cmcompare - Webserver

Florian Eggenhofer

Institute for Theoretical Chemistry
University of Vienna

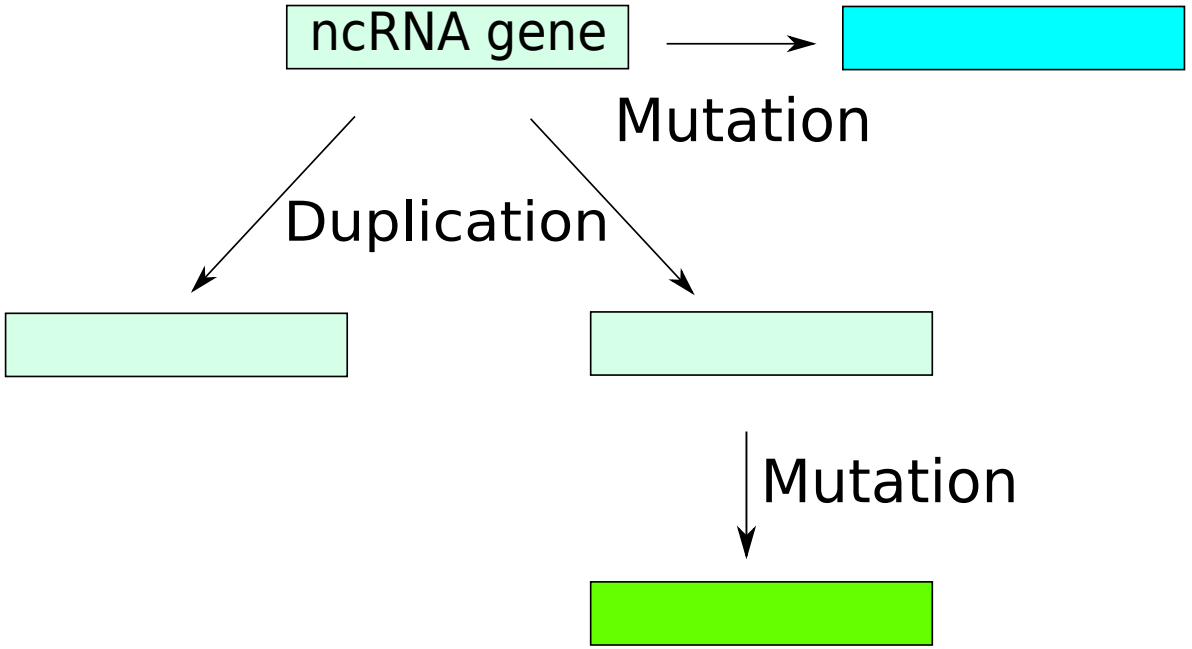
February 14, 2012

ncRNA Homology Search 1

- Diversification:

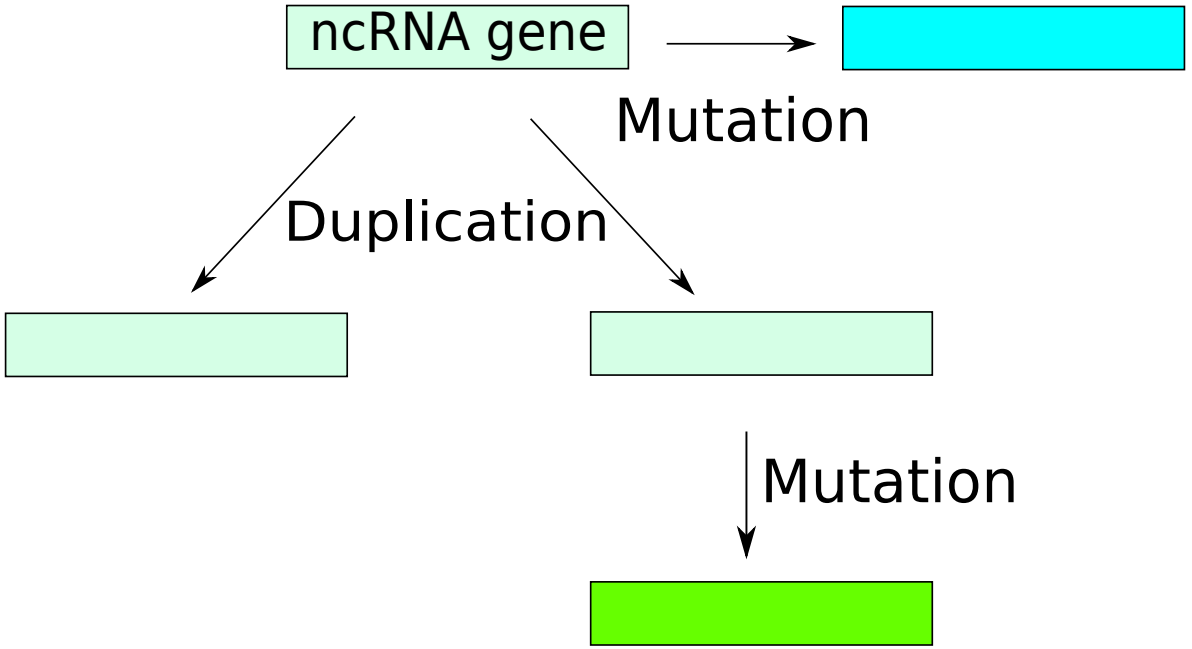
ncRNA Homology Search 1

- Diversification:



ncRNA Homology Search 1

- Diversification:



- Speciation event $\longrightarrow$ ortholog gene
- Gene duplication $\longrightarrow$ paralog gene

ncRNA Homology Search 2

- Helpful in finding related genes
- Simple case: $\longrightarrow$ conserved sequence $\longrightarrow$ profile HMM
- used for protein families

ncRNA Homology Search 2

- Helpful in finding related genes
- Simple case: $\longrightarrow$ conserved sequence $\longrightarrow$ profile HMM
- used for protein families
- what about genes with low sequence conservation?

Structure and Function 1

- More distantly related genes.. Sequence weakly conserved
- but function is conserved
- function $\longleftrightarrow$ structure
- secondary RNA structure $\longrightarrow$ basepairing

Structure and Function 2

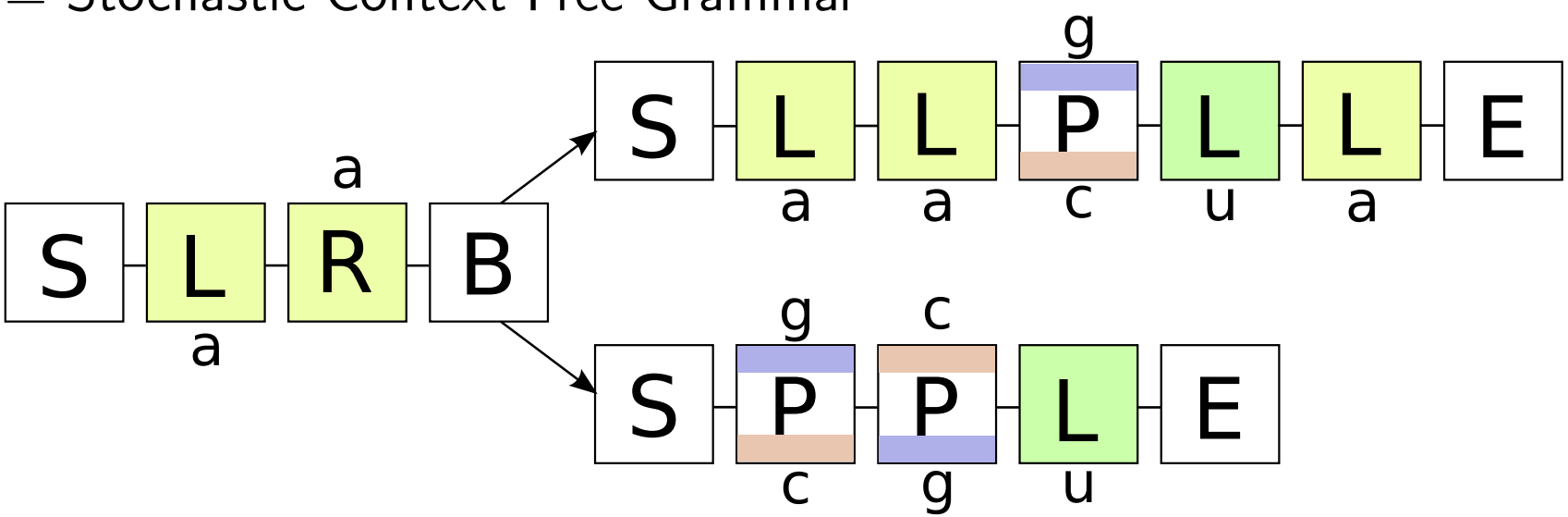
- ncRNA gene finding $\longrightarrow$ covariance models (cm)
- considers both basepairing and sequence
- what is a covariance model?

Covariance models

- represent ncRNA families with profile SCFGs
- = Stochastic Context Free Grammar

Covariance models

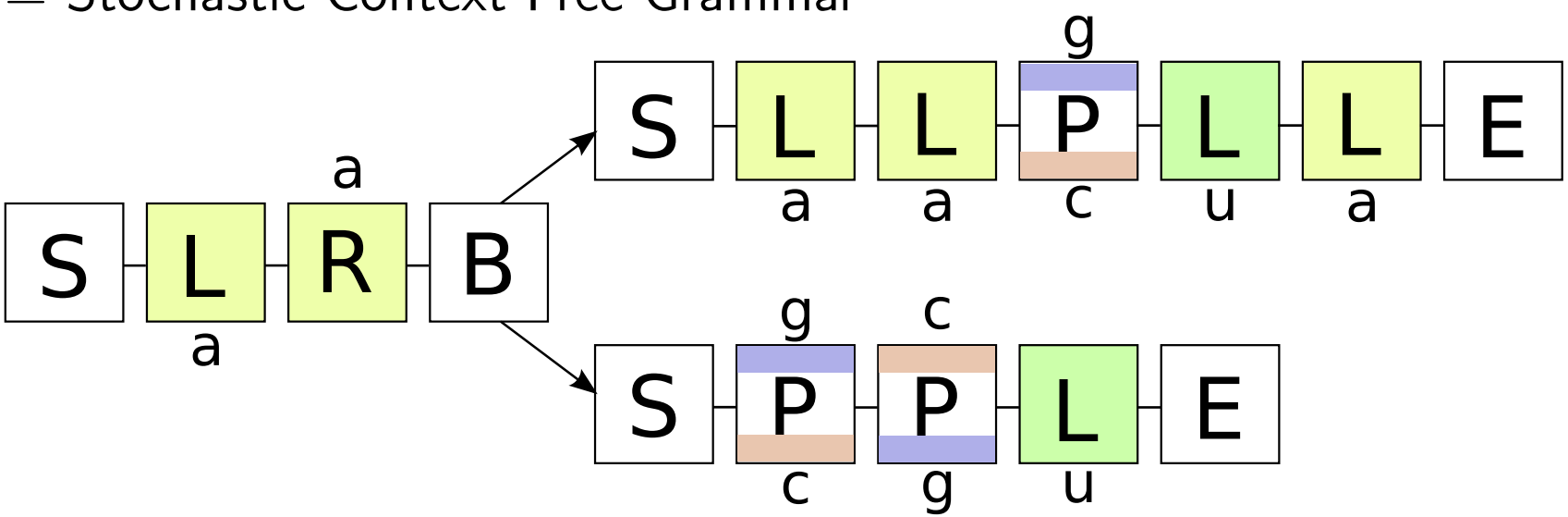
- represent ncRNA families with profile SCFGs
- = Stochastic Context Free Grammar



- SCFG is very general, cm specific
- abstract representation of RNA families

Covariance models

- represent ncRNA families with profile SCFGs
- = Stochastic Context Free Grammar



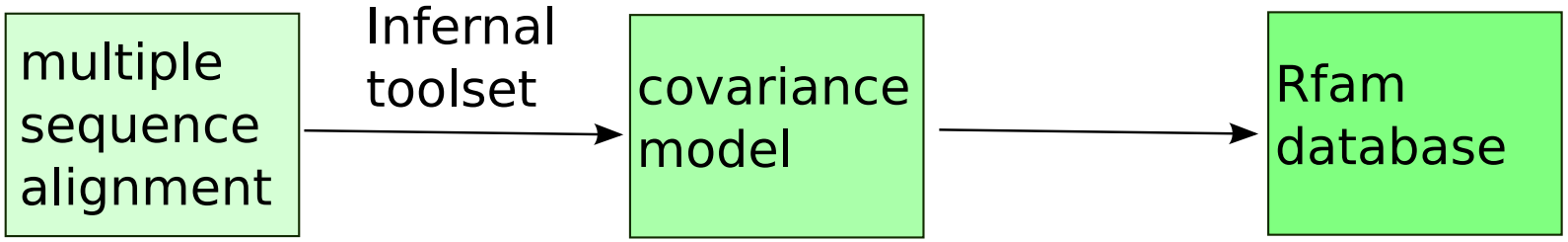
- SCFG is very general, cm specific
- abstract representation of RNA families
- How to build a covariance model?

Infernal + Rfam

- cm construction pipeline:

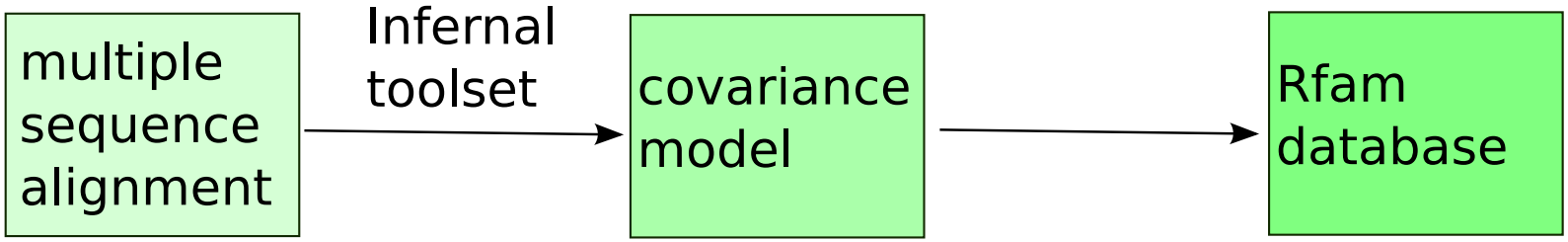
Infernal + Rfam

- cm construction pipeline:



Infernal + Rfam

- cm construction pipeline:



- Rfam 10.0 = 1446 RNA families, $>$ 3M genes
- cm quality? $\longrightarrow$ cmsearch

cm quality

- Search new genes with a cm:
- all substrings of given genome
- transition, transmission $\longrightarrow$ score

cm quality

- Search new genes with a cm:
- all substrings of given genome
- transition, transmission $\longrightarrow$ score

- Specificity
- 2 cms score high for same sequence
- $\longrightarrow$ specificity is low
- $\longrightarrow$ cmcompare

cmcompare

- Input $\longrightarrow$ 2 cms
- MaxiMin algorithm
- Output $\longrightarrow$ Link score and Link sequence

cmcompare

- Input $\longrightarrow$ 2 cms
- MaxiMin algorithm
- Output $\longrightarrow$ Link score and Link sequence

- highest scoring string in both models (suboptimals)
- link score is bit-score describing similarity
- shows relatedness of primary and secondary structure

Motivation

- Make cmcompare more accessible to rfam users
- Use power of gui to visualize cm relationships
- Improve cm quality
- $\longrightarrow$ Clear separation of models

Features 1

- Provide features of commandline-tool and more
- Start comparisons with multiple sequence alignments

Features 1

- Provide features of commandline-tool and more
- Start comparisons with multiple sequence alignments
- compare cm against all other cms in rfam
- comparison of a provided set of models

Features 2

- Other models for the same family exist?

Features 2

- Other models for the same family exist?
- cm submodel of other cm?

Features 2

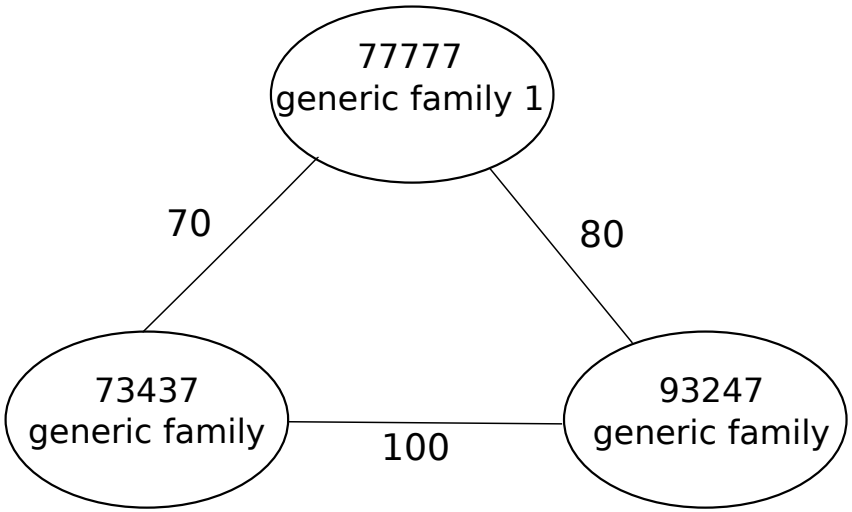
- Other models for the same family exist?
- cm submodel of other cm?
- cm supermodel of other cm?

Features 2

- Other models for the same family exist?
- cm submodel of other cm?
- cm supermodel of other cm?
- Model duplications?

Features 2

- Other models for the same family exist?
- cm submodel of other cm?
- cm supermodel of other cm?
- Model duplications?
- Result: Map of relationships



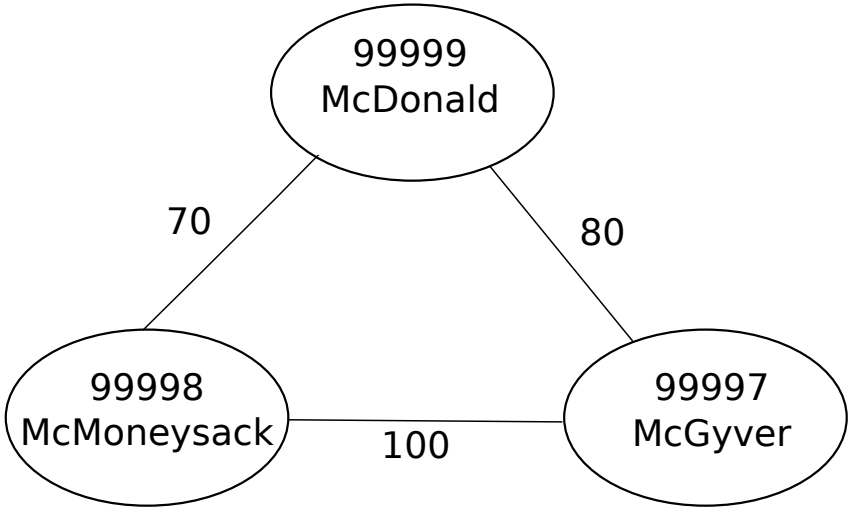
Looking for clans 1

Looking for clans 1

- Clans group biologically related RNA families
- Rfam 10.0: 99 clans, e.g. RNase P

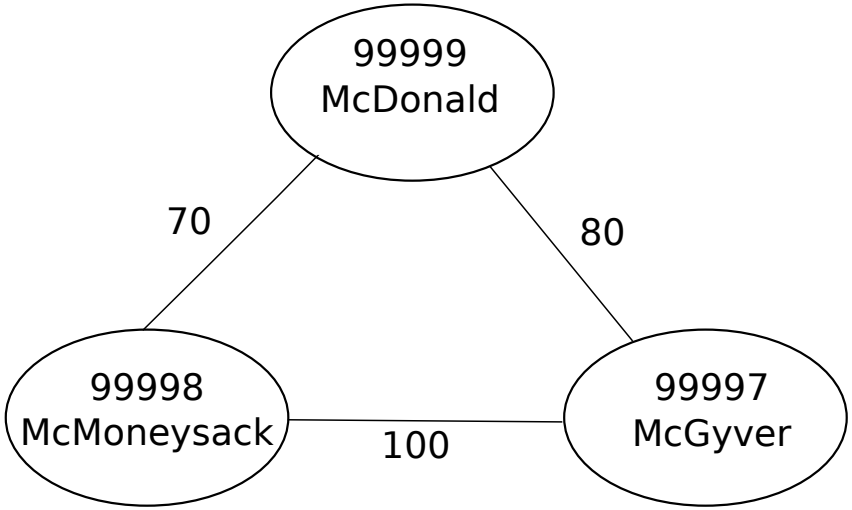
Looking for clans 2

- Problem: group of families with high link score



Looking for clans 2

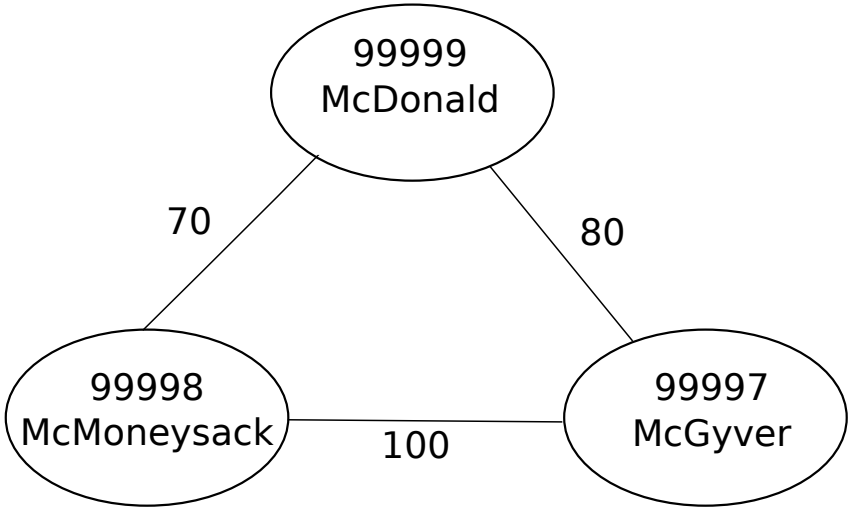
- Problem: group of families with high link score



- Biological relation?
- High link score = primary and secondary structure related

Looking for clans 2

- Problem: group of families with high link score



- Biological relation?
- High link score = primary and secondary structure related
- GO-terms

GO-terms

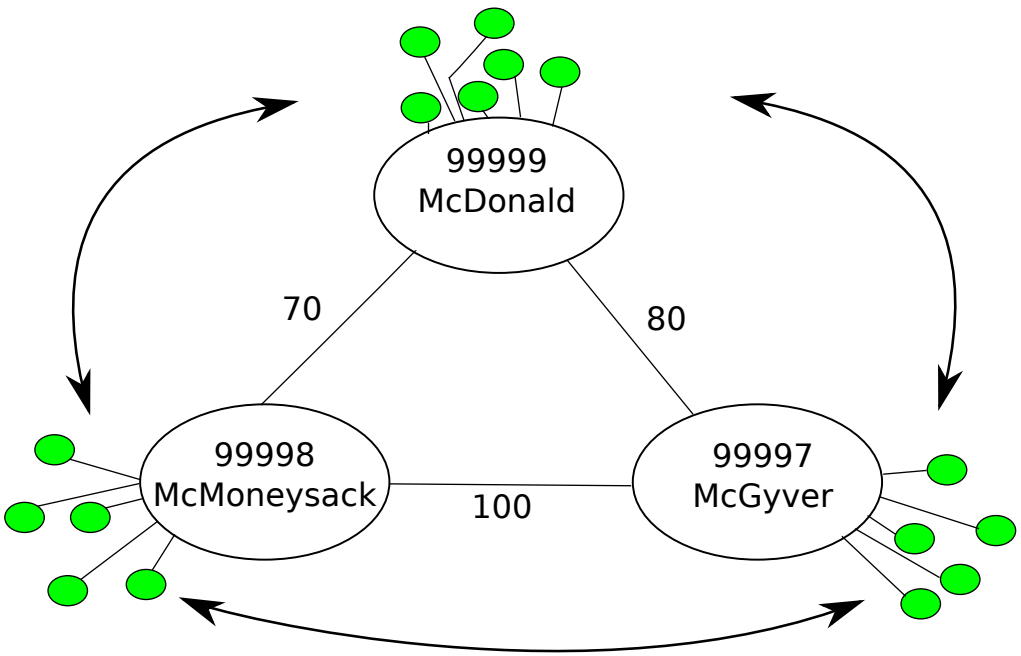
- Similarity in GO terms?
- GO = Gene Onthology
- Associates terms from 3 categories with genes
- Biological Process, Cellular Component, Molecular Function
- Sugar import, Membrane, Transporter

GO-terms 2

- Strategy

GO-terms 2

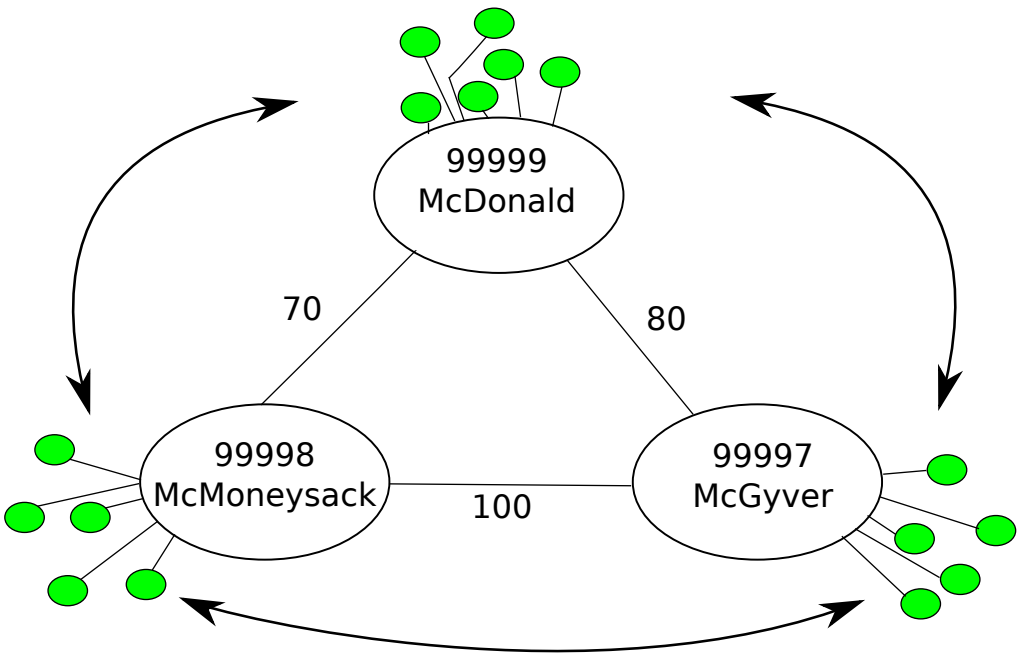
- Strategy



- looking for significant overlaps in associated terms

GO-terms 2

- Strategy



- looking for significant overlaps in associated terms
- obstacle: GO-annotation

Thanks

- Thanks for your attention!