Combinatorial group testing and signal recovery

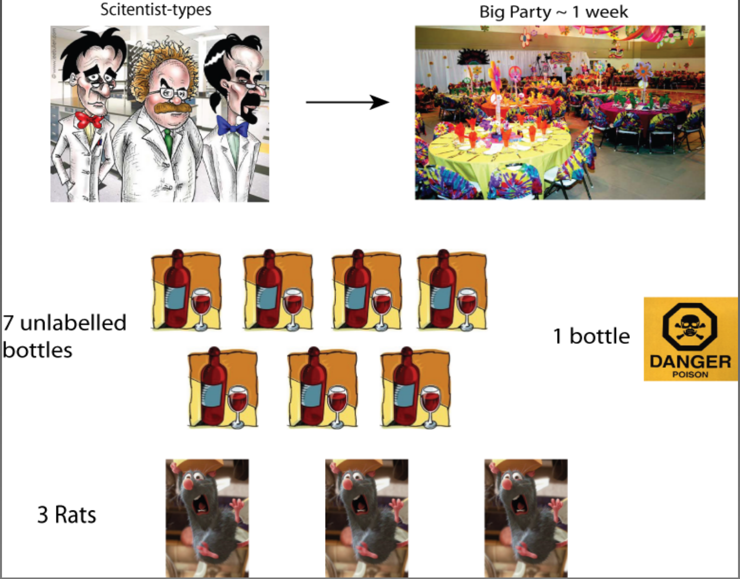
Anna C. Gilbert

University of Michigan

Computer Science: R. Berinde (MIT), P. Indyk (MIT),
H. Karloff (AT&T), M. Strauss (Univ. of Michigan),

Chemical Eng: R. Kainkaryam (Univ. of Michigan),
P. Woolf (Univ. of Michigan)

Combinatorial group testing



Scitentist-types

Big Party ~ 1 week

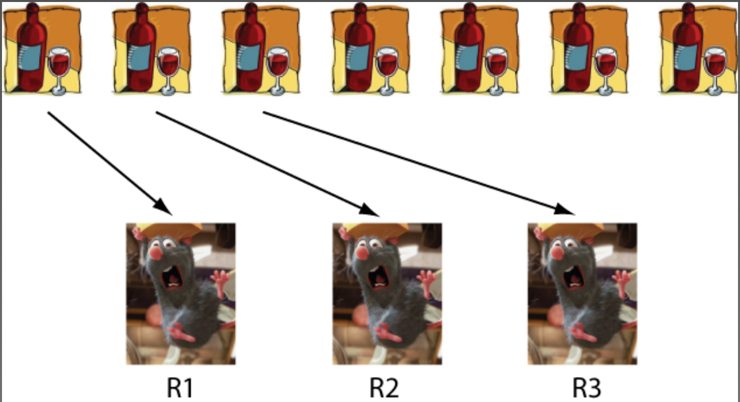
7 unlabelled bottles

1 bottle

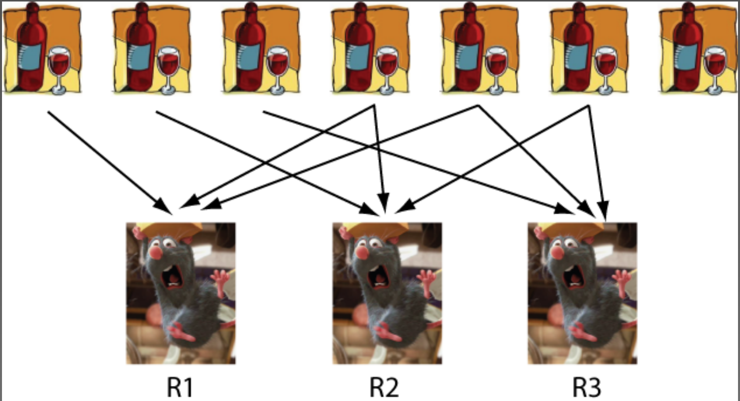
3 Rats

Rat dies only 1 week *after* drinking poisoned wine

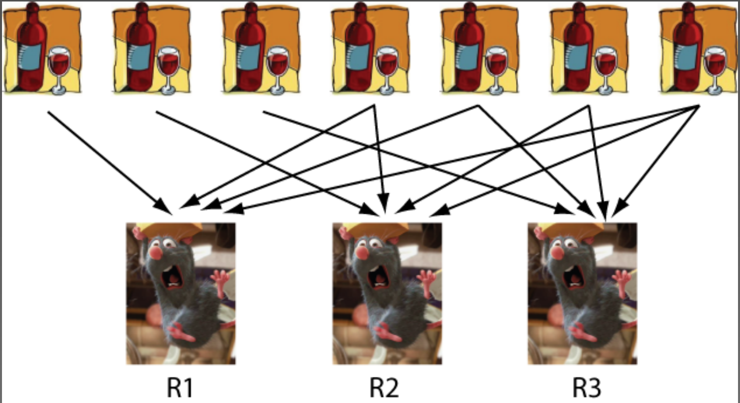
Being good (computer) scientists, they do the following:



	B1	B2	B3	B4	B5	B6	B7
R1	1	0	0				
R2	0	1	0				
R3	0	0	1				



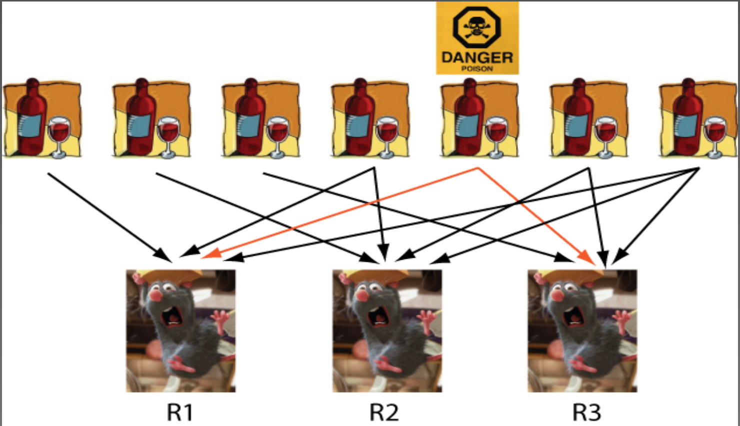
	B1	B2	B3	B4	B5	B6	B7
R1	1	0	0	1	1	0	
R2	0	1	0	1	0	1	
R3	0	0	1	0	1	1	



	B1	B2	B3	B4	B5	B6	B7
R1	1	0	0	1	1	0	1
R2	0	1	0	1	0	1	1
R3	0	0	1	0	1	1	1

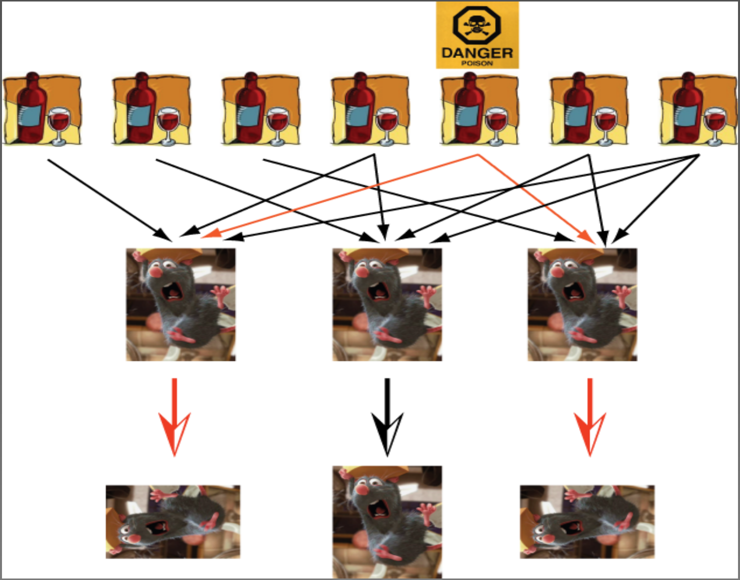
Unique encoding of each bottle

If bottle 5 were poison...



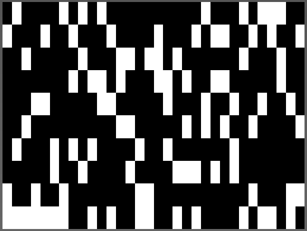
	B1	B2	B3	B4	B5	B6	B7
R1	1	0	0	1	1	0	1
R2	0	1	0	1	0	1	1
R3	0	0	1	0	1	1	1

...after 1 week



Problem statement: CGT

m as small
as possible



Assume x has
low complexity:
x has k-defects
the rest are zero

Construct matrix A: $\mathbb{B}^n \rightarrow \mathbb{B}^m$

Given Ax for any signal $x \in \mathbb{B}^n$, we can quickly recover k defects present in x. Note: arithmetic is boolean and result from pooled test is $\{0, 1\}$.

Parameters

Number of measurements m

Recovery time

Recovery of all k defects

One matrix vs. distribution over matrices

Explicit construction of matrix

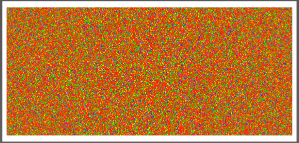
Tolerance to measurement errors (bits flipped, missing bits)

Number of replicates (number of times test each item)

Number of items in each pool

Problem statement: Sparse signal recovery

m as small as possible



Assume x has low complexity: x is k-sparse (with noise)

Construct matrix $A\colon \mathbb{R}^n \to \mathbb{R}^m$

Given Ax for any signal $x \in \mathbb{R}^n$, we can quickly recover $\widehat{x}$ with

$$\|x - \widehat{x}\|_p \leq C \min_{y\ k-sparse} \|x - y\|_q$$

Parameters

Number of measurements *m*

Recovery time

Approximation guarantee (norms, mixed)

One matrix vs. distribution over matrices

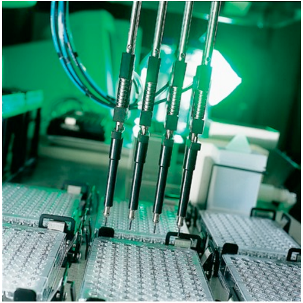
Explicit construction

Tolerance to measurement noise

High Throughput Screening (HTS)

HTS is an essential step in drug discovery (and elsewhere in biology)

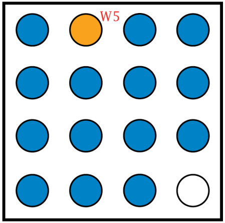
- Large chemical libraries screened on a biological target for activity
- Basic $\{0, 1\}$ type biological assays to find active compounds
- Usually a small number of compounds found
- One-at-a-time screening: automation and miniaturization
- Noisy assays with false positives and negative errors



Current HTS uses one-at-a-time testing scheme (with repeated trials).

NAIVE DESIGN

COMPOUNDS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
WELLS	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15



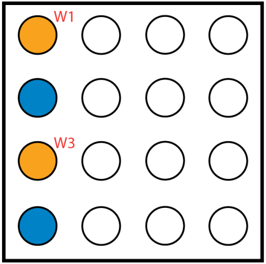
🟠 Wells with compound/s and a positive test result

🔵 Wells with compound/s and a negative test result

○ Wells with no compound/s

Pooled HTS design

POOLING DESIGN



WELLS	COMPOUNDS							
W1	1	3	**5**	7	9	11	13	15
W2	2	3	6	7	10	11	14	15
W3	4	**5**	6	7	12	13	14	15
W4	8	9	10	11	12	13	14	15

Propose using pooled testing of compounds

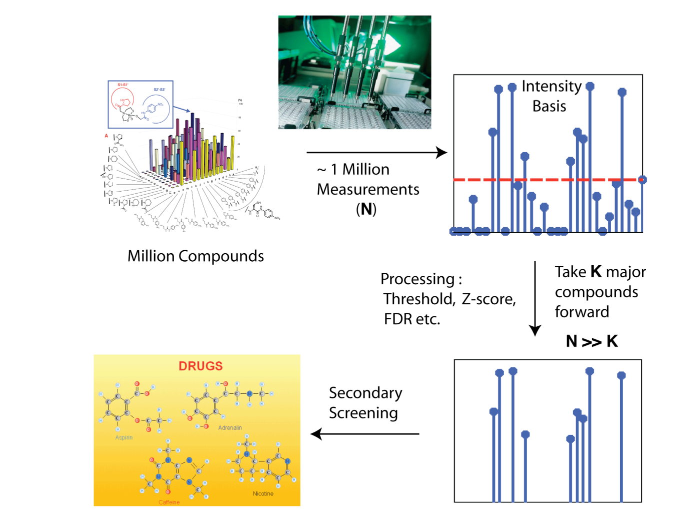
Uses fewer tests

Work moved from testing (costly) to computational analysis (cheap)

Handles errors in testing better due to built-in replication

Additional quantitative information

HTS and signal recovery



Million Compounds

~ 1 Million Measurements (**N**)

Processing : Threshold, Z-score, FDR etc.

Take **K** major compounds forward

N >> K

Secondary Screening

Quantitative analysis of pooling in HTS

Constraints

linearity: measured quantities map linearly to compound activities

sparsity: most compounds inactive

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 & 1 \\ 0 & 1 & \dots & 0 & 1 \\ & \vdots & & \vdots & \\ 1 & 0 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix}$$

Challenges

choosing a good mixing scheme

enforcing a mixing constraint

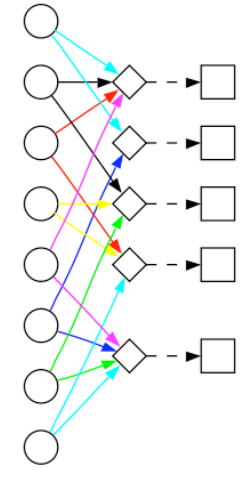
recovery algorithm tolerant to measurement noise + errors

Our approach

Binary measurement matrix: adjacency matrix of unbalanced expander graph

Appropriate linear biochemical model

Decoding via linear programming



Compressed sensing: sparse matrices

LP decoding using sparse matrices

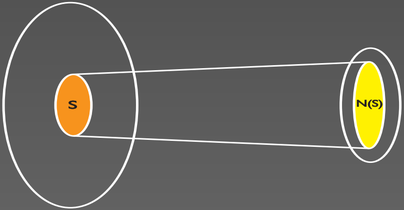
Deterministic (explicit) constructions

Control over number of replicates, number of compounds per pool

LP decoding robust to measurement noise

Recall: Piotr Indyk's talk Thursday

Sparse matrices: Expander graphs



Adjacency matrix A of a d regular $(1, \epsilon)$ expander graph
Graph $G = (X, Y, E)$, $|X| = n$, $|Y| = m$
For any $S \subset X$, $|S| \leq k$, the neighbor set

$$|N(S)| \geq (1 - \epsilon)d|S|$$

Probabilistic construction:

$$d = O(\log(n/k)/\epsilon), m = O(k \log(n/k)/\epsilon^2)$$

Deterministic construction:

$$d = O(2^{O(\log^3(\log(n)/\epsilon))}), m = k/\epsilon\, 2^{O(\log^3(\log(n)/\epsilon))}$$

RIP(p)

A measurement matrix A satisfies RIP(p, k, δ) property if for any k-sparse vector x,

$$(1-\delta)\|x\|_p \leq \|Ax\|_p \leq (1+\delta)\|x\|_p.$$

RIP(p) $\iff$ expander

Theorem

(k, ϵ) expansion implies

$$(1 - 2\epsilon) d \|x\|_1 \leq \|Ax\|_1 \leq d \|x\|_1$$

for any k-sparse x. Get RIP(p) for $1 \leq p \leq 1 + 1/\log n$.

Theorem

RIP(1) + binary sparse matrix implies (k, ϵ) expander for

$$\epsilon = \frac{1 - 1/(1 + \delta)}{2 - \sqrt{2}}.$$

Expansion $\Longrightarrow$ LP decoding

Theorem

Φ adjacency matrix of $(2k, \epsilon)$ expander. Consider two vectors x, x_* such that $\Phi x = \Phi x_*$ and $\|x_*\|_1 \leq \|x\|_1$. Then

$$\|x - x_*\|_1 \leq \frac{2}{1 - 2\alpha(\epsilon)} \|x - x_k\|_1$$

where x_k is the optimal k-term representation for x and $\alpha(\epsilon) = (2\epsilon)/(1 - 2\epsilon)$.

- Guarantees that Linear Program recovers good sparse approximation
- Robust to noisy measurements too

RIP(1) $\implies$ LP decoding

ℓ_1 uncertainty principle

Lemma

Let y satisfy $Ay = 0$. Let S the set of k largest coordinates of y. Then

$$\|y_S\|_1 \le \alpha(\epsilon)\|y\|_1.$$

LP guarantee

Theorem

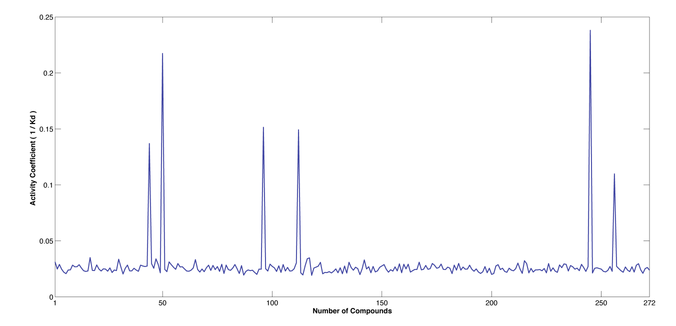
Consider any two vectors u, v such that for $y = u - v$ we have $Ay = 0$, $\|v\|_1 \le \|u\|_1$. S set of k largest entries of u. Then

$$\|y\|_1 \le \frac{2}{1 - 2\alpha(\epsilon)}\|u_{S^c}\|_1.$$

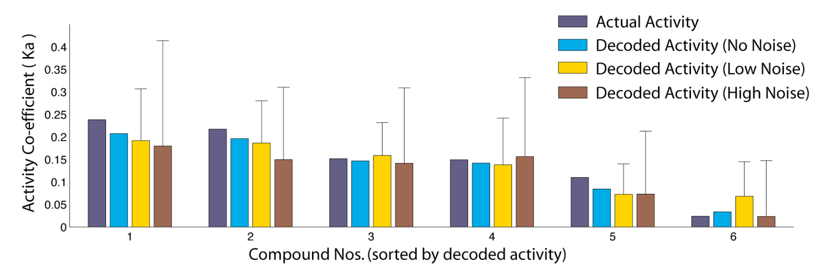
Small library

Synthetic screen: small molecule ligands for formylpeptide receptor, 6 active [Edwards, et al., Nature Protocols '06]

$n = 272, k = 6$, using deterministic STD matrix, $m = 116$



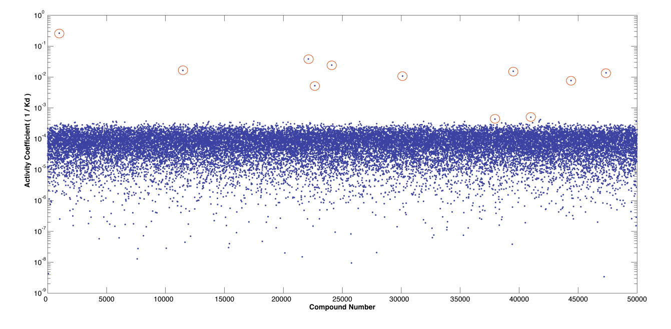
In silico


Actual Activity
Decoded Activity (No Noise)
Decoded Activity (Low Noise)
Decoded Activity (High Noise)
Activity Co-efficient (Ka)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
Compound Nos. (sorted by decoded activity)

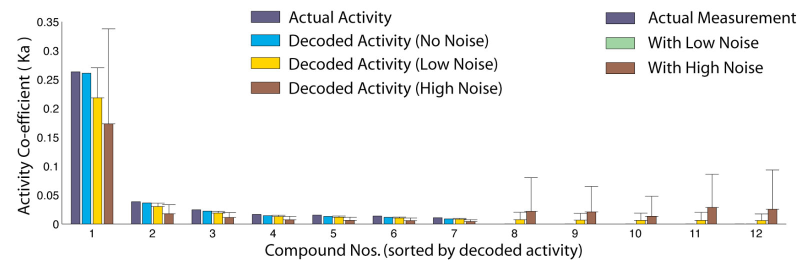
Large library

Actual screen: 50,000 compounds screened against E. coli dihydrofolate reductase (DHFR), 12 active [McMaster HTS Lab Data Mining and Docking Competition '05]

$n = 50,000$, $k = 12$ screened in 122 blocks of 410 compounds using STD deterministic matrix, $m = 10,004$



In silico



Actual Activity
Decoded Activity (No Noise)
Decoded Activity (Low Noise)
Decoded Activity (High Noise)
Actual Measurement
With Low Noise
With High Noise
Activity Co-efficient (Ka)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9
10
11
12
Compound Nos. (sorted by decoded activity)

Current/Future work

Computer Science:

greedy algorithms in place of LP decoding
decoding with noise + missing measurements
refined error analysis
decoding algorithms to rank compounds

Chemical Engineering:

good/best explicit constructions which meet experimental constraints
refine error analysis, algorithm output for cultural interpretations of biologists
design and implementation of several in vitro experiments (HTS, differential gene expression)