Annotations 289
- 10.4 Integrating Third-Party Gene Expression Data in GeneExpress 291
 - 10.4.1 Data Exchange Formats 291
 - 10.4.2 Structural Data Transformation Issues 293
 - 10.4.3 Semantic Data Mapping Issues 293
 - 10.4.4 Data Loading Issues 296
 - 10.4.5 Update Issues 297
- 10.5 Summary 298
 - Acknowledgments 299
 - Trademarks 299
 - References 300

11 DiscoveryLink 303

*Laura M. Haas, Barbara A. Eckman, Prasad Kodali,
Eileen T. Lin, Julia E. Rice, and Peter M. Schwarz*

- 11.1 Approach 306
 - 11.1.1 Architecture 309
 - 11.1.2 Registration 313

- 11.2 Query Processing Overview 316
 - 11.2.1 Query Optimization 317
 - 11.2.2 An Example 319
 - 11.2.3 Determining Costs 322
- 11.3 Ease of Use, Scalability, and Performance 327
- 11.4 Conclusions 329
 - References 331

12 A Model-Based Mediator System for Scientific Data Management 335

*Bertram Ludäscher, Amarnath Gupta,
and Maryann E. Martone*

- 12.1 Background 336
- 12.2 Scientific Data Integration Across Multiple Worlds: Examples and Challenges from the Neurosciences 338
 - 12.2.1 From Terminology and Static Knowledge to Process Context 340
- 12.3 Model-Based Mediation 343
 - 12.3.1 Model-Based Mediation: The Protagonists 343
 - 12.3.2 Conceptual Models and Registration of Sources at the Mediator 344
 - 12.3.3 Interplay Between Mediator and Sources 349
- 12.4 Knowledge Representation for Model-Based Mediation 351
 - 12.4.1 Domain Maps 352
 - 12.4.2 Process Maps 357
- 12.5 Model-Based Mediator System and Tools 360
 - 12.5.1 The KIND Mediator Prototype 360
 - 12.5.2 The Cell-Centered Database and SMART Atlas: Retrieval and Navigation Through Multi-Scale Data 362
- 12.6 Related Work and Conclusion 364
 - 12.6.1 Related Work 364
 - 12.6.2 Summary: Model-Based Mediation and Reason-Able Meta-Data 365
- Acknowledgments 366
- References 366

13	Compared Evaluation of Scientific Data Management Systems	371
	<i>Zoé Lacroix and Terence Critchlow</i>	
13.1	Performance Model	371
13.1.1	Evaluation Matrix	372
13.1.2	Cost Model	372
13.1.3	Benchmarks	374
13.1.4	User Survey	375
13.2	Evaluation Criteria	376
13.2.1	The Implementation Perspective	377
13.2.2	The User Perspective	382
13.3	Tradeoffs	385
13.3.1	Materialized vs. Non-Materialized	385
13.3.2	Data Distribution and Heterogeneity	386
13.3.3	Semi-Structured Data vs. Fully Structured Data	387
13.3.4	Text Retrieval	388
13.3.5	Integrating Applications	389
13.4	Summary	389
	References	390
	Concluding Remarks	393
	Summary	393
	Looking Toward the Future	394
	Appendix: Biological Resources	397
	Glossary	407
	System Information	425
	SRS	425
	Kleisli	425
	TAMBIS	426
	K2	426
	P/FDM Mediator	427
	GeneExpress	427
	DiscoveryLink	428
	KIND	428
	Index	431