Dottorato di Ricerca in Informatica - Ciclo XXII
Dipartimento di Informatica, Sistemistica e Comunicazione
Facoltà di Scienze Matematiche, Fisiche e Naturali
Università degli Studi di Milano–Bicocca

Combinatorial Problems in Studies of Genetic Variations

Progress Report

September 11, 2008

Student:	Yuri Pirola
Supervisor:	Prof. Paola Bonizzoni
Tutor:	Prof. Lucia Pomello

Outline

1. Aims and Motivations
2. Results
   - Haplotype Inference - PPXH
   - Expressed Sequence Alignment
3. Future Work and Conclusions
   - Future Work

Aims and Motivations

Aim

Analysis and design of combinatorial methods to perform large-scale studies of genetic variations.

Motivation: new sequencing technologies produce a lot of data.

Problems:

- Haplotype Inference (HI)
- Expressed Sequence Alignment

Work Done

Haplotype Inference → *Pure-Parsimony Xor-Haplotyping*

- Study and analysis of several complexity aspects
- Resolution by means of different techniques

Expressed Sequence Alignment → *Sequence Factorization*

- Algorithm design
- Factorization Agreement ("multiple expressed sequence alignment")

Haplotype Inference

Haplotype Inference Problem

For each individual in a population, distinguish the genome inherited from each parent accordingly to a reference genetic model.

- Well-known problem, studied under different assumptions.
- Pure-Parsimony Xor-Haplotyping (PPXH)

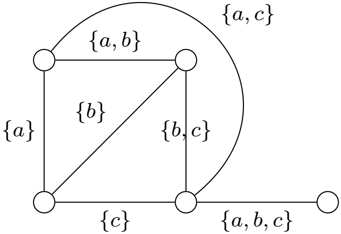
PPXH - Summary

Pure-Parsimony Xor-Haplotyping (PPXH):

- Modeling by combinatorial structures
- Analysis of the computational complexity
- Analysis of the parametrized complexity
- Design of exact algorithms
- Design of heuristic algorithms

Problem Modeling

- Modeling of the solutions as labelled graphs satisfying certain algebraic properties → Xor-Graph
- Study of the properties of the Xor-Graph



A Xor-Graph

Computational Complexity

- Investigation of the computational complexity of PPXH:
    - NP-hard? ongoing work

- Proof by L-reduction from Min-Vertex-Cover
    - APX-hard?

Parametrized Complexity

Parametrized Complexity: “A framework for sistematically confronting computational intractability” (Downey and Fellows, 1997)

- PPXH is *fixed parameter tractable* (with regard to the optimum)
- Proof by “kernelization”

Polynomial Exact Algorithms for restrictions of PPXH

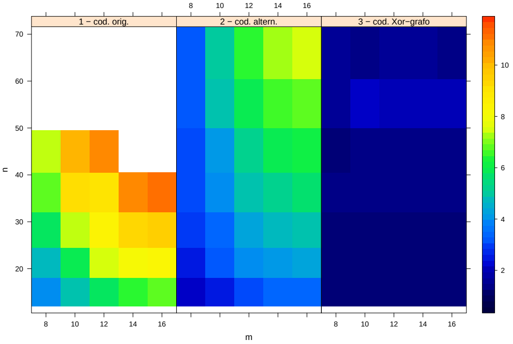
Design of polynomial exact algorithms to solve specific (and motivated) restrictions of PPXH:

- PPXH(*,2)
- PPXH(2,*)
- some instances with a particular "structure"

Heuristics

PPXH was heuristically solved by Genetic Algorithms during the PhD course of Dr. Vanneschi.

Avg. approximation factor of several coding techniques



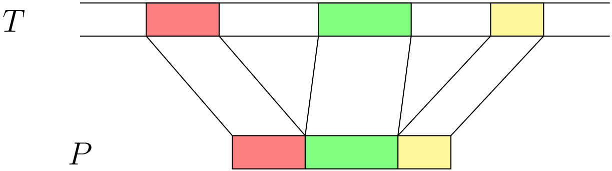
Using the Xor-Graph model increased the performance!

Expressed Sequence Alignment

Sequence Factorization Problem

Given two sequences P and T, partition P into a list of factors such that they occur in T in the same order.

Example:



Different factorizations can exist!

Sequence Factorization Problem

Design of an algorithm to find all the “maximal” factorizations of a pair of sequences

- efficient (it uses suffix trees!)
- compact representation of the set of factorizations

How to choose the “right” factorization?

- Idea: exploiting the redundancy of the libraries of expressed sequences

→ *definition of a new optimization problem!*

Factorization Agreement Problem

Factorization Agreement Problem

Given all the factorizations of a set S of sequences w.r.t. a sequence T, choose the minimum cardinality set F of factors of T such that each sequence of S can be factorized by using only factors that belong to F.

Results:

- NP-hard (by reduction from Min-Set-Cover)
- Algorithm that should perform well on real data

Future Work

PPXH:

- completion of the work on computational complexity
- design of approximation algorithms
- experimental assessment of the model

Expressed Sequence Alignment:

- implementation of the algorithms (*ongoing work*)
- experimentation on real sequences
- application on gene structure and alternative splicing prediction, gene clustering. . .

Courses and Summer Schools

PhD Courses:

- “Multilevel Models”, Blangiardo
- “Reti Bayesiane”, Fagiuoli
- “Teoria e Applicazioni del Calcolo Evoluzionistico”, Vanneschi
- “Biomolecular Computing: Theory and Experiments”, Jonoska

Summer Schools:

- International Summer School on Bioinformatics and Computational Biology, 2008

English course: from October 2007 to April 2008 (2 editions)