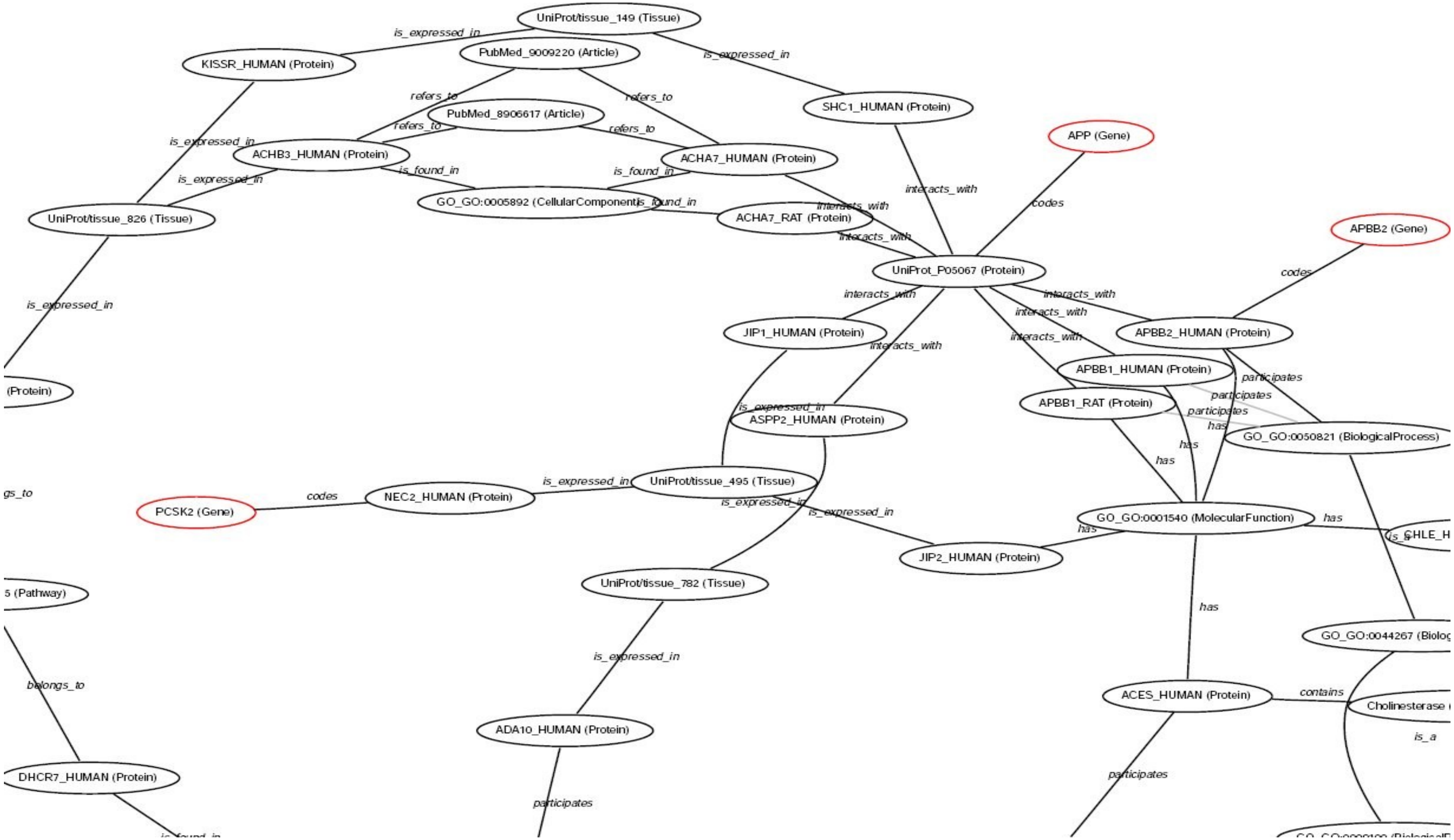
Discovery group: Data mining for pattern and link discovery

Hannu Toivonen
Petteri Sevon

University of Helsinki and HIIT

Structured and heterogeneous data



Mission

We develop novel methods and tools for pattern and link discovery

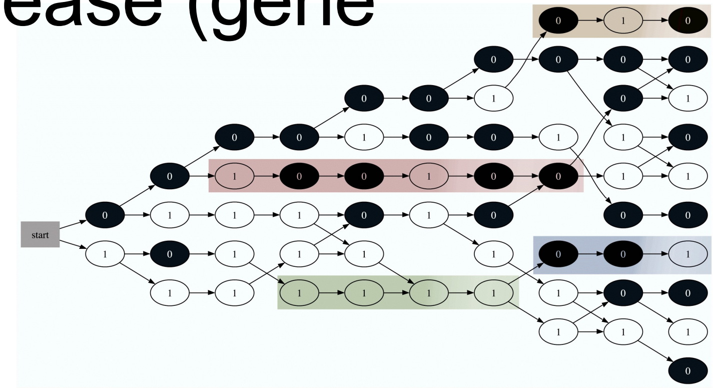
- Focus on structured and heterogeneous data (graphs, sequences)
- Applications in bioinformatics, genetics and ubicomp
- Collaboration with applied scientists and companies

Projects

- Analysis of genetic marker data
  - haplotyping of unrelated genotypes (HaploRec)
  - association-based gene mapping (TreeDT, HPM)
- Analysis of heterogeneous biological networks (Biomine)
- Context recognition by user situation data analysis

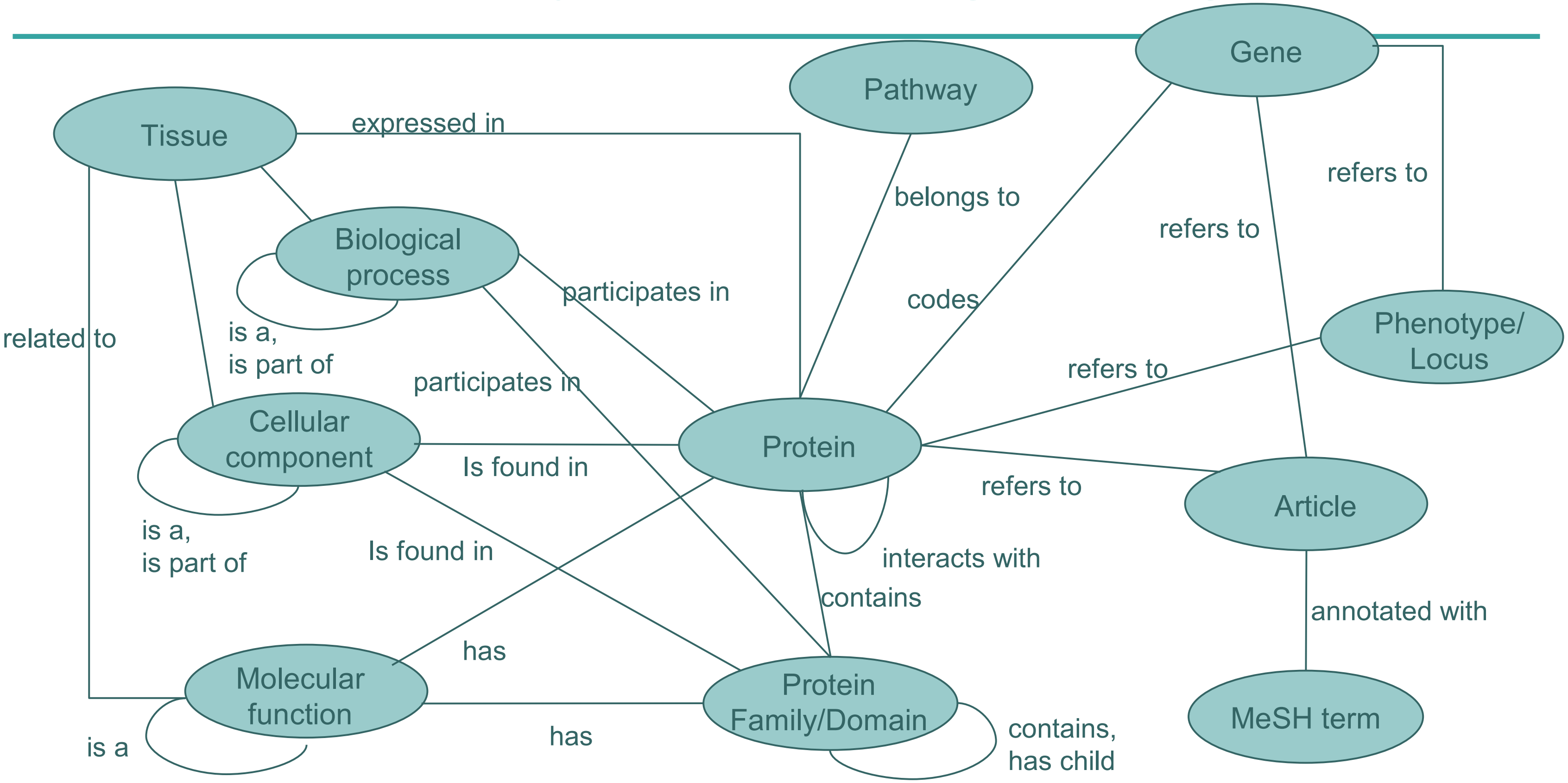
Analysis of genetic marker data

- Goal: Discovery of associations between genomic regions and a disease (gene mapping)



- HaploRec: Haplotyping unrelated individuals using variable-length Markov models (Eronen, Geerts, Toivonen, BMC Bioinformatics 2006)
- TreeDT: gene mapping method measuring disease-association in estimated genealogical trees (Sevon, Toivonen, Ollikainen, IEEE/ACM Transactions on Comp. Biol. and Bioinformatics 2006)
- A medical result: discovery of a novel asthma gene (Science 2004)

Biomine: Analysis of weighted graphs

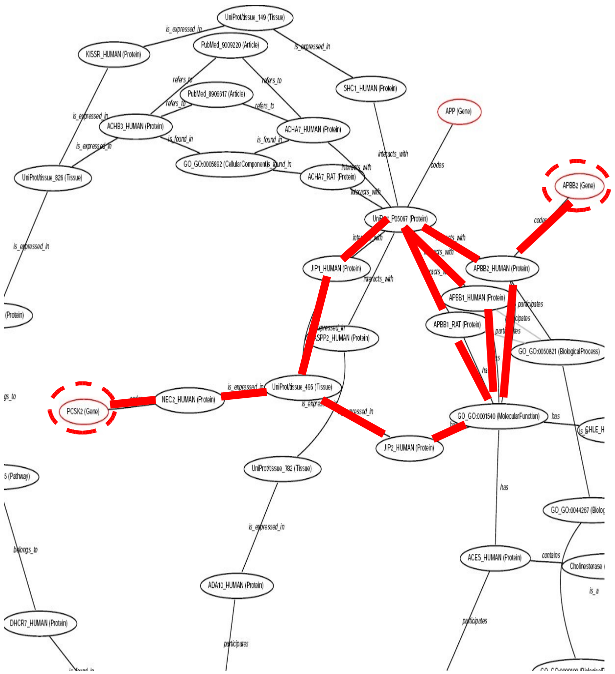


- Millions of heterogeneous nodes and edges in public biological databases
- Edge weight = probability

Example 1: Link discovery and analysis

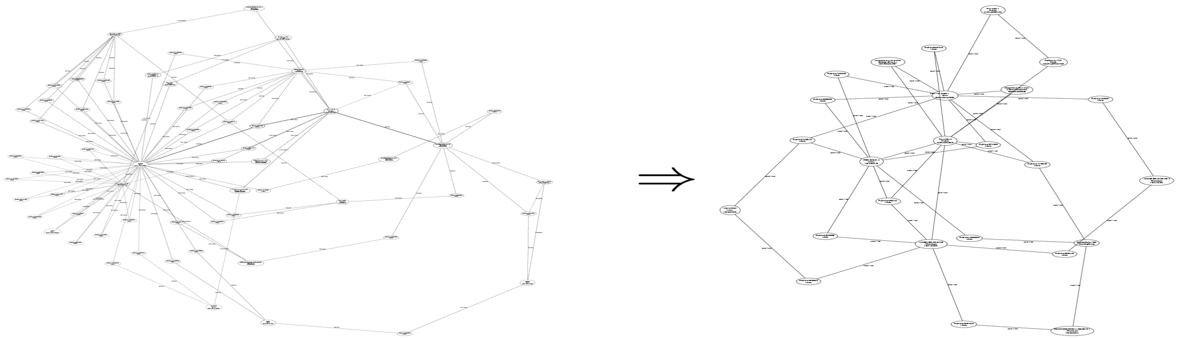
- Goal: discovery and evaluation of non-trivial connections between nodes, e.g., a gene and a disease

- A true link between two nodes is manifested as a subgraph connecting them
- We measure the strength of such a subgraph using two-terminal network reliability (Sevon et al., DILS 2006)
- On-going/future work: use of context-free grammar in complex queries (queries by specific path classes and production rule specific relevance coefficients)



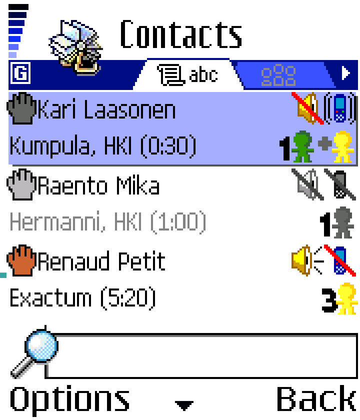
Example 2: Extraction of subgraphs

- Goal: extraction of a maximally relevant subgraph for given nodes or connections



- Concepts and methods for probabilistic graphs:
- Removal of edges that contribute least to the two-terminal network reliability (Hintsanen 2007, submitted)
- Compression of probabilistic Prolog theories with minimal reduction to likelihood of given positive and negative examples (De Raedt et al., ILP 2006)

Context



Contacts

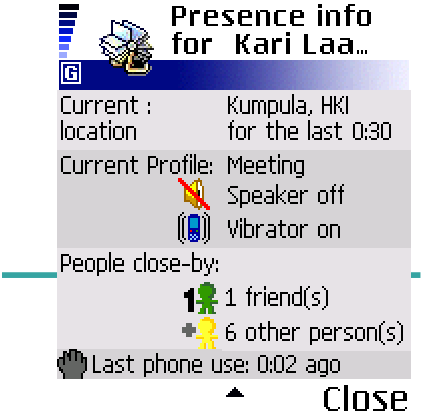
abc

Kari Laasonen
Kumpula, HKI (0:30) 1

Raento Mika
Hermanni, HKI (1:00) 1

Renaud Petit
Exactum (5:20) 3

Options Back



Presence info for Kari Laa…

Current : location	Kumpula, HKI for the last 0:30
Current Profile:	Meeting
	Speaker off
	Vibrator on
People close-by:	
1	1 friend(s)
	6 other person(s)

Last phone use: 0:02 ago

Close

- Goal: discovery of context patterns, studies of privacy issues in social awareness
- ContextPhone (IEEE Pervasive Computing 2005, Raento et al)
  - context data collection, distribution, display
  - the first public toolbox for standard mobile phones
- Applications: presence-augmented phonebook, automatic media annotation and distribution, logger
  - used by MIT, Berkeley, Nokia, ...
  - covered e.g. in The New York Times, New Scientist
- Provides unique data for finding patterns and models of interaction
- Research into algorithms, privacy issues, social networks, ...

Vision

- The importance of data mining in heterogeneous and structured data will grow, especially in scientific applications such as bioinformatics

- We develop novel methods and apply them succesfully to challenging and important applications problems, in collaboration with applied scientists