Contents

Chapter 1

Introduction

1.1 Context

- bioinformatics pipeline
- type of data
- more integrative approaches

1.2 Preliminary notions

1.2.1 Small biology overview

1.2.1.1 From genome to proteins, the Central Dogma

1.2.1.1.1 Genome

1.2.1.1.2 Transcriptome

1.2.1.1.3 Proteome

1.2.1.2 Modeling the living, biological networks

Biological networks are abstract representations of biological entities interconnected over some criteria. It can represent for example the relationships between species inside an ecosystem, or interconnections between cell types in any multicellular organism.

In this work, we are mostly interested in biological networks that pertain to the living, i.e. networks of components of an organism. Many such networks exists, to name a few:

- *Metabolic networks* represent biochemical reactions between substrates, enzymes and metabolites, and cluster them into pathways,
- *Gene co-expression networks* represent the similarity of expression between genes in some biological setup, by interconnecting pairs of genes similarly expressed,

- *Protein-protein interaction networks* represent the interaction between two proteins, usually of the same species.

On the one hand biological networks can be seen as observed or inferred facts, where the network is the knowledge in itself ; e.g. a known pathway that connect chemical reactants and products through enzymes. On the other hand they can be seen as an abstract representation of knowledge where nodes represent entities and edges represent some form of deduced connection ; e.g. a gene co-expression network which can be constructed from the control and condition expression profiles of the genes, with a statistical inference over the two samples resulting in the presence or absence of the edges.

Protein-protein interaction (PPI) networks play an important role in this work, and we will present them in more detail. But first let us stress the importance of biological networks in modern biology.

They structure our understanding of biological systems in such ways that both allow a comprehension of biological processes, and permit automated processing of the knowledge that they represent. As automated processing enabling tools, they can serve as both knowledge bases for local decisions and as global networks that can serve as substrate for integrated analysis.

XXX.

1.2.1.2.1 Protein-Protein Interactions

1.2.1.3 Measuring the state of the living

1.2.1.3.1 Measuring gene expression levels

1.2.1.3.2 Differential analysis

- better understanding of cellular processes
- biomarkers discovery

1.2.2 Some computer science elements

1.2.2.1 Graphs

1.2.2.2 Combinatorial optimization

1.2.2.2.1 Dynamic programming

1.2.2.2.2 Decision trees, Branch and bound, Branch and cut

1.2.2.2.3 Linear programming, Mixed integer linear programming

1.2.2.3 Complexity

1.2.2.3.1 APX-hardness

1.2.2.3.2 Pseudo-polynomial time

1.2.2.4 String

1.2.2.4.1 Suffix trees and array

Chapter 2

State of the art

2.1 Protein-protein interaction networks

- String: [SFW$^+$14]

2.2 Module discovery

- context, differential analysis
- traditionnally, gene centric (cf. next subsection)

2.2.1 Module as a set of genes

- First: [GST$^+$99]

2.2.2 Module as a connected cluster of genes

- Heuristic: [IOSS02], [MCRI13], ...
- Exact: [DKR$^+$08]

2.2.3 Cross-species discovery

- single gene conservation: [vNSH03]

2.3 The Maximum-Weight Connected Subgraph problem

2.3.1 Problem overview

2.3.1.1 The Prize-Collecting Steiner-Tree problem

2.3.2 Solving the problem, optimization methods

2.3.2.1 Through the Prize-Collecting Steiner Tree problem

2.3.2.2 Through a direct method (Miranda-Alvarez)

- rooted: [ÁMLM13b]

- unrooted: [ÁMLM13a]

Chapter 3

Cross-species biological module discovery

3.1 Weight assignments

3.1.1 Gene weight through expression profiles

3.1.2 Peptides mappings

3.2 Cross-species modules

3.3 Maximum-Weight Cross-Connected Subgraph

Chapter 4

Difficulty of the Maximum-Weight Cross-Connected Subgraph

4.1 APX-difficulty of the MWCCS problem

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

4.2 An MWCCS subproblem: the Ratio-Bounded MWCS

4.2.1 A more general variant of the Budget-Constrained MWCS

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

Chapter 5

Conclusion

Bibliography

[ÁMLM13a] Eduardo Álvarez-Miranda, Ivana Ljubić, and Petra Mutzel. The maximum weight connected subgraph problem. In *Facets of Combinatorial Optimization*, pages 245–270. Springer, 2013.

[ÁMLM13b] Eduardo Álvarez-Miranda, Ivana Ljubić, and Petra Mutzel. The rooted maximum node-weight connected subgraph problem. In *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*, pages 300–315. Springer, 2013.

[DKR$^+$08] Marcus T Dittrich, Gunnar W Klau, Andreas Rosenwald, Thomas Dandekar, and Tobias Müller. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–i231, 2008.

[GST$^+$99] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

[IOSS02] Trey Ideker, Owen Ozier, Benno Schwikowski, and Andrew F Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl 1):S233–S240, 2002.

[MCRI13] Koyel Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732, 2013.

[SFW$^+$14] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, page gku1003, 2014.

[vNSH03] Vera van Noort, Berend Snel, and Martijn A Huynen. Predicting gene function by conserved co-expression. *TRENDS in Genetics*, 19(5):238–242, 2003.