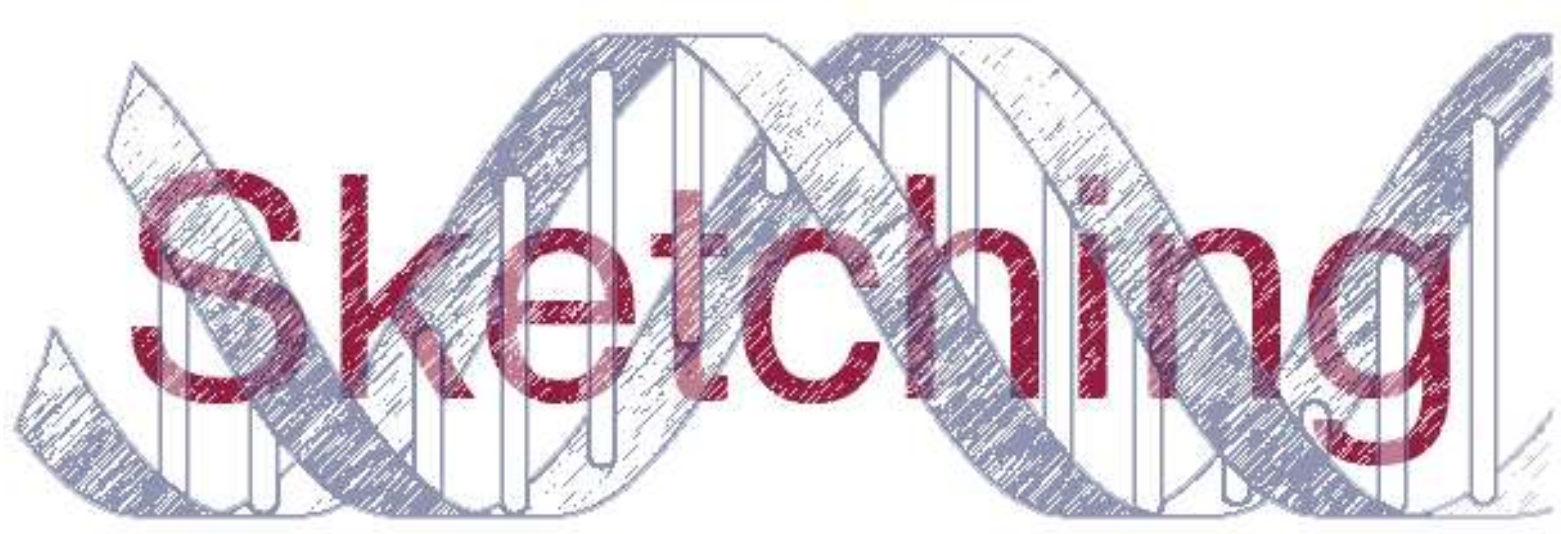
Sketching Genomic Data



Jordan Boyd-Graber, Adrian de Froment
Alex Golovinskiy, Jesse Levinson

Introduction

- Sketching: a more efficient way of storing genomic data for similarity comparisons
- Instead of storing entire data, store compact representations that preserve distances
- Useful for large data sets
  - Combining gene array experiments

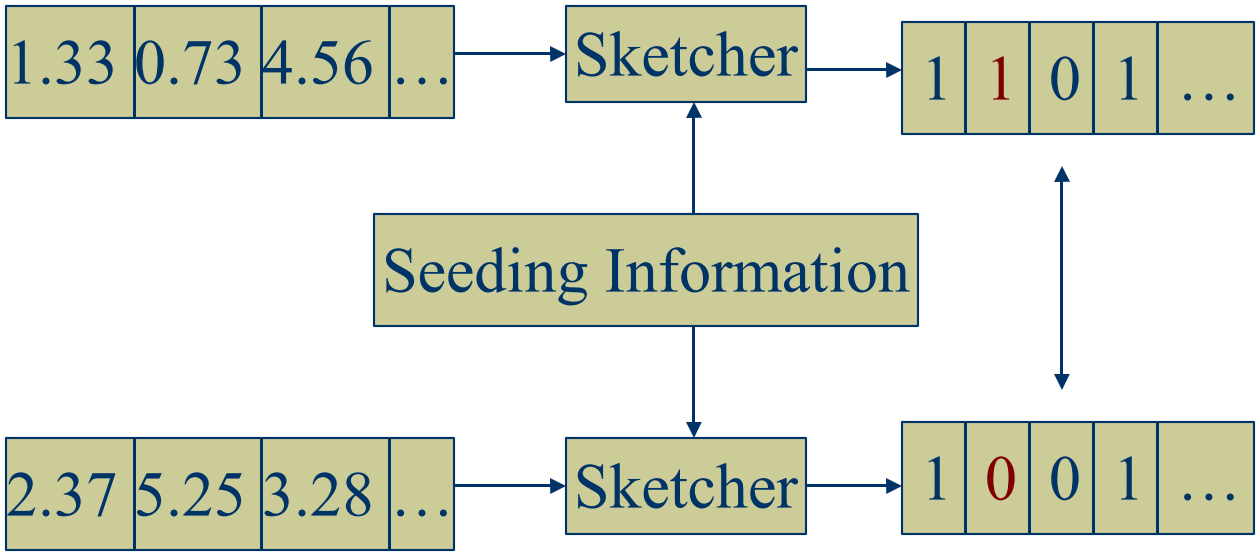
Outline

- Theory
- System/Implementation
- Accuracy test: correlation and neighbors
- Accuracy test: GO prediction
- Proposed applications and extensions

Theory

- Metric embedding: given data in a complex metric space, form a distance-preserving embedding into a simpler metric space
- Sketching: turn real number vector to vector of bits, where Hamming distance approximates some distance measure
- Used for image similarity search: (***Lv, Charikar, Li 2004)***

Theory


1.33
0.73
4.56
…
Sketcher
1
1
0
1
…
Seeding Information
2.37
5.25
3.28
…
Sketcher
1
0
0
1
…

Theory

- Work with the L1 distance:

$$d(v,w) = \sum |v_i - w_i|$$

Theory

- Work with the L1 distance:

$$d_{L_1}(v,w) = \sum |v_i - w_i|$$

- Need a function to convert to bits:

$$f : \Re^n, r \rightarrow \{0,1\}^m \text{ such that :}$$
$$d_{L_1}(v,w) \approx d_H(f(v,r), f(w,r))$$

Theory

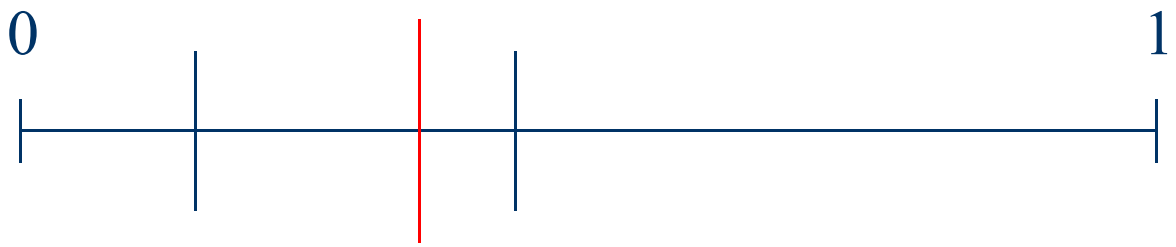
- Given two numbers in [0, 1], how can we estimate their distance?

Theory

- Given two numbers in [0, 1], how can we estimate their distance?

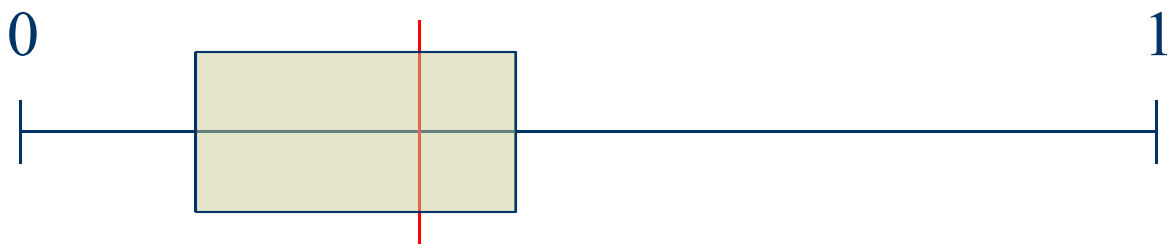
Theory

- Given two numbers in [0, 1], how can we estimate their distance?



Theory

- Given two numbers in [0, 1], how can we estimate their distance?



Theory

- Algorithm: if we want to go from number vector of dimension n to bit vector of dimension m, create seeding information:
  - Choose a random dimension d_i from n
  - Choose a random threshold t_i from [0, 1]
  - Seeding information is $\{d_i, t_i\}$ for i = 1 to m
- To create m-bit sketch b given seeding information $\{d_i, t_i\}$ and n-dimensional vector v:

$$b_i = \begin{cases} 1 \text{ if } v_{d_i} \geq t_{d_i} \\ 0 \text{ otherwise} \end{cases}$$

Theory

- So, if a and b are sketches of v and w of size m,

$$\Pr[a_i \neq b_i] \propto d_{L_1}(v, w)$$

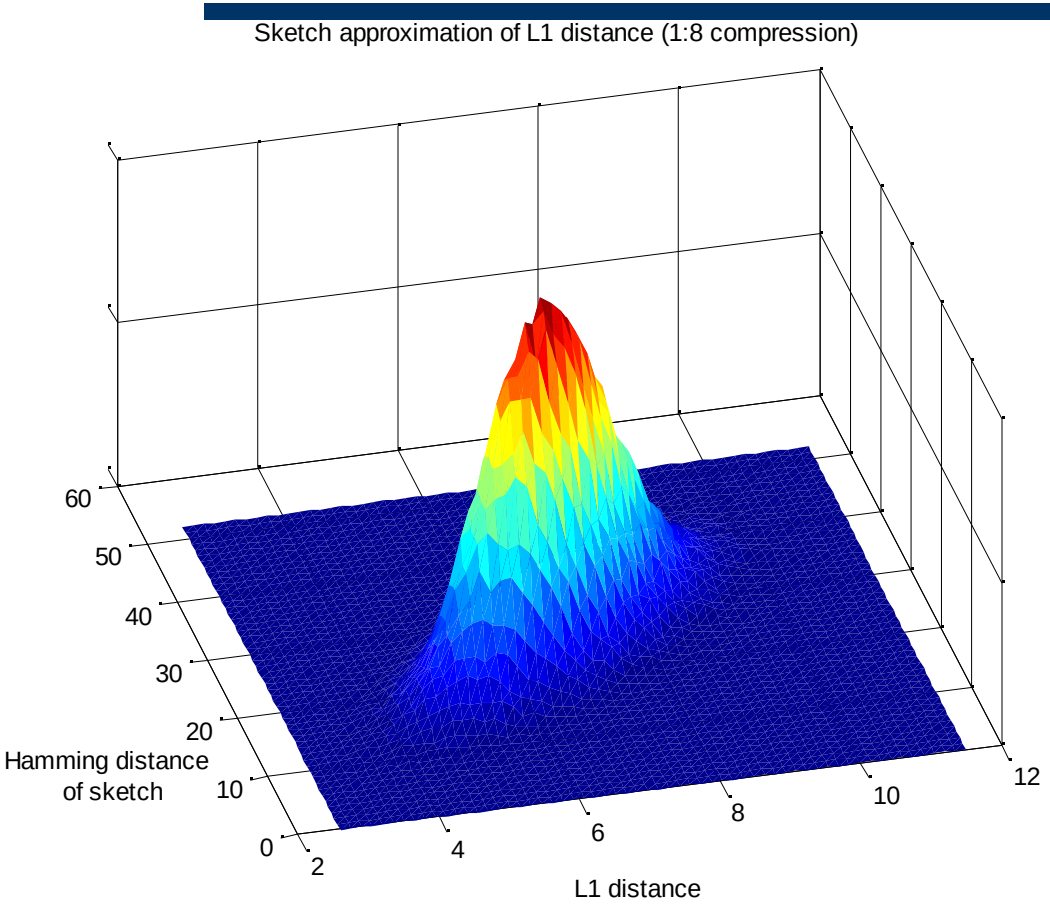
- Distribution of Hamming Distance is binomial:

$$p(d_h) = \binom{m}{d_h} \left(\frac{d_{L_1}}{T}\right)^{d_h} \left(1 - \frac{d_{L_1}}{T}\right)^{m - d_h}$$

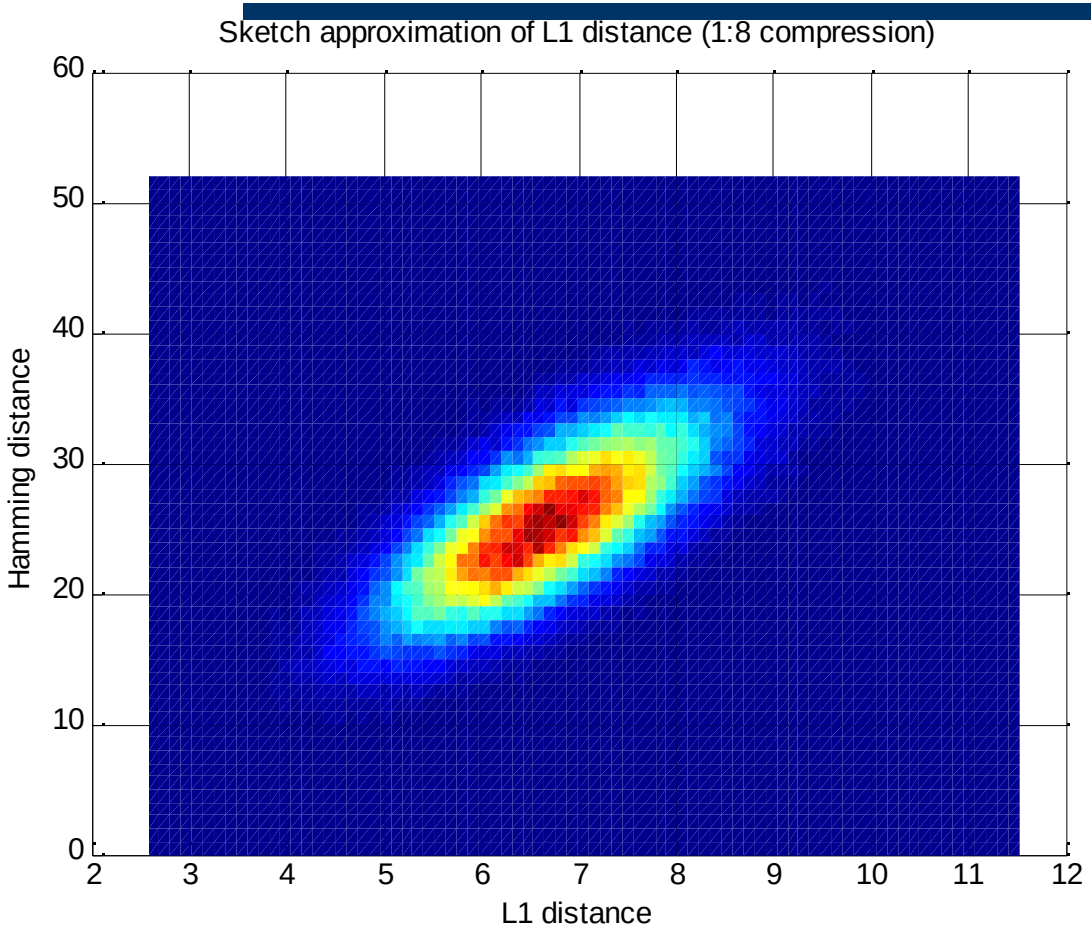
$$\mathrm{E}[d_h] \propto m d_{L_1}$$

$$\sigma_{d_h} \propto \sqrt{m}$$

Theory


Sketch approximation of L1 distance (1:8 compression)
Hamming distance of sketch
60
50
40
30
20
10
0
2
4
6
8
10
12
L1 distance

Theory

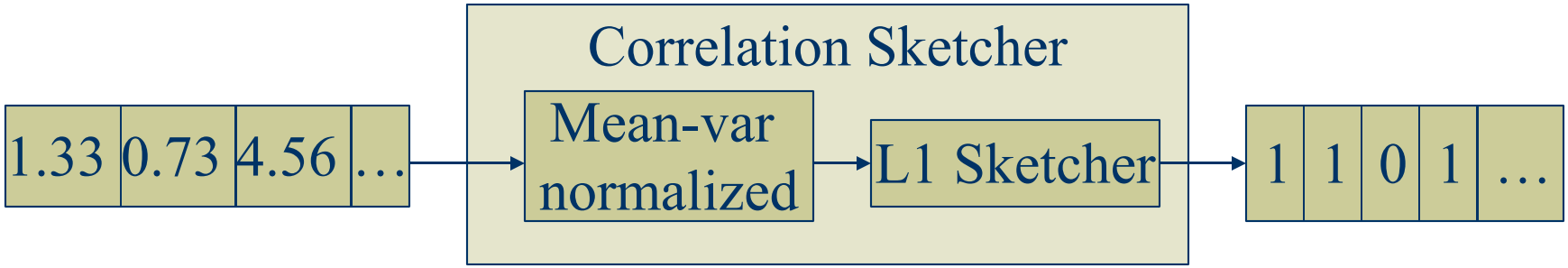

Sketch approximation of L1 distance (1:8 compression)
Hamming distance
L1 distance
0
10
20
30
40
50
60
2
3
4
5
6
7
8
9
10
11
12

Theory

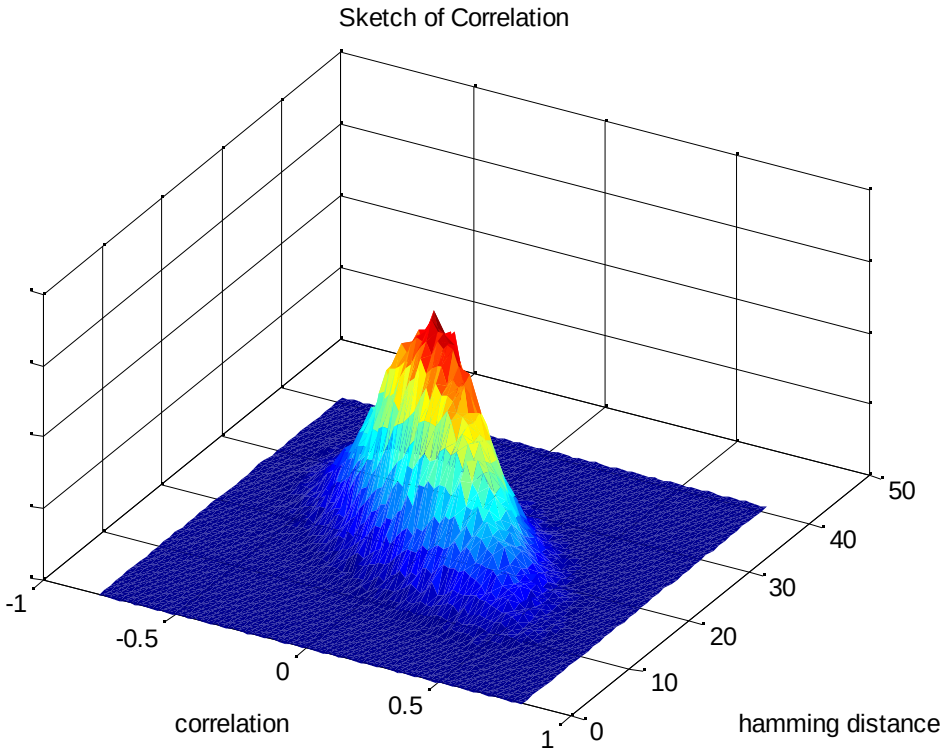
- What if we want another distance measure, ie correlation?
  - Come up with a new sketcher
  - Use old sketcher with modified input

Theory

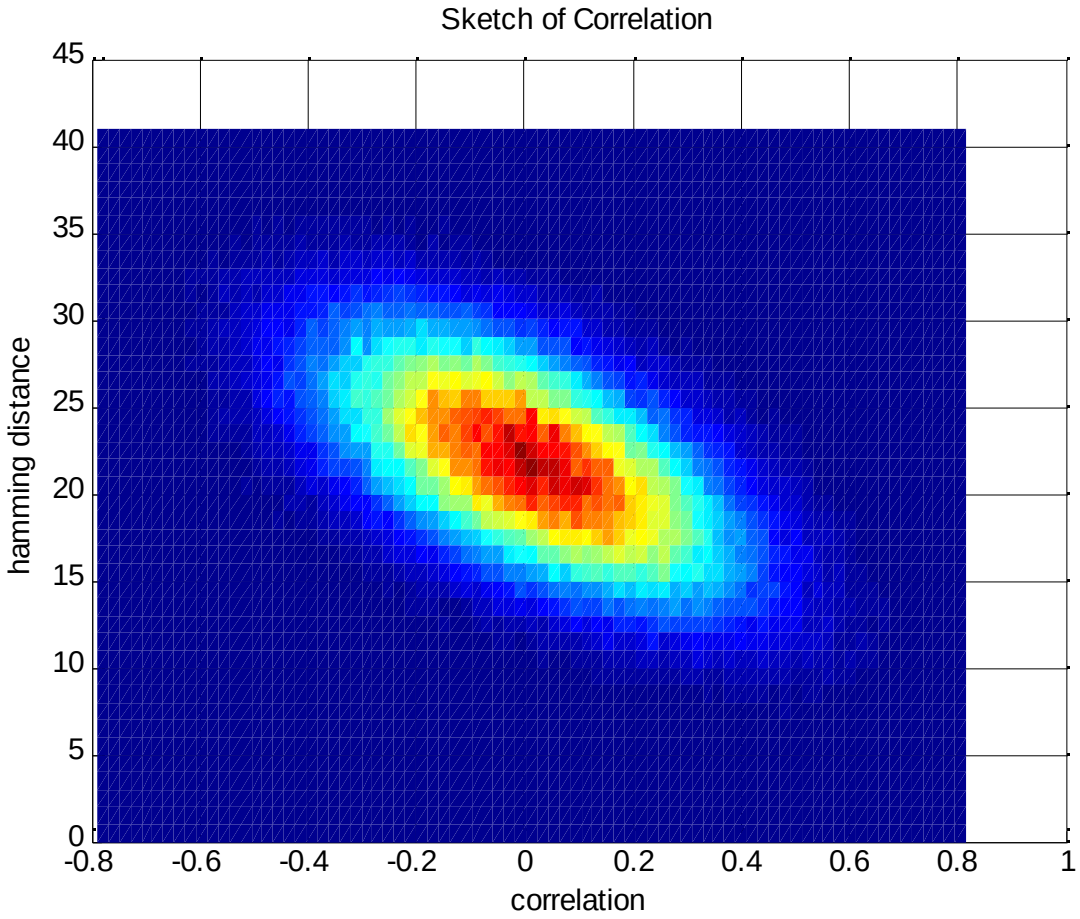

1.33
0.73
4.56
…
L1 Sketcher
1
1
0
1
…


Correlation Sketcher
1.33
0.73
4.56
…
Mean-var normalized
L1 Sketcher
1
1
0
1
…

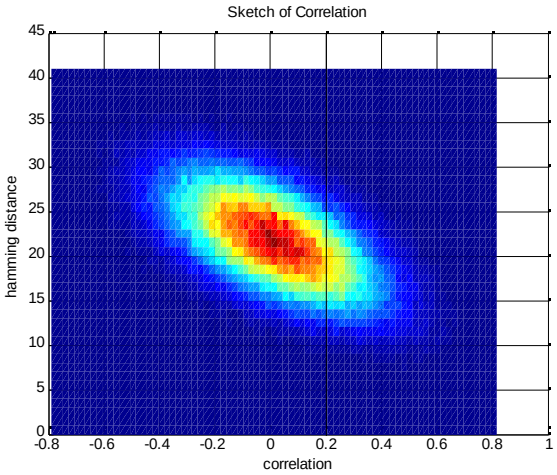
Theory

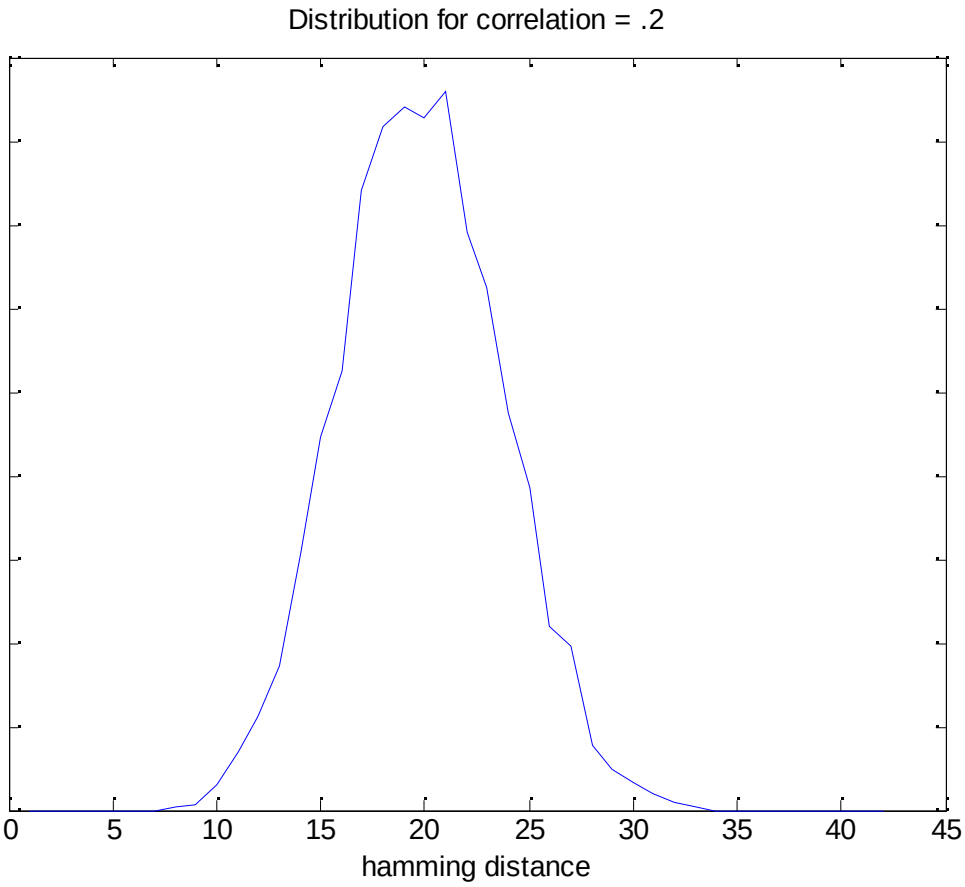

Sketch of Correlation
-1
-0.5
0
0.5
1
correlation
0
10
20
30
40
50
hamming distance

Theory


Sketch of Correlation
hamming distance
correlation
45
40
35
30
25
20
15
10
5
0
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

Theory


Sketch of Correlation
hamming distance
correlation
45
40
35
30
25
20
15
10
5
0
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1


Distribution for correlation = .2
0
5
10
15
20
25
30
35
40
45
hamming distance

Theory

- Claim: useful for genomic data
  - Lots of data
  - Data has noise
  - Distance measures are uncertain
- Implementing a prototype system…

Infrastructure

- MySQL Database
- Java representations of original real-valued vectors, sketch profiles, sketches
- Code to implement sketching algorithm
- Database accessors to take these representations and store/retrieve them from the database

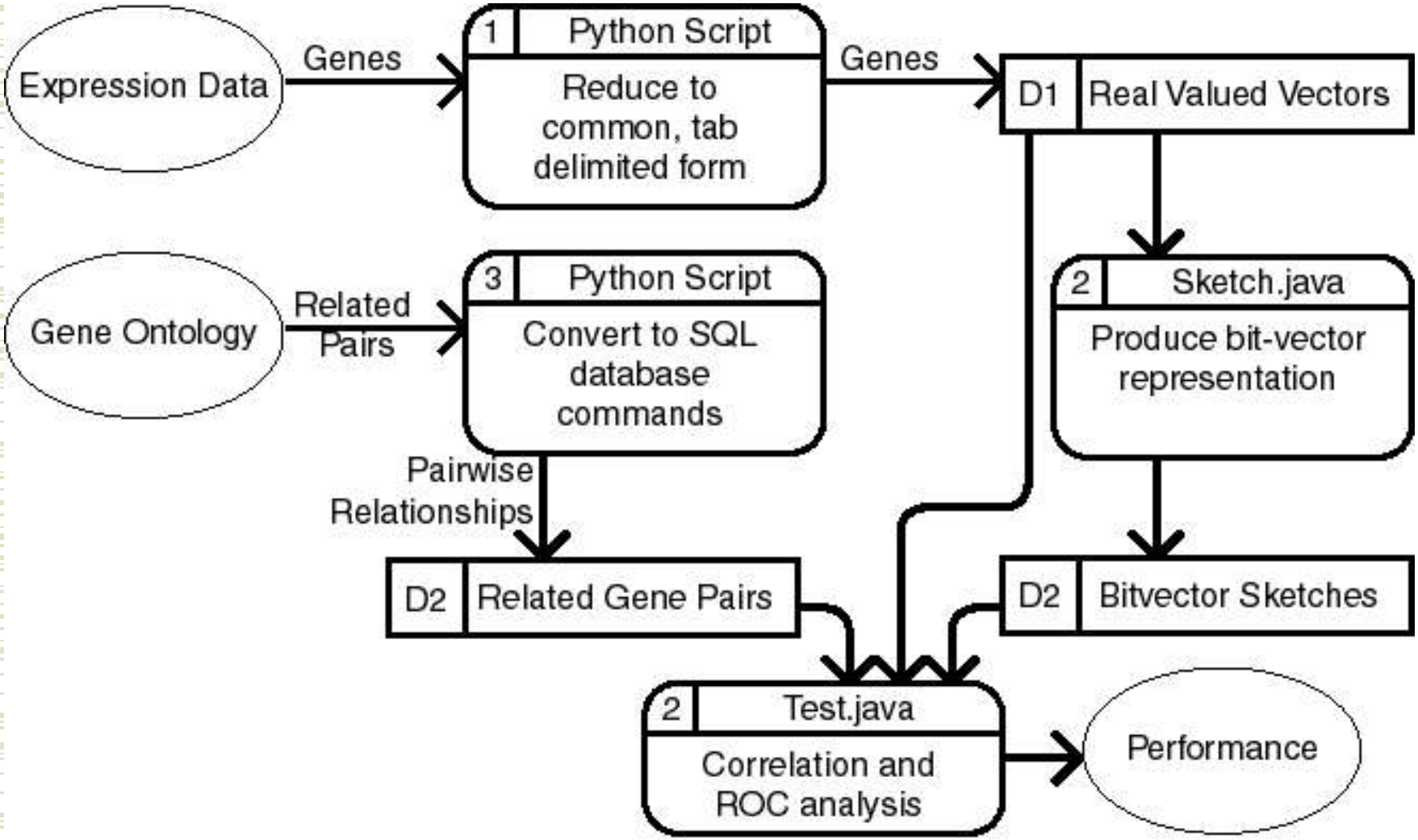
Database Design

- Original Data
- Sketch Profile
- Sketches
- GO

EXP	EXPERIMENT NAME	#COLS	BITS
12	[FULL]DED:ReporterDimens	20	500
11	DED:ReporterDimension:ME	0	20
13	[DUPE]DED:ReporterDimens	20	40
17	Hoheisel-Hauser_Heatshoc	11	80

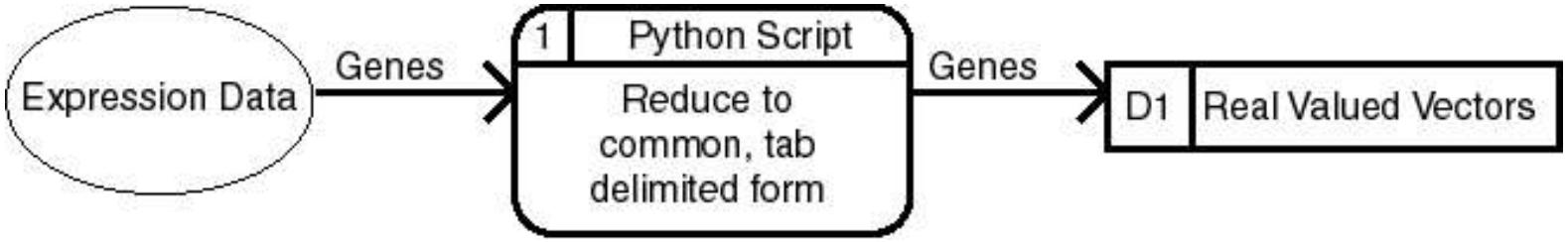
ID	COL	VALUE
0	3	0.1111111
0	1	0.1314234
0	2	0.423232
0	7	0.80125093

EXP	DATA SOURCE	GENE	ID
11	test.txt:6:R:A-SNGR-11:5504	SPBC1105.04C	8
11	test.txt:7:R:A-SNGR-11:5224	SPBC1861.02	9
11	test.txt:10:R:A-SNGR-11:4000	SPBC106.04	10
11	test.txt:13:R:A-SNGR-11:5391	SPAC144.03	11


Expression Data
Genes
1
Python Script
Reduce to common, tab delimited form
Genes
D1
Real Valued Vectors
Gene Ontology
Related Pairs
3
Python Script
Convert to SQL database commands
2
Sketch.java
Produce bit-vector representation
Pairwise Relationships
D2
Related Gene Pairs
D2
Bitvector Sketches
2
Test.java
Correlation and ROC analysis
Performance

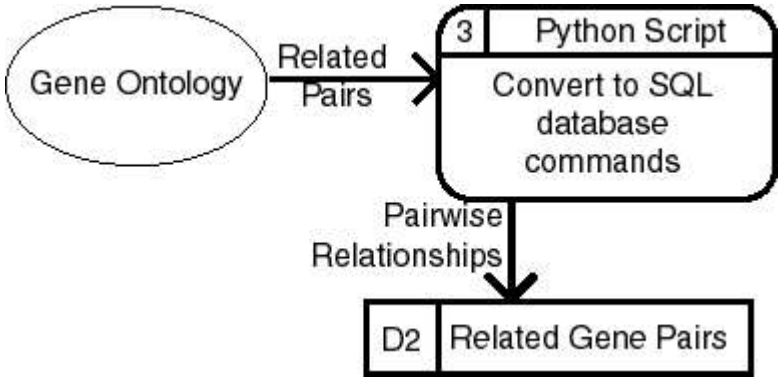
Data Sources

- S. cerevisiae timeseries data (Hauser & Hoheisel) from heatshock and saltshock response
- S. pombe cell cycle (Rustici)



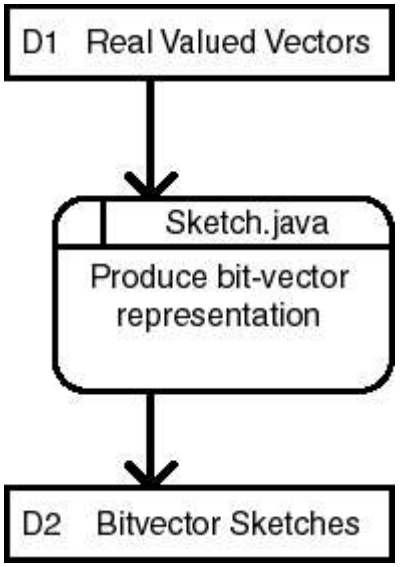
Gene Ontology Information

- Yeast Gene Ontology: +1 for matching depth-7 process ontology, -1 for non-matching, 0 for ambiguous annotation
- Each non-ambiguous pair is stored in the database



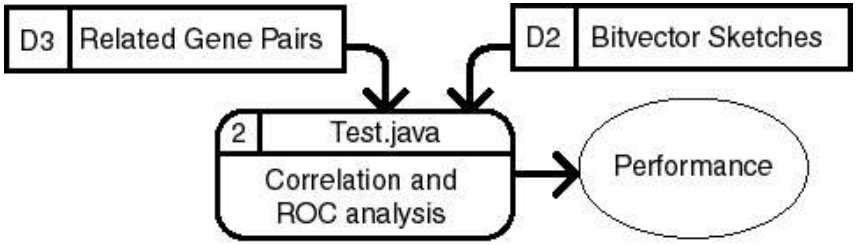
Creating Sketches

- Code to implement L1 sketch
- Stores information used to create sketches so new query genes can be analyzed



Methods for Analysis

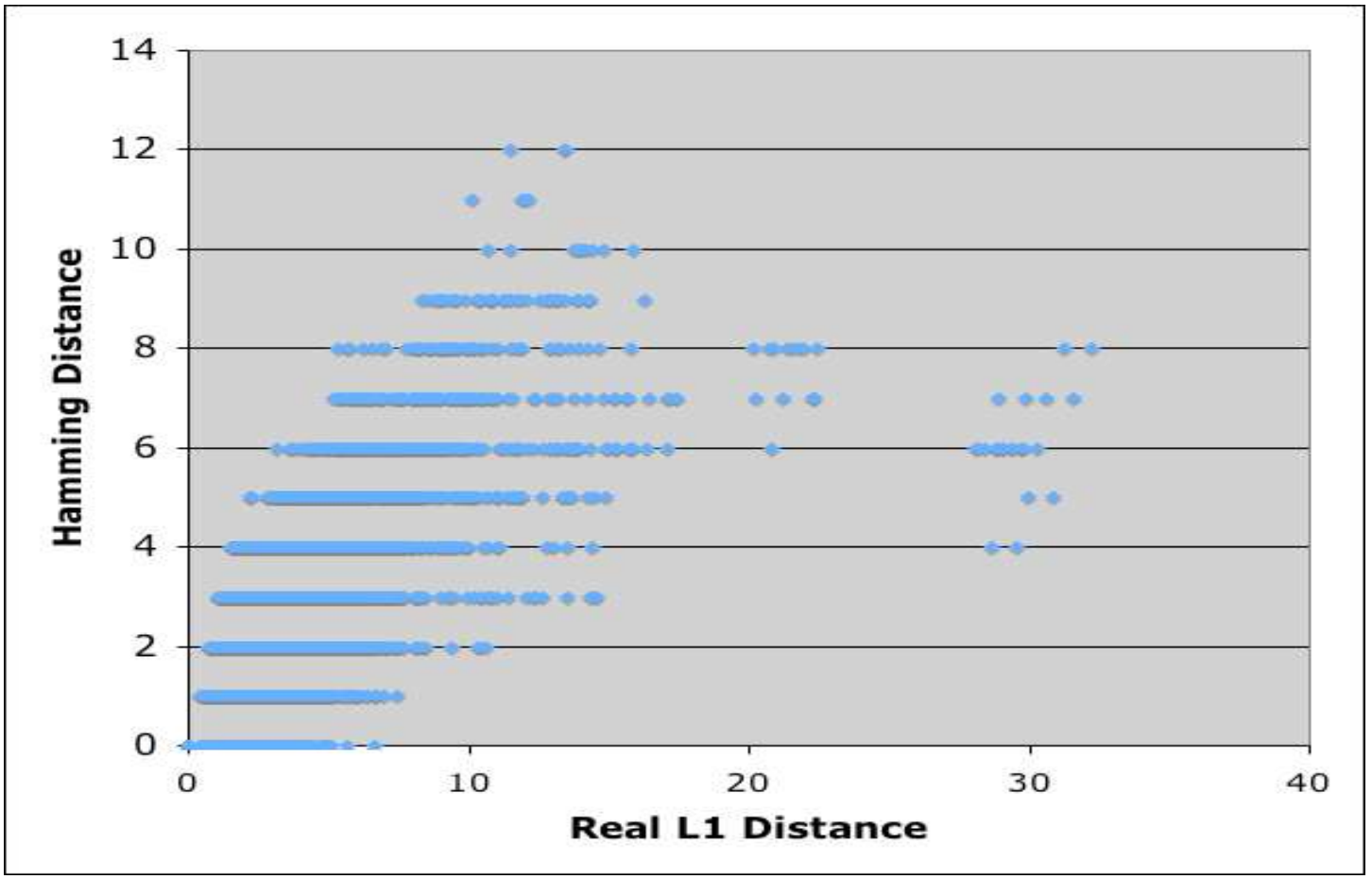
- Once we've stored the sketches, we can compare how well the sketches reflect standard distance measures
- In a real implementation, we would not have this redundancy



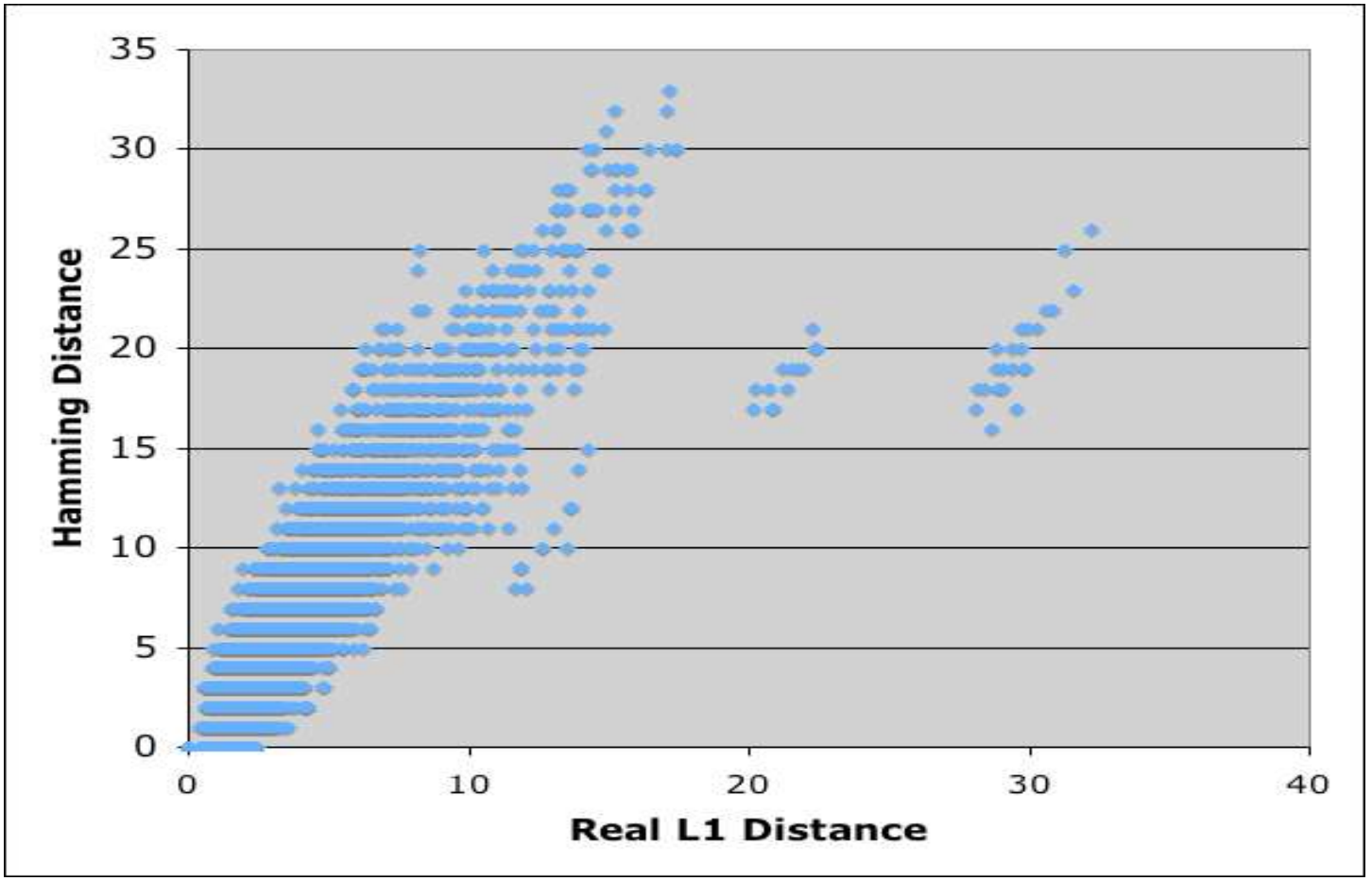
Assessing Accuracy

- Sketch is supposed to be a reasonable estimate of original function
- Similarity to original function should increase as size of bit vector increases
- Tradeoff between size/speed and accuracy

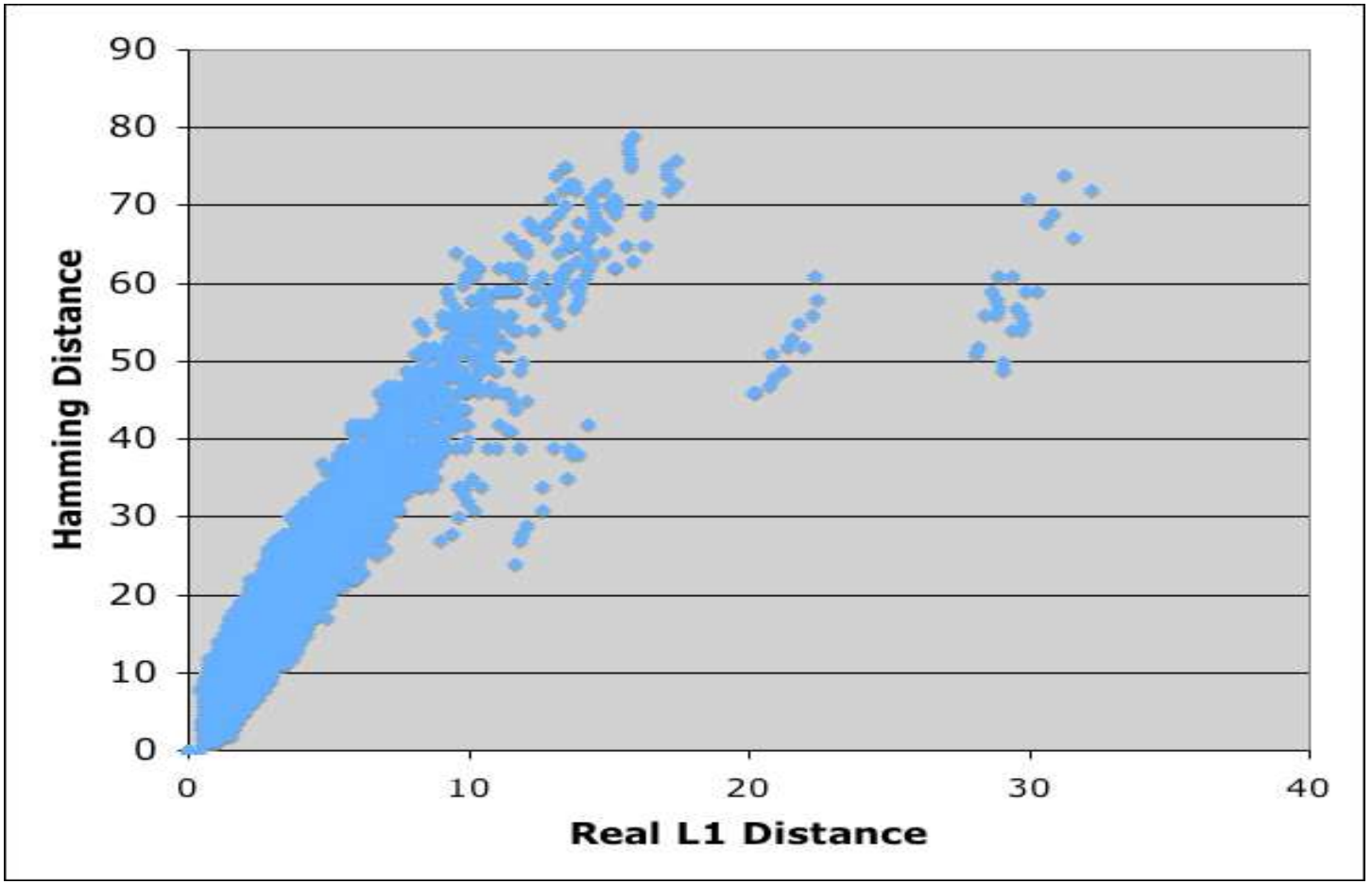
Real vs. Hamming Distance
20 bit vector --> r=.70



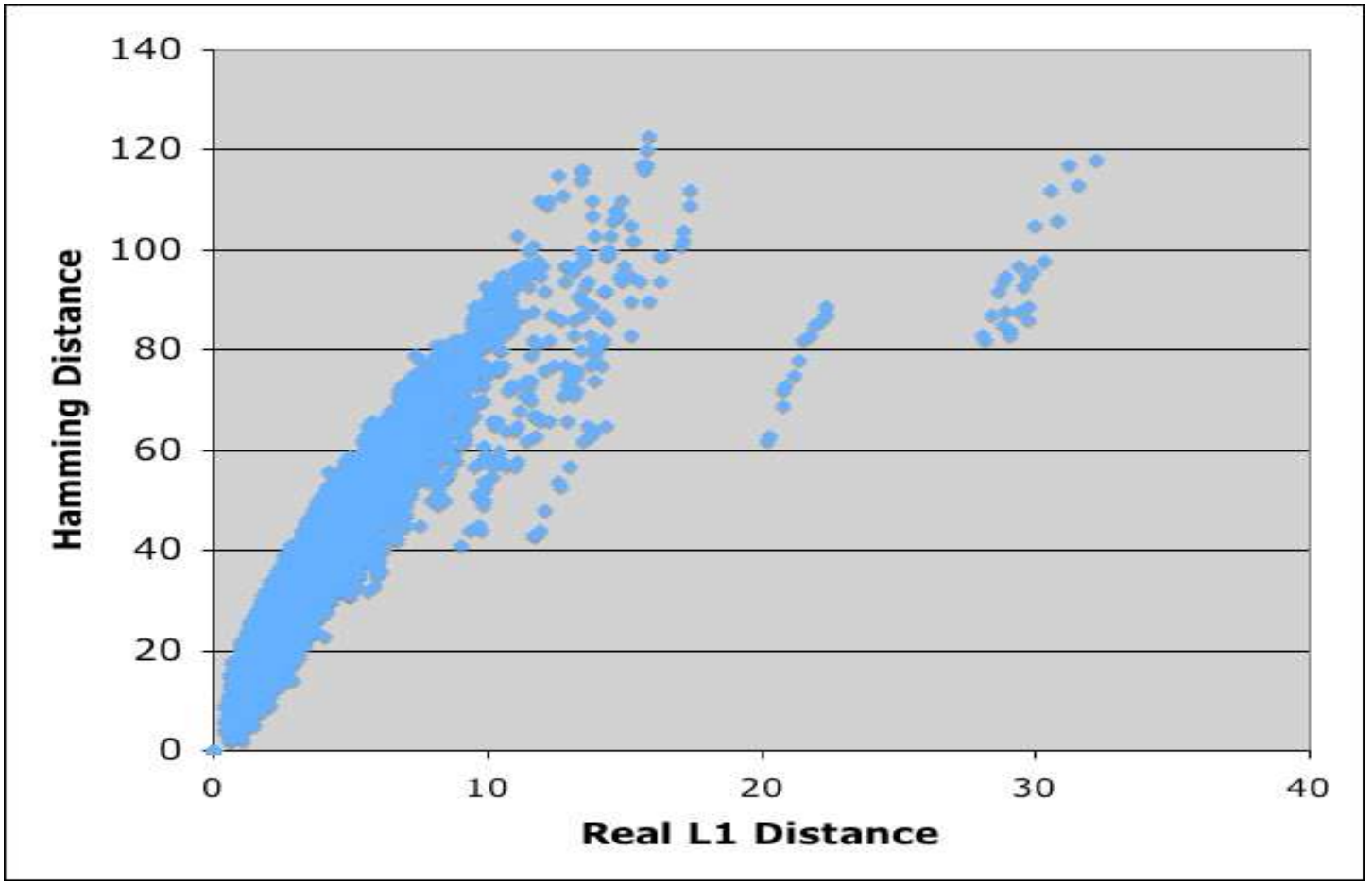
Real vs. Hamming Distance
80 bit vector --> r=.85



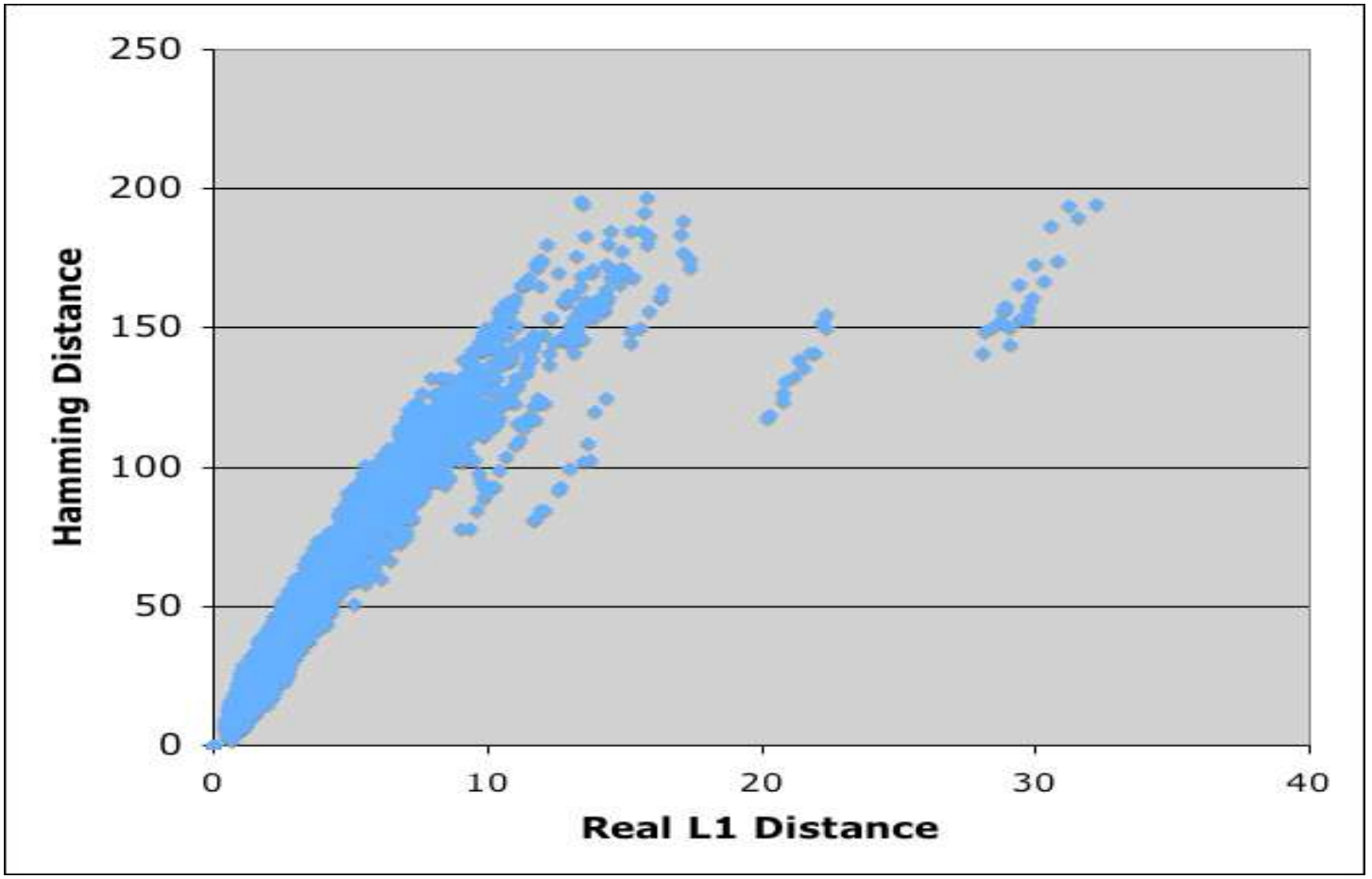
Real vs. Hamming Distance
180 bit vector --> r=.85



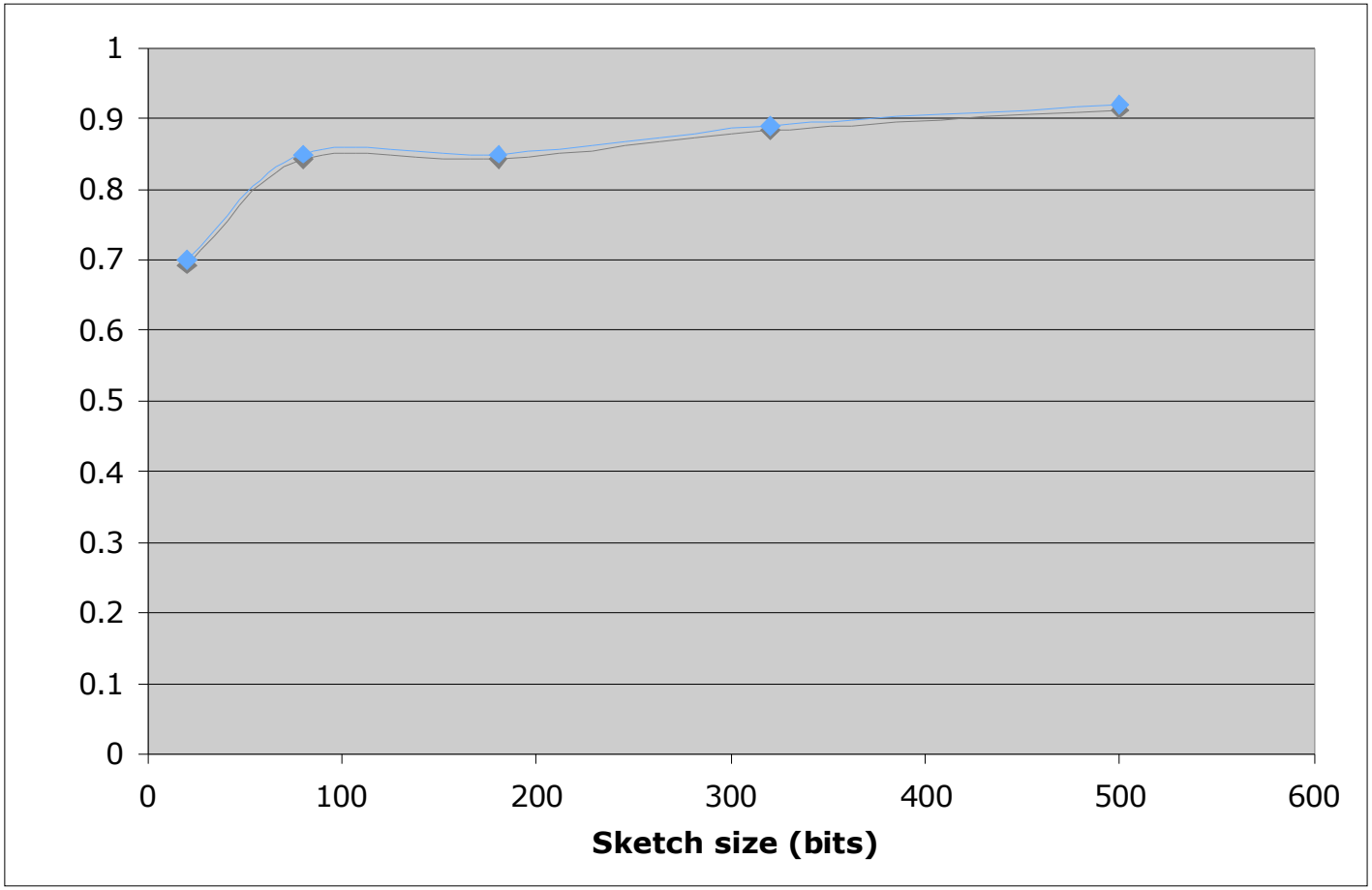
Real vs. Hamming Distance
320 bit vector --> r=.89



Real vs. Hamming Distance
500 bit vector --> r=.92



Correlation of Hamming and Real Distance



Results are reasonable

- Correlation between hamming distance and true L1 distance increases as sketch size increases
- Sketch size of 80 represents 640:80 = 8:1 compression ratio and achieves correlation of .85 with true L1 distance

Can we use sketches to find similar genes?

- We have seen that sketches afford reasonably good correlation with original distance function
- Correlation is nice, but it's not actually useful. What if our task is to find the most similar genes in the database - how effective are sketches?

Task: find k% most similar genes

- Given a query gene, find the k% most similar genes based on L1 distance
- Can we save time and space by using sketch to find the k% most similar genes?
- We wrote a program to sort and assess overlap between two sets of results

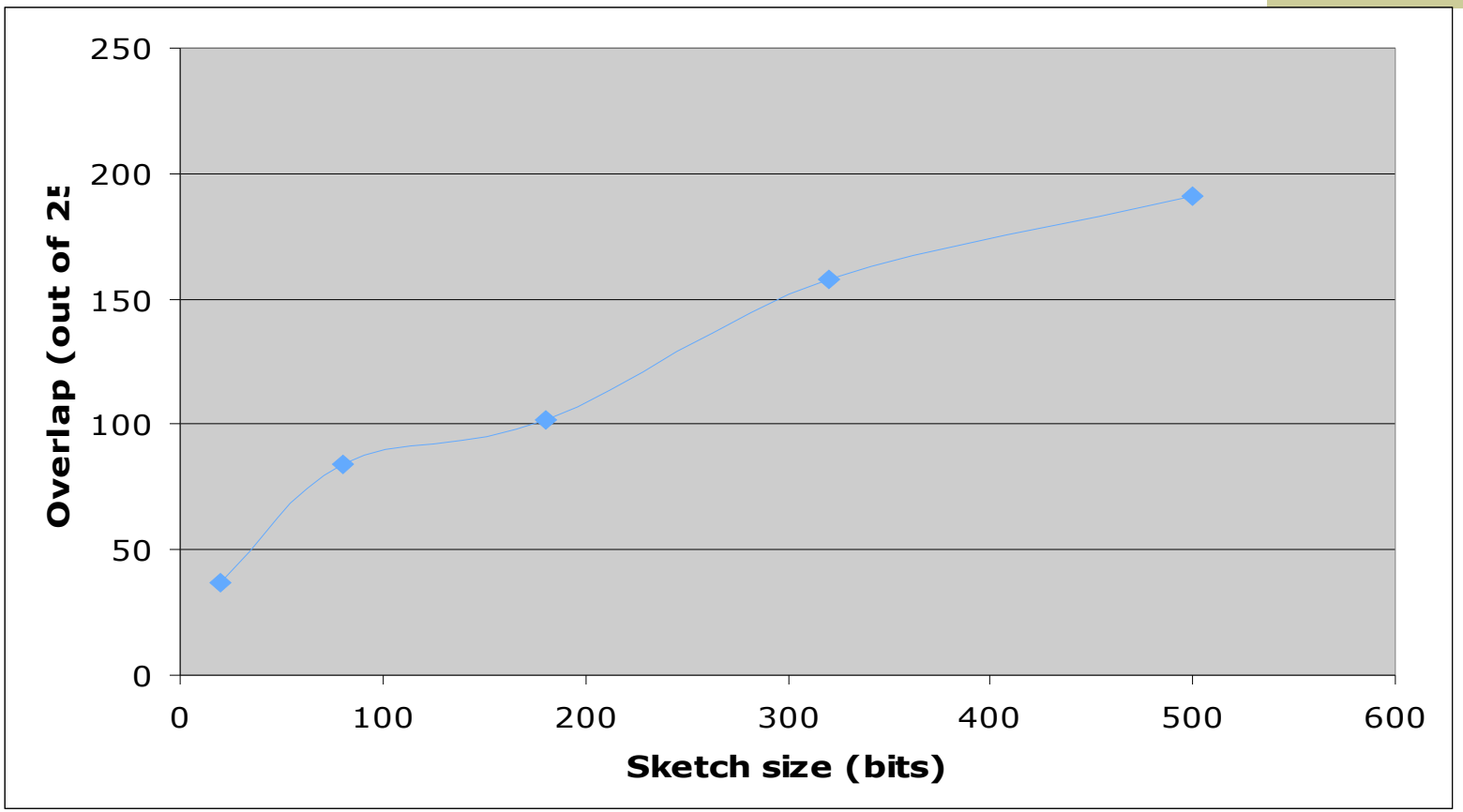
Task: find k% most similar genes

- Compute distance between query gene and every other gene
- Sort by distance
- Choose the k% with the least distances
- Perform above for true L1 distance and sketch approximation
- Compare results (want high overlap)

Experiment: Find 10% closest genes

- Out of 2650 genes, find the nearest 265 to a query gene
- See how many overlap between two distance methods.
- Only expect about 26 overlapping genes by chance… 265 would be perfect

Overlap between two methods



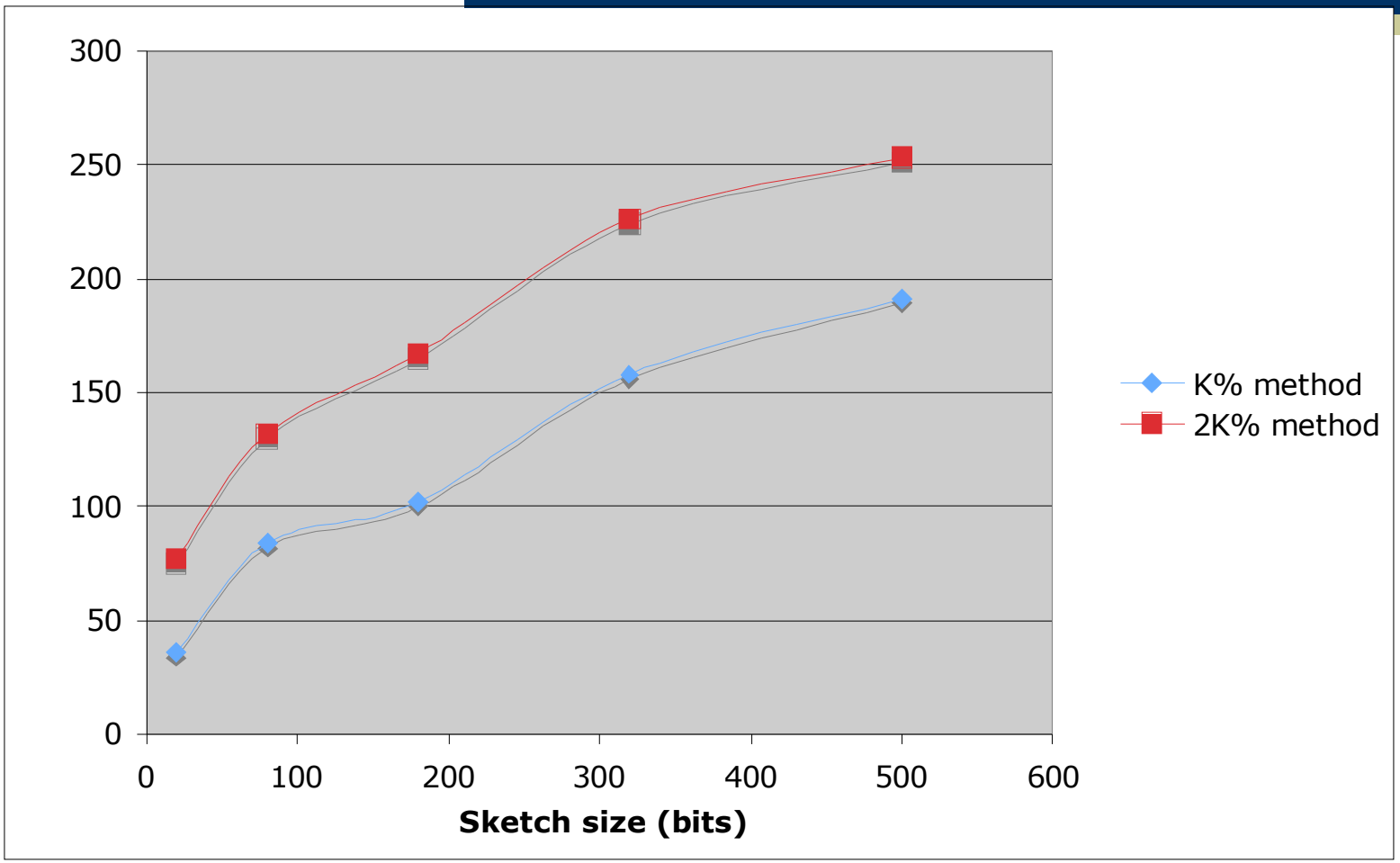
Encouraging results

- Sketch size of 80 bits (8:1 compression) generates 84/265 matches
- Sketch size of 320 bits (2:1 compression) generates 158/265 matches
- Much, much better than random

Can we do even better?

- Yes - we can find the nearest 2k% genes using sketch distance and then use true L1 distance to narrow down to nearest k%
- Assuming k << 50% this method still saves significant time compared to traditional approach. Does it work?

Overlap between two methods using 2k% sketch trick



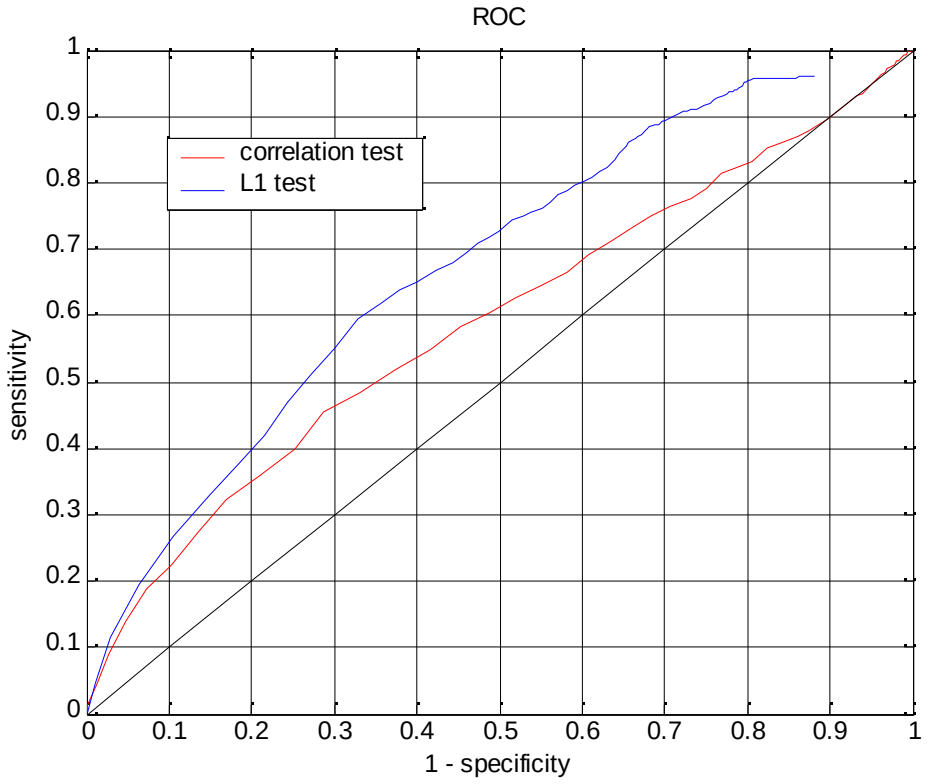
Much better!

- Sketch size of 80 bits (8:1 compression) generates 132/265 matches instead of 84/265
- Sketch size of 320 bits (2:1 compression) generates 226/265 matches instead of 158/265
- Significant improvement in accuracy with minimal performance loss for small k

Prediction of GO

- Gene array experiments often used to predict biological function
- How well do L1 distance and correlation predict GO in our case?
- (We used heat shock experiment, sampled 600 genes, which produced 167038 unknowns, 2582 matches, 10080 non-matches in the GO)

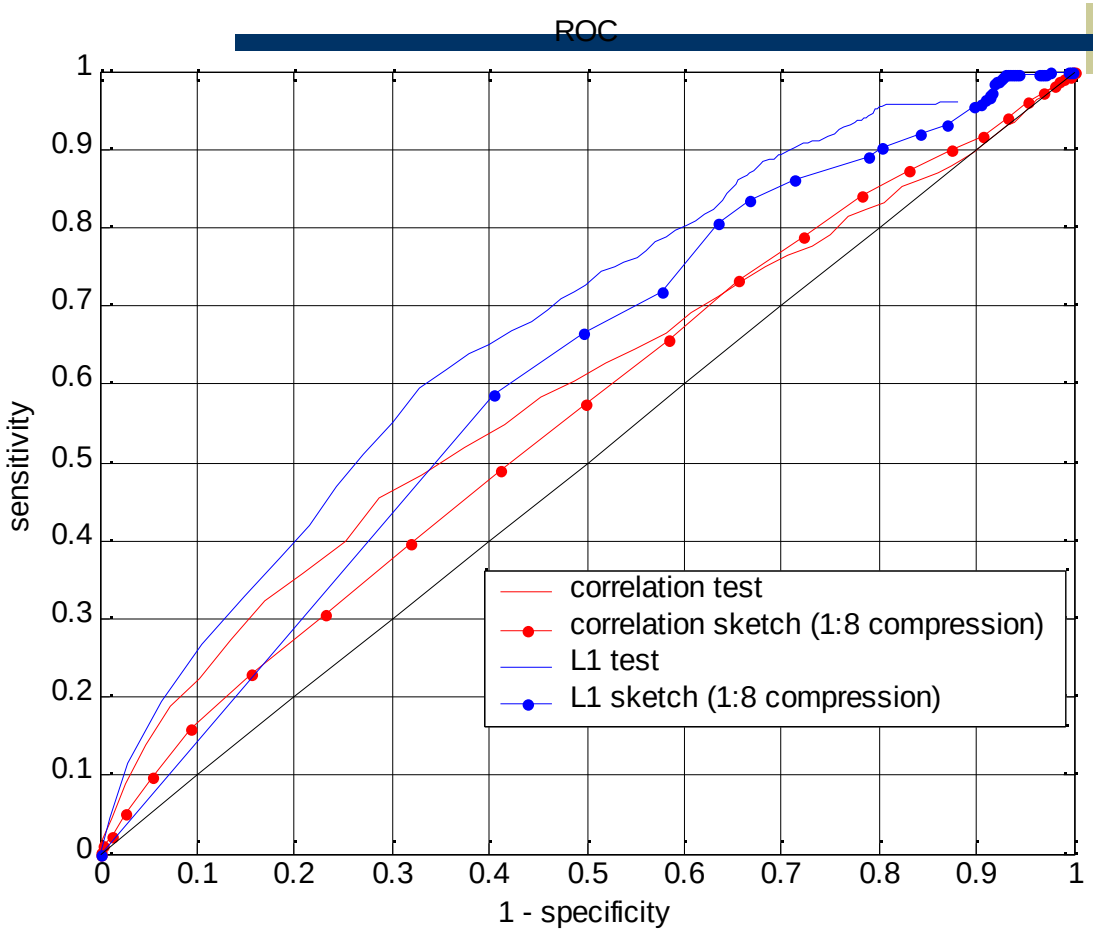
Prediction of GO



Prediction of GO

- Make sketches, and predictors of GO based on sketches

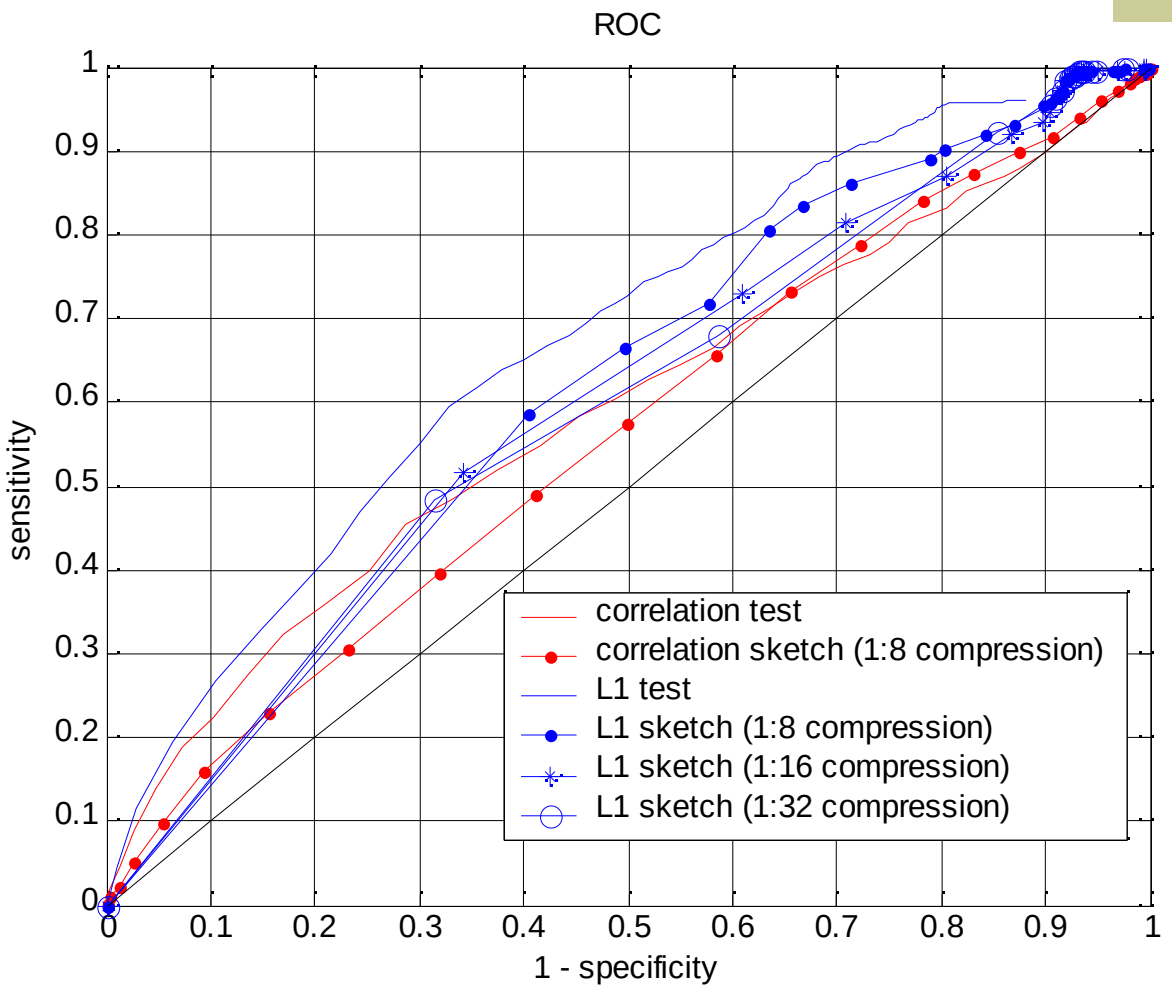
Prediction of GO



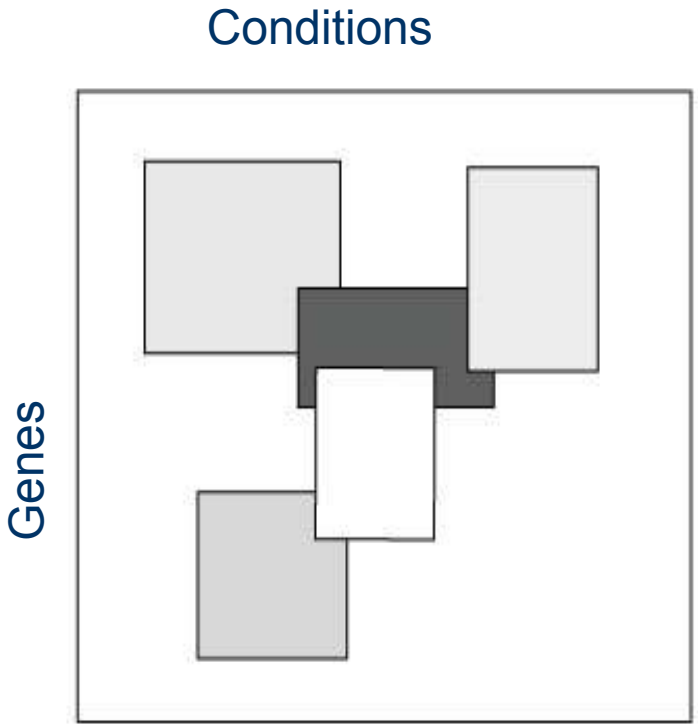
Prediction of GO

- How do the predictors deteriorate if we use fewer bits for sketches?

Prediction of GO

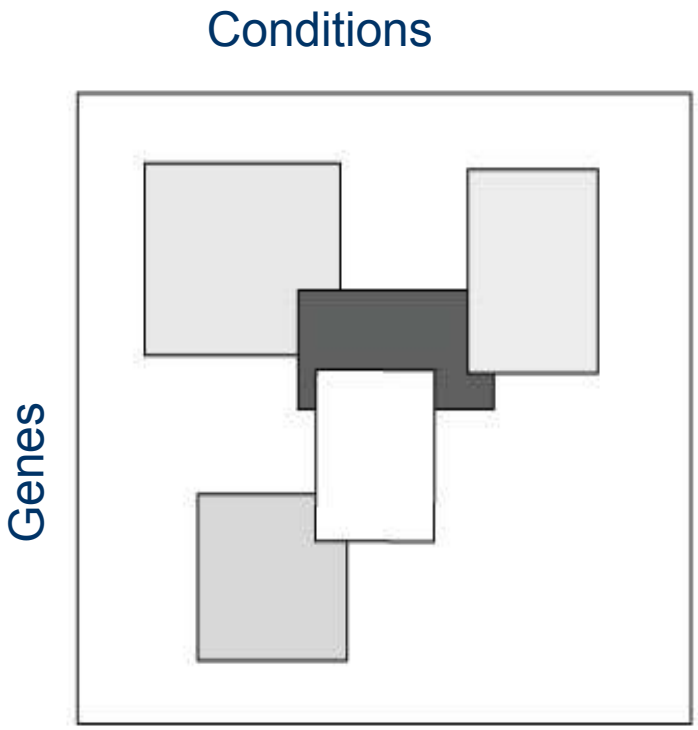


Biclustering



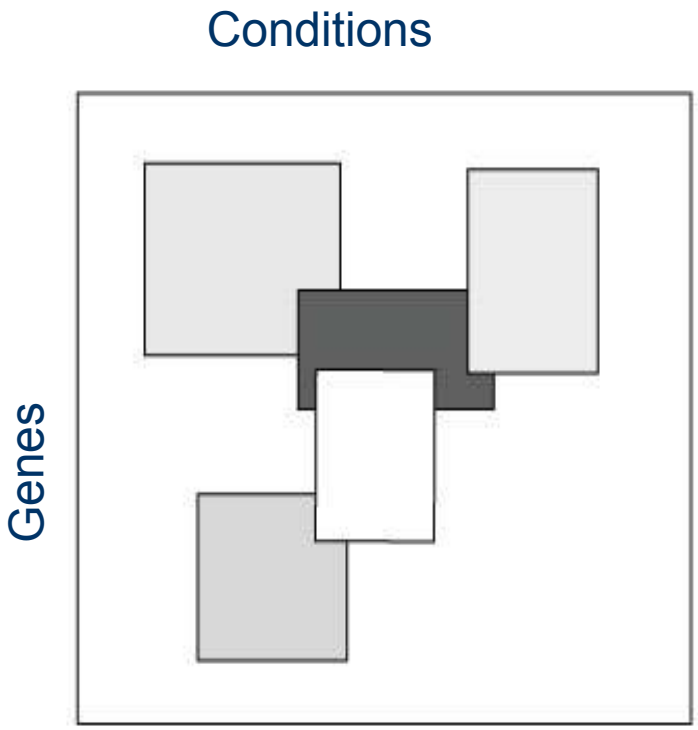
- Clustering:
  - Correlation between gene expression levels across all conditions in an experiment.
- Biclustering:
  - Clustering by row and column simultaneously.
- Output:
  - List of biclusters for each gene.

Advantages of Biclustering



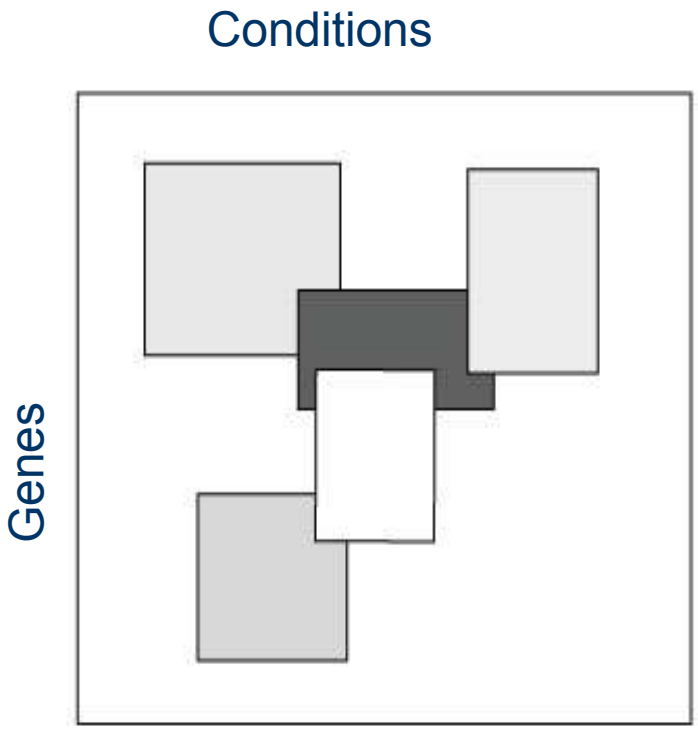
- Integrate data from many sources.
  - Individual biclusters can include conditions from different experiments.
- Grouping of conditions.
  - By algorithm, rather than experimenter.
- Noise reduction.
  - Microarray data inherently noisy.
  - Uninformative conditions are dropped out of cluster.

The Sketch Advantage



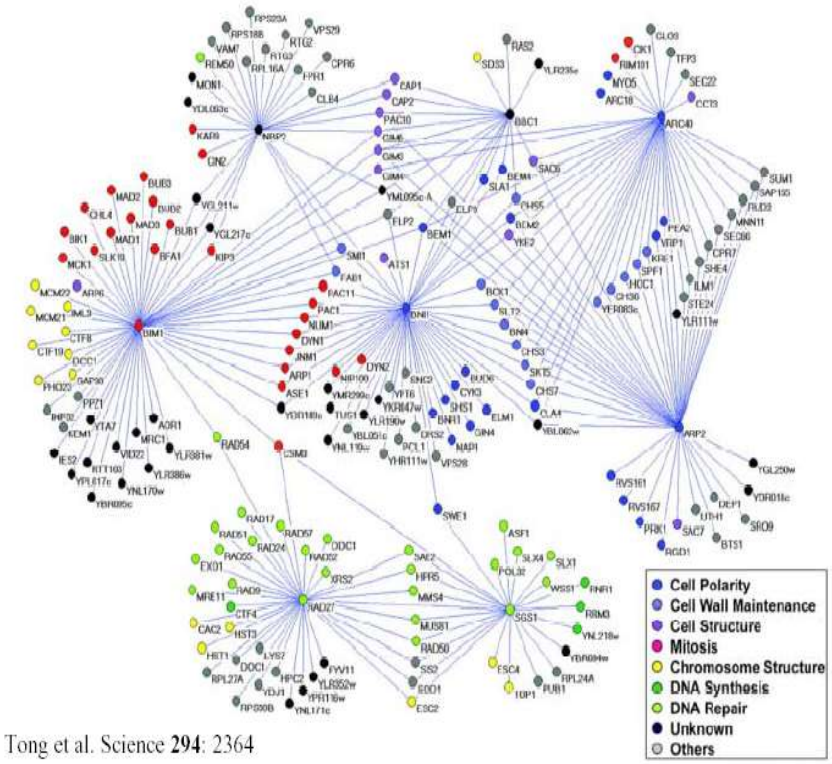
- Biclustering is an NP complete problem.
- Approximation algorithms are used
- Sketches will allow:
  1. Faster searching
  2. Increased accuracy for the time investment
  3. **Working with larger datasets**

Biological Realism



- Genes can be in multiple clusters.
  - Real genes often participate in different processes (pleiotropy).
- Clusters may exist only for subsets of conditions
  - Biological processes often operate under defined conditions.
  - Under other conditions genes may be uncorrelated/involved in other processes.

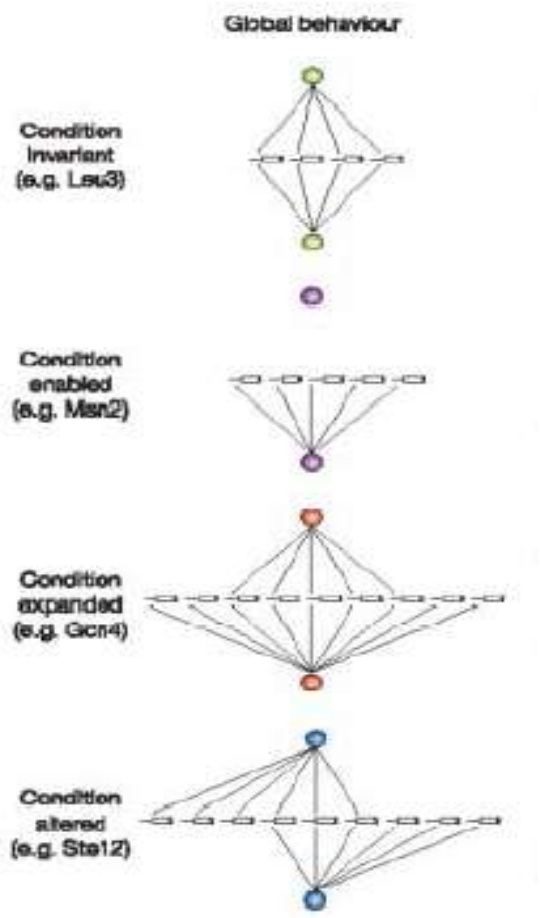
Biological Realism



Tong et al. Science **294**: 2364

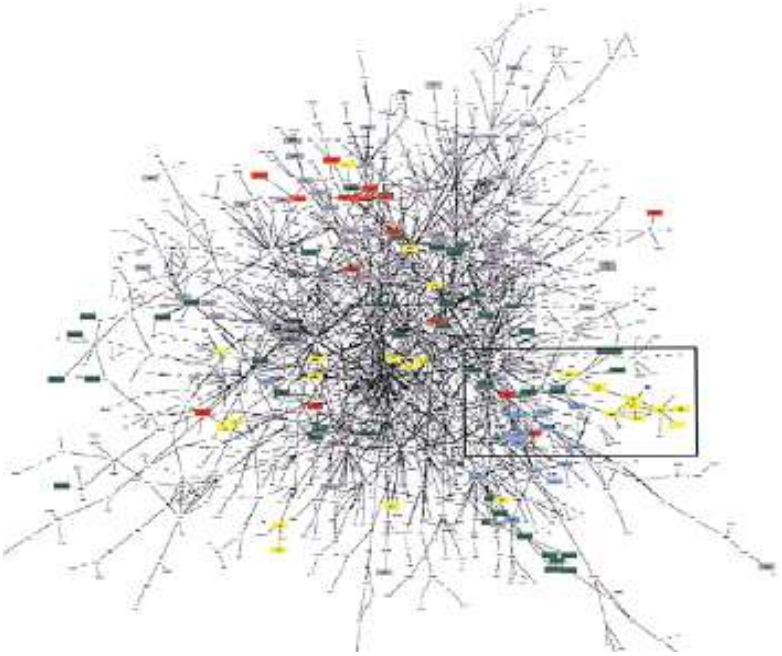
- Gene regulation is dynamic
  - Toolkits deployed in response to changes.
    - Internal.
      - *segmentation, DNA repair*
    - External environment.
      - *heatshock*
- Transcription Factor binding often depends on physiological conditions.
  - Differential gene regulation in response to environmental change is crucial for life.

Application: Coregulation



- Finding coregulated genes.
  - Correlation of expression in defined context points to coregulation in that context.
  - Genes' expression may be under the control of the same condition-sensitive TFs.
  - Transcription Factor binding depends on physiological conditions.
    - Differential gene regulation in response to environmental change is crucial for life.

Annotation



- Correlation points to similar function under conditions included in bicluster.
- Annotate genes of unknown function.
  - Significantly enriched with genes of particular function?
  - Assign same function to unknowns.
- Assign context dependent function.
  - Parallels our understanding of the dynamic nature of regulatory networks.

Extensions

- Optimized low level comparison functions
- Optimized database (structure, indexing)
- Detailed timing and memory analysis
- Heterogeneous data (e.g. combining expression datasets)
- More varied and thorough tests against GO

Conclusion

- Sketching is a promising method for applications that require similarity search on large genomic datasets
- We created a usable implementation that highlights advantages and possibilities afforded by sketching
- The use of sketching for genomic data merits further development

Acknowledgements

- Olga Troyanskaya
- Kai Li
- Qin Lv
- Chad Myers