

Combinatorial Problems in Studies of Genetic Variations: Haplotyping and Transcript Analysis

September 8, 2009

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

Outline

- 1 Aims and Motivations
- 2 Results
 - Haplotype Inference
 - Pure Parsimony Xor Haplotyping
 - Haplotyping on Pedigrees
 - Transcript Analysis
 - Gene Structure Prediction
- 3 Conclusions

Aims and Motivations

Studies of genetic variations are one of the most important task in the post-genomic era.

But...

- Lot of data are needed
- Cost/technological reasons limit data availability

Aim

Analysis and design of combinatorial methods that enable large-scale studies of genetic variations.

Original Contributions

Haplotype Inference:

- Exact and approximate algorithms for two haplotyping problems:
 - Pure Parsimony Xor Haplotyping (PPXH)
 - Haplotyping on Pedigrees with Mutations and Recombinations (MEHC)

Transcript Analysis:

- Efficient algorithm which exploits redundancy to perform gene structure prediction

Haplotype Inference

Haplotype Inference Problem

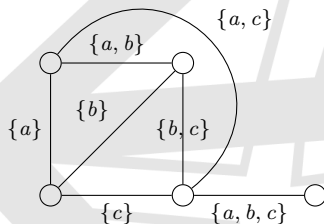
For each individual in a population, distinguish the genome inherited from each parent accordingly to a reference genetic model.

- Well-known problem, studied under different assumptions.
- Pure Parsimony Xor Haplotyping (PPXH)
- Minimum Events Haplotype Configuration (MEHC)

Pure Parsimony Xor Haplotyping

Pure Parsimony Xor Haplotyping

Characterization of solutions as graphs \rightarrow Xor graph



A Xor-Graph

PPXH - Exact Algorithms

PPXH is *fixed parameter tractable*

- $O(2^{k^2} nm)$ time algorithm
(parameter k = size of a optimal solution)

Polynomial-time algorithms for specific (and motivated) restrictions:

- PPXH(*,2)
- PPXH(2,*)

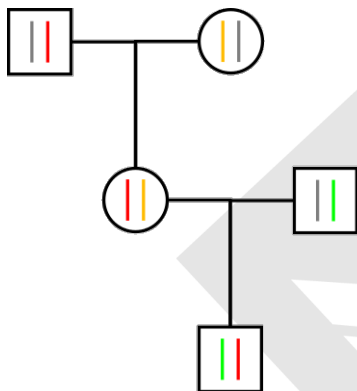
PPXH - Heuristic Algorithm

Heuristic Algorithm:

- Based on Xor-graph reconstruction
- Efficient: $O(\alpha(n, m)n^3m)$ time complexity
- Experimental validation on various kinds of instances
- Experimental observations: **performs well**
 - approximation factor ≤ 1.57 (often close to 1)
 - time $\leq 1h$ (on big instances)

Haplotyping on Pedigrees

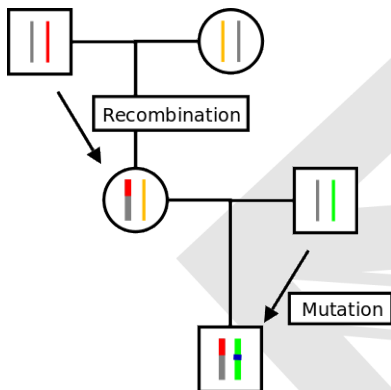
Pedigree \rightarrow parental relationships



Polynomial if “pure”
Mendelian Inheritance

Haplotyping on Pedigrees

Pedigree \rightarrow parental relationships



Polynomial if “pure”
 Mendelian Inheritance

Intractable if we admit
genetic variation events

Minimum Events Haplotype Configuration

		Recombinations	
		NO	YES
Mutations	NO	Polynomial	<i>APX-hard</i> Randomized algorithm ILP formulation
	YES	<i>NP-hard</i> Exponential algorithm	

Minimum Events Haplotype Configuration

		Recombinations	
		NO	YES
Mutations	NO	Polynomial	<i>APX-hard</i> Randomized algorithm ILP formulation
	YES	<i>NP-hard</i> Exponential algorithm APX-hardness	NP-hardness Heuristic

Original contributions

MEHC - Heuristic Algorithm

Minimum Events Haplotype Configuration (MEHC):

- connected (via L-reduction) to a well-known Information Theory problem (DECODING OF LINEAR CODES)

Heuristic Algorithm (based on the L-reduction):

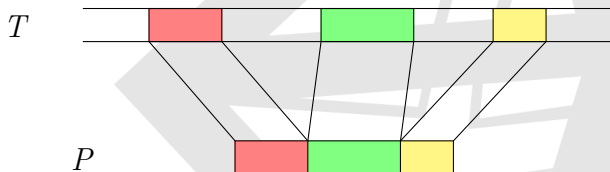
- Efficient: $O(n^3 m) + O(n^3 m^3 \cdot k)$ time complexity
- Experimentally validated: **extremely good performances**
 - 99.1% success rate (mutations and recombinations)
 - 100% success rate (only mutations on a real pedigree)
 - 99.8% success rate (only recombinations on a real ped.)
 - time per instance $\leq 16m$

Transcript Sequence Factorization

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 transcripts
→ *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 MINSETCOVER)
- Size-reduction algorithm + enumeration
 - performs well on some significant genes

Conclusions

Haplotype Inference:

- haplotyping under two different models, PPXH and MEHC
- coping computational intractability using different techniques
 - restrictions, FPT, heuristics, ...

Transcript Analysis:

- algorithm to find alternative factorizations
- gene structure prediction via factorization agreement

Publications

Bonizzoni, Della Vedova, Dondi, **Pirola**, and Rizzi.

“Pure Parsimony Xor Haplotyping”.

In *Proceedings ISBRA 2009*, 186–197, 2009.

(An extended version has been submitted to *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.)

Bonizzoni, Della Vedova, Dondi, **Pirola**, and Rizzi.

“Minimum Factorization Agreement of Spliced ESTs”.

In *Proceedings WABI 2009*, to appear, 2009.

The work about the *Minimum Events Haplotype Configuration* problem has been carried out while I was a visiting student at *University of California, Riverside* under the supervision of Prof. Tao Jiang (feb–jul 2009).

A manuscript is in preparation.

