Small Molecules: From Mass Spectral Fragmentation Data to Structural Elucidation

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der Fakultät für Mathematik und Informatik

der Friedrich-Schiller-Universität Jena

von Dipl.-Bioinf. Kerstin Scheubert

geboren am 10. Nov. 1984 **in** Weida, Deutschland

Contents

1 Introduction 1

2 Background 5
- 2.1 Molecules and Metabolites . . . 5
- 2.2 Mass Spectrometry . . . 8
  - 2.2.1 Components of a Mass Spectrometer . . . 9
  - 2.2.2 Separation Methods . . . 12
  - 2.2.3 Tandem and Multistage Mass Spectrometry . . . 13
- 2.3 Mass Spectrometry of Small Molecules and Natural Products . . . 14
- 2.4 Basic Concepts of Graph and Computational Complexity Theory . . . 16
  - 2.4.1 Graph Theory . . . 16
  - 2.4.2 Computational Complexity Classes . . . 17

3 Related Work: Computational Methods for Analyzing Small Molecule MS Data 19
- 3.1 Molecular Formula Identification . . . 19
- 3.2 Searching in Spectral Libraries . . . 20
- 3.3 Searching in Molecular Structure Databases . . . 21
- 3.4 Fragmentation Trees and Fragmentation Tree Alignments . . . 23

4 Significance of Metabolite Identifications from Searching Mass Spectral Libraries 27
- 4.1 Experimental Data . . . 28
- 4.2 Removing Noise Peaks from Mass Spectra . . . 30
- 4.3 Choosing a Suitable Scoring Scheme to Separate True from False Hits . . . 31
- 4.4 Estimating False Discovery Rates . . . 33
  - 4.4.1 Empirical Bayes Approach . . . 35
  - 4.4.2 Target-Decoy Approach . . . 38
- 4.5 Evaluating p-values . . . 40
- 4.6 Evaluating False Discovery Rates and q-values . . . 41

5 Estimating Compound Similarity Using Tandem Mass Spectrometry 45
- 5.1 Predicting Tanimoto Scores from Platt Probabilities . . . 45
- 5.2 Experimental Data . . . 48
- 5.3 Evaluation of Running Times and Quality of Predicted Tanimoto Scores . . . 48
  - 5.3.1 Running Times . . . 48
  - 5.3.2 Quality of Predicted Tanimoto Scores . . . 50
- 5.4 Applications . . . 52
  - 5.4.1 Dereplication . . . 52
  - 5.4.2 Compound Classification . . . 54

6 Computing Fragmentation Trees from Metabolite Multiple Mass Spectrometry Data **57**
6.1 Constructing Fragmentation Trees from Multiple MS Data 57
6.1.1 Problem Definition: The Combined Colorful Subtree Problem 57
6.1.2 Scoring 59
6.2 Hardness Results 60
6.2.1 NP-hardness of the COLORFUL SUBTREE CLOSURE Problem 60
6.2.2 Inapproximability of the COMBINED COLORFUL SUBTREE Problem 62
6.3 Exact Algorithms for the COMBINED COLORFUL SUBTREE Problem 63
6.3.1 Fixed Parameter Algorithm 64
6.3.2 Integer Linear Program 65
6.4 Evaluation of Running Times and Fragmentation Tree Quality 67
6.4.1 Experimental Data 67
6.4.2 Running Time Analysis 68
6.4.3 Fragmentation Tree Quality 69
6.4.4 Evaluation Against Expert Knowledge 74

7 Conclusion **77**

A Supplementary Figures and Tables of Chapter 4 **97**

B Supplementary Figures and Tables of Chapter 5 **105**

C Supplementary Proofs of Chapter 6 **107**
C.1 Proof of Theorem 2 107
C.2 Proof of Theorem 3 108