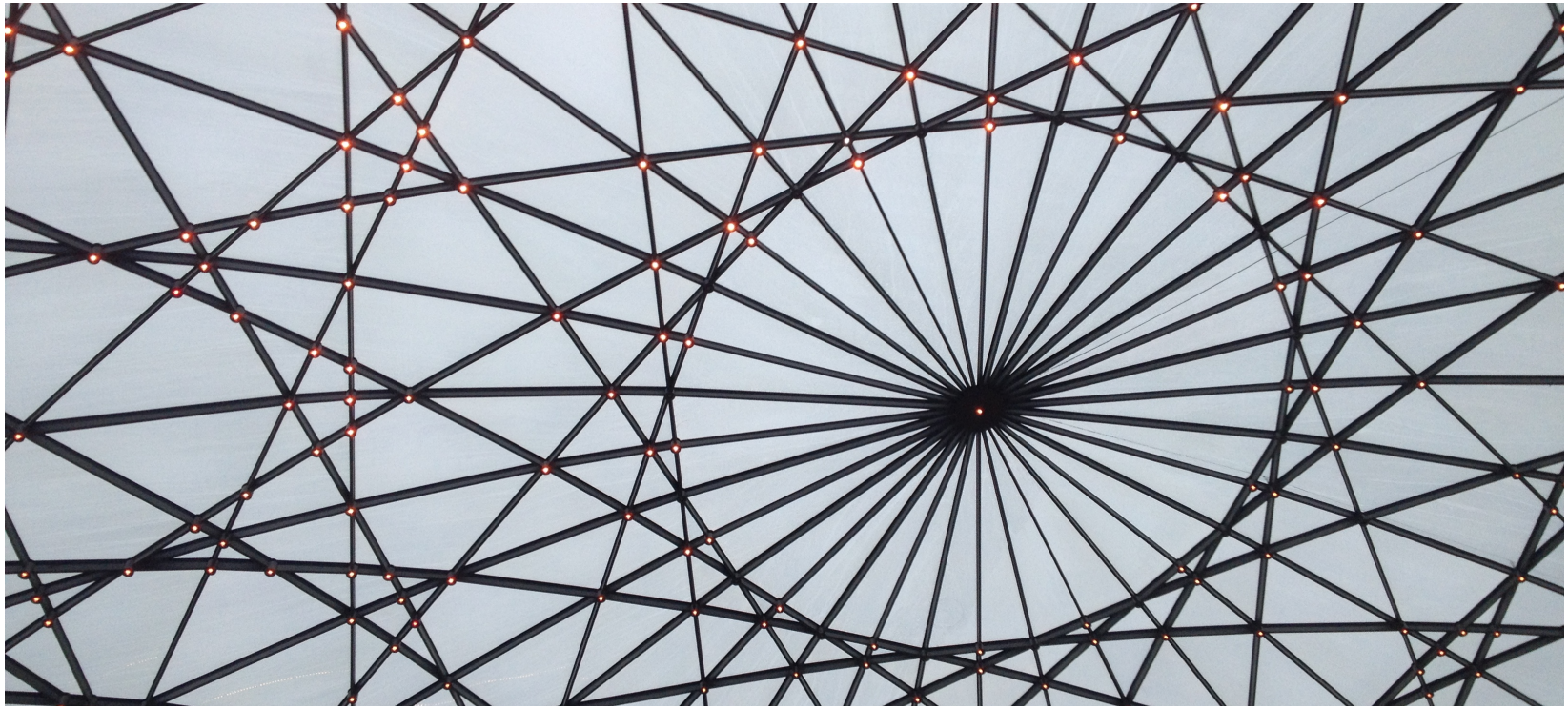
HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF



Seminar Kombinatorische Optimierung in der Bioinformatik

SS 2017 – Prof. Dr. G. Klau

Organisatorisches

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Über mich

- seit 2/2017 an der HHU
- Gebäude 25.12, Raum 01.46
- Sprechstunde nach Vereinbarung (e-Mail oder Sekretariat)
- Web: http://albi.hhu.de

Über das Seminar

- Sprache: Englisch gewünscht

Prozedur

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Erfolgreiche Teilnahme =

- Wähle Thema
- Lese Literatur, finde relevante weitere Literatur
- Bereite Vortrag vor
  - 2 Wochen vor Tag X. Draft zu Prof. Klau – Feedback
  - 1 Woche vor Tag X: Treffen mit Prof. Klau – Vortrag durchsprechen
- Tag X = xx.xx.: Vorträge. Aktive Mitarbeit.
  - Jeder Teilnehmer liest vorher **alle** Paper!
- Tag X + 2 Wochen. Schriftliche Ausarbeitung abgeben.

Def.: Combinatorial optimization problem (E, I, c), where E is a finite ground set, $I \subseteq 2^E$ is the set of feasible solutions, and $c: E \to \mathbb{R}$ is a cost function. The cost of a feasible solution $F \subseteq E$ is

$$c(F) = \sum_{e \in F} c(e).$$

An optimal solution is $F^* = \arg\max_{F \in I} c(F)$

Minimization can be achieved by adapting c.

Example: TSP

Instance: (C, d)

ground set $E = \{\{i,j\} \mid i,j \in C, i \neq j\} =: \binom{C}{2}$

Feas. solutions $I = \{F \subseteq E \mid F \text{ is tour}\}$

cost function $c(\{i,j\}) = -d(i,j)$

↑ we maximize in above def.

Example: Shortest path

Instance: $(G=(V,E), s, t)$ Find the shortest path from s to t in G.

here: number of edges

ground set: E

Feas. solutions $I = \{P \subseteq E \mid P \text{ is } s\text{-}t\text{-path}\}$

cost function $c: E \to -\mathbb{1}$

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Cancer Genomics

- MultiDendrix
  - [Leiserson MDM, Blokh D, Sharan R, Raphael BJ. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol*. 2013;9(5):e1003054. doi:10.1371/journal.pcbi.1003054.]
  - Distinguishing the somatic mutations responsible for cancer (driver mutations) from random, passenger mutations is a key challenge in cancer genomics. Driver mutations generally target cellular signaling and regulatory pathways consisting of multiple genes. This heterogeneity complicates the identification of driver mutations by their recurrence across samples, as different combinations of mutations in driver pathways are observed in different samples. We introduce the Multi-Dendrix algorithm for the simultaneous identification of multiple driver pathways de novo in somatic mutation data from a cohort of cancer samples. The algorithm relies on two combinatorial properties of mutations in a driver pathway: high coverage and mutual exclusivity. We derive an integer linear program that finds set of mutations exhibiting these properties. We apply Multi-Dendrix to somatic mutations from glioblastoma, breast cancer, and lung cancer samples. Multi-Dendrix identifies sets of mutations in genes that overlap with known pathways – including Rb, p53, PI(3)K, and cell cycle pathways – and also novel sets of mutually exclusive mutations, including mutations in several transcription factors or other genes involved in transcriptional regulation. These sets are discovered directly from mutation data with no prior knowledge of pathways or gene interactions. We show that Multi-Dendrix outperforms other algorithms for identifying combinations of mutations and is also orders of magnitude faster on genome-scale data. Software available at: http://compbio.cs.brown.edu/software.

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Cancer Genomics

- REVEALER
  - [Kim JW, Botvinnik OB, Abudayyeh O, et al. Characterizing genomic alterations in cancer by complementary functional associations. *Nat Biotechnol*. 2016;34(5):539-546. doi:10.1038/nbt.3527.]
  - Systematic efforts to sequence the cancer genome have identified large numbers of mutations and copy number alterations in human cancers. However, elucidating the functional consequences of these variants, and their interactions to drive or maintain oncogenic states, remains a challenge in cancer research. We developed REVEALER, a computational method that identifies combinations of mutually exclusive genomic alterations correlated with functional phenotypes, such as the activation or gene dependency of oncogenic pathways or sensitivity to a drug treatment. We used REVEALER to uncover complementary genomic alterations associated with the transcriptional activation of b-catenin and NRF2, MEK-inhibitor sensitivity, and KRAS dependency. REVEALER successfully identified both known and new associations, demonstrating the power of combining functional profiles with extensive characterization of genomic alterations in cancer genomes.

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Cancer Genomics

- HotNet2
  - [Leiserson MDM, Vandin F, Wu HT, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet*. 2014. doi:10.1038/ng.3168.]
  - Cancers exhibit extensive mutational heterogeneity, and the resulting long-tail phenomenon complicates the discovery of genes and pathways that are significantly mutated in cancer. We perform a pan-cancer analysis of mutated networks in 3,28 samples from 2 cancer types from The Cancer Genome Atlas (TCGA) using HotNet2, a new algorithm to find mutated subnetworks that overcomes the limitations of existing single-gene, pathway and network approaches. We identify 6 significantly mutated subnetworks that comprise well-known cancer signaling pathways as well as subnetworks with less characterized roles in cancer, including cohesin, condensin and others. Many of these subnetworks exhibit co-occurring mutations across samples. These subnetworks contain dozens of genes with rare somatic mutations across multiple cancers; many of these genes have additional evidence supporting a role in cancer. By illuminating these rare combinations of mutations, pan-cancer network analyses provide a roadmap to investigate new diagnostic and therapeutic opportunities across cancer types.

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Cancer Genomics

- OptDis
  - [Phuong Dao, Kendric Wang, Colin Collins, Martin Ester, Anna Lapuk, S. Cenk Sahinalp; Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics* 2011; 27 (13): i205-i213. doi: 10.1093/bioinformatics/btr245]
  - **Motivation:** Molecular profiles of tumour samples have been widely and successfully used for classification problems. A number of algorithms have been proposed to predict classes of tumor samples based on expression profiles with relatively high performance. However, prediction of response to cancer treatment has proved to be more challenging and novel approaches with improved generalizability are still highly needed. Recent studies have clearly demonstrated the advantages of integrating protein–protein interaction (PPI) data with gene expression profiles for the development of subnetwork markers in classification problems. **Results:** We describe a novel network-based classification algorithm (OptDis) using color coding technique to identify optimally discriminative subnetwork markers. Focusing on PPI networks, we apply our algorithm to drug response studies: we evaluate our algorithm using published cohorts of breast cancer patients treated with combination chemotherapy. We show that our OptDis method improves over previously published subnetwork methods and provides better and more stable performance compared with other subnetwork and single gene methods. We also show that our subnetwork method produces predictive markers that are more reproducible across independent cohorts and offer valuable insight into biological processes underlying response to therapy. **Availability:** The implementation is available at: http://www.cs.sfu .ca/~pdao/personal/OptDis.html

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Cancer Genomics

- An Efficient Branch and Cut Algorithm to Find Frequently Mutated Subnetworks in Cancer
  - Cancer is a disease driven mostly by somatic mutations appearing in an individual's genome. One of the main challenges in large cancer studies is to identify the handful of driver mutations responsible for cancer among the hundreds or thousands mutations present in a tumour genome. Recent approaches have shown that analyzing mutations in the context of interaction networks increases the power to identify driver mutations. In this work we propose an ILP formulation for the exact solution of the combinatorial problem of finding subnetworks mutated in a large fraction of cancer patients, a problem previously proposed to identify important mutations in cancer. We show that a branch and cut algorithm provides exact solutions and is faster than previously proposed greedy and approximation algorithms. We test our algorithm on real cancer data and show that our approach is viable and allows for the identification of subnetworks containing known cancer genes.

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Haploptype Assembly

- [Si, Hongbo, Haris Vikalo, and Sriram Vishwanath. 2014. “Haplotype Assembly: an Information Theoretic View.” In, 182–86. IEEE. doi:10.1109/ITW.2014.6970817]
  - This paper studies the haplotype assembly problem from an information-theoretic perspective. A haplotype is a sequence of nucleotide bases on a chromosome, often conveniently represented by a binary string, that differ from the bases in the corresponding positions on the other chromosome in a homologous pair. Information about the order of bases in a genome is readily inferred using short reads provided by high-throughput DNA sequencing technologies. Associating reads that cover variant positions with specific chromosomes in a homologous pairs, which enables haplotype assembly, is challenging due to limited lengths of the reads and presence of sequencing errors. In this paper, the recovery of the target pair of haplotype sequences using short reads is rephrased as a joint source-channel coding problem. Two messages, representing haplotypes and chromosome memberships of reads, are encoded and transmitted over a channel with erasures and errors, where the channel model reflects salient features of high-throughput sequencing. The focus of this paper is on determining the required number of reads for reliable haplotype reconstruction, and both the necessary and sufficient conditions are presented with order- wise optimal bounds.

Phylogenetic trees

- [Chimani, Rahmann, Böcker. Exact ILP Solutions for Phylogenetic Minimum Flip Problems. Proc. ACM-BCB 2010, Niagara Falls, NY, USA. ISBN 978-1-4503-0192-3]

  In computational phylogenetics, the problem of constructing a consensus tree or supertree of a given set of rooted input trees can be formalized in different ways. We consider the Minimum Flip Consensus Tree and Minimum Flip Supertree problem, where input trees are transformed into a 0/1/?-matrix, such that each row represents a taxon, and each column represents a subtree membership. For the consensus tree problem, all input trees contain the same set of taxa, and no ?-entries occur. For the supertree problem, the input trees may contain different subsets of the taxa, and unrepresented taxa are coded with ?-entries. In both cases, the goal is to find a perfect phylogeny for the input matrix requiring a minimum number of 0/1-flips, i.e., matrix entry corrections. Both optimization problems are NP-hard.

  We present the first efficient Integer Linear Programming (ILP) formulations for both problems, using three distinct characterizations of a perfect phylogeny. Although these three formulations seem to differ considerably at first glance, we show that they are in fact polytope-wise equivalent. In- troducing a novel column generation scheme, it turns out that the simplest, purely combinatorial formulation is the most efficient one in practice. Using our framework, it is possible to find exact solutions for instances with ~100 taxa.

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Motif finding

- [1. Zaslavsky E, Singh M. A combinatorial optimization approach for diverse motif finding applications. *Algorithms for molecular biology : AMB*. 2006;1(1):13. doi:10.1186/1748-7188-1-13]

  - **Background:** Discovering approximately repeated patterns, or motifs, in biological sequences is an important and widely-studied problem in computational molecular biology. Most frequently, motif finding applications arise when identifying shared regulatory signals within DNA sequences or shared functional and structural elements within protein sequences. Due to the diversity of contexts in which motif finding is applied, several variations of the problem are commonly studied. **Results:** We introduce a versatile combinatorial optimization framework for motif finding that couples graph pruning techniques with a novel integer linear programming formulation. Our approach is flexible and robust enough to model several variants of the motif finding problem, including those incorporating substitution matrices and phylogenetic distances. Additionally, we give an approach for determining statistical significance of uncovered motifs. In testing on numerous DNA and protein datasets, we demonstrate that our approach typically identifies statistically significant motifs corresponding to either known motifs or other motifs of high conservation. Moreover, in most cases, our approach finds provably optimal solutions to the underlying optimization problem. **Conclusion:** Our results demonstrate that a combined graph theoretic and mathematical programming approach can be the basis for effective and powerful techniques for diverse motif finding applications.