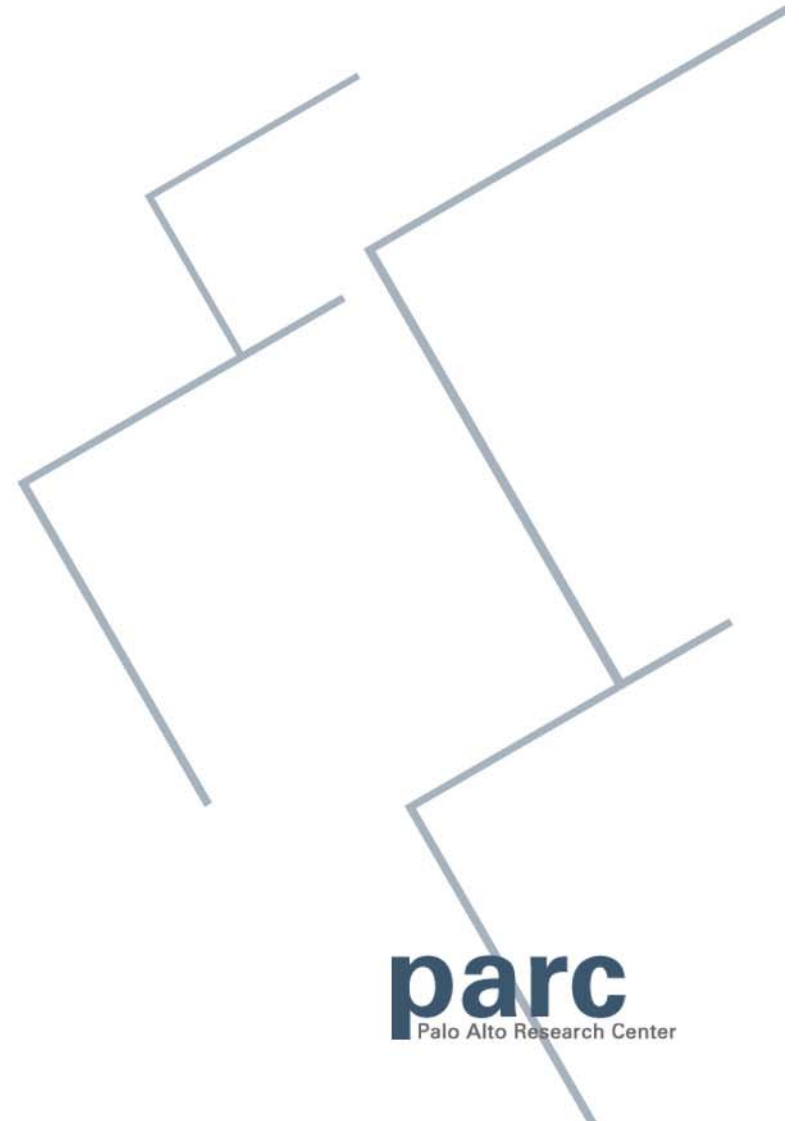


Doing more with Single MS

David Goldberg
PARC



Palo Alto Research Center

- Formerly research division of Xerox
- Now wholly owned subsidiary
- This gave us opportunity branch out



Motivation: Gene vs. Genomics

- Whole genome sequencing followed by computational gene finding
 - Misses genes
 - Even when it finds a gene, the location may be off
- But still very useful!
 - Gives a sense of the genome
 - Speeds up biochemical validation
- Genomics and biochemical validation are complementary

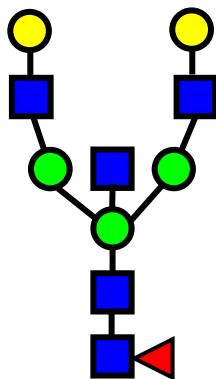
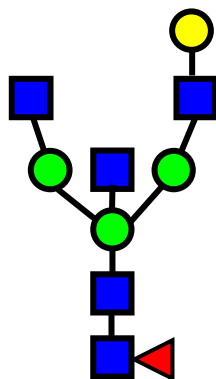
Analogy with Glycomics

- MALDI-TOF spectrum like a gene sequence
- Finding cartoons from spectrum like finding genes from a genome sequence
- Complements more detailed studies
 - MS/MS and MSⁿ
 - Glycosidase digestions
 - Gas chromatography linkage analysis

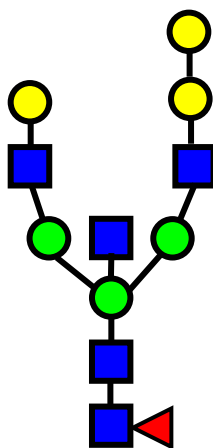
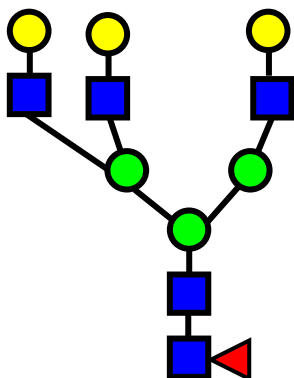
Doing more with Single MS

- In this spirit, how can you maximize the information in a single MS MALDI profile?
- Will exploit biosynthetic pathways
- Single MS gets (often unique) composition
 - But many possible cartoons
 - Don't pick cartoon for a peak in isolation
 - Pick cartoons that form families

Example:



If peaks 2286 and 2490 are known to have these structures

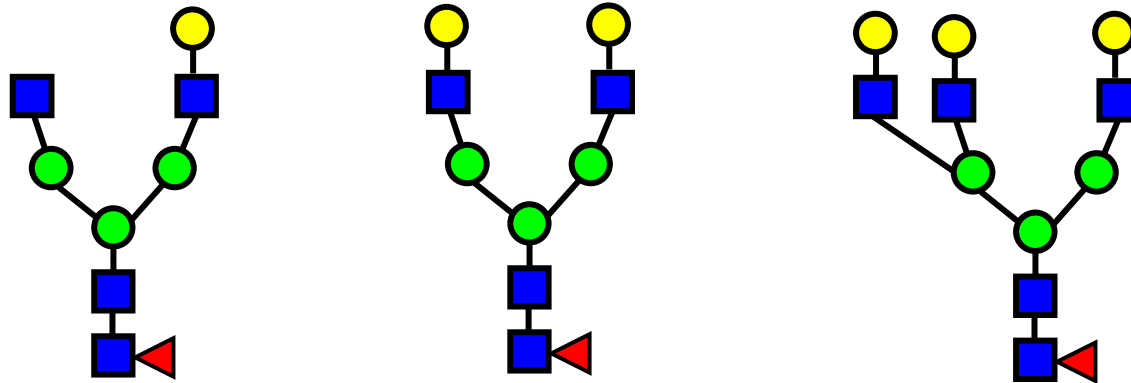


Which of these cartoons should be chosen for peak at 2694?

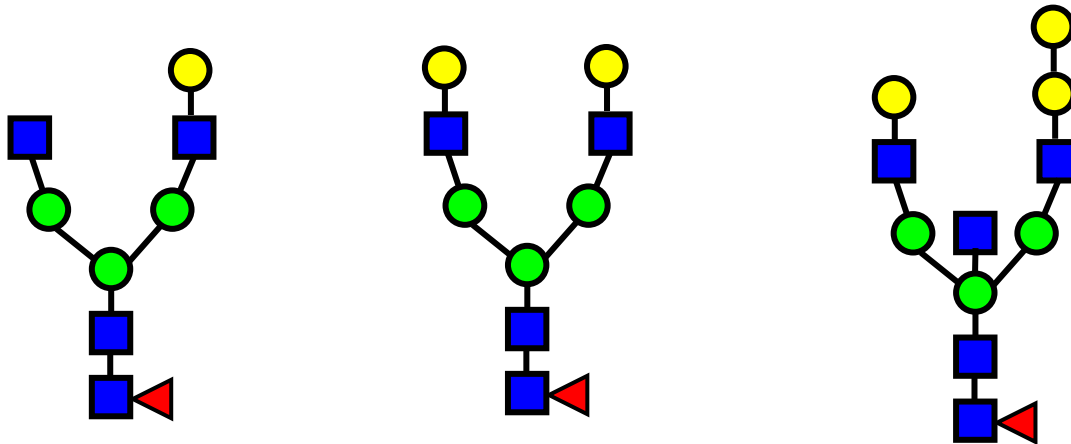
Testing the Idea

- For ground truth, take spectra on CFG website
- Compare 3 different ways of capturing 'biosynthetic pathways' to ground truth
 - Parsimony
 - Max subgraph
 - Random Walk
- These are mathematical, not biological models
 - Much as gene finding is done with HMM's

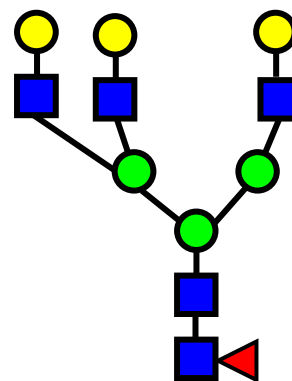
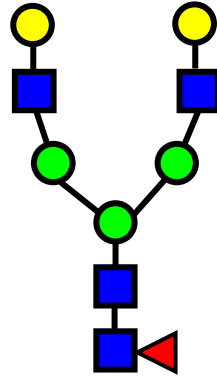
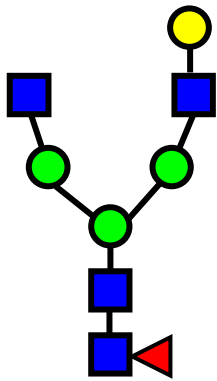
Parsimony



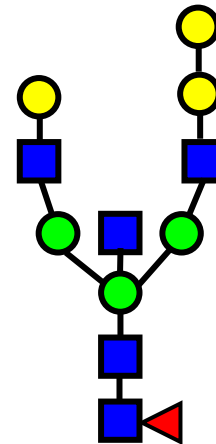
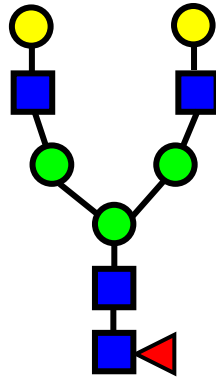
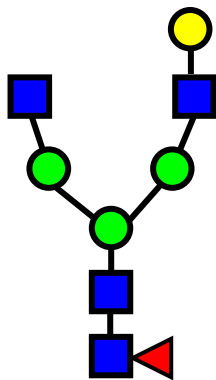
Want to minimize number of antennae



Parsimony



2 antennae

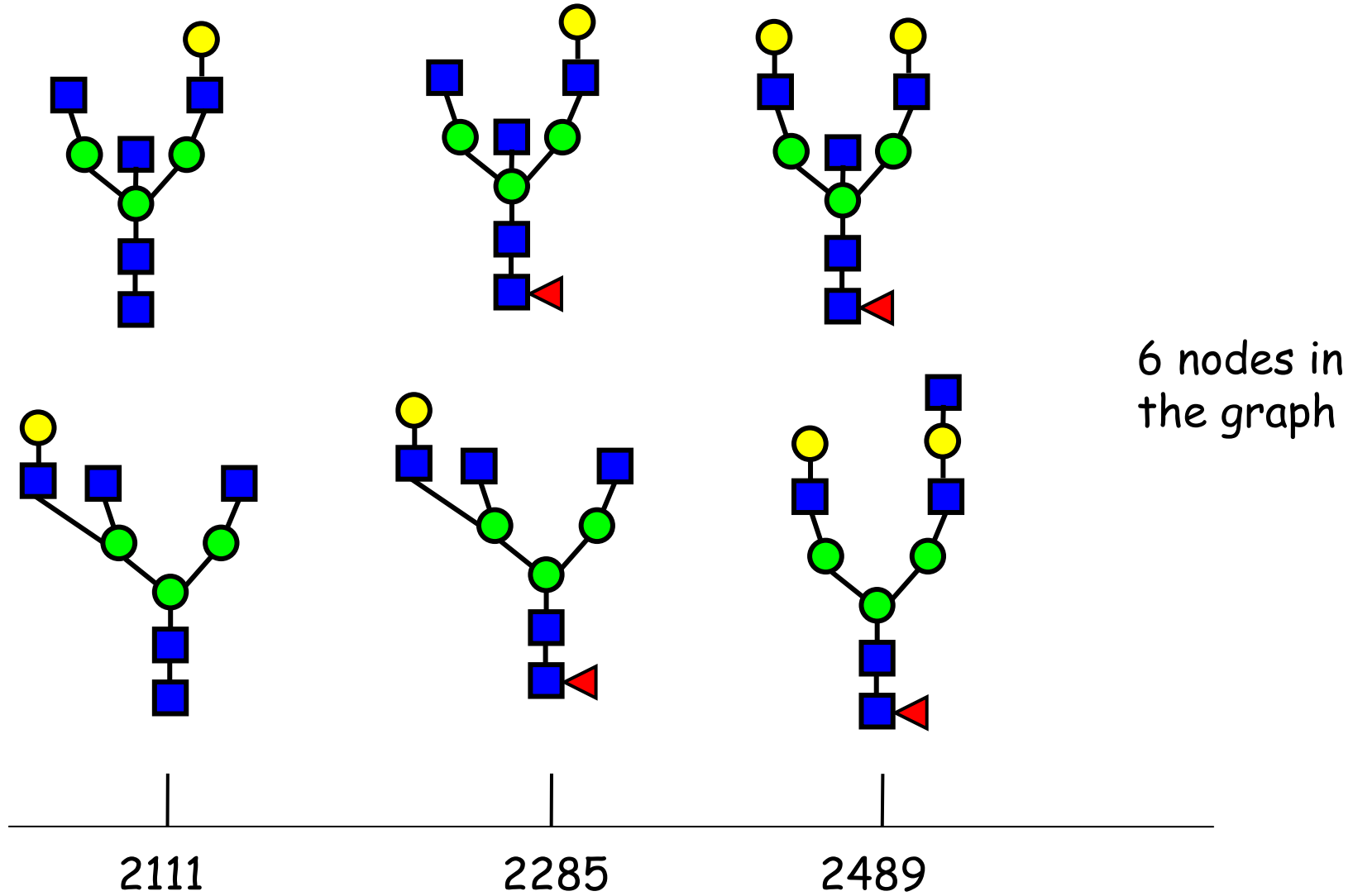


3 antennae

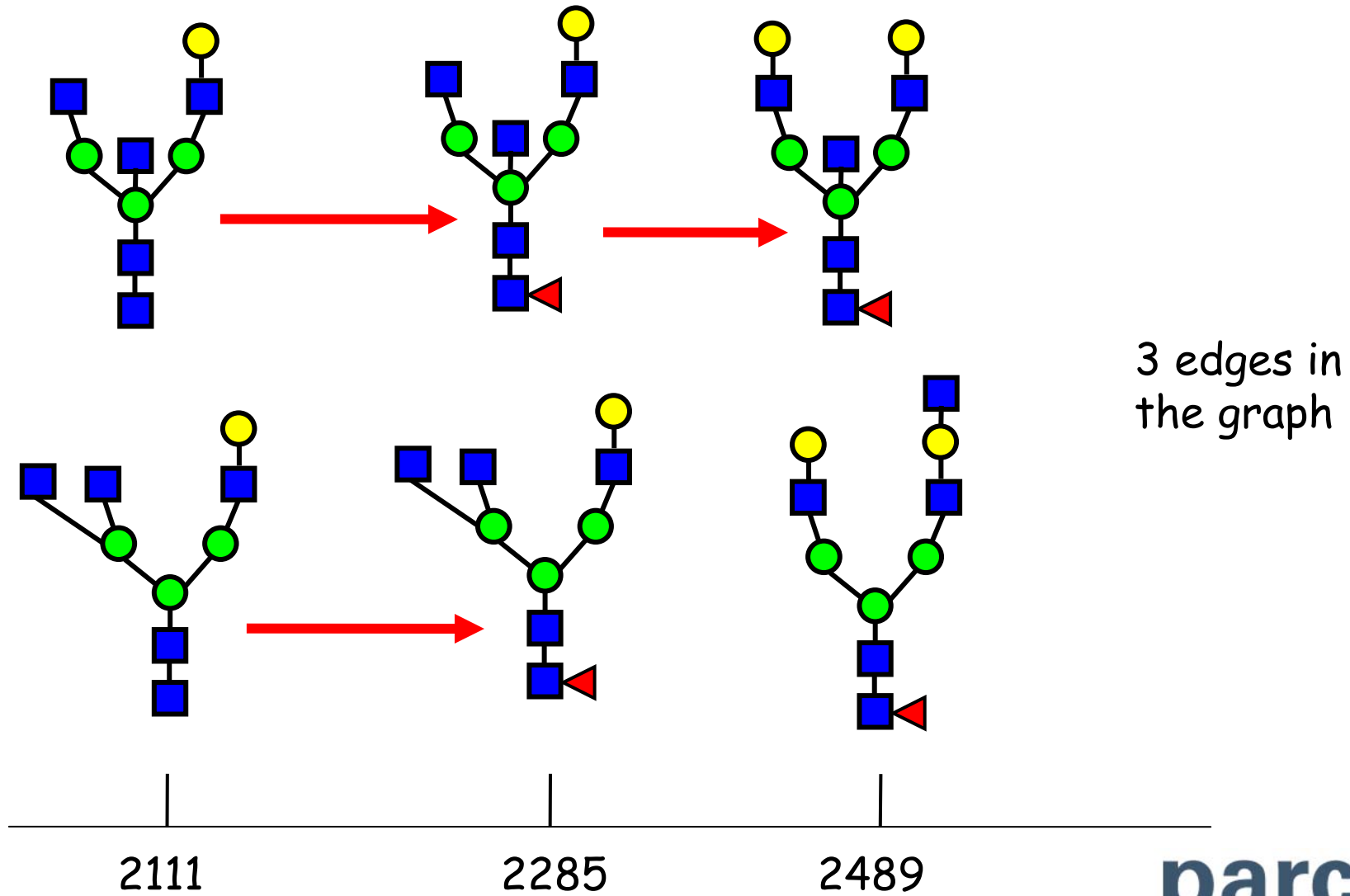
Max Subgraph and Random Walk

- Use "Cartoon Graph"
- The nodes of the graph are cartoons
 - All possible cartoons for all peaks in the spectrum
- The edges of the graph are biosynthetic paths
 - Two cartoons are connected if they are on the same biosynthetic pathway

Nodes of cartoon graph



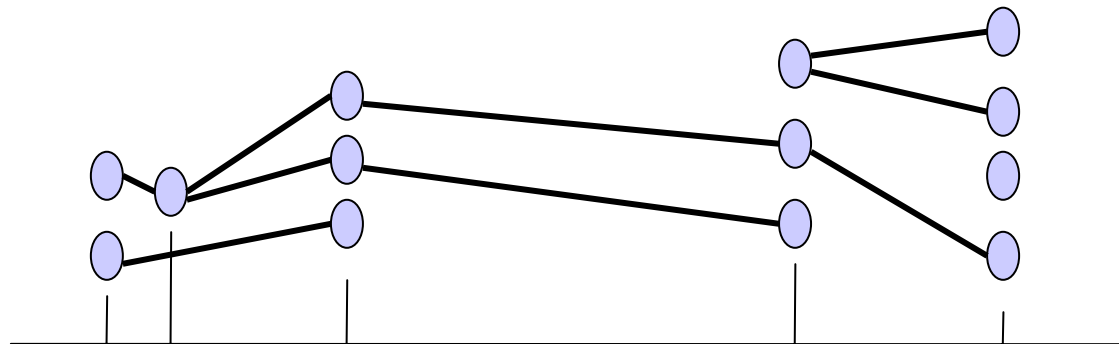
Edges of cartoon graph



3 edges in the graph

Max Subgraph

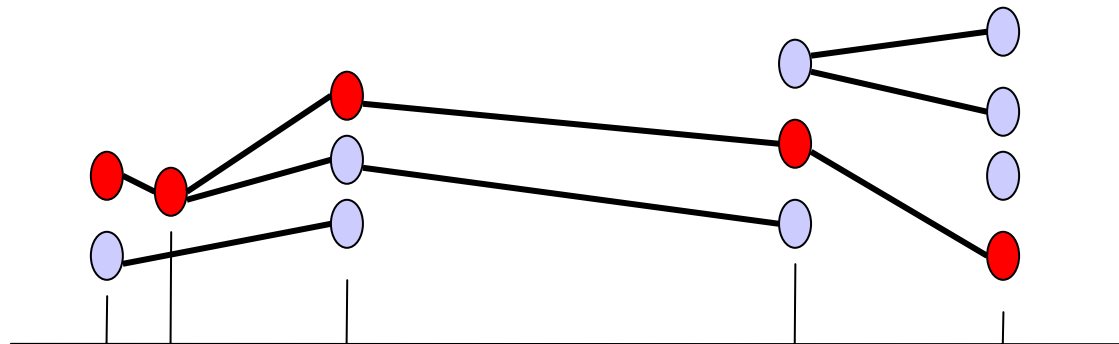
- Select cartoon that results in the maximum number of edges
 - Select one cartoon for each peak
 - Only keep edges linking these cartoons



Each oval  represents a cartoon

Max Subgraph

- Select cartoon that results in the maximum number of edges
 - Select one cartoon for each peak
 - Only keep edges linking these cartoons



Each oval  represents a cartoon

Random Walk

- Start a random walk on the Cartoon graph
- For each peak
 - select the cartoon that is most often visited
 - Or compute the fraction of time each cartoon is visited, use this as a 'probability'

Comparing the methods

- Random Walk ranks the cartoons
 - Other methods simply give a list
- Max subgraph performs best
 - Compared to the CFG annotations
- All the methods improve their performance when additional information is known
 - The cartoons at a few peaks are determined by further experiments
 - Extra knowledge, e.g. no NeuGc or Gal-Gal

Details

- Parsimony and Random Graph actually propose multiple cartoons
 - There may be more than one assignment with a parsimonious (minimal) set of antennae
 - Or multiple graphs with maximal number of edges
- Multiple possible definitions of 'edge' in the cartoon graph
 - Link if differ by one sugar or many?
 - Special case of lactosamine?

Bottom Line

Family analysis is a flexible way to provisionally assign cartoons to single MS

- If no biochemical information is known, does much better than random selection
- The assignments continually improve as more is known about the sample

Thanks to my Collaborators



- Anne Dell FRS
Imperial College,
London



- Simon North
 - Imperial
College, London



- Stuart Haslam
 - Imperial
College, London



- Marshall Bern
Palo Alto Research
Center



- Jim Paulson
PI of the CFG
Scripps Research
Institute