

Contents

List of Figures	ix
List of Tables	xi
List of Abbreviations	xiii
List of Publications	xv
1 Introduction	1
1.1 Scientific Workflows in e-Science	1
1.2 Challenges for Scientists Using Life Science Workflows	3
1.3 Goals of the Thesis	5
2 Concept Development for State-of-the-Art Workflow Optimization	9
2.1 General Aspects of Scientific Workflows	9
2.1.1 Scientific Workflows	9
2.1.2 Scientific Workflow Management Systems	12
2.1.3 e-Science Collaborations	12
2.2 General Aspects of Optimization and Learning	14
2.2.1 Mathematical Background and Notations	14
2.2.2 Different Optimization Algorithms	15
2.2.3 Design Optimization Frameworks	17
2.3 State-of-the-Art Scientific Workflow Optimization	18
2.3.1 Runtime Performance Optimization	18
2.3.2 Output Performance Optimization	19
2.3.3 Other Concepts of Workflow Modification	20
2.4 A Concept for Scientific Workflow Optimization	21

3	Enabling Parallel Execution in Scientific Workflow Management Systems	27
3.1	Investigation of Scientific Workflow Management Systems in e-Science	28
3.2	Extension of a Workflow Management System	32
3.2.1	The Taverna Workflow Management System	33
3.2.2	UNICORE Middleware	35
3.2.3	Architecture of the Grid Plugin	35
3.2.4	Development of the Grid Plugin	36
3.2.5	Enhanced Parallel Application Execution	38
3.3	Evaluation by Life Science Use Cases	39
3.4	Discussion	43
3.5	Conclusion	44
4	A Framework for Scientific Workflow Optimization	47
4.1	The Approach of Scientific Workflow Optimization	48
4.1.1	A New Optimization Phase in the Scientific Workflow Life Cycle	48
4.1.2	Investigation of Different Optimization Levels	51
4.1.3	Definition of the Optimization Target	54
4.2	The Usability Compliance of Workflow Optimization	55
4.3	The Taverna Optimization Framework	56
4.4	Enabling Optimization on Distributed Computing Infrastructures	60
4.4.1	Three Tier Execution Architecture	61
4.4.2	Implementation of Parallel Workflow Execution	62
4.4.3	Parallel Optimization Use Case	64
4.5	Discussion	64
4.6	Conclusion	65
5	Optimization Techniques for Scientific Workflow Optimization	67
5.1	Optimization Techniques for Scientific Workflow Parameters	67
5.1.1	Genetic Algorithms	70
5.1.2	A Genetic Algorithm for Scientific Workflows	71
5.2	The Parameter Optimization Plugin	72
5.2.1	Development of the Parameter Optimization Plugin	73
5.2.2	Discussion	76
5.3	Evaluation of the Parameter Optimization Plugin	76
5.3.1	Proteomics Workflows	77

5.3.2	Ecological Niche Modeling Workflows	84
5.3.3	Biomarker Identification Workflows	89
5.3.4	Protein Structure Similarity Workflows	92
5.3.5	Discussion	94
5.4	Simulation of Workflow Structure Optimization	96
5.4.1	The Component Level	96
5.4.2	Discussion	99
5.4.3	The Topology Level	100
5.4.4	Discussion	102
5.5	Conclusion	102
6	Discussion: Scientific Workflow Optimization in e-Science	105
6.1	Examination of Scientific Workflow Optimization	106
6.1.1	General Aspects of Optimization	106
6.1.2	Addressing Optimization Complexity	111
6.2	Provenance-based Optimization	115
7	Conclusion	123
7.1	Summary of the Work	123
7.2	Future Work	126
A	Additional Figures and Descriptions	131
	Bibliography	149