Contents

Chapter 1

Introduction

1.1 A few biology notions

1.1.1 From genome to proteins, the Central Dogma

1.1.1.1 ADN, genome

1.1.1.2 ARN, transcriptome

1.1.1.3 Proteins, peptides, proteome, and interactions

1.1.2 Measuring the state of the living, sequencing

1.1.2.1 Genome sequencing

1.1.2.2 ARN sequencing, expression levels

1.1.2.3 ARN levels as a proxy for proteins expression levels

1.1.2.4 Differential analysis

1.1.3 ??? Phage display ???

1.1.4 ??? Mass spectrometry ???

1.2 Some computer science elements

1.2.1 String

1.2.1.1 suffix tree

1.2.1.2 generalized suffix tree

1.2.1.3 generalized suffix array

1.2.2 Graph

1.2.2.1 Connectivity

1.2.2.2 Minimum cut

1.2.3 Combinatorial optimization

1.2.3.1 Dynamic programming

1.2.3.2 Decision trees, Branch and bound, Branch and cut

1.2.3.3 Linear programming, Mixed integer linear programming

1.2.4 Complexity

1.2.4.1 APX-difficulty

Chapter 2

String selection

2.1 Intro

2.1.1 Sequences

2.1.1.1 Two sets of sequences (PhD or ?? + proteome)

2.1.1.2 Equals two sets of small sequences (kmers)

2.1.2 Distance measure of two AA sequences

2.2 Algorithm

2.2.1 Encoding the two sets as generalized suffix arrays

2.2.2 Minimum-Maximum distance over length, and pruning

2.2.3 Mappings, profiles, and significance

2.3 Analysis

2.3.1 Performance

2.3.2 Significance

2.3.2.1 Pathways test

2.4 Biology, pretty please ?!

Chapter 3

Biological module discovery

3.1 Module introduction (topological module –connected– vs. aggregation module –stats–)

3.2 Mapping biological weights onto PPI networks

3.3 Maximum-Weight Connected Subgraph

3.3.1 Prize-Collecting Steiner Tree

3.3.2 Direct methods (Miranda-Alvarez)

Chapter 4

??? Multi-modules discovery ???

4.1 Some methods find multiple modules

4.2 We want exact method

4.2.1 Linear program

Chapter 5

Cross-species biological module discovery

5.1 Cross-species modules

5.2 Maximum-Weight Cross-Connected Subgraph

Chapter 6

Difficulty of the Maximum-Weight Cross-Connected Subgraph

6.1 APX-difficulty of the MWCCS problem

6.1.1 Frontier of solvability: tree, tree, 1-to-1

6.2 An MWCCS subproblem: the Ratio-Bounded MWCS

6.2.1 A more general variant of the Budget-Constrained MWCS

6.2.2 Producing pseudo-P algorithms for BC-MWCS and RB-MWCS from P algorithms for MWCS

Chapter 7

??? Module recognition, cancer classification ???

Chapter 8

Conclusion